Models for Binary Outcomes

Logistic regression and variations

Contents

Configuring R

Functions from these packages will be used throughout this document:

```
library(conflicted) # check for conflicting function definitions
# library(printr) # inserts help-file output into markdown output
library(rmarkdown) # Convert R Markdown documents into a variety of formats.
library(pander) # format tables for markdown
library(ggplot2) # graphics
library(ggfortify) # help with graphics
library(dplyr) # manipulate data
library(tibble) # `tibble`s extend `data.frame`s
library(magrittr) # `%>%` and other additional piping tools
library(haven) # import Stata files
library(knitr) # format R output for markdown
library(tidyr) # Tools to help to create tidy data
library(plotly) # interactive graphics
library(dobson) # datasets from Dobson and Barnett 2018
library(parameters) # format model output tables for markdown
library(haven) # import Stata files
library(latex2exp) # use LaTeX in R code (for figures and tables)
library(fs) # filesystem path manipulations
library(survival) # survival analysis
library(survminer) # survival analysis graphics
library(KMsurv) # datasets from Klein and Moeschberger
library(parameters) # format model output tables for
library(webshot2) # convert interactive content to static for pdf
library(forcats) # functions for categorical variables ("factors")
library(stringr) # functions for dealing with strings
library(lubridate) # functions for dealing with dates and times
library(broom) # Summarizes key information about statistical objects in tidy tibbles
library(broom.helpers) # Provides suite of functions to work with regression model 'broom::tidy()' t
```

Here are some R settings I use in this document:

```
rm(list = ls()) # delete any data that's already loaded into R

conflicts_prefer(dplyr::filter)
ggplot2::theme_set(
  ggplot2::theme_bw() +
        # ggplot2::labs(col = "") +
    ggplot2::theme(
      legend.position = "bottom",
      text = ggplot2::element_text(size = 12, family = "serif")))

knitr::opts_chunk$set(message = FALSE)
options('digits' = 6)

panderOptions("big.mark", ",")
pander::panderOptions("table.emphasize.rownames", FALSE)
pander::panderOptions("table.split.table", Inf)
conflicts_prefer(dplyr::filter) # use the `filter()` function from dplyr() by default
legend_text_size = 9
run_graphs = TRUE
```

```
options(digits = 6)
```

Acknowledgements

This content is adapted from:

- Dobson and Barnett (2018), Chapter 7
- Vittinghoff et al. (2012), Chapter 5
- David Rocke[1]'s materials from the 2021 edition of Epi 204[2]
- Nahhas (2024) Chapter 6[3]

1 Introduction

Exercise 1.1. What is logistic regression?

Solution

Solution 1.1.

Def

Definition 1.1 (Logistic regression model). **Logistic regression** is a framework for modeling binary[a] outcomes, conditional on one or more *predictors* (a.k.a. *covariates*).

[a] data.qmd#def-binary

Exercise 1.2 (Examples of binary outcomes)**.** What are some examples of binary outcomes in the health sciences?

Solution

Solution 1.2. Examples of binary outcomes include:

- exposure (exposed vs unexposed)
- disease (diseased vs healthy)
- recovery (recovered vs unrecovered)
- relapse (relapse vs remission)
- return to hospital (returned vs not)
- vital status (dead vs alive)

Logistic regression uses the Bernoulli[4] distribution to model the outcome variable, conditional on one or more covariates.

Exercise 1.3. Write down a mathematical definition of the Bernoulli distribution.

Solution

Solution 1.3. The **Bernoulli distribution** family for a random variable X is defined as:

$$\begin{aligned}\Pr(X = x) &\overset{\text{def}}{=} 1_{x\in\{0,1\}}\pi^x(1-\pi)^{1-x} \\ &= \begin{cases}\pi, & x = 1 \\ 1-\pi, & x = 0\end{cases}\end{aligned}$$

[1] https://dmrocke.ucdavis.edu/
[2] https://dmrocke.ucdavis.edu/Class/EPI204-Spring-2021/EPI204-Spring-2021.html
[3] https://www.bookdown.org/rwnahhas/RMPH/blr.html
[4] probability.qmd#def-bernoulli

1.0.1 Logistic regression versus linear regression

Logistic regression differs from linear regression, which uses the Gaussian ("normal") distribution to model the outcome variable, conditional on the covariates.

Exercise 1.4. Recall: what kinds of outcomes is linear regression used for?

Solution

Solution 1.4. Linear regression is typically used for numerical outcomes that aren't event counts or waiting times for an event.
Examples of outcomes that are often analyzed using linear regression include:

- weight
- height
- income
- prices

This chapter assumes familiarity with the measures of association for binary outcomes — risks, risk differences and ratios, odds, odds ratios, and the logit and expit functions — which are reviewed in Measures of Association for Binary Outcomes[5].

2 Introduction to logistic regression

- In the OC-MI example[6], we estimated the risk and the odds of MI for two groups, defined by oral contraceptive use.
- If the predictor is quantitative (dose) or there is more than one predictor, the task becomes more difficult.
- In this case, we will use logistic regression, which is a generalization of the linear regression models you have been using that can account for a binary response instead of a continuous one.

2.0.1 Independent binary outcomes - general

Exercise 2.1. Let $\tilde{y}$ represent a data set of mutually independent binary outcomes, each with a potentially different event probability π_i:

$$\tilde{y} = (y_1, ..., y_n)$$
$$y_i \sim_{\perp\!\!\!\perp} \text{Ber}(\pi_i)$$

Write the likelihood of $\tilde{y}$.

[5] binary-outcome-associations.qmd
[6] binary-outcome-associations.qmd#exm-oc-mi

Solution

Solution 2.1.

$$\begin{aligned}
\pi_i &\stackrel{\text{def}}{=} \mathrm{P}(Y_i = 1) \\
\mathrm{P}(Y_i = 0) &= 1 - \pi_i \\
\mathrm{P}(Y_i = y_i) &= \mathrm{P}(Y_i = 1)^{y_i}\, \mathrm{P}(Y_i = 0)^{1-y_i} \\
&= (\pi_i)^{y_i}(1-\pi_i)^{1-y_i} \\
\mathcal{L}_i(\pi_i) &\stackrel{\text{def}}{=} \mathrm{P}(Y_i = y_i) \\
\mathcal{L}(\tilde{\pi}) &\stackrel{\text{def}}{=} \mathrm{P}(Y_1 = y_1, \ldots, Y_n = y_n) \\
&= \prod_{i=1}^{n} \mathrm{P}(Y_i = y_i) \\
&= \prod_{i=1}^{n} \mathcal{L}_i(\pi_i) \\
&= \prod_{i=1}^{n} (\pi_i)^{y_i}(1-\pi_i)^{1-y_i}
\end{aligned}$$

Exercise 2.2. Write the log-likelihood of $\tilde{y}$.

Solution

Solution 2.2.

$$\begin{aligned}
\ell(\tilde{\pi}) &\stackrel{\text{def}}{=} \log\{\mathcal{L}(\tilde{\pi})\} \\
&= \log\left\{\prod_{i=1}^{n} \mathcal{L}_i(\pi_i)\right\} \\
&= \sum_{i=1}^{n} \log\{\mathcal{L}_i(\pi_i)\} \\
&= \sum_{i=1}^{n} \ell_i(\pi_i) \\
\ell_i(\pi_i) &\stackrel{\text{def}}{=} \log\{\mathcal{L}_i(\pi_i)\} \\
&= y_i \log\{\pi_i\} + (1-y_i)\log\{1-\pi_i\}
\end{aligned}$$

2.0.2 Modeling π_i as a function of X_i

If there are only a few distinct X_i values, we can model π_i separately for each value of X_i. Otherwise, we need regression.

$$\begin{aligned}
\pi(x) &\equiv \mathrm{E}(Y = 1 | X = x) \\
&= f(x^\top \beta)
\end{aligned}$$

Typically, we use the expit inverse-link:

$$\pi(\tilde{x}) = \mathrm{expit}(\tilde{x}'\beta) \tag{1}$$

2.0.3 Meet the beetles

Table 1: Mortality rates of adult flour beetles after five hours' exposure to gaseous carbon disulphide (Bliss 1935)

```
library(glmx)
library(dplyr)
data(BeetleMortality, package = "glmx")
beetles <- BeetleMortality |>
  mutate(
    pct = died / n,
    survived = n - died,
    dose_c = dose - mean(dose)
  )
beetles
#> # A tibble: 8 x 6
#>    dose  died     n   pct survived   dose_c
#>   <dbl> <int> <int> <dbl>    <int>    <dbl>
#> 1  1.69     6    59 0.102       53 -0.103
#> 2  1.72    13    60 0.217       47 -0.0692
#> 3  1.76    18    62 0.290       44 -0.0382
#> 4  1.78    28    56 0.5         28 -0.00923
#> 5  1.81    52    63 0.825       11  0.0179
#> 6  1.84    53    59 0.898        6  0.0435
#> 7  1.86    61    62 0.984        1  0.0676
#> 8  1.88    60    60 1            0  0.0905
```

```
library(ggplot2)
plot1 <-
  beetles |>
  ggplot(aes(x = dose, y = pct)) +
  geom_point(aes(size = n)) +
  xlab("Dose (log mg/L)") +
  ylab("Mortality rate (%)") +
  scale_y_continuous(labels = scales::percent) +
  scale_size(range = c(1, 2)) +
  theme_bw(base_size = 18)

print(plot1)
```

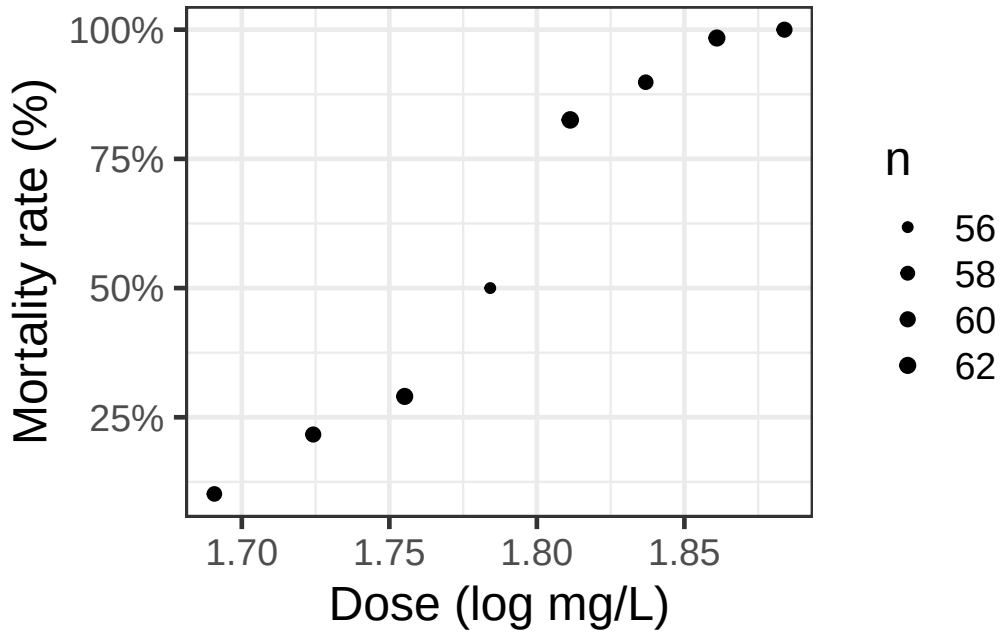


Figure 1: Mortality rates of adult flour beetles after five hours' exposure to gaseous carbon disulphide (Bliss 1935)

2.0.4 Why don't we use linear regression?

```
beetles_long <- beetles |>
  reframe(
    .by = everything(),
    outcome = c(
      rep(1, times = died),
      rep(0, times = survived)
    )
  ) |>
  as_tibble()

lm1 <- beetles_long |> lm(formula = outcome ~ dose)
f_linear <- function(x) predict(lm1, newdata = data.frame(dose = x))

range1 <- range(beetles$dose) + c(-.2, .2)

plot2 <-
  plot1 +
  geom_function(
    fun = f_linear,
    aes(col = "Straight line")
  ) +
  labs(colour = "Model", size = "")

plot2 |> print()
```

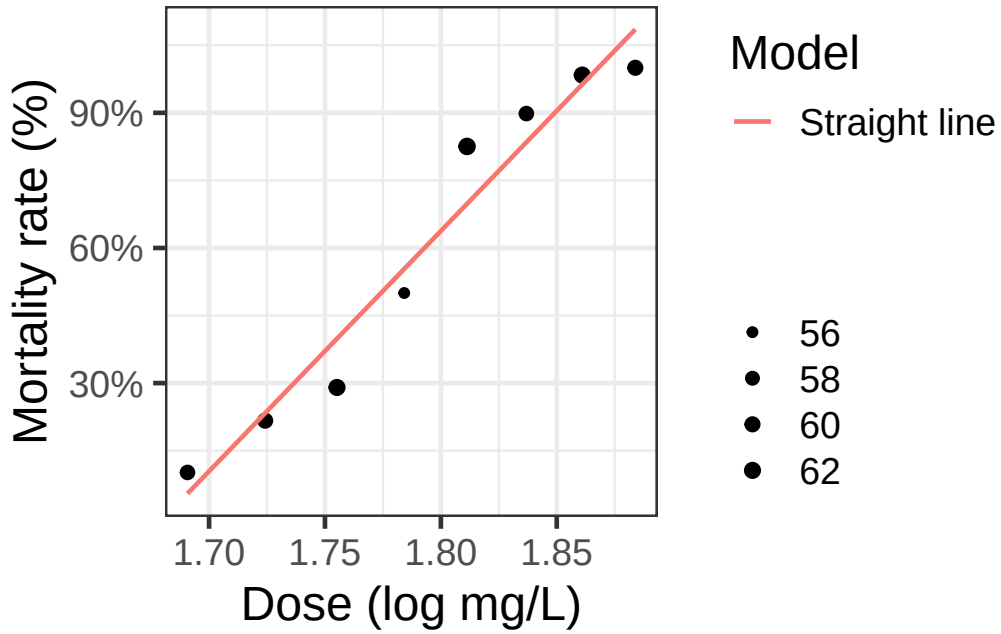


Figure 2: Mortality rates of adult flour beetles after five hours' exposure to gaseous carbon disulphide (Bliss 1935)

2.0.5 Zoom out

```
(plot2 + expand_limits(x = c(1.6, 2))) |> print()
```

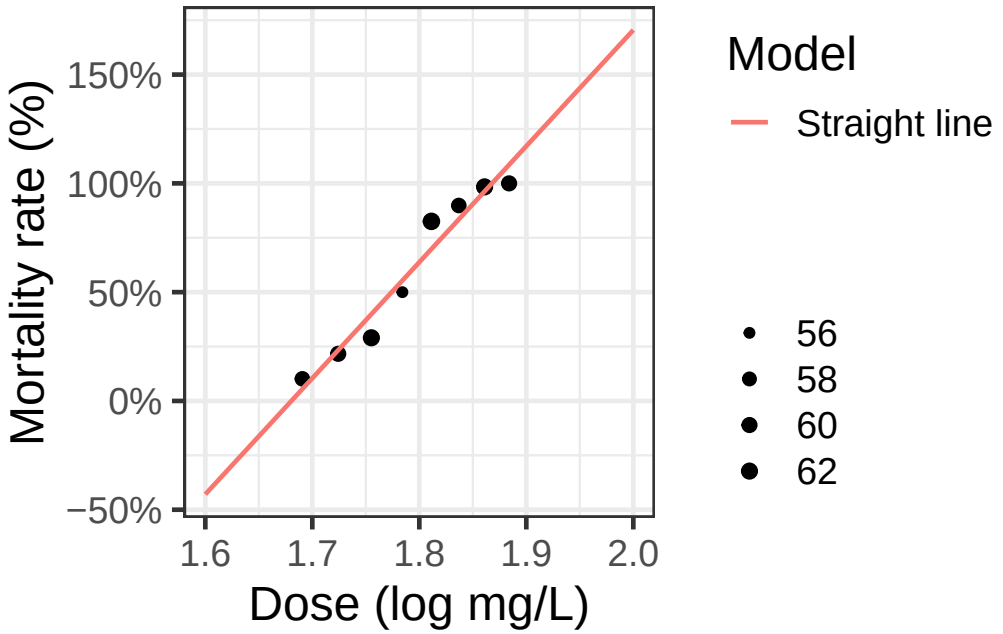


Figure 3: Mortality rates of adult flour beetles after five hours' exposure to gaseous carbon disulphide (Bliss 1935)

2.0.6 log transformation of dose?

```
lm2 <- beetles_long |> lm(formula = outcome ~ log(dose))
f_linearlog <- function(x) predict(lm2, newdata = data.frame(dose = x))

plot3 <- plot2 +
  expand_limits(x = c(1.6, 2)) +
  geom_function(fun = f_linearlog, aes(col = "Log-transform dose"))
(plot3 + expand_limits(x = c(1.6, 2))) |> print()
```

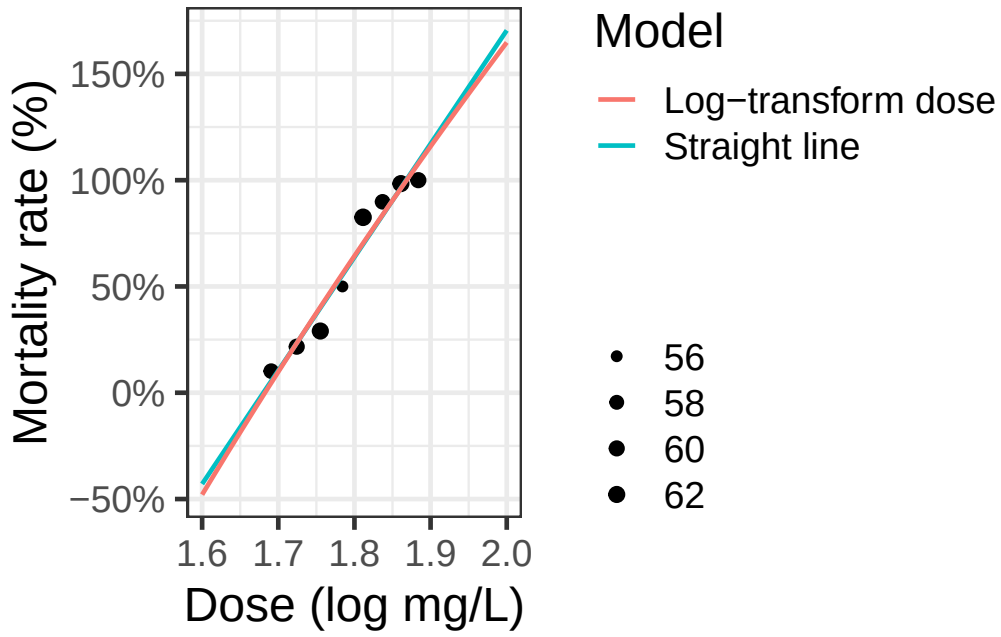


Figure 4: Mortality rates of adult flour beetles after five hours' exposure to gaseous carbon disulphide (Bliss 1935)

2.0.7 Logistic regression

```
beetles_glm_grouped <- beetles |>
  glm(formula = cbind(died, survived) ~ dose, family = "binomial")
f <- function(x) {
  beetles_glm_grouped |>
    predict(newdata = data.frame(dose = x), type = "response")
}

plot4 <- plot3 + geom_function(fun = f, aes(col = "Logistic regression"))
plot4 |> print()
```

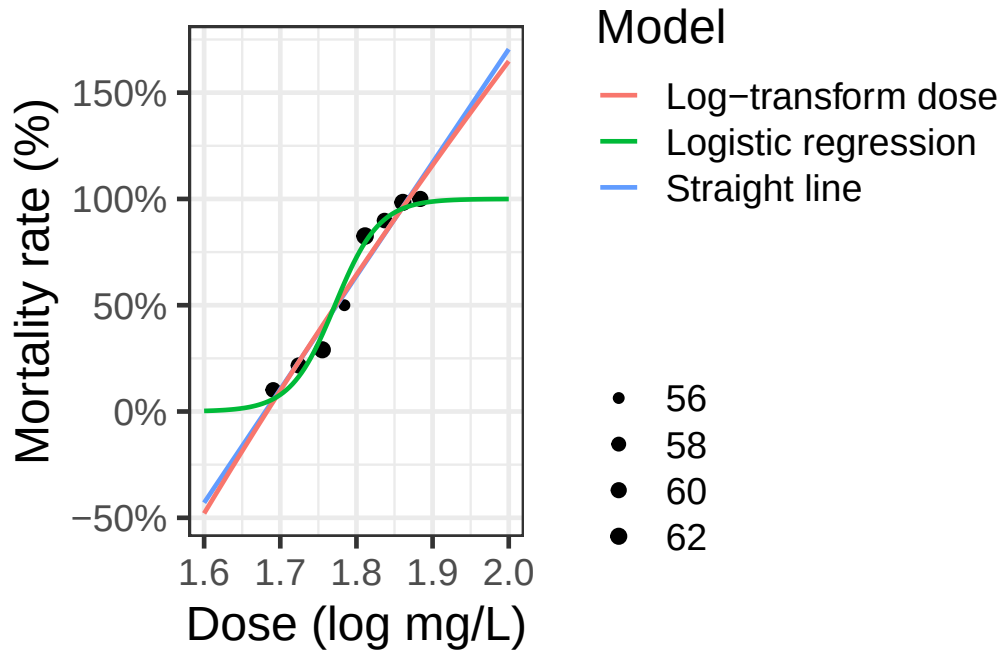


Figure 5: Mortality rates of adult flour beetles after five hours' exposure to gaseous carbon disulphide (Bliss 1935).

2.0.8 Three parts to regression models

- What distribution does the outcome have for a specific subpopulation defined by covariates? (outcome model)
- How does the combination of covariates relate to the mean? (link function)
- How do the covariates combine? (linear predictor, interactions)

2.0.9 Fitting and manipulating logistic regression models in R

```
library(glmx)
library(dplyr)
data(BeetleMortality)
beetles <- BeetleMortality |>
  mutate(
    pct = died / n,
    survived = n - died
  )

beetles_glm_grouped <-
  beetles |>
```

Table 3

```
fitted(beetles_glm_grouped)
#>        1        2        3        4        5        6        7        8
#> 0.058601 0.164028 0.362119 0.605315 0.795172 0.903236 0.955196 0.979049
predict(beetles_glm_grouped, type = "response")
#>        1        2        3        4        5        6        7        8
#> 0.058601 0.164028 0.362119 0.605315 0.795172 0.903236 0.955196 0.979049
```

```
  glm(
    formula = cbind(died, survived) ~ dose,
    family = "binomial"
  )

library(parameters)
beetles_glm_grouped |>
  parameters() |>
  print_md()
```

Table 2: logistic regression model for beetles data with grouped (binomial) data

Parameter	Log-Odds	SE	95% CI	z	p
(Intercept)	-60.72	5.18	(-71.44, -51.08)	-11.72	< .001
dose	34.27	2.91	(28.85, 40.30)	11.77	< .001

Fitted values

Fitted values are provided on the probability scale (Table 3)

Count scale

For grouped data, we can convert to the count scale by multiplying by the group size:

```
beetles$n * fitted(beetles_glm_grouped)
#>        1        2        3        4        5        6        7        8
#>  3.45746  9.84167 22.45138 33.89763 50.09582 53.29091 59.22216 58.74296
```

Logit scale

```
predict(beetles_glm_grouped, type = "link")
#>         1         2         3         4         5         6         7         8
#> -2.776615 -1.628559 -0.566179  0.427661  1.356386  2.233707  3.059622  3.844412
```

Converting between logit and probability scales works as expected:

```
predict(beetles_glm_grouped, type = "link") |> arm::invlogit()
#>        1        2        3        4        5        6        7        8
#> 0.058601 0.164028 0.362119 0.605315 0.795172 0.903236 0.955196 0.979049
predict(beetles_glm_grouped, type = "response")
#>        1        2        3        4        5        6        7        8
#> 0.058601 0.164028 0.362119 0.605315 0.795172 0.903236 0.955196 0.979049

predict(beetles_glm_grouped, type = "response") |> arm::logit()
#>         1         2         3         4         5         6         7         8
#> -2.776615 -1.628559 -0.566179  0.427661  1.356386  2.233707  3.059622  3.844412
predict(beetles_glm_grouped, type = "link")
#>         1         2         3         4         5         6         7         8
#> -2.776615 -1.628559 -0.566179  0.427661  1.356386  2.233707  3.059622  3.844412
```

Table 4: `beetles` data in long format

```
beetles_long
#> # A tibble: 481 x 7
#>     dose  died     n   pct survived dose_c outcome
#>    <dbl> <int> <int> <dbl>    <int>  <dbl>   <dbl>
#>  1  1.69     6    59 0.102       53 -0.103       1
#>  2  1.69     6    59 0.102       53 -0.103       1
#>  3  1.69     6    59 0.102       53 -0.103       1
#>  4  1.69     6    59 0.102       53 -0.103       1
#>  5  1.69     6    59 0.102       53 -0.103       1
#>  6  1.69     6    59 0.102       53 -0.103       1
#>  7  1.69     6    59 0.102       53 -0.103       0
#>  8  1.69     6    59 0.102       53 -0.103       0
#>  9  1.69     6    59 0.102       53 -0.103       0
#> 10  1.69     6    59 0.102       53 -0.103       0
#> # i 471 more rows
```

`type = "terms"` is confusing, because the variables get centered:

```
predict(beetles_glm_grouped, type = "terms")
#>        dose
#> 1 -3.520419
#> 2 -2.372363
#> 3 -1.309983
#> 4 -0.316144
#> 5  0.612582
#> 6  1.489902
#> 7  2.315817
#> 8  3.100608
#> attr(,"constant")
#> [1] 0.743804
coef(beetles_glm_grouped)["dose"] * beetles$dose
#> [1] 57.9408 59.0889 60.1513 61.1451 62.0738 62.9512 63.7771 64.5619
```

We can construct the link-scale predictions from the terms:

```
terms_pred <- predict(beetles_glm_grouped, type = "terms")
terms_pred + attr(terms_pred, "constant")
#>        dose
#> 1 -2.776615
#> 2 -1.628559
#> 3 -0.566179
#> 4  0.427661
#> 5  1.356386
#> 6  2.233707
#> 7  3.059622
#> 8  3.844412
#> attr(,"constant")
#> [1] 0.743804
predict(beetles_glm_grouped, type = "link")
#>         1         2         3         4         5         6         7         8
#> -2.776615 -1.628559 -0.566179  0.427661  1.356386  2.233707  3.059622  3.844412
```

Individual observations

```
beetles_glm_ungrouped <-
  beetles_long |>
  glm(
```

```
    formula = outcome ~ dose,
    family = "binomial"
  )

beetles_glm_ungrouped |>
  parameters() |>
  print_md()
```

Table 5: logistic regression model for beetles data with individual Bernoulli data

Parameter	Log-Odds	SE	95% CI	z	p
(Intercept)	-60.72	5.18	(-71.44, -51.08)	-11.72	$< .001$
dose	34.27	2.91	(28.85, 40.30)	11.77	$< .001$

Exercise 2.3. Compare this model with the grouped-observations model (Table 2).

Solution

Solution 2.3.
They seem the same! But not quite:

```
logLik(beetles_glm_grouped)
#> 'log Lik.' -18.7151 (df=2)
logLik(beetles_glm_ungrouped)
#> 'log Lik.' -186.235 (df=2)
```

The difference is due to the binomial coefficient $\binom{n}{x}$ which isn't included in the individual-observations (Bernoulli) version of the model.

3 Derivatives of logistic regression functions

In order to interpret logistic regression models and find their MLEs, we will need to compute various derivatives. This section compiles some useful results.

3.0.1 Derivatives of odds function

Theorem 3.1 (Derivative of odds function)**.**

$$\text{odds}'\{\pi\} = \frac{\partial \omega}{\partial \pi} = \frac{1}{(1-\pi)^2}$$

Proof

Proof. We can use odds as a function of probability[a] and the quotient rule[b]:

$$
\begin{aligned}
\frac{\partial \omega}{\partial \pi} &= \frac{\partial}{\partial \pi}\left(\frac{\pi}{1-\pi}\right) \\
&= \frac{\frac{\partial}{\partial \pi}\pi}{1-\pi} - \left(\frac{\pi}{(1-\pi)^2} \cdot \frac{\partial}{\partial \pi}(1-\pi)\right) \\
&= \frac{1}{1-\pi} - \frac{\pi}{(1-\pi)^2} \cdot (-1) \\
&= \frac{1}{1-\pi} + \frac{\pi}{(1-\pi)^2} \\
&= \frac{1-\pi}{(1-\pi)^2} + \frac{\pi}{(1-\pi)^2} \\
&= \frac{1-\pi+\pi}{(1-\pi)^2} \\
&= \frac{1}{(1-\pi)^2}
\end{aligned}
$$

□

[a] binary-outcome-associations.qmd#thm-prob-to-odds
[b] math-prereqs.qmd#thm-quotient-rule

Corollary 3.1 (Derivative of odds function in terms of odds)**.**

$$\frac{\partial \omega}{\partial \pi} = (1+\omega)^2$$

Proof

Proof. Using Theorem 3.1 and one plus odds in terms of non-event probability[a]:

$$
\begin{aligned}
\frac{\partial \omega}{\partial \pi} &= \frac{1}{(1-\pi)^2} && \text{(derivative of odds function)} \\
&= \left(\frac{1}{1-\pi}\right)^2 && \text{(the reciprocal of a square is the square of the reciprocal)} \\
&= (1+\omega)^2 && \text{(substitute } \frac{1}{1-\pi} = 1+\omega\text{)}
\end{aligned}
$$

□

[a] binary-outcome-associations.qmd#cor-inverse-odds-nonevent

3.0.2 Derivatives of inverse-odds function

Theorem 3.2 (Derivative of inverse odds function)**.**

$$\text{invodds}'\{\omega\} = \frac{\partial \pi}{\partial \omega} = (1-\pi)^2 = \frac{1}{(1+\omega)^2} \tag{2}$$

Proof

Proof. Using the inverse function rule[a], Theorem 3.1, and one minus inverse-odds[b]:

$$
\begin{aligned}
\frac{\partial \pi}{\partial \omega} &= \left(\frac{\partial \omega}{\partial \pi}\right)^{-1} && \text{(inverse function rule)} \\
&= \left(\frac{1}{(1-\pi)^2}\right)^{-1} && \text{(derivative of odds function)} \\
&= (1-\pi)^2 && \text{(the reciprocal of a reciprocal is the original quantity)} \\
&= \left(\frac{1}{1+\omega}\right)^2 && \text{(substitute } 1-\pi = \frac{1}{1+\omega}\text{)} \\
&= \frac{1}{(1+\omega)^2} && \text{(the square of a reciprocal is the reciprocal of the square)}
\end{aligned}
$$

Or for a direct approach, use the quotient rule[c] again:

$$
\begin{aligned}
\frac{\partial \pi}{\partial \omega} &= \frac{\partial}{\partial \omega}\frac{\omega}{1+\omega} \\
&= \frac{\frac{\partial}{\partial \omega}\omega}{1+\omega} - \frac{\omega}{(1+\omega)^2} \cdot \frac{\partial}{\partial \omega}(1+\omega) \\
&= \frac{1}{1+\omega} - \frac{\omega}{(1+\omega)^2} \cdot 1 \\
&= \frac{1}{1+\omega} - \frac{\omega}{(1+\omega)^2} \\
&= \frac{1+\omega}{(1+\omega)^2} - \frac{\omega}{(1+\omega)^2} \\
&= \frac{1+\omega-\omega}{(1+\omega)^2} \\
&= \frac{1}{(1+\omega)^2}
\end{aligned}
$$

□

[a] https://en.wikipedia.org/wiki/Inverse_function_rule
[b] binary-outcome-associations.qmd#lem-one-minus-odds-inv
[c] math-prereqs.qmd#thm-quotient-rule

3.0.3 Derivatives of logit function

Lemma 3.1 (Derivative of log-odds by odds)**.**

$$\frac{\partial \eta}{\partial \omega} = \omega^{-1}$$

Proof

Proof. Using the definition of log-odds[a]:

$$
\begin{aligned}
\frac{\partial \eta}{\partial \omega} &= \frac{\partial}{\partial \omega}\log \omega \\
&= \omega^{-1}
\end{aligned}
$$

□

[a] binary-outcome-associations.qmd#def-logodds

Theorem 3.3 (Derivative of log-odds by odds)**.**

$$\frac{\partial \eta}{\partial \omega} = \frac{1-\pi}{\pi}$$

Proof

Proof. Using odds as a function of probability[a] and Lemma 3.1:

$$\begin{aligned}\frac{\partial \eta}{\partial \omega} &= \omega^{-1} \\ &= \frac{1-\pi}{\pi}\end{aligned}$$

□

[a] binary-outcome-associations.qmd#thm-prob-to-odds

Theorem 3.4 (Derivative of log-odds by probability)**.**

$$\frac{\partial \eta}{\partial \pi} = \frac{1}{(\pi)(1-\pi)}$$

Proof

Proof. Using Theorem 3.3, Theorem 3.1, and the chain rule[a]:

$$\begin{aligned}\frac{\partial \eta}{\partial \pi} &= \frac{\partial \eta}{\partial \omega}\frac{\partial \omega}{\partial \pi} \\ &= \frac{1-\pi}{\pi}\frac{1}{(1-\pi)^2} \\ &= \frac{1}{(\pi)(1-\pi)}\end{aligned}$$

□

[a] math-prereqs.qmd#thm-chain-rule

Corollary 3.2 (Derivative of logit function)**.**

$$\text{logit}'(\pi) = \frac{1}{(\pi)(1-\pi)}$$

Proof

Proof. Using log-odds via the logit function[a] and Theorem 3.4:

$$\begin{aligned}\text{logit}'(\pi) &= \frac{\partial}{\partial \pi}\text{logit}\{\pi\} && \text{(Lagrange notation rewritten in Leibniz notation)} \\ &= \frac{\partial}{\partial \pi}\eta && \text{(log-odds via the logit function: } \text{logit}\{\pi\} = \eta\text{)} \\ &= \frac{\partial \eta}{\partial \pi} && \text{(Leibniz notation)} \\ &= \frac{1}{(\pi)(1-\pi)} && \text{(derivative of log-odds by probability)}\end{aligned}$$

□

[a] binary-outcome-associations.qmd#cor-logodds-logit

3.0.4 Derivatives of expit function

Lemma 3.2 (Derivative of odds w.r.t. log-odds)**.**

$$\frac{\partial \omega}{\partial \eta} = \omega$$

Proof

Proof. Using the odds-from-log-odds lemma[a] and the derivative of the exponential function[b]:

$$\begin{aligned} \frac{\partial \omega}{\partial \eta} &= \frac{\partial}{\partial \eta} \exp\{\eta\} \\ &= \exp\{\eta\} \\ &= \omega \end{aligned}$$

□

[a] binary-outcome-associations.qmd#lem-odds-from-logodds
[b] math-prereqs.qmd#thm-deriv-exp

Theorem 3.5 (Derivative of odds in terms of probability)**.**

$$\frac{\partial \omega}{\partial \eta} = \frac{\pi}{1-\pi} \tag{3}$$

Proof

Proof. Using Lemma 3.2 and odds as a function of probability[a]:

$$\begin{aligned} \frac{\partial \omega}{\partial \eta} &= \omega && \text{(derivative of odds w.r.t. log-odds)} \\ &= \frac{\pi}{1-\pi} && \text{(odds as a function of probability)} \end{aligned}$$

□

[a] binary-outcome-associations.qmd#thm-prob-to-odds

Theorem 3.6 (Derivative of probability w.r.t. log-odds)**.**

$$\frac{\partial \pi}{\partial \eta} = \pi(1-\pi)$$

Proof

Proof. By the chain rule[a], Theorem 3.5, and Theorem 3.2:

$$\begin{aligned} \frac{\partial \pi}{\partial \eta} &= \frac{\partial \omega}{\partial \eta} \frac{\partial \pi}{\partial \omega} \\ &= \frac{\pi}{1-\pi}(1-\pi)^2 \\ &= \pi(1-\pi) \end{aligned}$$

Alternatively, by Theorem 3.4:

$$\begin{aligned}\frac{\partial \pi}{\partial \eta} &= \left(\frac{\partial \eta}{\partial \pi}\right)^{-1} \\ &= \left(\frac{1}{(\pi)(1-\pi)}\right)^{-1} \\ &= \pi(1-\pi)\end{aligned}$$

□

[a] math-prereqs.qmd#thm-chain-rule

Corollary 3.3 (Derivative of probability w.r.t. linear predictor as a variance)**.** *If* $\pi = \Pr(Y = 1|\tilde{X} = \tilde{x})$*, then:*

$$\frac{\partial \pi}{\partial \eta} = \text{Var}(Y|X = x)$$

Proof

Proof. Since Y is a binary (0/1) outcome with $\Pr(Y = 1|\tilde{X} = \tilde{x}) = \pi$, the conditional distribution of Y given $\tilde{X} = \tilde{x}$ is Bernoulli[a] with parameter π.
First, we compute $\text{Var}(Y|X = x)$, using the simplified expression for variance[b], the expected value of the square of a Bernoulli variable[c], and the expectation of the Bernoulli distribution[d], all applied to the conditional distribution of Y given $X = x$:

$$\begin{aligned}\text{Var}(Y|X = x) &= \text{E}[Y^2|X = x] - (\text{E}[Y|X = x])^2 && \text{(simplified expression for variance)} \\ &= \pi - (\text{E}[Y|X = x])^2 && \text{(expected square of a Bernoulli variable: } \text{E}[Y^2|X = x] = \pi\text{)} \\ &= \pi - (\pi)^2 && \text{(Bernoulli mean: } \text{E}[Y|X = x] = \pi\text{)} \\ &= \pi(1-\pi) && \text{(factor out } \pi\text{)}\end{aligned} \tag{4}$$

Then, using Theorem 3.6 and Equation 4:

$$\begin{aligned}\frac{\partial \pi}{\partial \eta} &= \pi(1-\pi) && \text{(derivative of probability w.r.t. log-odds)} \\ &= \text{Var}(Y|X = x) && \text{(Bernoulli conditional variance, read right-to-left)}\end{aligned}$$

□

[a] probability.qmd#def-bernoulli
[b] probability.qmd#thm-variance
[c] probability.qmd#exm-lotus
[d] probability.qmd#thm-bernoulli-mean

Corollary 3.4 (Derivative of expit)**.**

$$\text{expit}'\{\eta\} = \text{expit}\{\eta\}(1 - \text{expit}\{\eta\})$$

Proof

Proof. Using probability via the expit function[a] and Theorem 3.6:

$$
\begin{aligned}
\text{expit}'\{\eta\} = \frac{\partial \pi}{\partial \eta} & \qquad \text{(probability via the expit function: } \pi = \text{expit}\{\eta\}) \\
= \pi(1-\pi) & \qquad \text{(derivative of probability w.r.t. log-odds)} \\
= \text{expit}\{\eta\}(1-\text{expit}\{\eta\}) & \qquad \text{(substitute } \pi = \text{expit}\{\eta\})
\end{aligned}
$$

□

[a] binary-outcome-associations.qmd#thm-expit-prob-logodds

Exm

Example 3.1 (Slope of expit for the beetles model)**.**

```
beetles_mid_row <- ceiling(nrow(beetles) / 2)
beetles_dose_mid <- beetles$dose[beetles_mid_row]
beetles_eta_mid <-
  predict(beetles_glm_grouped, type = "link")[beetles_mid_row] |>
  unname()
beetles_pi_mid <-
  predict(beetles_glm_grouped, type = "response")[beetles_mid_row] |>
  unname()
beetles_slope_mid <- beetles_pi_mid * (1 - beetles_pi_mid)
```

Consider the grouped-data logistic regression model for the beetles data (Table 2). At the middle (4th of 8) dose in the dataset, $x = 1.7842$, the fitted log-odds of death is $\hat{\eta} = 0.4277$, and the fitted probability of death is $\hat{\pi} = \text{expit}\{0.4277\} = 0.6053$. By Corollary 3.4, the slope of the expit function at this fitted log-odds value is:

$$\text{expit}'\{0.4277\} = 0.6053 \times (1 - 0.6053) = 0.2389$$

Specifically, near dose $x = 1.7842$, a small increase of $\Delta\eta$ in the fitted log-odds corresponds to an increase of approximately $0.2389 \times \Delta\eta$ in the fitted probability of death.

3.0.5 Summary matrix of derivatives

Each entry gives the derivative of the **column quantity** with respect to the **row quantity**; that is, the entry in row i and column j is $\frac{\partial(\text{col } j)}{\partial(\text{row } i)}$.

Dimensions of each entry depend on the types of the row and column quantities:

- **Scalar/scalar** (upper-left 3×3 block): row and column are both scalars (π, ω, or η); derivative is a scalar.
- **Gradient** (lower-left 2×3 block): row is a p-vector ($\tilde{x}$ or $\tilde{\beta}$), column is a scalar; derivative is a p-vector (gradient).
- **Jacobian** (lower-right 2×2 block): both row and column are p-vectors; derivative is a $p \times p$ matrix.
- **Undefined** (upper-right 3×2 block): row is a scalar (π, ω, or η), column is $\tilde{x}$ or $\tilde{\beta}$. Each of these entries would be the derivative of a p-vector with respect to a scalar, which would require expressing $\tilde{x}$ (or $\tilde{\beta}$) as a function of that scalar. No such function exists, because the map from the vector to the scalar is **not invertible**: for $p > 1$, many distinct vectors produce the same value of $\eta = \tilde{x} \cdot \tilde{\beta}$ (and hence the same ω and π), so the scalar does not determine the vector, and these derivatives are undefined.

Table 6: Matrix of logistic regression derivatives (p = number of predictors, $\eta \overset{\text{def}}{=} \tilde{x} \cdot \tilde{\beta}$, $\omega \overset{\text{def}}{=} \exp\{\eta\}$, $\pi \overset{\text{def}}{=} \omega/(1+\omega)$). Each entry is the derivative of the **column quantity** with respect to the **row quantity**. Entries marked "undef" are not defined because the vector-to-scalar map is not invertible: the scalar row quantity does not determine the vector column quantity.

	π	ω	η	$\tilde{x}$	$\tilde{\beta}$
π	1	$(1+\omega)^2$	$\frac{(1+\omega)^2}{\omega}$	undef	undef
ω	$(1-\pi)^2$	1	$\frac{1}{\omega}$	undef	undef
η	$\pi(1-\pi)$	ω	1	undef	undef
$\tilde{x}$	$\underbrace{\tilde{\beta}\pi(1-\pi)}_{p\times 1}$	$\underbrace{\tilde{\beta}\omega}_{p\times 1}$	$\underbrace{\tilde{\beta}}_{p\times 1}$	$\underbrace{\mathbb{I}}_{p\times p}$	$\mathbf{0}_{p\times p}$
$\tilde{\beta}$	$\underbrace{\tilde{x}\pi(1-\pi)}_{p\times 1}$	$\underbrace{\tilde{x}\omega}_{p\times 1}$	$\underbrace{\tilde{x}}_{p\times 1}$	$\mathbf{0}_{p\times p}$	$\underbrace{\mathbb{I}}_{p\times p}$

4 Understanding logistic regression models

Lemma 4.1 (Derivative of log-odds w.r.t. predictor)**.** *By the derivative of a linear combination*[a]*:*

$$\begin{aligned}\frac{\partial \eta}{\partial \tilde{x}} &= \frac{\partial}{\partial \tilde{x}} \tilde{x} \cdot \tilde{\beta} \\ &= \tilde{\beta}\end{aligned}$$

[a] math-prereqs.qmd#thm-deriv-lincom

Exercise 4.1. Consider a logistic regression model with a single predictor, X:

$$\begin{aligned} Y_i | X_i &\sim_{\perp\!\!\!\perp} \text{Ber}(\pi(X_i)) \\ \pi(x) &= \text{expit}\{\eta(x)\} = \pi(\omega(\eta(x))) \\ \eta(x) &= \alpha + \beta x \end{aligned} \tag{5}$$

Find the derivative of $\pi(x) = \text{E}[Y|X = x]$ with respect to x:

$$\frac{\partial \pi}{\partial x} = ?$$

Solution

Solution 4.1. By Theorem 3.6, Lemma 4.1, and the chain rule[a]:

$$\begin{aligned}\frac{\partial \pi}{\partial x} &= \frac{\partial \pi}{\partial \eta}\frac{\partial \eta}{\partial x} \\ &= \pi(1-\pi)\beta \\ &= \text{Var}(Y|X=x) \cdot \beta\end{aligned}$$

The slope is steepest at $\pi = 0.5$, i.e., at $\eta = 0$, which for a unipredictor model occurs at $x = -\alpha/\beta$. The slope goes to 0 as x goes to $-\infty$ or $+\infty$ (compare with the graph of the expit function[b]).

[a] math-prereqs.qmd#thm-chain-rule
[b] binary-outcome-associations.qmd#fig-expit-plot

Note

See the general procedure for interpreting a regression coefficient[a] for full details, including how to handle interaction terms and how to set remaining covariates to their reference values. **In order to interpret** β_j: differentiate or difference $\eta(\tilde{x})$ with respect to x_j (depending on whether x_j is continuous or discrete, respectively):

$$\frac{\partial}{\partial x_j}\eta(\tilde{x})$$

In order to find the MLE $\hat{\tilde{\beta}}$: differentiate the log-likelihood function $\ell(\tilde{\beta})$ with respect to $\tilde{\beta}$:

$$\frac{\partial}{\partial \tilde{\beta}}\ell(\tilde{\beta})$$

[a] Linear-models-overview.qmd#def-coef-interp-procedure

Exercise 4.2 (General formula for odds ratios in logistic regression)**.** Consider the generic logistic regression model:

- $Y_i|\tilde{X}_i \sim_{\perp\!\!\!\perp} \text{Ber}(\pi(\tilde{X}_i))$
- $\text{logit}\{\pi(\tilde{x})\} = \eta(\tilde{x})$
- $\eta(\tilde{x}) = \tilde{x}'\tilde{\beta}$

Let $\tilde{x}$ and $\tilde{x}^*$ be two covariate patterns, representing two individuals or two subpopulations. Find a concise formula to compute the odds ratio comparing covariate patterns $\tilde{x}$ and $\tilde{x}^*$:

$$\theta_\omega(\tilde{x}, \tilde{x}^*) \stackrel{\text{def}}{=} \frac{\omega(\tilde{x})}{\omega(\tilde{x}^*)} \tag{6}$$

Solution (General formula for odds ratios in logistic regression)

Solution 4.2 (General formula for odds ratios in logistic regression).

$$\begin{aligned}\theta_\omega(\tilde{x}, \tilde{x}^*) &\stackrel{\text{def}}{=} \frac{\omega(\tilde{x})}{\omega(\tilde{x}^*)} \\ &= \frac{\exp\{\eta(\tilde{x})\}}{\exp\{\eta(\tilde{x}^*)\}} \\ &= \exp\{\eta(\tilde{x}) - \eta(\tilde{x}^*)\}\end{aligned}$$

Solution 4.2 is more concrete than Equation 6, but it doesn't yet completely explain how to compute $\theta_\omega(\tilde{x}, \tilde{x}^*)$, so let's mark it as a lemma:

Lemma 4.2 (General formula for odds ratios in logistic regression)**.**

$$\theta_\omega(\tilde{x}, \tilde{x}^*) = \exp\{\eta(\tilde{x}) - \eta(\tilde{x}^*)\} \tag{7}$$

Proof

Proof. By Solution 4.2. □

Definition 4.1 (Difference in log-odds)**.**
Let $\tilde{x}$ and $\tilde{x}^*$ be two covariate patterns, representing two individuals or two subpopulations. Then we can define the difference in log-odds between $\tilde{x}$ and $\tilde{x}^*$, denoted $\Delta\eta(\tilde{x}, \tilde{x}^*)$ or $\Delta\eta$ for

short, as:

$$\Delta\eta \stackrel{\text{def}}{=} \eta(\tilde{x}) - \eta(\tilde{x}^*)$$

Corollary 4.1 (Shorthand general formula for odds ratios in logistic regression)**.**

$$\theta_\omega(\tilde{x}, \tilde{x}^*) = \exp\{\Delta\eta\} \tag{8}$$

Proof

Proof. By Lemma 4.2 and Definition 4.1. □

Exercise 4.3 (Difference in log-odds)**.** Find a concise expression for the difference in log-odds:

$$\Delta\eta \stackrel{\text{def}}{=} \eta(\tilde{x}) - \eta(\tilde{x}^*)$$

Solution (Difference in log-odds)

Solution 4.3 (Difference in log-odds).

$$\begin{aligned}
\Delta\eta &\stackrel{\text{def}}{=} \eta(\tilde{x}) - \eta(\tilde{x}^*) \\
&= (\tilde{x} \cdot \tilde{\beta}) - (\tilde{x}^* \cdot \tilde{\beta}) \\
&= (\tilde{x}^\top \tilde{\beta}) - ((\tilde{x}^*)^\top \tilde{\beta}) \\
&= (\tilde{x}^\top - (\tilde{x}^*)^\top)\tilde{\beta} \\
&= (\tilde{x} - \tilde{x}^*)^\top \tilde{\beta} \\
&= (\tilde{x} - \tilde{x}^*) \cdot \tilde{\beta}
\end{aligned}$$

Lemma 4.3 (Difference in log-odds)**.**

$$\Delta\eta = (\tilde{x} - \tilde{x}^*) \cdot \tilde{\beta}$$

Proof

Proof. By Solution 4.3. □

Definition 4.2 (Difference in covariate patterns)**.**
Let $\tilde{x}$ and $\tilde{x}^*$ be two covariate patterns, representing two individuals or two subpopulations. The difference in covariate patterns, denoted $\Delta\tilde{x}$, is defined as:

$$\Delta\tilde{x} \stackrel{\text{def}}{=} \tilde{x} - \tilde{x}^*$$

Corollary 4.2 (Difference in log-odds)**.**

$$\Delta\eta = (\Delta\tilde{x}) \cdot \tilde{\beta}$$

Proof

Proof. By Lemma 4.3 and Definition 4.2. □

Exercise 4.4. Find an expression for the odds ratio $\theta_\omega(\tilde{x}, \tilde{x}^*)$ in terms of $\Delta\tilde{x}$ and $\tilde{\beta}$.

Solution

Solution 4.4. Combine Corollary 4.1 and Corollary 4.2:

$$\begin{aligned}\theta_\omega(\tilde{x}, \tilde{x}^*) &= \exp\{\Delta\eta\} \\ &= \exp\Big\{\Delta\tilde{x} \cdot \tilde{\beta}\Big\}\end{aligned}$$

Theorem 4.1 (Odds ratio from difference in covariate patterns)**.** *The odds ratio comparing covariate patterns $\tilde{x}$ and $\tilde{x}^*$ is:*

$$\theta_\omega(\tilde{x}, \tilde{x}^*) = \exp\Big\{(\Delta\tilde{x}) \cdot \tilde{\beta}\Big\} \tag{9}$$

Proof

Proof. By Solution 4.4. □

Exm

Example 4.1 (Odds ratio between adjacent beetle dose levels)**.**
We can apply Theorem 4.1 to the `beetles` logistic regression model (Table 2).
The estimated dose coefficient for that model is $\hat{\beta}_1 = 34.270326$.
Consider the two covariate patterns given by the two lowest dose levels in the `beetles` data: $\tilde{x} = (1, 1.7242)^\top$ and $\tilde{x}^* = (1, 1.6907)^\top$ (intercept and dose). Their difference is $\Delta\tilde{x} = \tilde{x} - \tilde{x}^* = (0, 0.0335)^\top$.
By Theorem 4.1, the estimated odds ratio comparing these two dose levels is:

$$\begin{aligned}\hat{\theta_\omega}(\tilde{x}, \tilde{x}^*) &= \exp\Big\{(\Delta\tilde{x}) \cdot \hat{\beta}\Big\} && \text{(@thm-logistic-OR)} \\ &= \exp\Big\{0 \cdot \hat{\beta}_0 + (0.0335) \cdot \hat{\beta}_1\Big\} && \text{(expand the dot product)} \\ &= \exp\{(0.0335) \cdot (34.270326)\} && \text{(substitute } \hat{\beta}_1\text{)} \\ &= 3.152059 && \text{(evaluate the exponential)}\end{aligned}$$

So an increase of 0.0335 units in dose (the gap between the two lowest dose levels) multiplies the estimated odds of death by about 3.15.

Corollary 4.3 (Log odds ratio equals the difference in log-odds)**.**

$$\log\{\theta_\omega(\tilde{x}, \tilde{x}^*)\} = \Delta\eta$$

5 Estimating logistic regression models

5.0.1 Model

Assume:

- $Y_i|\tilde{X}_i \sim_{\perp\!\!\!\perp} \text{Ber}(\pi(X_i))$

- $\pi(\tilde{x}) = \text{expit}\{\eta(\tilde{x})\}$
- $\eta(\tilde{x}) = \tilde{x} \cdot \tilde{\beta}$

5.0.2 Likelihood function

Exercise 5.1. Compute and graph the likelihood for the `beetles` data model:

Table 7: Mortality rates of adult flour beetles after five hours' exposure to gaseous carbon disulphide (Bliss 1935)

```
library(glmx)
library(dplyr)
data(BeetleMortality)
beetles <- BeetleMortality |>
  mutate(
    pct = died / n,
    survived = n - died,
    dose_c = dose - mean(dose)
  )
beetles_long <-
  beetles |>
  reframe(
    .by = everything(),
    outcome = c(
      rep(1, times = died),
      rep(0, times = survived)
    )
  )
beetles
#> # A tibble: 8 x 6
#>    dose  died     n   pct survived   dose_c
#>   <dbl> <int> <int> <dbl>    <int>    <dbl>
#> 1  1.69     6    59 0.102       53 -0.103
#> 2  1.72    13    60 0.217       47 -0.0692
#> 3  1.76    18    62 0.290       44 -0.0382
#> 4  1.78    28    56 0.5         28 -0.00923
#> 5  1.81    52    63 0.825       11  0.0179
#> 6  1.84    53    59 0.898        6  0.0435
#> 7  1.86    61    62 0.984        1  0.0676
#> 8  1.88    60    60 1            0  0.0905
```

```
beetles_glm <-
  beetles |>
  glm(
    formula = cbind(died, survived) ~ dose,
    family = "binomial"
  )
equatiomatic::extract_eq(beetles_glm)
```

$$\log\left[\frac{P(\text{died} = 60)}{1 - P(\text{died} = 60)}\right] = \alpha + \beta_1(\text{dose}) \tag{10}$$

```
beetles_glm |>
  parameters::parameters() |>
  parameters::print_md()
```

Table 8: Fitted logistic regression model for `beetles` data

Parameter	Log-Odds	SE	95% CI	z	p
(Intercept)	-60.72	5.18	(-71.44, -51.08)	-11.72	$< .001$
dose	34.27	2.91	(28.85, 40.30)	11.77	$< .001$

Solution

Solution 5.1.

```
odds_inv <- function(omega) (1 + omega^-1)^-1
lik_beetles0 <- function(beta_0, beta_1) {
  beetles |>
    mutate(
      eta = beta_0 + beta_1 * dose,
      omega = exp(eta),
      pi = odds_inv(omega),
      Lik = pi^died * (1 - pi)^survived,
      # llik = died*eta + n*log(1 - pi)
    ) |>
    pull(Lik) |>
    prod()
}

lik_beetles <- Vectorize(lik_beetles0)
```

5.0.3 Log-likelihood function

Exercise 5.2. Find the log-likelihood function for the general logistic regression model.

Solution

Solution 5.2.

$$
\begin{aligned}
\ell(\tilde{\beta}, \tilde{y}) &\stackrel{\text{def}}{=} \log\Big\{\mathcal{L}(\tilde{\beta}, \tilde{y})\Big\} && \text{(definition of log-likelihood)} \\
&= \log\left\{\prod_{i=1}^{n} \mathcal{L}_i(\tilde{\beta}, y_i)\right\} && \text{(likelihood factors, by independence)} \\
&= \sum_{i=1}^{n} \log\Big\{\mathcal{L}_i(\tilde{\beta}, y_i)\Big\} && \text{(log of a product is the sum of the logs)} \\
&= \sum_{i=1}^{n} \ell_i(\pi(\tilde{x}_i)) && \text{(definition: } \ell_i \stackrel{\text{def}}{=} \log\{\mathcal{L}_i\}\text{)}
\end{aligned}
\tag{11}
$$

Using log-odds as a function of probability[a] and one plus odds in terms of non-event probability[b]:

$$
\begin{aligned}
\ell_i(\pi_i) &= y_i \log\{\pi_i\} + (1 - y_i) \log\{1 - \pi_i\} && \text{(log of the Bernoulli likelihood } \pi_i^{y_i}(1-\pi_i)^{1-y_i}\text{)} \\
&= y_i \log\{\pi_i\} + 1 \cdot \log\{1 - \pi_i\} - y_i \cdot \log\{1 - \pi_i\} && \text{(distribute } (1 - y_i)\text{)} \\
&= y_i \log\{\pi_i\} + \log\{1 - \pi_i\} - y_i \log\{1 - \pi_i\} && \text{(multiplicative identity: } 1 \cdot \log\{1 - \pi_i\} = \log\{1 - \pi_i\}\text{)} \\
&= y_i \log\{\pi_i\} - y_i \log\{1 - \pi_i\} + \log\{1 - \pi_i\} && \text{(commute the addition)} \\
&= y_i (\log\{\pi_i\} - \log\{1 - \pi_i\}) + \log\{1 - \pi_i\} && \text{(factor out } y_i\text{)} \\
&= y_i \log\left\{ \frac{\pi_i}{1 - \pi_i} \right\} + \log\{1 - \pi_i\} && \text{(difference of logs is the log of the quotient)} \\
&= y_i \operatorname{logit}(\pi_i) + \log\{1 - \pi_i\} && \text{(log-odds identity: } \operatorname{logit}(\pi) = \log\{\pi/(1-\pi)\}\text{)} \\
&= y_i \eta_i + \log\{1 - \pi_i\} && \text{(definition: } \eta_i \stackrel{\text{def}}{=} \operatorname{logit}(\pi_i)\text{)} \\
&= y_i \eta_i + \log\left\{ (1 + \omega_i)^{-1} \right\} && \text{(non-event probability: } 1 - \pi_i = (1 + \omega_i)^{-1}\text{)} \\
&= y_i \eta_i - \log\{1 + \omega_i\} && \text{(log of a reciprocal)}
\end{aligned}
$$

[a] binary-outcome-associations.qmd#thm-logodds-pi
[b] binary-outcome-associations.qmd#cor-inverse-odds-nonevent

Lemma 5.1 (Per-observation log-likelihood component)**.**

$$\ell_i(\pi_i) = y_i \eta_i - \log\{1 + \omega_i\}$$

Exercise 5.3. Compute and graph the log-likelihood for the `beetles` data.

Solution

Solution 5.3.

```
odds_inv <- function(omega) (1 + omega^-1)^-1
llik_beetles0 <- function(beta_0, beta_1) {
  beetles |>
    mutate(
      eta = beta_0 + beta_1 * dose,
      omega = exp(eta),
      pi = odds_inv(omega), # need for next line:
      llik = died*eta + n*log(1 - pi)
    ) |>
    pull(llik) |>
    sum()
}

llik_beetles <- Vectorize(llik_beetles0)

# to check that we implemented it correctly:
# ests = coef(beetles_glm_ungrouped)
# logLik(beetles_glm_ungrouped)
# llik_beetles(ests[1], ests[2])
```

Let's center dose:

```
beetles_glm_grouped_centered <- beetles |>
  glm(
    formula = cbind(died, survived) ~ dose_c,
    family = "binomial"
  )
```

```r
beetles_glm_ungrouped_centered <- beetles_long |>
  mutate(died = outcome) |>
  glm(
    formula = died ~ dose_c,
    family = "binomial"
  )

equatiomatic::extract_eq(beetles_glm_ungrouped_centered)
```

$$\log\left[\frac{P(\text{died} = 1)}{1 - P(\text{died} = 1)}\right] = \alpha + \beta_1(\text{dose\_c}) \tag{12}$$

```r
beetles_glm_grouped_centered |>
  parameters::parameters() |>
  parameters::print_md()
```

Table 9: Fitted logistic regression model for `beetles` data, with `dose` centered

Parameter	Log-Odds	SE	95% CI	z	p
(Intercept)	0.74	0.14	(0.48, 1.02)	5.40	$< .001$
dose c	34.27	2.91	(28.85, 40.30)	11.77	$< .001$

```r
odds_inv <- function(omega) (1 + omega^-1)^-1
lik_beetles0 <- function(beta_0, beta_1) {
  beetles |>
    mutate(
      eta = beta_0 + beta_1 * dose_c,
      omega = exp(eta),
      pi = odds_inv(omega),
      Lik = (pi^died) * (1 - pi)^(survived)
    ) |>
    pull(Lik) |>
    prod()
}

lik_beetles <- Vectorize(lik_beetles0)
```

```r
odds_inv <- function(omega) (1 + omega^-1)^-1
llik_beetles0 <- function(beta_0, beta_1) {
  beetles |>
    mutate(
      eta = beta_0 + beta_1 * dose_c,
      omega = exp(eta),
      pi = odds_inv(omega),
      llik = died * eta + n*log(1 - pi)
    ) |>
    pull(llik) |>
    sum()
}

llik_beetles <- Vectorize(llik_beetles0)
```

5.0.4 Score function

As usual, by independence, we have:

Lemma 5.2 (Score function decomposes over observations)**.**

$$
\begin{aligned}
\tilde{\ell}'(\tilde{\beta}) &\overset{def}{=} \frac{\partial}{\partial \tilde{\beta}} \ell(\tilde{\beta}) && \textit{(definition of the score function)} \\
&= \frac{\partial}{\partial \tilde{\beta}} \sum_{i=1}^{n} \ell_i(\tilde{\beta}) && \textit{(the log-likelihood decomposes over observations, by independence)} \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \tilde{\beta}} \ell_i(\tilde{\beta}) && \textit{(linearity of differentiation)} \\
&= \sum_{i=1}^{n} \tilde{\ell}_i'(\tilde{\beta}) && \textit{(definition: } \tilde{\ell}_i' \overset{def}{=} \frac{\partial}{\partial \tilde{\beta}} \ell_i \textit{)}
\end{aligned}
$$

Starting from Lemma 5.1, we can apply the vector chain rule[7]:

Lemma 5.3 (Chain rule applied to the score component)**.**

$$
\begin{aligned}
\tilde{\ell}_i'(\tilde{\beta}) &\overset{def}{=} \frac{\partial}{\partial \tilde{\beta}} \ell_i(\tilde{\beta}) && \textit{(definition of the score component)} \\
&= \frac{\partial}{\partial \tilde{\beta}} (y_i \eta_i - \log\{1 + \omega_i\}) && \textit{(per-observation log-likelihood)} \\
&= \frac{\partial}{\partial \tilde{\beta}} (y_i \eta_i) - \frac{\partial}{\partial \tilde{\beta}} \log\{1 + \omega_i\} && \textit{(linearity of differentiation)} \\
&= \frac{\partial \eta_i}{\partial \tilde{\beta}} y_i - \frac{\partial}{\partial \tilde{\beta}} \log\{1 + \omega_i\} && \textit{(constant-factor rule: } y_i \textit{ does not depend on } \tilde{\beta}\textit{)} \\
&= \frac{\partial \eta_i}{\partial \tilde{\beta}} y_i - \frac{\partial u_i}{\partial \tilde{\beta}} \left(\frac{\partial}{\partial u_i} \log\{u_i\} \right) && \textit{(vector chain rule, with } u_i \overset{def}{=} 1 + \omega_i\textit{)} \\
&= \frac{\partial \eta_i}{\partial \tilde{\beta}} y_i - \frac{\partial u_i}{\partial \tilde{\beta}} \frac{1}{u_i} && \textit{(derivative of the natural log: } \frac{\partial}{\partial u} \log\{u\} = 1/u\textit{)} \\
&= \frac{\partial \eta_i}{\partial \tilde{\beta}} y_i - \frac{\partial \omega_i}{\partial \tilde{\beta}} \frac{1}{u_i} && \textit{(} \frac{\partial u_i}{\partial \tilde{\beta}} = \frac{\partial \omega_i}{\partial \tilde{\beta}} \textit{: the constant 1 has derivative } \mathbf{0}_{(p+1)\times 1}\textit{)} \\
&= \frac{\partial \eta_i}{\partial \tilde{\beta}} y_i - \frac{\partial \omega_i}{\partial \tilde{\beta}} \frac{1}{1 + \omega_i} && \textit{(substitute back } u_i = 1 + \omega_i\textit{)}
\end{aligned}
$$

Lemma 5.4 (Derivative of log-odds with respect to coefficients)**.** *By the derivative of a linear combination*[a]*:*

$$
\begin{aligned}
\frac{\partial \eta}{\partial \tilde{\beta}} &= \frac{\partial}{\partial \tilde{\beta}} (\tilde{x} \cdot \tilde{\beta}) \\
&= \tilde{x}
\end{aligned}
\tag{13}
$$

[a] math-prereqs.qmd#thm-deriv-lincom

Lemma 5.4 is very similar to Lemma 4.1, but not quite the same; Lemma 4.1 differentiates by $\tilde{x}$, whereas Lemma 5.4 differentiates by $\tilde{\beta}$.

Theorem 5.1 (Gradient of odds w.r.t. coefficients)**.**
To derive $\frac{\partial \omega}{\partial \tilde{\beta}}$*, we can apply the vector chain rule*[a] *again along with Lemma 3.2 and Lemma 5.4:*

[7] math-prereqs.qmd#thm-chain-vec

$$
\begin{aligned}
\frac{\partial \omega}{\partial \tilde{\beta}} &= \frac{\partial \eta}{\partial \tilde{\beta}} \frac{\partial \omega}{\partial \eta} && \textit{(vector chain rule, through } \eta = \eta\textit{)} \\
&= \tilde{x} \frac{\partial \omega}{\partial \eta} && \textit{(gradient of log-odds w.r.t. coefficients: } \frac{\partial \eta}{\partial \tilde{\beta}} = \tilde{x}\textit{)} \\
&= \tilde{x} \omega && \textit{(derivative of odds w.r.t. log-odds: } \frac{\partial \omega}{\partial \eta} = \omega\textit{)}
\end{aligned}
$$

[a] math-prereqs.qmd#thm-chain-vec

Corollary 5.1 (Gradient of odds w.r.t. coefficients in terms of probability)**.**

$$
\begin{aligned}
\frac{\partial \omega}{\partial \tilde{\beta}} &= \tilde{x} \omega && \textit{(gradient of odds w.r.t. coefficients)} \\
&= \tilde{x} \frac{\pi}{1 - \pi} && \textit{(odds in terms of probability: } \omega = \frac{\pi}{1 - \pi}\textit{)}
\end{aligned}
$$

Now we can combine Lemma 5.3, Lemma 5.4, and Theorem 5.1:

$$
\begin{aligned}
\ell_i'(\tilde{\beta}) &= \frac{\partial \eta_i}{\partial \tilde{\beta}} y_i - \frac{\partial \omega_i}{\partial \tilde{\beta}} \frac{1}{1 + \omega_i} && \text{(score component lemma)} \\
&= \tilde{x}_i y_i - \frac{\partial \omega_i}{\partial \tilde{\beta}} \frac{1}{1 + \omega_i} && \text{(gradient of log-odds w.r.t. coefficients: } \frac{\partial \eta_i}{\partial \tilde{\beta}} = \tilde{x}_i\text{)} \\
&= \tilde{x}_i y_i - \tilde{x}_i \omega_i \frac{1}{1 + \omega_i} && \text{(gradient of odds w.r.t. coefficients: } \frac{\partial \omega_i}{\partial \tilde{\beta}} = \tilde{x}_i \omega_i\text{)} \\
&= \tilde{x}_i y_i - \tilde{x}_i \frac{\omega_i}{1 + \omega_i} && \text{(combine the scalar factors into one fraction)} \\
&= \tilde{x}_i y_i - \tilde{x}_i \pi_i && \text{(inverse-odds identity: } \frac{\omega_i}{1 + \omega_i} = \pi_i\text{)} \\
&= \tilde{x}_i (y_i - \pi_i) && \text{(factor out } \tilde{x}_i\text{)} \\
&= \tilde{x}_i (y_i - \mu_i) && \text{(Bernoulli conditional mean: } \pi_i = \mu_i\text{)} \\
&= \tilde{x}_i (y_i - \mathrm{E}[Y_i | \tilde{X}_i = \tilde{x}_i]) && \text{(definition: } \mu_i \stackrel{\text{def}}{=} \mathrm{E}[Y_i | \tilde{X}_i = \tilde{x}_i]\text{)} \\
&= \tilde{x}_i \, \varepsilon(y_i | \tilde{X}_i = \tilde{x}_i) && \text{(definition of the error: } \varepsilon(y | \tilde{X} = \tilde{x}) \stackrel{\text{def}}{=} y - \mathrm{E}[Y | \tilde{X} = \tilde{x}]\text{)} \\
&= \tilde{x}_i \varepsilon_i && \text{(abbreviation: } \varepsilon_i \stackrel{\text{def}}{=} \varepsilon(y_i | \tilde{X}_i = \tilde{x}_i)\text{)}
\end{aligned}
$$

Theorem 5.2 (Score component for one observation)**.**

$$
\ell_i'(\tilde{\beta}) = \tilde{x}_i \varepsilon_i \tag{14}
$$

This last expression is essentially the same as we found in linear regression[8].

Finally, combining Lemma 5.2 and Theorem 5.2, we have:

[8] Linear-models-overview.qmd#eq-scorefun-linreg

Theorem 5.3 (Logistic-model score function)**.**

$$\begin{aligned}
\tilde{\ell}'(\tilde{\beta}) &= \sum_{i=1}^{n} \ell_i'(\tilde{\beta}) && \textit{(the score decomposes over observations)} \\
&= \sum_{i=1}^{n} \underbrace{\tilde{x}_i}_{(p+1)\times 1} \underbrace{\varepsilon_i}_{1\times 1} && \textit{(score component for one observation)} \\
&= \underbrace{\mathbf{X}^\top}_{(p+1)\times n} \underbrace{\tilde{\varepsilon}}_{n\times 1} && \textit{(matrix form: } \sum_{i=1}^{n} \tilde{x}_i \varepsilon_i = \mathbf{X}^\top \tilde{\varepsilon}\textit{)}
\end{aligned} \tag{15}$$

The score function is vector-valued; its components are:

$$\frac{\partial \ell}{\partial \tilde{\beta}} = \begin{pmatrix} \frac{\partial \ell}{\partial \beta_0} \\ \frac{\partial \ell}{\partial \beta_1} \\ \vdots \\ \frac{\partial \ell}{\partial \beta_p} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} 1\varepsilon_i \\ \sum_{i=1}^{n} x_{i,1}\varepsilon_i \\ \vdots \\ \sum_{i=1}^{n} x_{i,p}\varepsilon_i \end{pmatrix} = \begin{pmatrix} \tilde{1} \cdot \tilde{\varepsilon} \\ \mathbf{X}_{\cdot 1} \cdot \tilde{\varepsilon} \\ \vdots \\ \mathbf{X}_{\cdot p} \cdot \tilde{\varepsilon} \end{pmatrix}$$

Here, $\mathbf{X}_{\cdot j} \stackrel{\text{def}}{=} (x_{1j}, \ldots, x_{nj})$ denotes the j^{th} column of the design matrix $\mathbf{X}$ (the n-vector of covariate j's values across all observations), in contrast with $\tilde{x}_i$, which denotes observation i's covariate row vector (a p-vector).

Thus, the score equation $\tilde{\ell}' = \mathbf{0}_{(p+1)\times 1}$ means that for the MLE $\hat{\tilde{\beta}}$:

1. the sum of the errors (aka deviations) equals 0:

$$\sum_{i=1}^{n} \varepsilon_i = 0$$

2. the sums of the errors times each covariate also equal 0:

$$\mathbf{X}_{\cdot j} \cdot \tilde{\varepsilon} = \sum_{i=1}^{n} x_{ij}\varepsilon_i = 0, \forall j \in \{1 : p\}$$

Exm

Example 5.1. In our model for the `beetles` data, we only have an intercept plus one covariate, gas concentration (c):

$$\tilde{x} = (1, c)$$

If c_i is the gas concentration for the beetle in observation i, and $\tilde{c} = (c_1, c_2, \ldots c_n)$, then the score equation $\tilde{\ell}' = \mathbf{0}_{2\times 1}$ means that for the MLE $\hat{\tilde{\beta}}$:

1. the sum of the errors (aka deviations) equals 0:

$$\sum_{i=1}^{n} \varepsilon_i = 0$$

2. the weighted sum of the error times the gas concentrations equals 0:

$$\sum_{i=1}^{n} c_i\varepsilon_i = 0$$

Exercise 5.4. Implement and graph the score function for the beetles data

Solution

Solution 5.4.

```
odds_inv <- function(omega) (1 + omega^-1)^-1

score_fn_beetles_beta0_0 <- function(beta_0, beta_1) {
  beetles |>
    mutate(
      eta = beta_0 + beta_1 * dose_c,
      omega = exp(eta),
      pi = odds_inv(omega),
      mu = pi * n,
      epsilon = died - mu,
      score = epsilon
    ) |>
    pull(score) |>
    sum()
}
score_fn_beetles_beta_0 <- Vectorize(score_fn_beetles_beta0_0)

score_fn_beetles_beta1_0 <- function(beta_0, beta_1) {
  beetles |>
    mutate(
      eta = beta_0 + beta_1 * dose_c,
      omega = exp(eta),
      pi = odds_inv(omega),
      mu = pi * n,
      epsilon = died - mu,
      score = dose_c * epsilon
    ) |>
    pull(score) |>
    sum()
}
score_fn_beetles_beta_1 <- Vectorize(score_fn_beetles_beta1_0)
```

5.0.5 Hessian function

$$\underbrace{\ell''(\tilde{\beta})}_{(p+1)\times(p+1)} = \sum_{i=1}^{n} \underbrace{\ell_i''(\tilde{\beta})}_{(p+1)\times(p+1)} \tag{16}$$

$$\begin{aligned}
\ell_i''(\tilde{\beta}) &\stackrel{\text{def}}{=} \frac{\partial}{\partial \tilde{\beta}^\top} \tilde{\ell}_i' && \text{(definition of the Hessian component)} \\
&= \frac{\partial}{\partial \tilde{\beta}^\top} (\tilde{x}_i \varepsilon_i) && \text{(score component for one observation)} \\
&= \tilde{x}_i \frac{\partial}{\partial \tilde{\beta}^\top} \varepsilon_i && \text{(constant-factor rule: } \tilde{x}_i \text{ does not depend on } \tilde{\beta}) \\
&= \tilde{x}_i \varepsilon_i' && \text{(definition: } \varepsilon_i' \stackrel{\text{def}}{=} \frac{\partial}{\partial \tilde{\beta}^\top} \varepsilon_i)
\end{aligned} \tag{17}$$

Theorem 5.4 (Gradient of fitted probability w.r.t. coefficients)**.** *Using Lemma 5.4 and Theorem 3.6:*

$$
\begin{aligned}
\frac{\partial \pi}{\partial \tilde{\beta}} &= \frac{\partial \eta}{\partial \tilde{\beta}} \frac{\partial \pi}{\partial \eta} && \textit{(vector chain rule, through } \eta\textit{)} \\
&= \tilde{x} \frac{\partial \pi}{\partial \eta} && \textit{(gradient of log-odds w.r.t. coefficients: } \frac{\partial \eta}{\partial \tilde{\beta}} = \tilde{x}\textit{)} \\
&= \tilde{x} \pi (1 - \pi) && \textit{(derivative of probability w.r.t. log-odds: } \frac{\partial \pi}{\partial \eta} = \pi(1-\pi)\textit{)}
\end{aligned}
$$

Using Theorem 5.4:

$$
\begin{aligned}
\varepsilon_i' &\stackrel{\text{def}}{=} \frac{\partial \varepsilon_i}{\partial \tilde{\beta}^\top} && \text{(definition of } \varepsilon_i'\text{)} \\
&= \frac{\partial}{\partial \tilde{\beta}^\top}(y_i - \mu_i) && \text{(definition of the error: } \varepsilon_i \stackrel{\text{def}}{=} y_i - \mu_i\text{)} \\
&= \frac{\partial}{\partial \tilde{\beta}^\top} y_i - \frac{\partial}{\partial \tilde{\beta}^\top} \mu_i && \text{(linearity of differentiation)} \\
&= \mathbf{0}_{1 \times (p+1)} - \frac{\partial}{\partial \tilde{\beta}^\top} \mu_i && (y_i \text{ does not depend on } \tilde{\beta}) \\
&= -\frac{\partial \mu_i}{\partial \tilde{\beta}^\top} && \text{(additive identity)} \\
&= -\frac{\partial \pi_i}{\partial \tilde{\beta}^\top} && \text{(Bernoulli conditional mean: } \mu_i = \pi_i\text{)} \\
&= -\pi_i (1 - \pi_i) \tilde{x}_i^\top && \text{(transpose of the gradient of } \pi \text{ w.r.t. coefficients)} \\
&= -\operatorname{Var}\left(Y_i | \tilde{X}_i = \tilde{x}_i\right) \tilde{x}_i^\top && \text{(Bernoulli conditional variance: } \operatorname{Var}\left(Y_i | \tilde{X}_i = \tilde{x}_i\right) = \pi_i(1-\pi_i)\text{)}
\end{aligned}
$$

Returning to Equation 17:

$$
\begin{aligned}
\ell_i''(\tilde{\beta}) &= \tilde{x}_i \varepsilon_i' && \text{(Hessian component identity)} \\
&= \tilde{x}_i \left(-\operatorname{Var}\left(Y_i | \tilde{X}_i = \tilde{x}_i\right) \tilde{x}_i^\top\right) && \text{(substitute } \varepsilon_i'\text{)} \\
&= - \underbrace{\tilde{x}_i}_{(p+1)\times 1} \underbrace{\operatorname{Var}\left(Y_i | \tilde{X}_i = \tilde{x}_i\right)}_{1 \times 1} \underbrace{\tilde{x}_i^\top}_{1 \times (p+1)} && \text{(move the scalar factor } -\operatorname{Var}\left(Y_i | \tilde{X}_i = \tilde{x}_i\right) \text{ between the vectors)}
\end{aligned}
\tag{18}
$$

Returning to Equation 16:

$$
\begin{aligned}
\ell''(\tilde{\beta}) &= \sum_{i=1}^{n} \ell_i''(\tilde{\beta}) && \text{(the Hessian decomposes over observations)} \\
&= -\sum_{i=1}^{n} \underbrace{\tilde{x}_i}_{(p+1)\times 1} \underbrace{\operatorname{Var}\left(Y_i | \tilde{X}_i = \tilde{x}_i\right)}_{1 \times 1} \underbrace{\tilde{x}_i^\top}_{1 \times (p+1)} && \text{(Hessian component for one observation)} \\
&= - \underbrace{\mathbf{X}^\top}_{(p+1)\times n} \underbrace{\mathbf{D}}_{n \times n} \underbrace{\mathbf{X}}_{n \times (p+1)} && \text{(matrix form: } \mathbf{D} \stackrel{\text{def}}{=} \operatorname{diag}(\operatorname{Var}\left(Y_i | \tilde{X}_i = \tilde{x}_i\right))\text{)}
\end{aligned}
\tag{19}
$$

where $\mathbf{D} \stackrel{\text{def}}{=} \operatorname{diag}(\operatorname{Var}\left(Y_i | \tilde{X}_i = \tilde{x}_i\right))$ is the diagonal matrix whose i^{th} diagonal element is $\operatorname{Var}\left(Y_i | \tilde{X}_i = \tilde{x}_i\right)$.

Remark 5.1 (The logistic log-likelihood is concave). The Hessian $\ell''(\tilde{\beta}) = -\mathbf{X}^\top \mathbf{D} \mathbf{X}$ (Equation 19) is negative semi-definite for every $\tilde{\beta}$, because the diagonal entries of $\mathbf{D}$ are

$\mathrm{Var}\big(Y_i|\tilde{X}_i = \tilde{x}_i\big) = \pi_i(1-\pi_i) \geq 0$. Since the diagonal entries are non-negative, $\mathbf{D}^{1/2} \overset{\text{def}}{=} \mathrm{diag}\Big(\sqrt{\pi_i(1-\pi_i)}\Big)$ is well-defined, and for any vector $\tilde{v} \in \mathbb{R}^{p+1}$:

$$
\begin{aligned}
\underset{1\times(p+1)}{\tilde{v}^\top}\left(-\underset{(p+1)\times n}{\mathbf{X}^\top}\underset{n\times n}{\mathbf{D}}\underset{n\times(p+1)}{\mathbf{X}}\right)\underset{(p+1)\times 1}{\tilde{v}} &= -\tilde{v}^\top\mathbf{X}^\top\mathbf{D}\mathbf{X}\tilde{v} && \text{(the scalar } -1 \text{ factors out)}\\
&= -\tilde{v}^\top\mathbf{X}^\top\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{X}\tilde{v} && (\mathbf{D} = \mathbf{D}^{1/2}\mathbf{D}^{1/2})\\
&= -\big(\mathbf{D}^{1/2}\mathbf{X}\tilde{v}\big)^\top\big(\mathbf{D}^{1/2}\mathbf{X}\tilde{v}\big) && \text{(transpose of a product; } \mathbf{D}^{1/2} \text{ is diagonal, hence sy}\\
&= -\big\|\mathbf{D}^{1/2}\mathbf{X}\tilde{v}\big\|^2 && \text{(definition of the squared Euclidean norm)}\\
&\leq 0 && \text{(squared norms are non-negative)}
\end{aligned}
$$

A twice-differentiable function whose Hessian is negative semi-definite everywhere is concave[a] (Boyd and Vandenberghe 2004, sec. 3.1.4), so the log-likelihood $\ell(\tilde{\beta})$ is concave, and every stationary point (every solution of the score equation $\tilde{\ell}'(\tilde{\beta}) = \mathbf{0}_{(p+1)\times 1}$) is a global maximum[b] (Boyd and Vandenberghe 2004, sec. 4.2.2). In fact, since $\pi_i = \mathrm{expit}\big\{\tilde{x}_i \cdot \tilde{\beta}\big\} \in (0,1)$ strictly, every diagonal entry $\pi_i(1-\pi_i)$ is strictly positive; so when $\mathbf{X}$ has full column rank, $\mathbf{D}^{1/2}\mathbf{X}\tilde{v} = \mathbf{0}_{n\times 1}$ only for $\tilde{v} = \mathbf{0}_{(p+1)\times 1}$, the Hessian is negative definite, the log-likelihood is strictly concave, and the maximum is unique. Consequently, when the Newton-Raphson algorithm[c] converges, it converges to the unique global maximum likelihood estimate, not merely to a local optimum.

[a] https://en.wikipedia.org/wiki/Concave_function
[b] https://en.wikipedia.org/wiki/Concave_function#Properties
[c] intro-MLEs.qmd#sec-newton-raphson

Compare with the linear regression Hessian[9]:

$$
\begin{aligned}
\ell''(\tilde{\beta}) &= -\frac{1}{\sigma^2}\sum_{i=1}^{n}\underset{(p+1)\times 1}{\tilde{x}_i}\underset{1\times(p+1)}{\tilde{x}_i'} && \text{(linear-model Hessian)}\\
&= -\underset{(p+1)\times n}{\mathbf{X}^\top}\underset{n\times n}{\mathbf{D}^{-1}}\underset{n\times(p+1)}{\mathbf{X}} && \text{(matrix form: here } \mathbf{D} = \sigma^2\mathbb{I}_{n\times n}\text{, so } \mathbf{D}^{-1} = \frac{1}{\sigma^2}\mathbb{I}_{n\times n})
\end{aligned}
\tag{20}
$$

Exercise 5.5. Determine the elements of the Hessian matrix for logistic regression.

Solution

Solution 5.5. The components of the Hessian are:

[9] Linear-models-overview.qmd#eq-lm-hess

$$
\begin{aligned}
\underbrace{\ell''(\beta)}_{(p+1)\times(p+1)} &= \frac{\partial^2}{\partial\beta^\top \partial\beta}\ell && \text{(definition of the Hessian)} \\
&= \frac{\partial}{\partial\beta^\top}\underbrace{\ell'}_{(p+1)\times 1} && \text{(the Hessian is the derivative of the score)} \\
&= \begin{bmatrix} \underbrace{\frac{\partial \ell'}{\partial \beta_0}}_{(p+1)\times 1} & \underbrace{\frac{\partial \ell'}{\partial \beta_1}}_{(p+1)\times 1} & \cdots & \underbrace{\frac{\partial \ell'}{\partial \beta_p}}_{(p+1)\times 1} \end{bmatrix} && \text{(differentiate column-by-column w.r.t. } \beta^\top\text{)} \\
&= \underbrace{\begin{bmatrix} \frac{\partial^2 \ell}{\partial \beta_0^2} & \frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1} & \cdots & \frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_p} \\ \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 \ell}{\partial \beta_1^2} & \cdots & \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_p} \\ \vdots & \ddots & \ddots & \vdots \\ \frac{\partial^2 \ell}{\partial \beta_p \partial \beta_0} & \frac{\partial^2 \ell}{\partial \beta_p \partial \beta_1} & \cdots & \frac{\partial^2 \ell}{\partial \beta_p^2} \end{bmatrix}}_{(p+1)\times(p+1)} && \text{(each score entry is itself a partial derivative of } \ell\text{)}
\end{aligned}
$$

Exercise 5.6. Determine the Hessian for the `beetles` model.

Solution

Solution 5.6. In the `beetles` model, the Hessian is:

$$
\begin{aligned}
\underbrace{\ell''(\beta)}_{2\times 2} &= \begin{bmatrix} \underbrace{\frac{\partial \ell'}{\partial \beta_0}}_{2\times 1} & \underbrace{\frac{\partial \ell'}{\partial \beta_1}}_{2\times 1} \end{bmatrix} && \text{(differentiate column-by-column; here } p+1=2\text{)} \\
&= \begin{bmatrix} \frac{\partial^2 \ell}{\partial \beta_0^2} & \frac{\partial^2 \ell}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 \ell}{\partial \beta_1^2} \end{bmatrix} && \text{(each score entry is a partial derivative of } \ell\text{)} \\
&= -\sum_{i=1}^{n} \underbrace{\begin{pmatrix} 1 \\ c_i \end{pmatrix}}_{2\times 1} \underbrace{\pi_i(1-\pi_i)}_{1\times 1} \underbrace{\begin{pmatrix} 1 & c_i \end{pmatrix}}_{1\times 2} && \text{(logistic Hessian with } \tilde{x}_i = (1, c_i)^\top\text{)} \\
&= \begin{bmatrix} -\sum_{i=1}^n \pi_i(1-\pi_i) & -\sum_{i=1}^n c_i \pi_i(1-\pi_i) \\ -\sum_{i=1}^n c_i \pi_i(1-\pi_i) & -\sum_{i=1}^n c_i^2 \pi_i(1-\pi_i) \end{bmatrix} && \text{(multiply out the outer products, entry by entry)}
\end{aligned}
$$

Setting $\ell'(\tilde{\beta}; \tilde{y}) = 0$ gives us:

$$
\sum_{i=1}^{n} \left\{ \underbrace{\tilde{x}_i}_{(p+1)\times 1} \underbrace{\left(y_i - \operatorname{expit}\left\{\tilde{x}_i^\top \tilde{\beta}\right\}\right)}_{1\times 1} \right\} = \mathbf{0}_{(p+1)\times 1} \tag{21}
$$

In general, the estimating equation $\ell'(\tilde{\beta}; \tilde{y}) = 0$ cannot be solved analytically.

Instead, we can use the Newton-Raphson method[10]:

$$
\hat{\theta}^* \leftarrow \hat{\theta}^* - \left(\ell''\left(\tilde{y}; \hat{\theta}^*\right)\right)^{-1} \ell'\left(\tilde{y}; \hat{\theta}^*\right)
$$

We make an iterative series of guesses, and each guess helps us make the next guess better (i.e., higher log-likelihood). You can see some information about this process like so:

```
beetles_glm_ungrouped <-
  beetles_long |>
  glm(
```

[10] intro-MLEs.qmd#sec-newton-raphson

Table 10: Fitted model for `beetles` data

```
beetles_glm_ungrouped |> summary()
#>
#> Call:
#> glm(formula = outcome ~ dose, family = "binomial", data = beetles_long,
#>     control = glm.control(trace = TRUE))
#>
#> Coefficients:
#>             Estimate Std. Error z value Pr(>|z|)
#> (Intercept)   -60.72       5.18   -11.7   <2e-16 ***
#> dose           34.27       2.91    11.8   <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#>     Null deviance: 645.44  on 480  degrees of freedom
#> Residual deviance: 372.47  on 479  degrees of freedom
#> AIC: 376.5
#>
#> Number of Fisher Scoring iterations: 5
```

Table 11: Parameter estimate covariance matrix for `beetles` data

```
beetles_glm_ungrouped |> vcov()
#>             (Intercept)      dose
#> (Intercept)     26.8393 -15.08189
#> dose           -15.0819   8.48041
```

```
    control = glm.control(trace = TRUE),
    formula = outcome ~ dose,
    family = "binomial"
  )
#> Deviance = 383.249 Iterations - 1
#> Deviance = 372.921 Iterations - 2
#> Deviance = 372.472 Iterations - 3
#> Deviance = 372.471 Iterations - 4
#> Deviance = 372.471 Iterations - 5
```

After each iteration of the fitting procedure, the deviance ($2(\ell_{\text{full}} - \ell(\widehat{\beta}))$) is printed. You can see that the algorithm took 5 iterations to converge to a solution where the likelihood wasn't changing much anymore.

Table 10 and Table 11 show the fitted model and the covariance matrix of the estimates, respectively.

6 Inference for logistic regression models

6.1 Inference for individual predictor coefficients

6.1.1 Wald tests and confidence intervals

By the central limit theorem for MLEs[11], the maximum likelihood estimates $\hat{\beta}_k$ are approximately Gaussian for large sample sizes; see also the table of Gaussian vs. MLE-based tests[12].

[11] intro-MLEs.qmd#thm-dist-mle
[12] intro-MLEs.qmd#tbl-gaussian-vs-mle-tests

Wald test statistic

To test $H_0 : \beta_k = \beta_{k,0}$ (typically $\beta_{k,0} = 0$):

$$z_k = \frac{\hat{\beta}_k - \beta_{k,0}}{\widehat{\mathrm{SE}}\big(\hat{\beta}_k\big)}$$

Under H_0, $z_k \dot{\sim} N(0,1)$ for large n.

Confidence intervals for regression coefficients

A 95% confidence interval for β_k is:

$$\hat{\beta}_k \pm 1.96 \cdot \widehat{\mathrm{SE}}\big(\hat{\beta}_k\big)$$

Confidence intervals for exponentiated coefficients

By the invariance property of MLEs, the MLE of e^{β_k} is $e^{\hat{\beta}_k}$.

A 95% confidence interval for e^{β_k} is obtained by exponentiating the endpoints of the CI for β_k:

$$e^{\beta_k} \in \left(e^{\hat{\beta}_k - 1.96 \cdot \widehat{\mathrm{SE}}(\hat{\beta}_k)},\ e^{\hat{\beta}_k + 1.96 \cdot \widehat{\mathrm{SE}}(\hat{\beta}_k)}\right)$$

For covariate coefficients ($k \neq 0$), e^{β_k} is an odds ratio (the multiplicative change in odds per 1-unit increase in x_k, holding all other covariates fixed). For the intercept ($k = 0$), e^{β_0} is the baseline odds (the odds when all covariates equal zero), not an odds ratio.

In R

In R, `parameters()` from the `parameters` package automatically computes Wald tests and confidence intervals for logistic regression model coefficients:

```
beetles_glm_ungrouped |>
  parameters() |>
  print_md()
#> Deviance = 372.714 Iterations - 1
#> Deviance = 372.714 Iterations - 2
#> Deviance = 373.417 Iterations - 1
#> Deviance = 373.417 Iterations - 2
#> Deviance = 374.544 Iterations - 1
#> Deviance = 374.544 Iterations - 2
#> Deviance = 376.063 Iterations - 1
#> Deviance = 376.063 Iterations - 2
#> Deviance = 377.943 Iterations - 1
#> Deviance = 377.943 Iterations - 2
#> Deviance = 380.156 Iterations - 1
#> Deviance = 380.156 Iterations - 2
#> Deviance = 372.727 Iterations - 1
#> Deviance = 372.727 Iterations - 2
#> Deviance = 373.526 Iterations - 1
#> Deviance = 373.526 Iterations - 2
#> Deviance = 374.912 Iterations - 1
#> Deviance = 374.912 Iterations - 2
#> Deviance = 376.937 Iterations - 1
#> Deviance = 376.937 Iterations - 2
#> Deviance = 376.937 Iterations - 3
#> Deviance = 379.654 Iterations - 1
#> Deviance = 379.654 Iterations - 2
```

```
#> Deviance = 379.654 Iterations - 3
#> Deviance = 372.727 Iterations - 1
#> Deviance = 372.727 Iterations - 2
#> Deviance = 373.526 Iterations - 1
#> Deviance = 373.526 Iterations - 2
#> Deviance = 374.913 Iterations - 1
#> Deviance = 374.913 Iterations - 2
#> Deviance = 376.94 Iterations - 1
#> Deviance = 376.94 Iterations - 2
#> Deviance = 379.66 Iterations - 1
#> Deviance = 379.66 Iterations - 2
#> Deviance = 379.66 Iterations - 3
#> Deviance = 372.714 Iterations - 1
#> Deviance = 372.714 Iterations - 2
#> Deviance = 373.417 Iterations - 1
#> Deviance = 373.417 Iterations - 2
#> Deviance = 374.543 Iterations - 1
#> Deviance = 374.543 Iterations - 2
#> Deviance = 376.061 Iterations - 1
#> Deviance = 376.061 Iterations - 2
#> Deviance = 377.939 Iterations - 1
#> Deviance = 377.939 Iterations - 2
#> Deviance = 380.151 Iterations - 1
#> Deviance = 380.151 Iterations - 2
```

Table 12: Wald tests and 95% CIs for `beetles` logistic regression

Parameter	Log-Odds	SE	95% CI	z	p
(Intercept)	-60.72	5.18	(-71.44, -51.08)	-11.72	$< .001$
dose	34.27	2.91	(28.85, 40.30)	11.77	$< .001$

Pass `exponentiate = TRUE` to `parameters()` to report exponentiated coefficients. The intercept is then interpreted as baseline odds, while the non-intercept coefficients are odds ratios:

```
beetles_glm_ungrouped |>
  parameters(exponentiate = TRUE) |>
  print_md()
#> Deviance = 372.714 Iterations - 1
#> Deviance = 372.714 Iterations - 2
#> Deviance = 373.417 Iterations - 1
#> Deviance = 373.417 Iterations - 2
#> Deviance = 374.544 Iterations - 1
#> Deviance = 374.544 Iterations - 2
#> Deviance = 376.063 Iterations - 1
#> Deviance = 376.063 Iterations - 2
#> Deviance = 377.943 Iterations - 1
#> Deviance = 377.943 Iterations - 2
#> Deviance = 380.156 Iterations - 1
#> Deviance = 380.156 Iterations - 2
#> Deviance = 372.727 Iterations - 1
#> Deviance = 372.727 Iterations - 2
#> Deviance = 373.526 Iterations - 1
#> Deviance = 373.526 Iterations - 2
#> Deviance = 374.912 Iterations - 1
#> Deviance = 374.912 Iterations - 2
#> Deviance = 376.937 Iterations - 1
#> Deviance = 376.937 Iterations - 2
```

```
#> Deviance = 376.937 Iterations - 3
#> Deviance = 379.654 Iterations - 1
#> Deviance = 379.654 Iterations - 2
#> Deviance = 379.654 Iterations - 3
#> Deviance = 372.727 Iterations - 1
#> Deviance = 372.727 Iterations - 2
#> Deviance = 373.526 Iterations - 1
#> Deviance = 373.526 Iterations - 2
#> Deviance = 374.913 Iterations - 1
#> Deviance = 374.913 Iterations - 2
#> Deviance = 376.94 Iterations - 1
#> Deviance = 376.94 Iterations - 2
#> Deviance = 379.66 Iterations - 1
#> Deviance = 379.66 Iterations - 2
#> Deviance = 379.66 Iterations - 3
#> Deviance = 372.714 Iterations - 1
#> Deviance = 372.714 Iterations - 2
#> Deviance = 373.417 Iterations - 1
#> Deviance = 373.417 Iterations - 2
#> Deviance = 374.543 Iterations - 1
#> Deviance = 374.543 Iterations - 2
#> Deviance = 376.061 Iterations - 1
#> Deviance = 376.061 Iterations - 2
#> Deviance = 377.939 Iterations - 1
#> Deviance = 377.939 Iterations - 2
#> Deviance = 380.151 Iterations - 1
#> Deviance = 380.151 Iterations - 2
```

Table 13: Exponentiated coefficients and 95% CIs for `beetles`

Parameter	Odds Ratio	SE	95% CI	z	p
(Intercept)	4.27e-27	2.21e-26	(9.39e-32, 6.56e-23)	-11.72	$< .001$
dose	7.65e+14	2.23e+15	(3.40e+12, 3.18e+17)	11.77	$< .001$

6.2 Inference for predicted probabilities

Exercise 6.1. Given a maximum likelihood estimate $\hat{\beta}$ and a corresponding estimated covariance matrix $\hat{\Sigma} \overset{\text{def}}{=} \widehat{\text{Cov}}(\hat{\beta})$, calculate a 95% confidence interval for the predicted probability $\pi(\tilde{x}) = \Pr(Y = 1 | \tilde{X} = \tilde{x})$.

Solution

Solution 6.1.
By the central limit theorem for MLEs[a], a 95% confidence interval for $\pi(\tilde{x})$ can be constructed as:

$$\hat{\pi}(\tilde{x}) \pm 1.96 * \widehat{\text{SE}}(\hat{\pi}(\tilde{x}))$$

However, $\widehat{\text{SE}}(\hat{\pi}(\tilde{x}))$ seems difficult to compute; doing so would require using the delta method[b]. Instead, using the invariance property of MLEs, we can first calculate a confidence interval for the log-odds,

$$\eta(\tilde{x}) = \tilde{x}' \tilde{\beta} \in (L, R)$$

and then apply expit to the endpoints of that log-odds-scale confidence interval:

$$\pi(\tilde{x}) \in (\text{expit}(L), \text{expit}(R)) \tag{22}$$

[a] intro-MLEs.qmd#thm-dist-mle
[b] https://en.wikipedia.org/wiki/Delta_method

Exercise 6.2. Find a 95% confidence interval for the log-odds $\eta(\tilde{x}) = \tilde{x}'\tilde{\beta}$.

Solution

Solution 6.2. By the central limit theorem for MLEs[a], a 95% confidence interval for $\eta(\tilde{x})$ can be constructed as:

$$\hat{\eta}(\tilde{x}) \pm 1.96 * \widehat{\text{SE}}(\hat{\eta}(\tilde{x}))$$

where $\hat{\eta}(\tilde{x}) = \tilde{x}'\hat{\beta}$.

[a] intro-MLEs.qmd#thm-dist-mle

Exercise 6.3.
How can we estimate the standard error of $\hat{\eta}(\tilde{x})$?

$$\widehat{\text{SE}}(\hat{\eta}(\tilde{x})) = ?$$

Solution

Solution 6.3.

$$\text{SE}(\hat{\eta}(\tilde{x})) = \sqrt{\text{Var}(\hat{\eta}(\tilde{x}))} \tag{23}$$

By the definition $\hat{\eta}(\tilde{x}) = \tilde{x}'\hat{\beta}$ and the variance of a linear combination[a]:

$$\begin{aligned} \text{Var}(\hat{\eta}(\tilde{x})) &= \text{Var}\left(\tilde{x}^\top \hat{\beta}\right) \\ &= \tilde{x}^\top \text{Cov}\left(\hat{\beta}\right)\tilde{x} \\ &= \tilde{x}^\top \Sigma \tilde{x} \end{aligned} \tag{24}$$

where $\Sigma \overset{\text{def}}{=} \text{Cov}(\hat{\beta})$.
Expanding Equation 24 out of matrix-vector notation, we have:

$$\begin{aligned} \tilde{x}^\top \Sigma \tilde{x} &= \sum_{i=1}^{p}\sum_{j=1}^{p} x_i \Sigma_{ij} x_j \\ &= \sum_{i=1}^{p}\sum_{j=1}^{p} x_i \,\text{Cov}(\hat{\beta}_i, \hat{\beta}_j) x_j \end{aligned}$$

[a] probability.qmd#thm-var-lincom

Combining Equation 24 with the invariance property of MLEs gives an estimator for the variance and standard error:

Theorem 6.1 (Estimated variance and standard error of log-odds)**.**

$$\widehat{\text{Var}}(\hat{\eta}(\tilde{x})) = \tilde{x}^\top \hat{\Sigma} \tilde{x} \tag{25}$$

$$\widehat{\text{SE}}(\hat{\eta}(\tilde{x})) = \sqrt{\tilde{x}^\top \hat{\Sigma} \tilde{x}} \tag{26}$$

Proof

Proof. By Solution 6.3, $\text{Var}(\hat{\eta}(\tilde{x})) = \tilde{x}^\top \Sigma \tilde{x}$ (Equation 24). By the invariance property of MLEs, we estimate Σ with $\hat{\Sigma} \overset{\text{def}}{=} \widehat{\text{Cov}}(\hat{\beta})$, which yields Equation 25; taking the square root gives Equation 26. □

Exm

Example 6.1 (Standard error and CI for a beetle log-odds)**.**
We instantiate Theorem 6.1 for the `beetles` logistic regression model (Table 5).
Consider the covariate pattern $\tilde{x} = (1, 1.8)^\top$, representing the intercept and a dose of 1.8.

```
x_vec <- c(1, 1.8)
sigma_hat <- vcov(beetles_glm_ungrouped)
logodds_hat <- as.numeric(x_vec %*% coef(beetles_glm_ungrouped))
se_logodds <- sqrt(as.numeric(t(x_vec) %*% sigma_hat %*% x_vec))
ci_logodds <- logodds_hat + c(-1, 1) * 1.96 * se_logodds
```

By Equation 26, the estimated standard error of the log-odds is $\widehat{\text{SE}}(\hat{\eta}(\tilde{x})) = \sqrt{\tilde{x}^\top \hat{\Sigma} \tilde{x}} = 0.145056$, while the estimated log-odds itself is $\hat{\eta}(\tilde{x}) = \tilde{x}' \hat{\beta} = 0.969132$.
The resulting 95% confidence interval for the log-odds is $(0.684823, 1.253441)$, and applying expit to its endpoints (Equation 22) gives a 95% confidence interval for $\pi(\tilde{x})$ of $(0.664814, 0.777895)$.
These values match the `dose = 1.8` row of Table 14, which uses `predict()` with `type = "link"` and `se.fit = TRUE` to compute the estimated log-odds $\hat{\eta}(\tilde{x}) = \tilde{x}' \hat{\beta}$ and its estimated standard error for each covariate pattern:

```
library(dplyr)
new_doses <- tibble(dose = c(1.7, 1.8, 1.9))

pred_logodds <-
  beetles_glm_ungrouped |>
  predict(
    newdata = new_doses,
    type = "link",
    se.fit = TRUE
  )

new_doses |>
  mutate(
    logodds_hat = pred_logodds$fit,
    se = pred_logodds$se.fit,
    ci_lower_logodds = logodds_hat - 1.96 * se,
    ci_upper_logodds = logodds_hat + 1.96 * se,
    pi_hat = plogis(logodds_hat),
    ci_lower_prob = plogis(ci_lower_logodds),
    ci_upper_prob = plogis(ci_upper_logodds)
  ) |>
  knitr::kable(digits = 3)
```

Table 14: Predicted log-odds and 95% CI for `beetles` logistic regression

dose	logodds_hat	se	ci_lower_logodds	ci_upper_logodds	pi_hat	ci_lower_prob	ci_upper_prob
1.7	-2.458	0.263	-2.974	-1.942	0.079	0.049	0.125
1.8	0.969	0.145	0.685	1.253	0.725	0.665	0.778
1.9	4.396	0.377	3.656	5.136	0.988	0.975	0.994

6.3 Inference for odds ratios

Exercise 6.4. Given a maximum likelihood estimate $\hat{\beta}$ and a corresponding estimated covariance matrix $\hat{\Sigma} \overset{\text{def}}{=} \widehat{\text{Cov}}(\hat{\beta})$, calculate a 95% confidence interval for the odds ratio comparing covariate patterns $\tilde{x}$ and $\tilde{x}^*$, $\theta_\omega(\tilde{x}, \tilde{x}^*)$.

Solution

Solution 6.4.
By the central limit theorem for MLEs[a], a 95% confidence interval for $\theta_\omega(\tilde{x}, \tilde{x}^*)$ can be constructed as:

$$\hat{\theta}_\omega \pm 1.96 * \widehat{\text{SE}}\left(\hat{\theta}_\omega\right) \tag{27}$$

However, $\widehat{\text{SE}}\left(\hat{\theta}_\omega\right)$ seems difficult to compute; doing so would require using the delta method[b]. Instead, using the invariance property of MLEs, we can first calculate a confidence interval for the logarithm of the odds ratio,

$$\log\{\theta_\omega(\tilde{x}, \tilde{x}^*)\} \in (L, R) \tag{28}$$

and then exponentiate the endpoints of that log-odds-scale confidence interval:

$$\theta_\omega(\tilde{x}, \tilde{x}^*) \in (e^L, e^R) \tag{29}$$

[a] intro-MLEs.qmd#thm-dist-mle
[b] https://en.wikipedia.org/wiki/Delta_method

Exercise 6.5. Find a 95% confidence interval for the natural logarithm of the odds ratio, $\log\{\theta_\omega(\tilde{x}, \tilde{x}^*)\}$

Solution

Solution 6.5. From Corollary 4.3, we know:

$$\log\{\theta_\omega(\tilde{x}, \tilde{x}^*)\} = \Delta\eta$$

By the central limit theorem for MLEs[a], a 95% confidence interval for $\Delta\eta$ can be constructed as:

$$\widehat{\Delta\eta} \pm 1.96 * \widehat{\text{SE}}\left(\widehat{\Delta\eta}\right)$$

[a] intro-MLEs.qmd#thm-dist-mle

Exercise 6.6.
How can we estimate the standard error of $\widehat{\Delta\eta}$?

$$\widehat{\text{SE}}\left(\widehat{\Delta\eta}\right) = ?$$

Solution

Solution 6.6.

$$\text{SE}\left(\widehat{\Delta\eta}\right) = \sqrt{\text{Var}\left(\widehat{\Delta\eta}\right)} \tag{30}$$

By Lemma 4.3 and the variance of a linear combination[a]:

$$\begin{aligned}\widehat{\mathrm{Var}\left(\Delta\eta\right)} &= \mathrm{Var}\left((\Delta\tilde{x})\cdot\hat{\beta}\right) \\ &= (\Delta\tilde{x})^\top \mathrm{Cov}\left(\hat{\beta}\right)(\Delta\tilde{x}) \\ &= (\Delta\tilde{x})^\top\Sigma(\Delta\tilde{x})\end{aligned} \tag{31}$$

where $\Sigma \overset{\text{def}}{=} \mathrm{Cov}(\hat{\beta})$.
Expanding Equation 31 out of matrix-vector notation, we have:

$$\begin{aligned}(\Delta\tilde{x})^\top\Sigma(\Delta\tilde{x}) &= \sum_{i=1}^{p}\sum_{j=1}^{p}(\Delta\tilde{x})_i\Sigma_{ij}(\Delta\tilde{x})_j \\ &= \sum_{i=1}^{p}\sum_{j=1}^{p}(\Delta x_i)\Sigma_{ij}(\Delta x_j) \\ &= \sum_{i=1}^{p}\sum_{j=1}^{p}(x_i - x_i^*)\,\mathrm{Cov}(\hat{\beta}_i,\hat{\beta}_j)(x_j - x_j^*)\end{aligned}$$

[a] probability.qmd#thm-var-lincom

Combining Equation 31 with the invariance property of MLEs gives an estimator for the variance and standard error:

Theorem 6.2 (Estimated variance and standard error of difference in log-odds)**.**

$$\widehat{\mathrm{Var}}(\Delta\hat{\eta}) = \Delta\tilde{x}^\top\hat{\Sigma}(\Delta\tilde{x}) \tag{32}$$

$$\widehat{\mathrm{SE}}(\Delta\hat{\eta}) = \sqrt{\Delta\tilde{x}^\top\hat{\Sigma}(\Delta\tilde{x})} \tag{33}$$

Proof

Proof. By Solution 6.6, $\widehat{\mathrm{Var}\left(\Delta\eta\right)} = \Delta\tilde{x}^\top\Sigma(\Delta\tilde{x})$ (Equation 31). By the invariance property of MLEs, we estimate Σ with $\hat{\Sigma} \overset{\text{def}}{=} \widehat{\mathrm{Cov}}(\hat{\beta})$, which yields Equation 32; taking the square root gives Equation 33. □

Compare this result with confidence intervals[13].

Exm

Example 6.2 (Standard error and CI for a beetle odds ratio)**.**
We instantiate Theorem 6.2 for the `beetles` logistic regression model (Table 5).
Compare two covariate patterns, a dose of 1.8 versus a dose of 1.7: $\tilde{x} = (1, 1.8)^\top$ and $\tilde{x}^* = (1, 1.7)^\top$, so that $\Delta\tilde{x} = \tilde{x} - \tilde{x}^* = (0, 0.1)^\top$.

```
delta_x <- c(0, 0.1)
sigma_hat <- vcov(beetles_glm_ungrouped)
diff_logodds_hat <- as.numeric(delta_x %*% coef(beetles_glm_ungrouped))
se_diff <- sqrt(as.numeric(t(delta_x) %*% sigma_hat %*% delta_x))
ci_diff <- diff_logodds_hat + c(-1, 1) * 1.96 * se_diff
```

By Equation 33, the estimated standard error of the difference in log-odds is $\widehat{\mathrm{SE}}(\Delta\hat{\eta}) = \sqrt{\Delta\tilde{x}^\top\hat{\Sigma}(\Delta\tilde{x})} = 0.291211$, and the estimated difference in log-odds is $\Delta\hat{\eta} = (\Delta\tilde{x})\cdot\hat{\beta} = 3.427033$. The resulting 95% confidence interval for $\Delta\eta$ is $(2.856258, 3.997807)$, so by Equation 29 the estimated odds ratio and its 95% confidence interval are $\hat{\theta}_\omega(\tilde{x}, \tilde{x}^*) = e^{3.427033} = 30.785154$,

[13] Linear-models-overview.qmd#sec-se-fitted

with CI $(17.396309, 54.478551)$.

7 Multiple logistic regression

7.0.1 Coronary heart disease (WCGS) study data

Let's use the data from the Western Collaborative Group Study (WCGS) (Rosenman et al. (1975)) to explore multiple logistic regression:

From Vittinghoff et al. (2012):

"The **Western Collaborative Group Study (WCGS)** was a large epidemiological study designed to investigate the association between the"type A" behavior pattern and coronary heart disease (CHD)".

Exercise 7.1. What is "type A" behavior?

Solution

Solution 7.1. *From Wikipedia, "Type A and Type B personality theory":*
"The hypothesis describes Type A individuals as outgoing, ambitious, rigidly organized, highly status-conscious, impatient, anxious, proactive, and concerned with time management….
The hypothesis describes Type B individuals as a contrast to those of Type A. Type B personalities, by definition, are noted to live at lower stress levels. They typically work steadily and may enjoy achievement, although they have a greater tendency to disregard physical or mental stress when they do not achieve."

Study design

from `?faraway::wcgs`*:*

3154 healthy young men aged 39-59 from the San Francisco area were assessed for their personality type. All were free from coronary heart disease at the start of the research. Eight and a half years later change in CHD status was recorded.

Details (from `faraway::wcgs`*)*

The WCGS began in 1960 with 3,524 male volunteers who were employed by 11 California companies. Subjects were 39 to 59 years old and free of heart disease as determined by electrocardiogram. After the initial screening, the study population dropped to 3,154 and the number of companies to 10 because of various exclusions. The cohort comprised both blue- and white-collar employees.

7.0.2 Baseline data collection

socio-demographic characteristics:

- age
- education
- marital status
- income
- occupation
- physical and physiological including:
- height
- weight
- blood pressure
- electrocardiogram
- corneal arcus

biochemical measurements:

Table 15: wcgs data

```
wcgs
#> # A tibble: 3,154 x 22
#>      age arcus behpat   bmi chd69  chol   dbp dibpat height    id lnsbp lnwght
#>    <dbl> <lgl> <fct>  <dbl> <fct> <dbl> <dbl> <fct>   <dbl> <dbl> <dbl>  <dbl>
#>  1    50 TRUE  A1      31.3 No      249    90 Type A     67  2343  4.88   5.30
#>  2    51 FALSE A1      25.3 No      194    74 Type A     73  3656  4.79   5.26
#>  3    59 TRUE  A1      28.7 No      258    94 Type A     70  3526  5.06   5.30
#>  4    51 TRUE  A1      22.1 No      173    80 Type A     69 22057  4.84   5.01
#>  5    44 FALSE A1      22.3 No      214    80 Type A     71 12927  4.84   5.08
#>  6    47 FALSE A1      27.1 No      206    76 Type A     64 16029  4.75   5.06
#>  7    40 FALSE A1      23.2 No      190    78 Type A     70  3894  4.80   5.09
#>  8    41 FALSE A1      23.0 No      212    84 Type A     70 11389  4.87   5.08
#>  9    50 TRUE  A1      27.2 No      130    70 Type A     71 12681  4.72   5.27
#> 10    43 FALSE A1      28.4 No      233    80 Type A     68 10005  4.79   5.23
#> # i 3,144 more rows
#> # i 10 more variables: ncigs <dbl>, sbp <dbl>, smoke <fct>, t1 <dbl>,
#> #   time169 <dbl>, typchd69 <fct>, uni <dbl>, weight <dbl>, wghtcat <fct>,
#> #   agec <fct>
```

- cholesterol and lipoprotein fractions;
- medical and family history and use of medications;

behavioral data:

- Type A interview,
- smoking,
- exercise
- alcohol use.

Later surveys added data on:

- anthropometry
- triglycerides
- Jenkins Activity Survey
- caffeine use

Average follow-up continued for 8.5 years with repeat examinations.

7.0.3 Load the data

Here, I load the data:

```
### load the data directly from a UCSF website:
library(haven)
url <- paste0(
  # I'm breaking up the url into two chunks for readability
  "https://regression.ucsf.edu/sites/g/files/",
  "tkssra6706/f/wysiwyg/home/data/wcgs.dta"
)
wcgs <- haven::read_dta(url)
```

A copy is also available on Kaggle[14].

7.0.4 Data cleaning

Now let's do some data cleaning

[14] https://www.kaggle.com/datasets/nmd2104/wcgs-data

```r
library(arsenal) # provides `set_labels()`
library(forcats) # provides `as_factor()`
library(haven)
library(plotly)
wcgs <- wcgs |>
  mutate(
    age = age |>
      arsenal::set_labels("Age (years)"),
    arcus = arcus |>
      as.logical() |>
      arsenal::set_labels("Arcus Senilis"),
    time169 = time169 |>
      as.numeric() |>
      arsenal::set_labels("Observation (follow up) time (days)"),
    dibpat = dibpat |>
      as_factor() |>
      relevel(ref = "Type B") |>
      arsenal::set_labels("Behavioral Pattern"),
    typchd69 = typchd69 |>
      labelled(
        label = "Type of CHD Event",
        labels =
          c(
            "None" = 0,
            "infdeath" = 1,
            "silent" = 2,
            "angina" = 3
          )
      ),

    # turn stata-style labelled variables in to R-style factors:
    across(
      where(is.labelled),
      haven::as_factor
    )
  )
```

7.0.5 What's in the data

Table 16 summarizes the data.

7.0.6 Data by age and personality type

For now, we will look at the interaction between age and personality type (`dibpat`). To make it easier to visualize the data, we summarize the event rates for each combination of age:

```r
library(dplyr)
odds <- function(pi) pi / (1 - pi)
chd_grouped_data <-
  wcgs |>
  summarize(
    .by = c(age, dibpat),
    n = sum(chd69 %in% c("Yes", "No")),
    x = sum(chd69 == "Yes")
  ) |>
  mutate(
    `n - x` = n - x,
    `p(chd)` = (x / n) |>
```

Table 16: Baseline characteristics by CHD status at end of follow-up

```
library(gtsummary)
wcgs |>
  dplyr::select(
    -dplyr::all_of(c("id", "uni", "t1"))
  ) |>
  gtsummary::tbl_summary(
    by = "chd69",
    missing_text = "Missing"
  ) |>
  gtsummary::add_p() |>
  gtsummary::add_overall() |>
  gtsummary::bold_labels() |>
  gtsummary::separate_p_footnotes()
```

Characteristic	**Overall** N = 3,154[1]	**No** N = 2,897[1]	**Yes** N = 2
Age (years)	45.0 (42.0, 50.0)	45.0 (41.0, 50.0)	49.0 (44.0, 5
Arcus Senilis	941 (30%)	839 (29%)	102 (40%
Missing	2	0	2
Behavioral Pattern			
A1	264 (8.4%)	234 (8.1%)	30 (12%
A2	1,325 (42%)	1,177 (41%)	148 (58%
B3	1,216 (39%)	1,155 (40%)	61 (24%
B4	349 (11%)	331 (11%)	18 (7.0%
Body Mass Index (kg/m2)	24.39 (22.96, 25.84)	24.39 (22.89, 25.84)	24.82 (23.63,
Total Cholesterol	223 (197, 253)	221 (195, 250)	245 (222, 2
Missing	12	12	0
Diastolic Blood Pressure	80 (76, 86)	80 (76, 86)	84 (80, 90
Behavioral Pattern			
Type B	1,565 (50%)	1,486 (51%)	79 (31%
Type A	1,589 (50%)	1,411 (49%)	178 (69%
Height (inches)	70.00 (68.00, 72.00)	70.00 (68.00, 72.00)	70.00 (68.00,
Ln of Systolic Blood Pressure	4.84 (4.79, 4.91)	4.84 (4.77, 4.91)	4.87 (4.82, 4
Ln of Weight	5.14 (5.04, 5.20)	5.13 (5.04, 5.20)	5.16 (5.09, 5
Cigarettes per day	0 (0, 20)	0 (0, 20)	20 (0, 30
Systolic Blood Pressure	126 (120, 136)	126 (118, 136)	130 (124, 1
Current smoking	1,502 (48%)	1,343 (46%)	159 (62%
Observation (follow up) time (days)	2,942 (2,842, 3,037)	2,952 (2,864, 3,048)	1,666 (934, 2
Type of CHD Event			
None	0 (0%)	0 (0%)	0 (0%)
infdeath	2,897 (92%)	2,897 (100%)	0 (0%)
silent	135 (4.3%)	0 (0%)	135 (53%
angina	71 (2.3%)	0 (0%)	71 (28%
4	51 (1.6%)	0 (0%)	51 (20%
Weight (lbs)	170 (155, 182)	169 (155, 182)	175 (162, 1
Weight Category			
< 140	232 (7.4%)	217 (7.5%)	15 (5.8%
140-170	1,538 (49%)	1,440 (50%)	98 (38%
170-200	1,171 (37%)	1,049 (36%)	122 (47%
> 200	213 (6.8%)	191 (6.6%)	22 (8.6%
RECODE of age (Age)			
35-40	543 (17%)	512 (18%)	31 (12%
41-45	1,091 (35%)	1,036 (36%)	55 (21%
46-50	750 (24%)	680 (23%)	70 (27%
51-55	528 (17%)	463 (16%)	65 (25%

```
      labelled(label = "CHD Event by 1969"),
    `odds(chd)` = `p(chd)` / (1 - `p(chd)`),
    `logit(chd)` = log(`odds(chd)`)
  )

chd_grouped_data
#> # A tibble: 42 x 8
#>      age dibpat     n     x `n - x` `p(chd)`  `odds(chd)` `logit(chd)`
#>    <dbl> <fct>  <int> <int>   <int> <dbl+lbl>       <dbl>        <dbl>
#>  1    50 Type A    76     8      68 0.105          0.118        -2.14
#>  2    51 Type A    67    11      56 0.164          0.196        -1.63
#>  3    59 Type A    30     7      23 0.233          0.304        -1.19
#>  4    44 Type A   113     9     104 0.0796         0.0865       -2.45
#>  5    47 Type A    72     7      65 0.0972         0.108        -2.23
#>  6    40 Type A   133     9     124 0.0677         0.0726       -2.62
#>  7    41 Type A   108     7     101 0.0648         0.0693       -2.67
#>  8    43 Type A    97     7      90 0.0722         0.0778       -2.55
#>  9    54 Type A    53     7      46 0.132          0.152        -1.88
#> 10    48 Type A    80    12      68 0.15           0.176        -1.73
#> # i 32 more rows
```

7.0.7 Graphical exploration

```
library(ggplot2)
library(scales)
chd_plot_probs <-
  chd_grouped_data |>
  ggplot() +
  aes(
    x = age,
    y = `p(chd)`,
    col = dibpat
  ) +
  geom_point(aes(size = n), alpha = .7) +
  scale_size(range = c(1, 4)) +
  geom_line() +
  theme_bw() +
  ylab("P(CHD Event by 1969)") +
  scale_y_continuous(
    labels = scales::label_percent(),
    sec.axis = sec_axis(
      ~ odds(.),
      name = "odds(CHD Event by 1969)"
    )
  ) +
  theme(legend.position = "bottom")

print(chd_plot_probs)
```

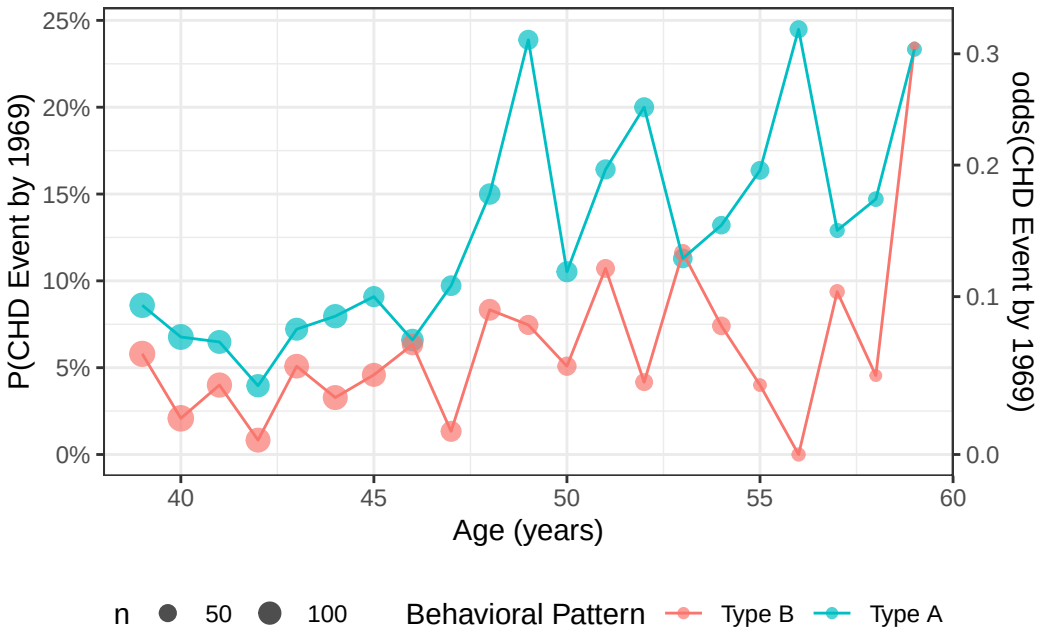


Figure 6: CHD rates by age group, probability scale

7.1 Odds scale

```
odds_inv <- function(omega) omega / (1 + omega)
trans_odds <- trans_new(
  name = "odds",
  transform = odds,
  inverse = odds_inv
)

chd_plot_odds <- chd_plot_probs +
  scale_y_continuous(
    trans = trans_odds, # this line changes the vertical spacing
    name = chd_plot_probs$labels$y,
    sec.axis = sec_axis(
      ~ odds(.),
      name = "odds(CHD Event by 1969)"
    )
  )

print(chd_plot_odds)
```

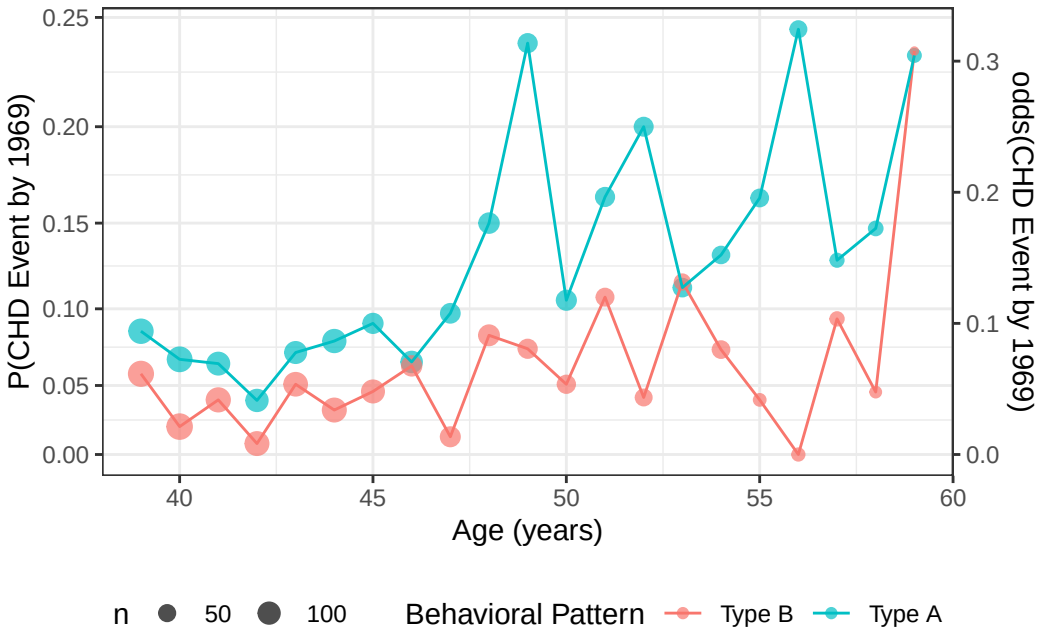


Figure 7: CHD rates by age group, odds spacing

7.2 Log-odds (logit) scale

```
logit <- function(pi) log(odds(pi))
expit <- function(eta) odds_inv(exp(eta))
trans_logit <- trans_new(
  name = "logit",
  transform = logit,
  inverse = expit
)

chd_plot_logit <-
  chd_plot_probs +
  scale_y_continuous(
    trans = trans_logit, # this line changes the vertical spacing
    name = chd_plot_probs$labels$y,
    breaks = c(seq(.01, .1, by = .01), .15, .2),
    minor_breaks = NULL,
    sec.axis = sec_axis(
      ~ logit(.),
      name = "log(odds(CHD Event by 1969))"
    )
  )

print(chd_plot_logit)
```

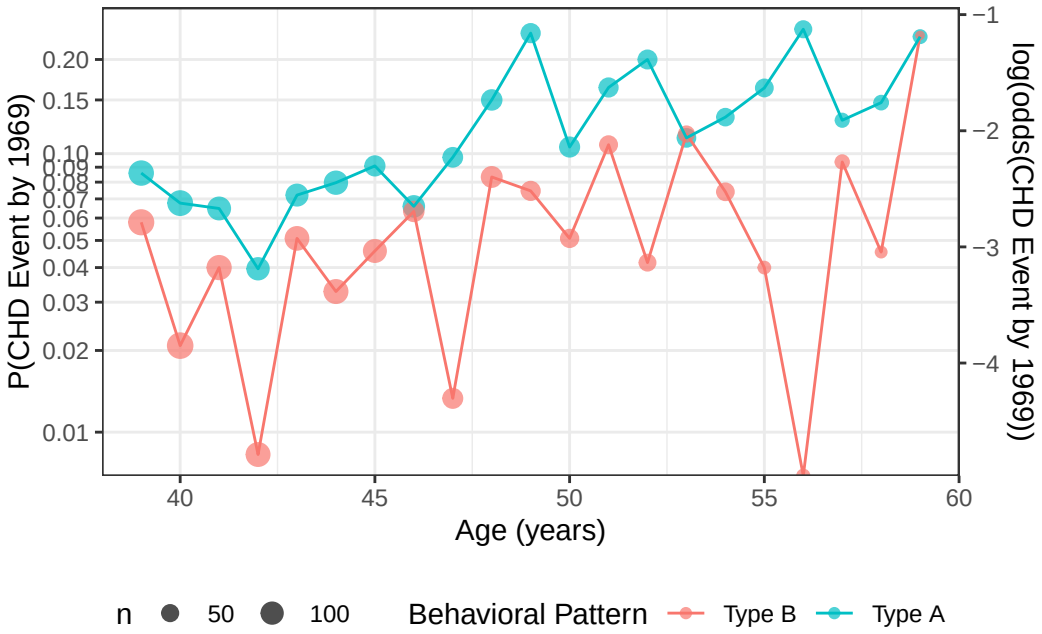


Figure 8: CHD data (logit-scale)

7.2.1 Logistic regression models for CHD data

For the `wcgs` dataset, let's consider a **logistic regression model** for the outcome of Coronary Heart Disease (Y; `chd` in computer output):

- $Y = 1$ if an individual developed CHD by the end of the study;
- $Y = 0$ if they have not developed CHD by the end of the study.

Let's include an intercept, two covariates, plus their interaction:

- A: age at study enrollment (`age`, recorded in years)
- P: personality type (`dibpat`):
  - $P = 1$ represents "Type A personality",
  - $P = 0$ represents "Type B personality".
- PA: the interaction of personality type and age (`dibpat:age`)
- $\tilde{X} = (1, A, P, PA)$

```
chd_glm_contrasts <-
  wcgs |>
  glm(
    "data" = _,
    "formula" = chd69 == "Yes" ~ dibpat * age,
    "family" = binomial(link = "logit")
  )

library(equatiomatic)
equatiomatic::extract_eq(chd_glm_contrasts)
```

$$\log\left[\frac{P(\text{chd69} = \text{Yes})}{1 - P(\text{chd69} = \text{Yes})}\right] = \alpha + \beta_1(\text{dibpat}_{\text{Type A}}) + \beta_2(\text{age}) + \beta_3(\text{dibpat}_{\text{Type A}} \times \text{age}) \quad (34)$$

Or in more formal notation:

$$
\begin{aligned}
Y_i|\tilde{X}_i &\sim_{\perp\!\!\!\perp} \text{Ber}(\pi(\tilde{X}_i)) \\
\pi(\tilde{x}) &= \text{expit}(\eta(\tilde{x})) \\
\eta(\tilde{x}) &= \beta_0 + \beta_P p + \beta_A a + \beta_{PA} pa
\end{aligned}
\tag{35}
$$

7.2.2 Models superimposed on data

We can graph our fitted models on each scale (probability, odds, log-odds).

probability scale

```r
curve_type_A <- function(x) { # nolint: object_name_linter
  chd_glm_contrasts |> predict(
    type = "response",
    newdata = tibble(age = x, dibpat = "Type A")
  )
}

curve_type_B <- function(x) { # nolint: object_name_linter
  chd_glm_contrasts |> predict(
    type = "response",
    newdata = tibble(age = x, dibpat = "Type B")
  )
}

chd_plot_probs_2 <-
  chd_plot_probs +
  geom_function(
    fun = curve_type_A,
    aes(col = "Type A")
  ) +
  geom_function(
    fun = curve_type_B,
    aes(col = "Type B")
  )
print(chd_plot_probs_2)
```

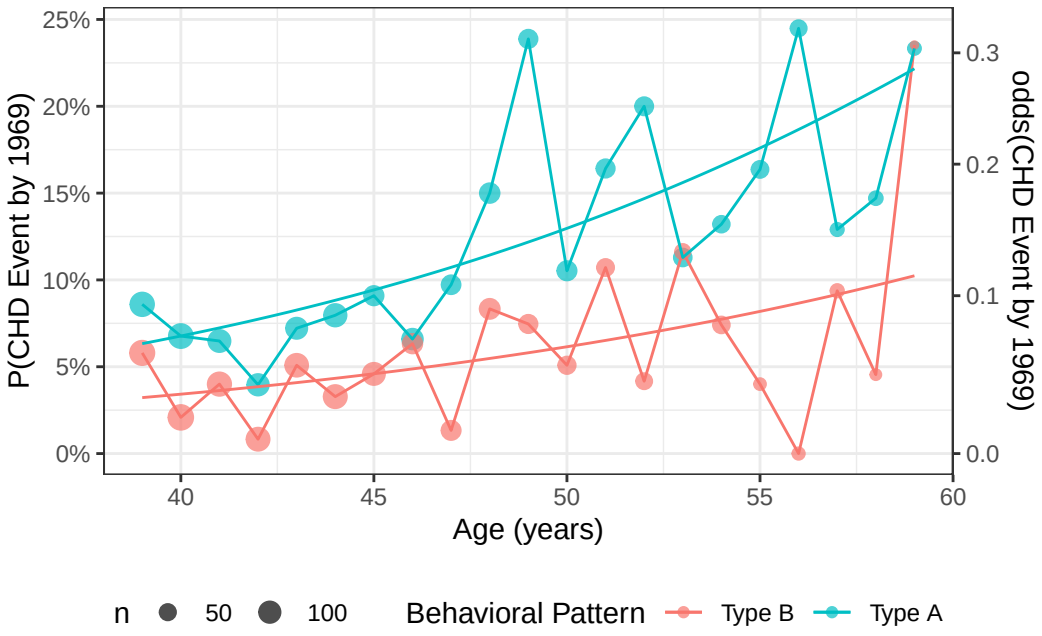


Figure 9: Fitted CHD risk by age and behavior pattern (probability scale)

odds scale

```r
chd_plot_odds_2 <-
  chd_plot_odds +
  geom_function(
    fun = curve_type_A,
    aes(col = "Type A")
  ) +
  geom_function(
    fun = curve_type_B,
    aes(col = "Type B")
  )
print(chd_plot_odds_2)
```

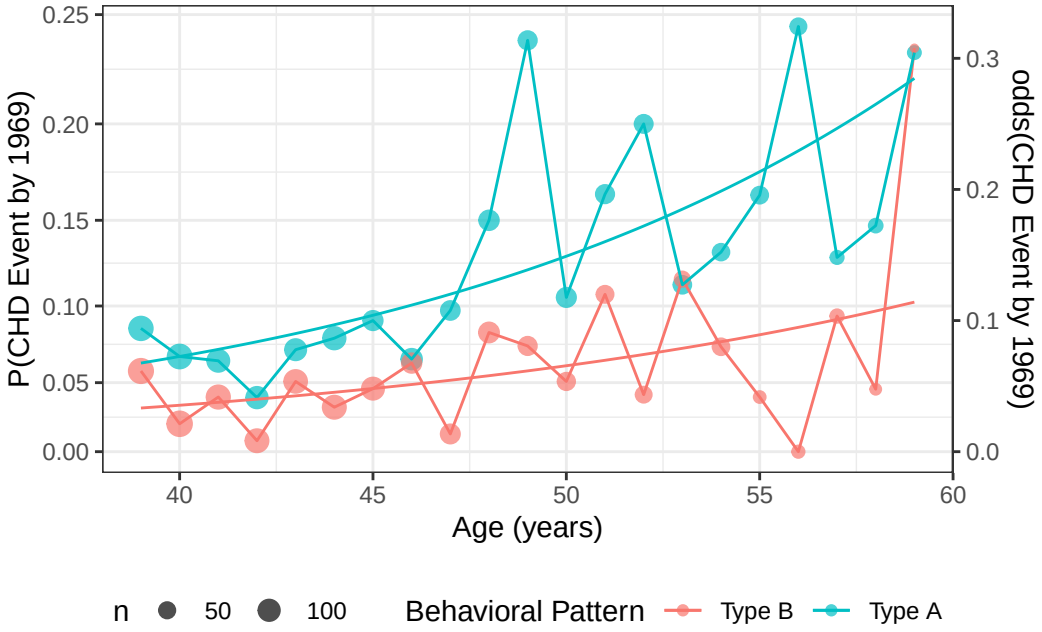


Figure 10: Fitted CHD odds by age and behavior pattern (odds scale)

log-odds (logit) scale

```r
chd_plot_logit_2 <-
  chd_plot_logit +
  geom_function(
    fun = curve_type_A,
    aes(col = "Type A")
  ) +
  geom_function(
    fun = curve_type_B,
    aes(col = "Type B")
  )

print(chd_plot_logit_2)
```

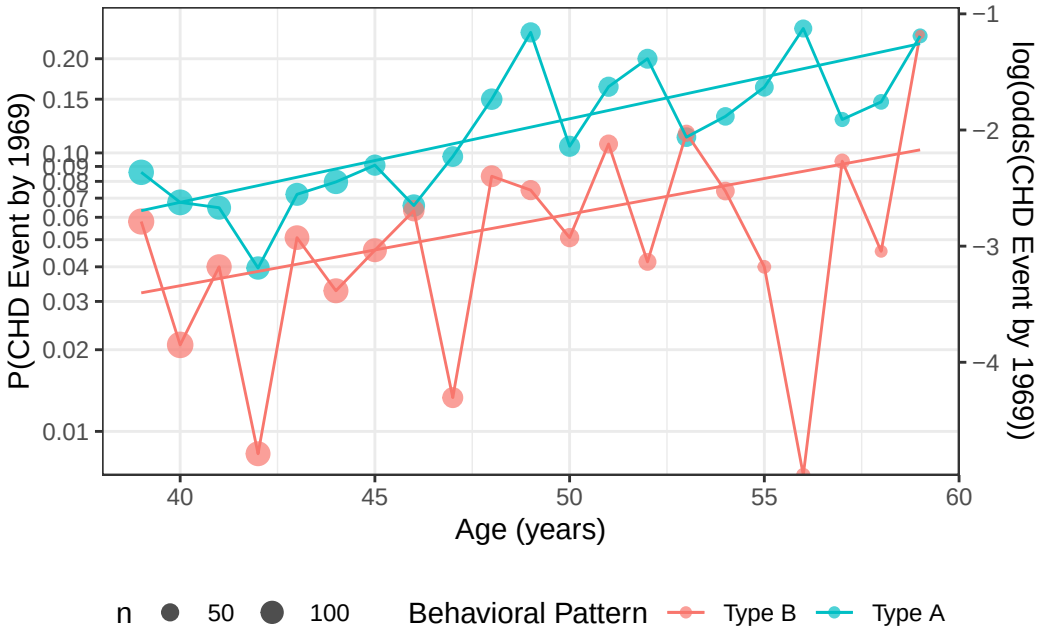


Figure 11: Fitted CHD log-odds by age and behavior pattern (log-odds scale)

7.2.3 Interpreting the model parameters

Exercise 7.2. For Equation 35, derive interpretations of β_0, β_P, β_A, and β_{PA} on the odds **and** log-odds scales, State the interpretations concisely in math **and** in words.
(Hint: apply the general procedure for interpreting a regression coefficient[a] *to the log-odds linear predictor* $\eta(\tilde{x}) = \text{logit}(\pi(\tilde{x}))$*.)*

[a] Linear-models-overview.qmd#def-coef-interp-procedure

Solution

Solution 7.2.

```r
# include: false
age_offset = 0L
```

$$
\begin{aligned}
\eta(P=0, A=0) &= \beta_0 + \beta_P \cdot 0 + \beta_A \cdot 0 + \beta_{PA}(0 \cdot 0) && \text{(model equation, with } p=0, a=0) \\
&= \beta_0 + 0 + 0 + 0 && \text{(multiplication by zero)} \\
&= \beta_0 && \text{(additive identity)}
\end{aligned}
$$

Therefore:

$$\beta_0 = \eta(P = 0, A = 0) \tag{36}$$

β_0 is the natural logarithm of the odds ("log-odds") of experiencing CHD for a 0 year-old person with a type B personality; that is,
The WCGS cohort ranges in age from 39 to 59 years, so a 0-year-old is far outside the observed data. This intercept is therefore an extrapolation with no in-sample referent. Centering age at a representative value (replacing A with $A - \bar{A}$) is the standard remedy: it makes the intercept the log-odds at the mean age, which does have in-sample support.
e^{β_0} is the odds of experiencing CHD for a 0 year-old with a type B personality. Here we write

$$\omega(p, a) \stackrel{\text{def}}{=} \text{odds}(Y = 1 | P = p, A = a)$$

for the odds of CHD at covariate values (p, a). Then:

$$\begin{aligned}
\exp\{\beta_0\} &= \exp\{\eta(P = 0, A = 0)\} && (\beta_0 \text{ as a log-odds}) \\
&= \omega(P = 0, A = 0) && (\text{exponential of log-odds is odds: } \omega = \exp\{\eta\}) \\
&= \frac{\Pr(Y = 1 | P = 0, A = 0)}{1 - \Pr(Y = 1 | P = 0, A = 0)} && (\text{definition of odds}) \\
&= \frac{\Pr(Y = 1 | P = 0, A = 0)}{\Pr(Y = 0 | P = 0, A = 0)} && (\text{complement rule: } 1 - \Pr(Y = 1|\cdot) = \Pr(Y = 0|\cdot))
\end{aligned}$$

As with the log-odds interpretation, this odds is evaluated at age 0, which lies outside the cohort's 39-to-59-year age range, so e^{β_0} has no in-sample referent. Centering age before fitting gives an intercept whose odds is evaluated at the mean age instead.

$$\begin{aligned}
\frac{\partial}{\partial a}\eta(P = 0, A = a) &= \frac{\partial}{\partial a}(\beta_0 + \beta_P \cdot 0 + \beta_A a + \beta_{PA}(0 \cdot a)) && (\text{model equation, with } p = 0) \\
&= \frac{\partial}{\partial a}(\beta_0 + 0 + \beta_A a + 0) && (\text{multiplication by zero}) \\
&= \frac{\partial}{\partial a}(\beta_0 + \beta_A a) && (\text{additive identity}) \\
&= \frac{\partial}{\partial a}\beta_0 + \frac{\partial}{\partial a}(\beta_A a) && (\text{linearity of differentiation}) \\
&= 0 + \beta_A && (\text{constants have derivative 0; derivative of a linear term}) \\
&= \beta_A && (\text{additive identity})
\end{aligned}$$

Therefore:

$$\beta_A = \frac{\partial}{\partial a}\eta(P = 0, A = a) \tag{37}$$

β_A is the slope of the log-odds of CHD with respect to age, for individuals with personality type B.
Alternatively:

$$\begin{aligned}
\eta(P = 0, A = a + 1) - \eta(P = 0, A = a) &= (\beta_0 + \beta_P \cdot 0 + \beta_A(a + 1) + \beta_{PA}(0 \cdot (a + 1))) - (\beta_0 + \beta_P \cdot 0 + \beta_A a + \beta_{PA}(0 \cdot \\
&= (\beta_0 + 0 + \beta_A(a + 1) + 0) - (\beta_0 + 0 + \beta_A a + 0) \\
&= (\beta_0 + \beta_A(a + 1)) - (\beta_0 + \beta_A a) \\
&= \beta_0 + \beta_A a + \beta_A - \beta_0 - \beta_A a \\
&= \beta_A
\end{aligned}$$

In terms of finite differences, β_A is the difference in log-odds of experiencing CHD per one-year difference in age between two individuals with type B personalities.

$$
\begin{aligned}
\exp\{\beta_A\} &= \exp\{\eta(P=0, A=a+1) - \eta(P=0, A=a)\} && (\beta_A \text{ as a difference in log-odds}) \\
&= \frac{\exp\{\eta(P=0, A=a+1)\}}{\exp\{\eta(P=0, A=a)\}} && (\text{exponential of a difference is a ratio}) \\
&= \frac{\omega(P=0, A=a+1)}{\omega(P=0, A=a)} && (\text{exponential of log-odds is odds: } \omega = \exp\{\eta\}) \\
&= \frac{\text{odds}(Y=1|P=0, A=a+1)}{\text{odds}(Y=1|P=0, A=a)} && (\text{the } \omega(p,a) \text{ notation defined at } \beta_0\text{'s interpretation}) \\
&= \theta_\omega(\Delta a = 1|P=0) && (\text{definition of the odds ratio})
\end{aligned}
$$

- The odds ratio of experiencing CHD (aka "the odds ratio") differs by a factor of e^{β_A} per one-year difference in age between individuals with type B personality.

β_P is the difference in log-odds of experiencing CHD for a 0 year-old person with type A personality compared to a 0 year-old person with type B personality; specifically,

$$\beta_P = \eta(P=1, A=0) - \eta(P=0, A=0) \tag{38}$$

- e^{β_P} is the ratio of the odds (aka "the odds ratio") of experiencing CHD, for a 0-year old individual with type A personality vs a 0-year old individual with type B personality; specifically,

$$\exp\{\beta_P\} = \frac{\text{odds}(Y=1|P=1, A=0)}{\text{odds}(Y=1|P=0, A=0)}$$

$$
\begin{aligned}
\frac{\partial}{\partial a}\eta(P=1, A=a) &= \frac{\partial}{\partial a}(\beta_0 + \beta_P \cdot 1 + \beta_A a + \beta_{PA}(1 \cdot a)) && (\text{model equation, with } p=1) \\
&= \frac{\partial}{\partial a}(\beta_0 + \beta_P + \beta_A a + \beta_{PA} a) && (\text{multiplicative identity}) \\
&= \frac{\partial}{\partial a}\beta_0 + \frac{\partial}{\partial a}\beta_P + \frac{\partial}{\partial a}(\beta_A a) + \frac{\partial}{\partial a}(\beta_{PA} a) && (\text{linearity of differentiation}) \\
&= 0 + 0 + \beta_A + \beta_{PA} && (\text{constants have derivative 0; derivatives of linear tern} \\
&= \beta_A + \beta_{PA} && (\text{additive identity}) \\
\frac{\partial}{\partial a}\eta(P=0, A=a) &= \beta_A && (\text{the type B slope, from the } \beta_A \text{ derivation})
\end{aligned}
$$

Therefore:

$$
\begin{aligned}
\frac{\partial}{\partial a}\eta(P=1, A=a) - \frac{\partial}{\partial a}\eta(P=0, A=a) &= (\beta_A + \beta_{PA}) - \beta_A && (\text{substitute the two slopes}) \\
&= \beta_A - \beta_A + \beta_{PA} && (\text{commute the addition}) \\
&= \beta_{PA} && (\text{cancel } \beta_A)
\end{aligned}
$$

That slope difference implies that

$$
\begin{aligned}
\beta_{PA} &= \frac{\partial}{\partial a}\eta(P=1, A=a) - \frac{\partial}{\partial a}\eta(P=0, A=a) && (\text{rearrange the difference of slopes}) \\
&= \frac{\partial}{\partial a}\eta(P=1, A=a) - \beta_A && (\text{substitute the type B slope, } \frac{\partial}{\partial a}\eta(P=0, A=a) = \beta_A)
\end{aligned}
$$

β_{PA} is the difference in the slopes of log-odds over age between participants with Type A personalities and participants with Type B personalities.
Accordingly, the odds ratio of experiencing CHD per one-year difference in age differs by a factor of $e^{\beta_{PA}}$ for participants with type A personality compared to participants with type B personality; specifically,

$$\theta_\omega(\Delta a = 1|P = 1) = \exp\{\beta_{PA}\} \times \theta_\omega(\Delta a = 1|P = 0)$$

or equivalently:

$$\exp\{\beta_{PA}\} = \frac{\theta_\omega(\Delta a = 1|P = 1)}{\theta_\omega(\Delta a = 1|P = 0)}$$

See Section 5.1.1[15] of Vittinghoff et al. (2012) for another perspective, also using the `wcgs` data as an example.

7.2.4 Interpreting the model parameter estimates

Table 17 shows the fitted model.

```
library(parameters)
chd_glm_contrasts |>
  parameters() |>
  print_md()
```

Table 17: CHD model (corner-point parametrization)

Parameter	Log-Odds	SE	95% CI	z	p
(Intercept)	-5.80	0.98	(-7.73, -3.90)	-5.95	< .001
dibpat (Type A)	0.30	1.18	(-2.02, 2.63)	0.26	0.797
age	0.06	0.02	(0.02, 0.10)	3.01	0.003
dibpat (Type A) × age	0.01	0.02	(-0.04, 0.06)	0.42	0.674

We can get the corresponding odds ratio estimates ($e^{\hat{\beta}}$s) by passing `exponentiate = TRUE` to `parameters()`:

```
chd_glm_contrasts |>
  parameters(exponentiate = TRUE) |>
  print_md()
```

Table 18: Odds ratio estimates for CHD model

Parameter	Odds Ratio	SE	95% CI	z	p
(Intercept)	3.02e-03	2.94e-03	(4.40e-04, 0.02)	-5.95	< .001
dibpat (Type A)	1.36	1.61	(0.13, 13.88)	0.26	0.797
age	1.06	0.02	(1.02, 1.11)	3.01	0.003
dibpat (Type A) × age	1.01	0.02	(0.96, 1.06)	0.42	0.674

7.2.5 Stratified parametrization

We could instead use a stratified parametrization:

```
chd_glm_strat <- glm(
  "formula" = chd69 == "Yes" ~ dibpat + dibpat:age - 1,
  "data" = wcgs,
  "family" = binomial(link = "logit")
)
equatiomatic::extract_eq(chd_glm_strat)
```

[15] https://link.springer.com/chapter/10.1007/978-1-4614-1353-0_5#Sec2_5

$$\log\left[\frac{P(\text{chd69} = \text{Yes})}{1 - P(\text{chd69} = \text{Yes})}\right] = \beta_1(\text{dibpat}_{\text{Type B}}) + \beta_2(\text{dibpat}_{\text{Type A}}) + \beta_3(\text{dibpat}_{\text{Type B}} \times \text{dibpat}_{\text{age}}) + \beta_4(\text{dibpat}_{\text{Type A}} \times \text{dibp} \tag{39}$$

```
chd_glm_strat |>
  parameters() |>
  print_md()
```

Table 19: CHD model, stratified parametrization

Parameter	Log-Odds	SE	95% CI	z	p
dibpat (Type B)	-5.80	0.98	(-7.73, -3.90)	-5.95	< .001
dibpat (Type A)	-5.50	0.67	(-6.83, -4.19)	-8.18	< .001
dibpat (Type B) × age	0.06	0.02	(0.02, 0.10)	3.01	0.003
dibpat (Type A) × age	0.07	0.01	(0.05, 0.10)	5.24	< .001

Again, we can get the corresponding odds ratios (e^{β}s) by passing `exponentiate = TRUE` to `parameters()`:

```
chd_glm_strat |>
  parameters(exponentiate = TRUE) |>
  print_md()
```

Table 20: Odds ratio estimates for CHD model

Parameter	Odds Ratio	SE	95% CI	z	p
dibpat (Type B)	3.02e-03	2.94e-03	(4.40e-04, 0.02)	-5.95	< .001
dibpat (Type A)	4.09e-03	2.75e-03	(1.08e-03, 0.02)	-8.18	< .001
dibpat (Type B) × age	1.06	0.02	(1.02, 1.11)	3.01	0.003
dibpat (Type A) × age	1.07	0.01	(1.05, 1.10)	5.24	< .001

Compare with Table 17.

Exercise 7.3. If I give you model 1, how would you get the coefficients of model 2?

Solution

Solution 7.3. Model 1 is the corner-point (reference/contrast) parametrization fitted in `chd_glm_contrasts`, with linear predictor (Equation 35):

$$\eta(p, a) = \beta_0 + \beta_P p + \beta_A a + \beta_{PA} pa$$

Model 2 is the stratified parametrization fitted in `chd_glm_strat`, which gives each personality type its own intercept and its own age slope. Writing γ_B, γ_A for the stratum-specific intercepts and δ_B, δ_A for the stratum-specific age slopes:

$$\eta(p, a) \overset{\text{def}}{=} \gamma_B(1 - p) + \gamma_A p + \delta_B(1 - p)a + \delta_A pa \tag{40}$$

The two parametrizations span the same set of linear-predictor functions, so both models have the same fitted linear predictor $\eta(p, a)$. We can therefore recover the model 2 coefficients by equating the two expressions for $\eta(p, a)$ within each personality-type stratum.

Type B stratum ($p = 0$):

In model 1 (Equation 35):

$$\begin{aligned}
\eta(0,a) &= \beta_0 + \beta_P \cdot 0 + \beta_A a + \beta_{PA} \cdot 0 \cdot a && \text{(substitute } p = 0\text{)} \\
&= \beta_0 + 0 + \beta_A a + 0 && \text{(multiply by 0)} \\
&= \beta_0 + \beta_A a && \text{(drop the zero terms)}
\end{aligned}$$

In model 2 (Equation 40):

$$\begin{aligned}
\eta(0,a) &= \gamma_B(1-0) + \gamma_A \cdot 0 + \delta_B(1-0)a + \delta_A \cdot 0 \cdot a && \text{(substitute } p = 0\text{)} \\
&= \gamma_B \cdot 1 + \gamma_A \cdot 0 + \delta_B \cdot 1 \cdot a + \delta_A \cdot 0 \cdot a && (1 - 0 = 1) \\
&= \gamma_B + 0 + \delta_B a + 0 && \text{(multiply by 1 and by 0)} \\
&= \gamma_B + \delta_B a && \text{(drop the zero terms)}
\end{aligned}$$

Equating the two expressions for $\eta(0,a)$, for all ages a:

$$\gamma_B + \delta_B a = \beta_0 + \beta_A a \tag{41}$$

Setting $a = 0$ in Equation 41:

$$\begin{aligned}
\gamma_B + \delta_B \cdot 0 &= \beta_0 + \beta_A \cdot 0 && \text{(substitute } a = 0\text{)} \\
\gamma_B + 0 &= \beta_0 + 0 && \text{(multiply by 0)} \\
\gamma_B &= \beta_0 && \text{(drop the zero terms)}
\end{aligned}$$

Subtracting $\gamma_B = \beta_0$ from both sides of Equation 41 and then setting $a = 1$:

$$\begin{aligned}
\delta_B a &= \beta_A a && \text{(subtract } \gamma_B = \beta_0 \text{ from both sides)} \\
\delta_B \cdot 1 &= \beta_A \cdot 1 && \text{(substitute } a = 1\text{)} \\
\delta_B &= \beta_A && \text{(multiply by 1)}
\end{aligned}$$

Type A stratum ($p = 1$):
In model 1 (Equation 35):

$$\begin{aligned}
\eta(1,a) &= \beta_0 + \beta_P \cdot 1 + \beta_A a + \beta_{PA} \cdot 1 \cdot a && \text{(substitute } p = 1\text{)} \\
&= \beta_0 + \beta_P + \beta_A a + \beta_{PA} a && \text{(multiply by 1)} \\
&= (\beta_0 + \beta_P) + (\beta_A + \beta_{PA})a && \text{(group the constant terms; factor out } a\text{)}
\end{aligned}$$

In model 2 (Equation 40):

$$\begin{aligned}
\eta(1,a) &= \gamma_B(1-1) + \gamma_A \cdot 1 + \delta_B(1-1)a + \delta_A \cdot 1 \cdot a && \text{(substitute } p = 1\text{)} \\
&= \gamma_B \cdot 0 + \gamma_A \cdot 1 + \delta_B \cdot 0 \cdot a + \delta_A \cdot 1 \cdot a && (1 - 1 = 0) \\
&= 0 + \gamma_A + 0 + \delta_A a && \text{(multiply by 0 and by 1)} \\
&= \gamma_A + \delta_A a && \text{(drop the zero terms)}
\end{aligned}$$

Equating the two expressions for $\eta(1,a)$ and applying the same two steps as in the Type B stratum (set $a = 0$; then subtract the intercepts and set $a = 1$):

$$\begin{aligned}
\gamma_A &= \beta_0 + \beta_P \\
\delta_A &= \beta_A + \beta_{PA}
\end{aligned}$$

Summary of the conversion:

$$\begin{aligned}
\gamma_B &= \beta_0 \\
\gamma_A &= \beta_0 + \beta_P \\
\delta_B &= \beta_A \\
\delta_A &= \beta_A + \beta_{PA}
\end{aligned} \tag{42}$$

Solving Equation 42 for the model 1 coefficients inverts the conversion, so either parametrization determines the other:

$$
\begin{aligned}
\beta_0 &= \gamma_B && \text{(first equation of the summary)} \\
\beta_P &= \gamma_A - \gamma_B && \text{(subtract the first equation from the second)} \\
\beta_A &= \delta_B && \text{(third equation)} \\
\beta_{PA} &= \delta_A - \delta_B && \text{(subtract the third equation from the fourth)}
\end{aligned}
$$

Numerical check: applying Equation 42 to the fitted model 1 coefficients (Table 17) reproduces the fitted model 2 coefficients (Table 19):

```
b <- coef(chd_glm_contrasts)

gamma_delta <- c(
  "dibpatType B" = unname(b["(Intercept)"]),
  "dibpatType A" = unname(b["(Intercept)"] + b["dibpatType A"]),
  "dibpatType B:age" = unname(b["age"]),
  "dibpatType A:age" = unname(b["age"] + b["dibpatType A:age"])
)

cbind(
  "converted from model 1" = gamma_delta,
  "model 2 (fitted directly)" = coef(chd_glm_strat)[names(gamma_delta)]
)
#>                   converted from model 1 model 2 (fitted directly)
#> dibpatType B                 -5.8032532                -5.8032532
#> dibpatType A                 -5.4988618                -5.4988618
#> dibpatType B:age              0.0615634                 0.0615634
#> dibpatType A:age              0.0719071                 0.0719071
```

8 Model comparisons for logistic models

8.0.1 Deviance test

Definition 8.1 (Deviance goodness-of-fit test)**.** We can compare the maximized log-likelihood of our model, $\ell(\hat{\beta}; \mathbf{x})$, versus the log-likelihood of the full model (aka saturated model aka maximal model), ℓ_{full}, which has one parameter per covariate pattern. The **deviance** statistic is:

$$D \stackrel{\text{def}}{=} 2(\ell_{\text{full}} - \ell(\hat{\beta}; \mathbf{x}))$$

With enough data, $D \dot{\sim} \chi^2(q-p)$, where q is the number of distinct covariate patterns and p is the number of β parameters in our model. The **deviance test** compares D to that $\chi^2(q-p)$ distribution; a significant p-value for this **deviance** statistic indicates that there's some detectable pattern in the data that our model isn't flexible enough to catch.

Caution

The deviance statistic needs to have a large amount of data **for each covariate pattern** for the χ^2 approximation to hold. A guideline from Dobson and Barnett (2018, chap. 5) is that if there are q distinct covariate patterns $x_1 ..., x_q$, with $n_1, ..., n_q$ observations per pattern, then the expected frequencies $n_k \cdot \pi(x_k)$ should be at least 1 for every pattern $k \in 1 : q$.

If you have covariates measured on a continuous scale, you may not be able to use the deviance tests to assess goodness of fit.

Exm

Example 8.1 (Deviance test for the WCGS CHD model)**.** Our CHD model `chd_glm_contrasts` was fit to individual-level observations, but its covariate patterns are just the combinations of personality type (`dibpat`) and whole-year `age`, so we can refit the same model to the grouped data and perform the deviance test:

```
wcgs_gof_grouped <-
  wcgs |>
  summarize(
    .by = c(dibpat, age),
    n = n(),
    chd = sum(chd69 == "Yes"),
    no_chd = sum(chd69 == "No")
  )

chd_glm_gof_grouped <- glm(
  "formula" = cbind(chd, no_chd) ~ dibpat * age,
  "data" = wcgs_gof_grouped,
  "family" = binomial(link = "logit")
)

deviance_gof <- deviance(chd_glm_gof_grouped)
q_gof <- nrow(wcgs_gof_grouped)
p_gof <- length(coef(chd_glm_gof_grouped))
pval_deviance_gof <- pchisq(
  deviance_gof,
  df = q_gof - p_gof,
  lower.tail = FALSE
)
min_exp_freq <- min(wcgs_gof_grouped$n * fitted(chd_glm_gof_grouped))
```

The smallest expected event frequency across the covariate patterns is $\min_k \{n_k \cdot \hat{\pi}(x_k)\} = 1.74$, which is at least 1, so these grouped data meet the minimum-expected-frequency guideline from Dobson and Barnett (2018, chap. 5) for using the χ^2 approximation.
The grouped data have $q = 42$ distinct covariate patterns, and the model has $p = 4$ parameters, so the deviance statistic $D = 45$ has $q - p = 38$ degrees of freedom. The corresponding p-value is $p(\chi^2(38) > 45) = 0.202$, so the deviance test finds no significant evidence of lack of fit for this model.

8.0.2 Hosmer-Lemeshow test

If our covariate patterns produce groups that are too small, a reasonable solution is to make bigger groups by merging some of the covariate-pattern groups together.

Definition 8.2 (Hosmer-Lemeshow test)**.** Hosmer and Lemeshow (1980) proposed that we group the patterns by their predicted probabilities according to the model of interest. For example, you could group all of the observations with predicted probabilities of 10% or less together, then group the observations with 11%-20% probability together, and so on; $g = 10$ categories in all.
Then we can construct a statistic

$$X^2 \stackrel{\text{def}}{=} \sum_{c=1}^{g} \frac{(o_c - e_c)^2}{e_c}$$

where o_c is the number of events *observed* in group c, and e_c is the number of events expected in group c (based on the sum of the fitted values $\hat{\pi}_i$ for observations in group c).
If each group has enough observations in it, you can compare X^2 to a χ^2 distribution; by simulation, the degrees of freedom has been found to be approximately $g - 2$.

Note

The Hosmer-Lemeshow statistic depends on how the observations are grouped. Changing the number of groups g, or the cut points used to form them, can change the value of X^2 and its p-value for the same fitted model, so the test can reach different conclusions depending on the grouping choice. Hosmer et al. (1997) compare the Hosmer-Lemeshow test against several alternative goodness-of-fit tests and document this sensitivity.

Exm

Example 8.2 (Hosmer-Lemeshow test for the WCGS CHD model)**.** For our CHD model, this procedure would be:

```
wcgs <-
  wcgs |>
  mutate(
    pred_probs_glm1 = chd_glm_contrasts |> fitted(),
    pred_prob_cats1 = pred_probs_glm1 |>
      cut(
        breaks = seq(0, 1, by = .1),
        include.lowest = TRUE
      )
  )

HL_table <- # nolint: object_name_linter
  wcgs |>
  summarize(
    .by = pred_prob_cats1,
    n = n(),
    o = sum(chd69 == "Yes"),
    e = sum(pred_probs_glm1)
  )

library(pander)
HL_table |> pander()
```

pred_prob_cats1	n	o	e
(0.1,0.2]	785	116	108
(0.2,0.3]	64	12	13.77
[0,0.1]	2,305	129	135.2

```
X2 <- HL_table |> # nolint: object_name_linter
  summarize(
    `X^2` = sum((o - e)^2 / e)
  ) |>
  pull(`X^2`)
print(X2)
#> [1] 1.11029

pval1 <- pchisq(X2, lower = FALSE, df = nrow(HL_table) - 2)
```

Our statistic is $X^2 = 1.110287$; $p(\chi^2(1) > 1.110287) = 0.29202$, which is our p-value for detecting a lack of goodness of fit.
Unfortunately that grouping plan left us with just three categories with any observations, so instead of grouping by 10% increments of predicted probability, typically analysts use deciles of the predicted probabilities:

```
wcgs <-
  wcgs |>
  mutate(
    pred_probs_glm1 = chd_glm_contrasts |> fitted(),
    pred_prob_cats1 = pred_probs_glm1 |>
      cut(
        breaks = quantile(pred_probs_glm1, seq(0, 1, by = .1)),
        include.lowest = TRUE
      )
  )

HL_table <- # nolint: object_name_linter
  wcgs |>
  summarize(
    .by = pred_prob_cats1,
    n = n(),
    o = sum(chd69 == "Yes"),
    e = sum(pred_probs_glm1)
  )

HL_table |> pander()
```

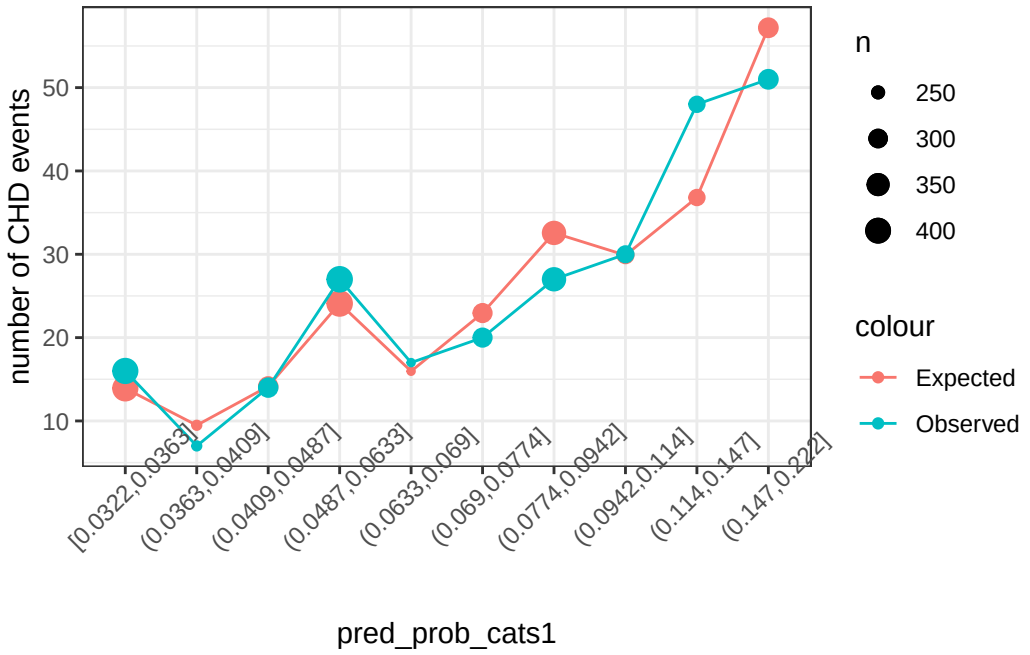
pred_prob_cats1	n	o	e
(0.114,0.147]	275	48	36.81
(0.147,0.222]	314	51	57.19
(0.0774,0.0942]	371	27	32.56
(0.0942,0.114]	282	30	29.89
(0.0633,0.069]	237	17	15.97
(0.069,0.0774]	306	20	22.95
(0.0487,0.0633]	413	27	24.1
(0.0409,0.0487]	310	14	14.15
[0.0322,0.0363]	407	16	13.91
(0.0363,0.0409]	239	7	9.48

```
X2 <- HL_table |> # nolint: object_name_linter
  summarize(
    `X^2` = sum((o - e)^2 / e)
  ) |>
  pull(`X^2`)

print(X2)
#> [1] 6.78114

pval1 <- pchisq(X2, lower = FALSE, df = nrow(HL_table) - 2)
```

Now we have more evenly split categories. The p-value is 0.56042, still not significant. Graphically, we have compared:

```
HL_plot <- # nolint: object_name_linter
  HL_table |>
  ggplot(aes(x = pred_prob_cats1)) +
  geom_line(
    aes(y = e, x = pred_prob_cats1, group = "Expected", col = "Expected")
  ) +
  geom_point(aes(y = e, size = n, col = "Expected")) +
  geom_point(aes(y = o, size = n, col = "Observed")) +
  geom_line(aes(y = o, col = "Observed", group = "Observed")) +
  scale_size(range = c(1, 4)) +
  theme_bw() +
  ylab("number of CHD events") +
  theme(axis.text.x = element_text(angle = 45))
```

```
print(HL_plot)
```



8.0.3 Comparing models

The deviance test (Definition 8.1) and the Hosmer-Lemeshow test (Definition 8.2) assess the fit of a single model in absolute terms. To choose between two competing models, we can instead compare their fits against each other, using criteria that trade off fit against complexity:

- AIC $= -2 * \ell(\widehat{\theta}) + 2 * p$ [lower is better] (Akaike 1974) (see the AIC definition[16])
- BIC $= -2 * \ell(\widehat{\theta}) + p * \log(n)$ [lower is better] (Schwarz 1978) (see the BIC definition[17])

A larger maximized log-likelihood is not, by itself, evidence that one model is better than another: adding parameters to a model can never decrease its maximized log-likelihood, so the more complex of two nested models always fits at least as well. For **nested** models, the likelihood ratio test accounts for this complexity difference by testing whether the improvement in fit is larger than chance alone would produce:

Definition 8.3 (Likelihood ratio test (LRT))**.** Suppose a **reduced** model with p_0 parameters is nested in a **full** model with $p_1 > p_0$ parameters; that is, the reduced model is a special case of the full model. The **likelihood ratio test statistic** is:

$$\Lambda \stackrel{\text{def}}{=} 2(\ell_{\text{full}}(\widehat{\theta}_1) - \ell_{\text{red}}(\widehat{\theta}_0))$$

where $\widehat{\theta}_1$ and $\widehat{\theta}_0$ are the maximum likelihood estimates of the full and reduced models, respectively. Under the null hypothesis that the reduced model is adequate, $\Lambda \dot{\sim} \chi^2(p_1 - p_0)$ by Wilks' theorem[a]. The **likelihood ratio test (LRT)** rejects the reduced model when Λ is large relative to that $\chi^2(p_1 - p_0)$ distribution.

[a] intro-MLEs.qmd#thm-wilks

See also the likelihood ratio test for linear models[18].

Exm

Example 8.3 (Testing the WCGS interaction term with a likelihood ratio test)**.** Our CHD model `chd_glm_contrasts` includes a `dibpat:age` interaction term. Does that interaction term improve the fit enough to justify its extra parameter? Let's fit the nested model without the interaction term and compare the two models with a likelihood ratio test:

[16] Linear-models-overview.qmd#def-aic
[17] Linear-models-overview.qmd#def-bic
[18] Linear-models-overview.qmd#sec-linreg-LRT

```
chd_glm_additive <-
  wcgs |>
  glm(
    "data" = _,
    "formula" = chd69 == "Yes" ~ dibpat + age,
    "family" = binomial(link = "logit")
  )

lrt_chd <- anova(
  chd_glm_additive,
  chd_glm_contrasts,
  test = "LRT"
)
lrt_chd
#> # A tibble: 2 x 5
#>   `Resid. Df` `Resid. Dev`    Df Deviance `Pr(>Chi)`
#>         <dbl>        <dbl> <dbl>    <dbl>      <dbl>
#> 1        3151        1704.    NA   NA         NA
#> 2        3150        1704.     1    0.177      0.674

llik_full_chd <- logLik(chd_glm_contrasts) |> as.numeric()
llik_red_chd <- logLik(chd_glm_additive) |> as.numeric()
lambda_chd <- lrt_chd$Deviance[2]
df_chd <- lrt_chd$Df[2]
pval_lrt_chd <- lrt_chd$`Pr(>Chi)`[2]
```

Here the test statistic is:

$$
\begin{aligned}
\Lambda &\overset{\text{def}}{=} 2(\ell_{\text{full}}(\hat{\theta}_1) - \ell_{\text{red}}(\hat{\theta}_0)) && \text{(definition of the LRT statistic)} \\
&= 2 \times (-851.99 - (-852.08)) && \text{(substituting the two maximized log-likelihoods)} \\
&= 2 \times 0.09 && \text{(subtraction)} \\
&= 0.18 && \text{(multiplication)}
\end{aligned}
$$

The full model has $p_1 = 4$ parameters and the reduced model has $p_0 = 3$, so the null distribution is $\chi^2(p_1 - p_0) = \chi^2(1)$, and the p-value is $p(\chi^2(1) > 0.18) = 0.674$. The LRT finds no significant evidence that the interaction term improves the fit.
The `lrtest()` function from the {lmtest}[a] package performs the same test:

```
lmtest::lrtest(chd_glm_contrasts, chd_glm_additive)
#> # A tibble: 2 x 5
#>   `#Df` LogLik    Df  Chisq `Pr(>Chisq)`
#>   <dbl>  <dbl> <dbl>  <dbl>        <dbl>
#> 1     4  -852.    NA NA           NA
#> 2     3  -852.    -1  0.177        0.674
```

AIC and BIC can also compare these two models:

```
AIC(chd_glm_additive, chd_glm_contrasts)
#> # A tibble: 2 x 2
#>      df   AIC
#>   <dbl> <dbl>
#> 1     3 1710.
#> 2     4 1712.
BIC(chd_glm_additive, chd_glm_contrasts)
#> # A tibble: 2 x 2
#>      df   BIC
#>   <dbl> <dbl>
#> 1     3 1728.
#> 2     4 1736.
```

Both criteria are lower for the additive model (AIC: 1710.16 vs 1711.98; BIC: 1728.33 vs 1736.21), agreeing with the LRT that the interaction term is not worth its extra parameter.

[a] https://cran.r-project.org/package=lmtest

9 Residual-based diagnostics

9.0.1 Logistic regression residuals only work for grouped data

The residual and leverage diagnostics in this section follow the logistic-regression diagnostics developed by Pregibon (1981); see also Dobson and Barnett (2018, chap. 7) and Vittinghoff et al. (2012, chap. 5) for textbook treatments.

```
library(haven)
url <- paste0(
  # I'm breaking up the url into two chunks for readability
  "https://regression.ucsf.edu/sites/g/files/",
  "tkssra6706/f/wysiwyg/home/data/wcgs.dta"
)
library(here) # provides the `here()` function
library(fs) # provides the `path()` function
here::here() |>
  fs::path("Data/wcgs.rda") |>
  load()
chd_glm_contrasts <-
  wcgs |>
  glm(
    "data" = _,
    "formula" = chd69 == "Yes" ~ dibpat * age,
    "family" = binomial(link = "logit")
  )
library(ggfortify)
chd_glm_contrasts |> autoplot()
```

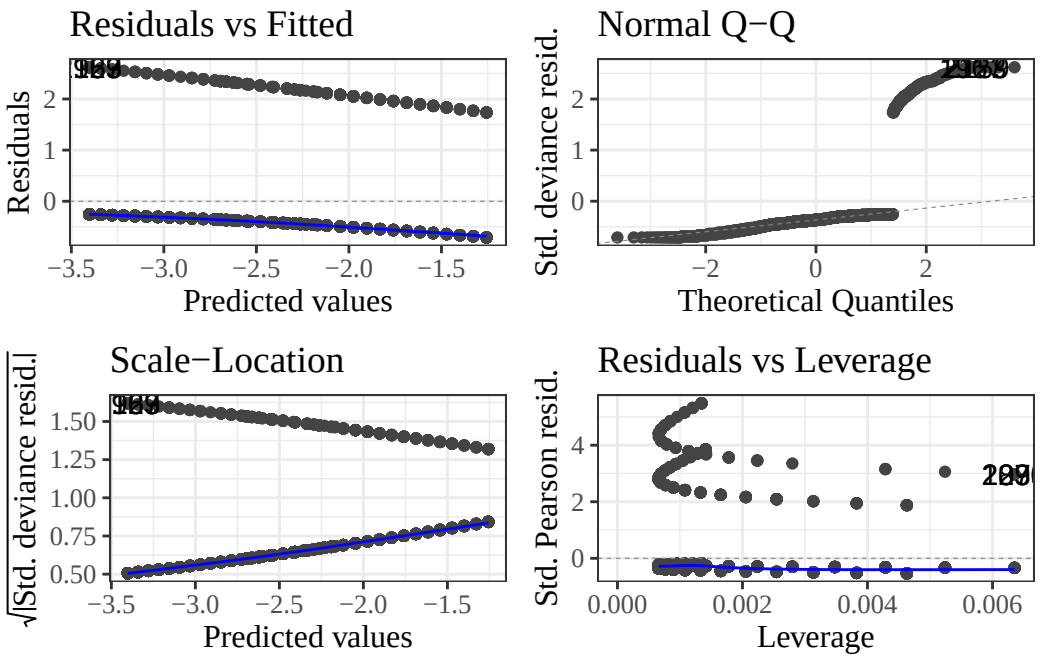


Figure 12: Residual diagnostics for WCGS model with individual-level observations

Residuals only work if there is more than one observation for most covariate patterns.

Here we will create the grouped-data version of our CHD model from the WCGS study:

```
library(dplyr)
wcgs_grouped <-
  wcgs |>
  summarize(
    .by = c(dibpat, age),
    n = n(),
    chd = sum(chd69 == "Yes"),
    no_chd = sum(chd69 == "No")
  ) |>
  mutate(p_chd = chd/n)

chd_glm_contrasts_grouped <- glm(
  "formula" = cbind(chd, no_chd) ~ dibpat*age,
  "data" = wcgs_grouped,
  "family" = binomial(link = "logit")
)
chd_glm_contrasts_grouped |> equatiomatic::extract_eq()
```

$$\log\left[\frac{P(\text{chd})}{1-P(\text{chd})}\right] = \alpha + \beta_1(\text{dibpat}_{\text{Type A}}) + \beta_2(\text{age}) + \beta_3(\text{dibpat}_{\text{Type A}} \times \text{age}) \tag{43}$$

```
library(parameters)
chd_glm_contrasts_grouped |>
  parameters() |>
  print_md()
```

Table 21: CHD model with grouped `wcgs` data

Parameter	Log-Odds	SE	95% CI	z	p
(Intercept)	-5.80	0.98	(-7.73, -3.90)	-5.95	$< .001$
dibpat (Type A)	0.30	1.18	(-2.02, 2.63)	0.26	0.797
age	0.06	0.02	(0.02, 0.10)	3.01	0.003
dibpat (Type A) × age	0.01	0.02	(-0.04, 0.06)	0.42	0.674

```
chd_glm_contrasts_grouped |>
  sjPlot::plot_model(type = "pred", terms = c("age", "dibpat")) +
  geom_point(data = wcgs_grouped |> mutate(group_col = dibpat),
             aes(x = age, y = p_chd))
```

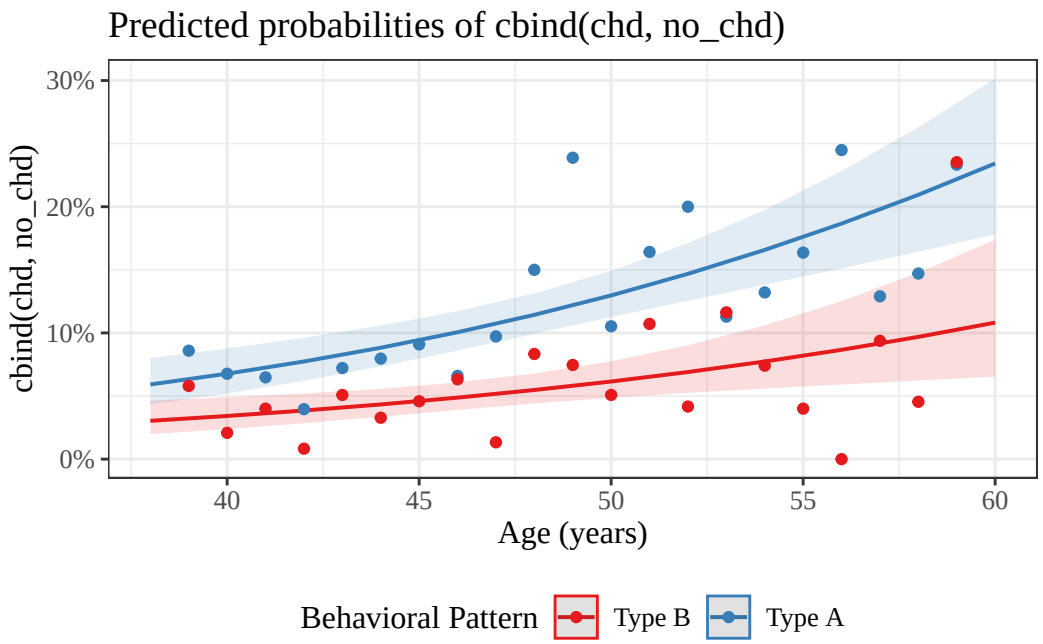
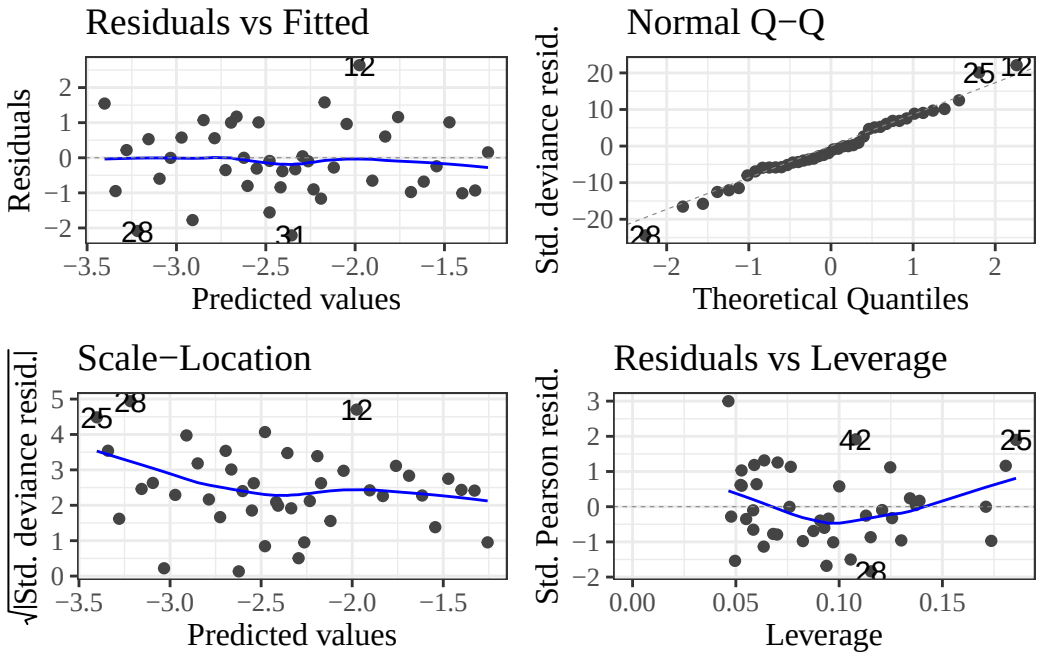


Figure 13: CHD model with grouped `wcgs` data

```
library(ggfortify)
chd_glm_contrasts_grouped |> autoplot()
```



9.0.2 (Response) residuals

Definition 9.1.

Response residual
The **response residual** for covariate pattern k is

$$e_k \stackrel{\text{def}}{=} \bar{y}_k - \hat{\pi}(x_k)$$

where k indexes the covariate patterns, $\bar{y}_k$ is the observed proportion of events in pattern k, and $\hat{\pi}(x_k)$ is the fitted probability for pattern k.

Exm

Example 9.1 (Response residuals for the WCGS CHD model)**.** We can graph these residuals e_k (Definition 9.1) against the fitted values $\hat{\pi}(x_k)$:

```
odds <- function(pi) pi/(1-pi)
logit <- function(pi) log(odds(pi))
wcgs_grouped <-
  wcgs_grouped |>
  mutate(
    fitted = chd_glm_contrasts_grouped |> fitted(),
    fitted_logit = fitted |> logit(),
    response_resids = chd_glm_contrasts_grouped |> resid(type = "response")
  )

wcgs_response_resid_plot <-
  wcgs_grouped |>
  ggplot(
    mapping = aes(
      x = fitted,
      y = response_resids
    )
  ) +
  geom_point(
    aes(col = dibpat)
  ) +
  geom_hline(yintercept = 0) +
  geom_smooth(                                                    ①
    se = TRUE,
    method.args = list(
      span = 2 / 3,
      degree = 1,
      family = "symmetric",
      iterations = 3
    ),
    method = stats::loess
  )
```

① Don't worry about these options for now; I chose them to match `autoplot()` as closely as I can. `plot.glm` and `autoplot` use `stats::lowess` instead of `stats::loess`; `stats::lowess` is older, hard to use with `geom_smooth`, and hard to match exactly with `stats::loess`; see https://support.bioconductor.org/p/2323/.]

```
wcgs_response_resid_plot |> print()
```

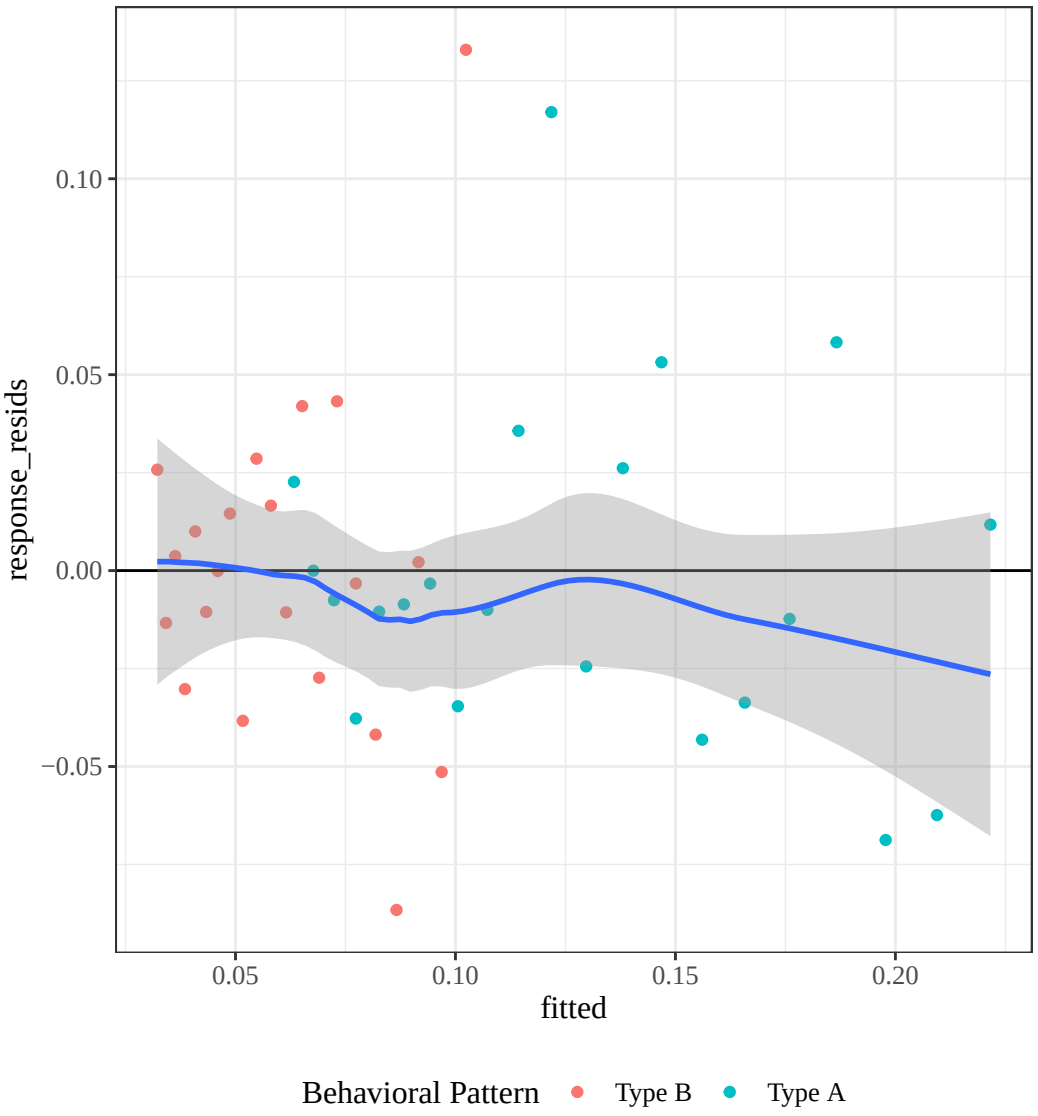


Figure 14: residuals plot for `wcgs` model

We can see a slight fan-shape here: observations on the right have larger variance (as expected since $var(\bar{y}) = \pi(1-\pi)/n$ is maximized when $\pi = 0.5$).

9.0.3 Pearson residuals

The fan-shape in the response residuals plot isn't necessarily a concern here, since we haven't made an assumption of constant residual variance, as we did for linear regression.

However, we might want to divide by the standard error in order to make the graph easier to interpret. Here's one way to do that:

Definition 9.2 (Pearson residual)**.** The Pearson (chi-squared) residual for covariate pattern k is:

$$X_k \stackrel{\text{def}}{=} \frac{\bar{y}_k - \hat{\pi}_k}{\sqrt{\hat{\pi}_k(1-\hat{\pi}_k)/n_k}}$$

where

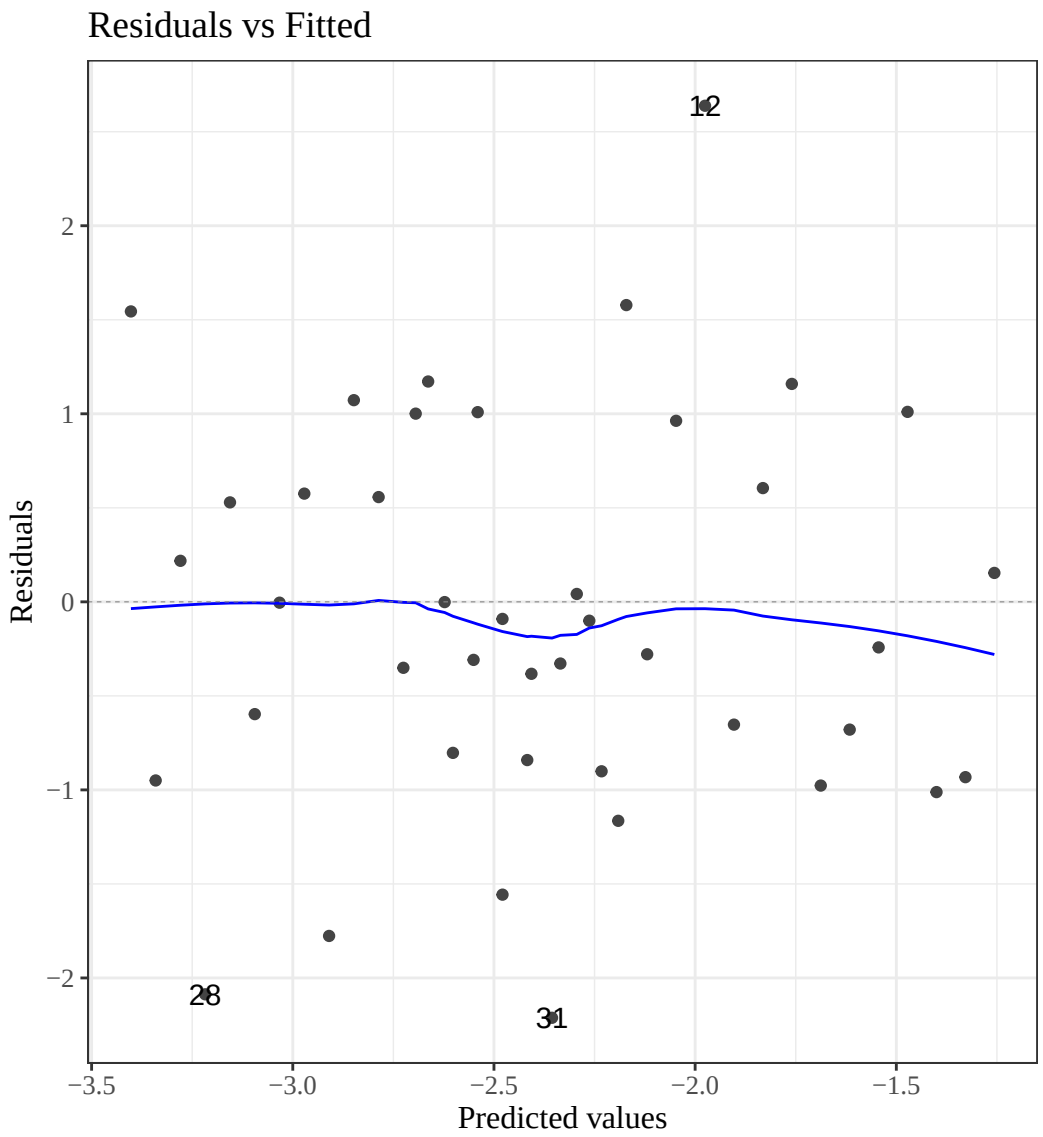
$$\begin{aligned}\hat{\pi}_k &\stackrel{\text{def}}{=} \hat{\pi}(x_k)\\ &\stackrel{\text{def}}{=} \hat{P}(Y=1|X=x_k)\\ &\stackrel{\text{def}}{=} \text{expit}(x_i'\hat{\beta})\\ &\stackrel{\text{def}}{=} \text{expit}(\hat{\beta}_0 + \sum_{j=1}^{p}\hat{\beta}_j x_{ij})\end{aligned}$$

Exm

Example 9.2 (Pearson residuals for the WCGS CHD model)**.** Let's take a look at the Pearson residuals (Definition 9.2) for our CHD model from the WCGS data (graphed against the fitted values on the logit scale):

```
library(ggfortify)
```

```
autoplot(chd_glm_contrasts_grouped, which = 1, ncol = 1) |> print()
```

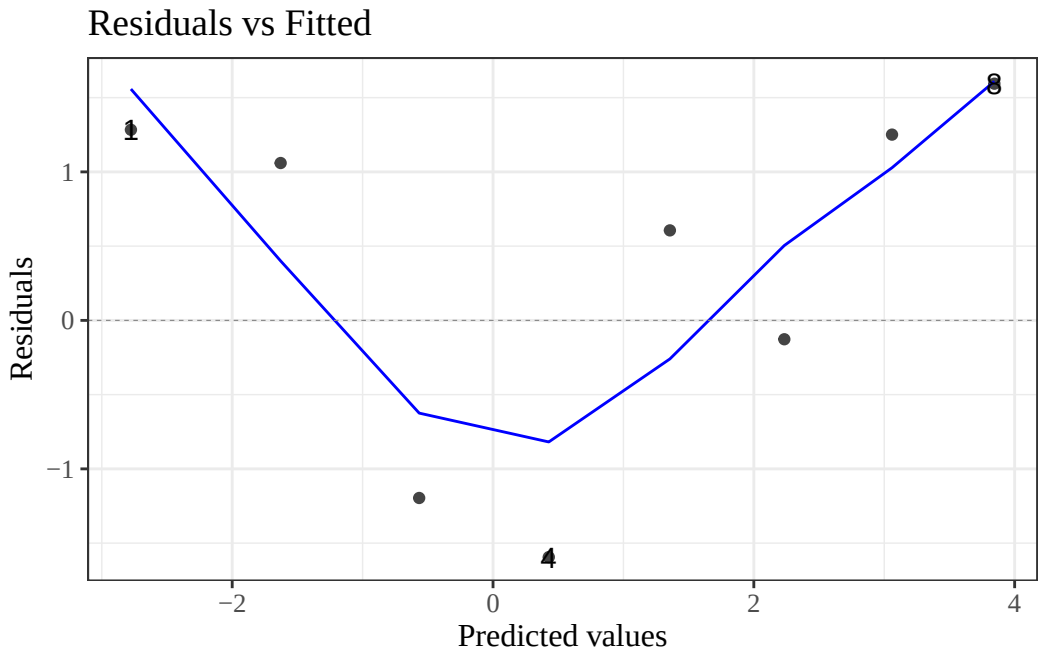


The fan-shape is gone, and these residuals don't show any obvious signs of model fit issues.

Exm

Example 9.3 (Counterexample: Pearson residuals for the beetles model)**.** If we create the same plot for the `beetles` model, we see some strong evidence of a lack of fit; this plot is a **counterexample** to a well-fitting model:

```
library(glmx)
library(dplyr)
data(BeetleMortality)
beetles <- BeetleMortality |>
  mutate(
    pct = died / n,
    survived = n - died,
    dose_c = dose - mean(dose)
  )
beetles_glm_grouped <- beetles |>
  glm(
    formula = cbind(died, survived) ~ dose,
    family = "binomial"
  )
autoplot(beetles_glm_grouped, which = 1, ncol = 1) |> print()
```



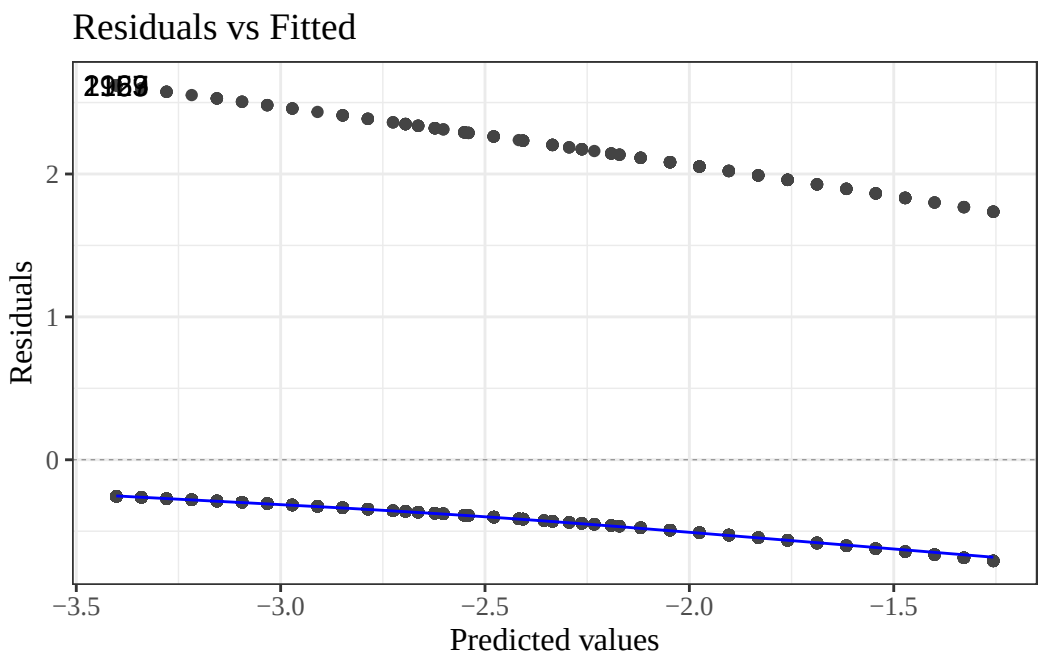
The clear pattern in the residuals signals that the straight-line logit model does not fit the beetle-mortality data well.

Pearson residuals with individual (ungrouped) data

What happens if we try to compute residuals without grouping the data by covariate pattern?

```
library(ggfortify)
```

```
chd_glm_strat <- glm(
  "formula" = chd69 == "Yes" ~ dibpat + dibpat:age - 1,
  "data" = wcgs,
  "family" = binomial(link = "logit")
)

autoplot(chd_glm_strat, which = 1, ncol = 1) |> print()
```



This residuals-vs-fitted plot is meaningless: with ungrouped (individual-level) data, each covariate pattern contains only one observation ($n_k = 1$), so each residual can take only one of two possible values, both determined by the fitted probability $\hat{\pi}(x_k)$ (one value when $y_k = 1$ and another when $y_k = 0$). With a single binary observation per covariate pattern, the residuals have no meaningful reference distribution to assess the model fit against; this lack of a reference distribution is why logistic regression residuals only work for grouped data, with multiple observations per covariate pattern.

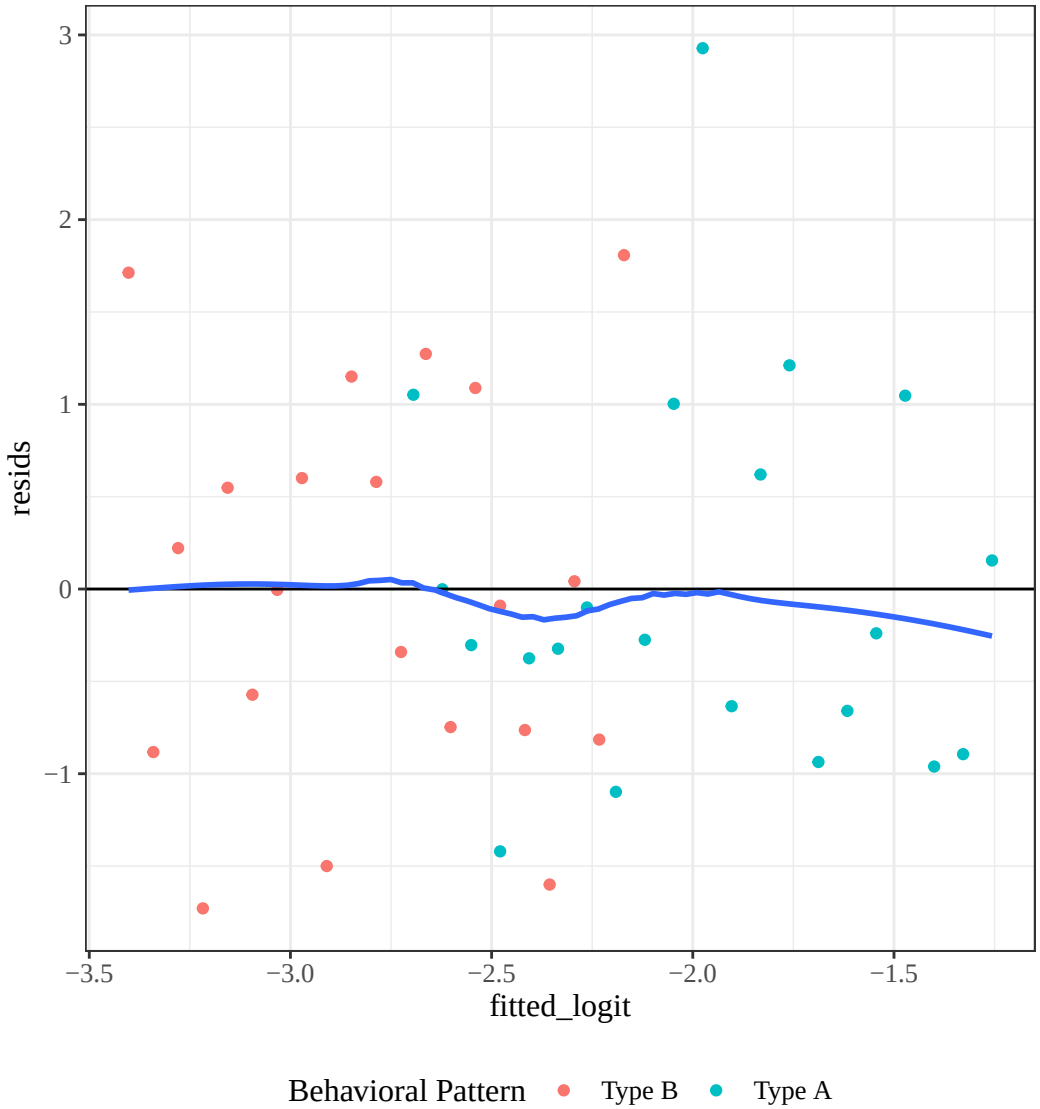
Residuals plot by hand

If you want to check your understanding of what these residual plots are, try building them yourself:

```
wcgs_grouped <-
  wcgs_grouped |>
  mutate(
    fitted = chd_glm_contrasts_grouped |> fitted(),
    fitted_logit = fitted |> logit(),
    resids = chd_glm_contrasts_grouped |> resid(type = "pearson")
  )

wcgs_resid_plot1 <-
  wcgs_grouped |>
  ggplot(
    mapping = aes(
      x = fitted_logit,
      y = resids
    )
  ) +
  geom_point(
    aes(col = dibpat)
  ) +
  geom_hline(yintercept = 0) +
  geom_smooth(
    se = FALSE,
    method.args = list(
      span = 2 / 3,
      degree = 1,
```

```
      family = "symmetric",
      iterations = 3,
      surface = "direct"
    ),
    method = stats::loess
  )
# plot.glm and autoplot use stats::lowess, which is hard to use with
# geom_smooth and hard to match exactly;
# see https://support.bioconductor.org/p/2323/
```

```
wcgs_resid_plot1 |> print()
```



Pearson chi-squared goodness of fit test

The Pearson chi-squared goodness of fit statistic is:

$$X^2 = \sum_{k=1}^{q} X_k^2$$

Under the null hypothesis that the model in question is correct (i.e., sufficiently complex), $X^2 \dot{\sim} \chi^2(q-p)$.

```r
x_pearson <- chd_glm_contrasts_grouped |>
  resid(type = "pearson")

chisq_stat <- sum(x_pearson^2)

pval <- pchisq(
  chisq_stat,
  lower = FALSE,
  df = length(x_pearson) - length(coef(chd_glm_contrasts_grouped))
)
```

For our CHD model, the p-value for this test is 0.265236; no significant evidence of a lack of fit at the 0.05 level.

Definition 9.3 (Leverage for a GLM)**.** For a logistic (or other) generalized linear model fit to q covariate patterns with p parameters, the **weighted hat matrix** is

$$\underset{q\times q}{\mathbf{H}} \overset{\text{def}}{=} \underbrace{\mathbf{W}^{1/2}}_{q\times q}\ \underset{q\times p}{\mathbf{X}} \left(\underbrace{\mathbf{X}^\top \mathbf{W}\, \mathbf{X}}_{p\times p} \right)^{-1} \underset{p\times q}{\mathbf{X}^\top}\ \underbrace{\mathbf{W}^{1/2}}_{q\times q}$$

where $\underset{q\times q}{\mathbf{W}} \overset{\text{def}}{=} \operatorname{diag}(n_k\, \hat{\pi}_k\, (1-\hat{\pi}_k))$ is the diagonal matrix of binomial weights, and the **leverage** h_k of covariate pattern k is the k-th diagonal element of $\mathbf{H}$.
This weighted hat matrix is the GLM analog of the linear-model hat matrix[a]; the weight matrix $\mathbf{W}$ reduces to the identity in the ordinary-least-squares case.

[a] Linear-models-overview.qmd#def-hat-matrix

Definition 9.4 (Standardized Pearson residual)**.** Especially for small data sets, we might want to adjust our residuals for leverage (since outliers in X add extra variance to the residuals):

$$r_{P_k} \overset{\text{def}}{=} \frac{X_k}{\sqrt{1-h_k}}$$

where X_k is the Pearson residual (Definition 9.2) and h_k is the leverage of covariate pattern k (Definition 9.3). The functions `autoplot()` and `plot.lm()` use these for some of their graphs.

9.0.4 Deviance residuals

For large sample sizes, the Pearson and deviance residuals will be approximately the same. For small sample sizes, the deviance residuals from covariate patterns with small sample sizes can be unreliable (high variance).

Definition 9.5 (Deviance residual)**.** Here $\ell_{\text{full}}(x_k)$ is the saturated model's log-likelihood contribution from covariate pattern k, so that summing over the covariate patterns gives the full log-likelihood:

$$\ell_{\text{full}} \overset{\text{def}}{=} \sum_k \ell_{\text{full}}(x_k)$$

The **deviance residual** for covariate pattern k is

$$d_k \overset{\text{def}}{=} \operatorname{sign}(y_k - n_k \hat{\pi}_k) \left\{ \sqrt{2[\ell_{\text{full}}(x_k) - \ell(\hat{\beta}; x_k)]} \right\}$$

Standardized deviance residuals

Analogously to the standardized Pearson residual (Definition 9.4), we can adjust the deviance

residual for leverage:

$$r_{D_k} \stackrel{\text{def}}{=} \frac{d_k}{\sqrt{1-h_k}}$$

where h_k is the leverage of covariate pattern k (Definition 9.3).

9.0.5 Diagnostic plots

Definition 9.6.

Cook's distance for a GLM
Some of the `autoplot()` diagnostic panels display **Cook's distance**, which measures how influential each covariate pattern is on the fit. For a generalized linear model, the Cook's distance for covariate pattern k combines the standardized Pearson residual (Definition 9.4) and the leverage (Definition 9.3):

$$D_k \stackrel{\text{def}}{=} \frac{(r_{P_k})^2}{p} \cdot \frac{h_k}{1-h_k}$$

where p is the number of parameters. This diagnostic is the GLM analog of Cook's distance[a] for linear models, and shares its closed-form structure.

[a] Linear-models-overview.qmd#def-cooks-distance

Exm

Example 9.4 (Diagnostics for the WCGS CHD model)**.** Let's take a look at the full set of `autoplot()` diagnostics now for our `CHD` model:

```
chd_glm_contrasts_grouped |>
  autoplot(which = 1:6) |>
  print()
```

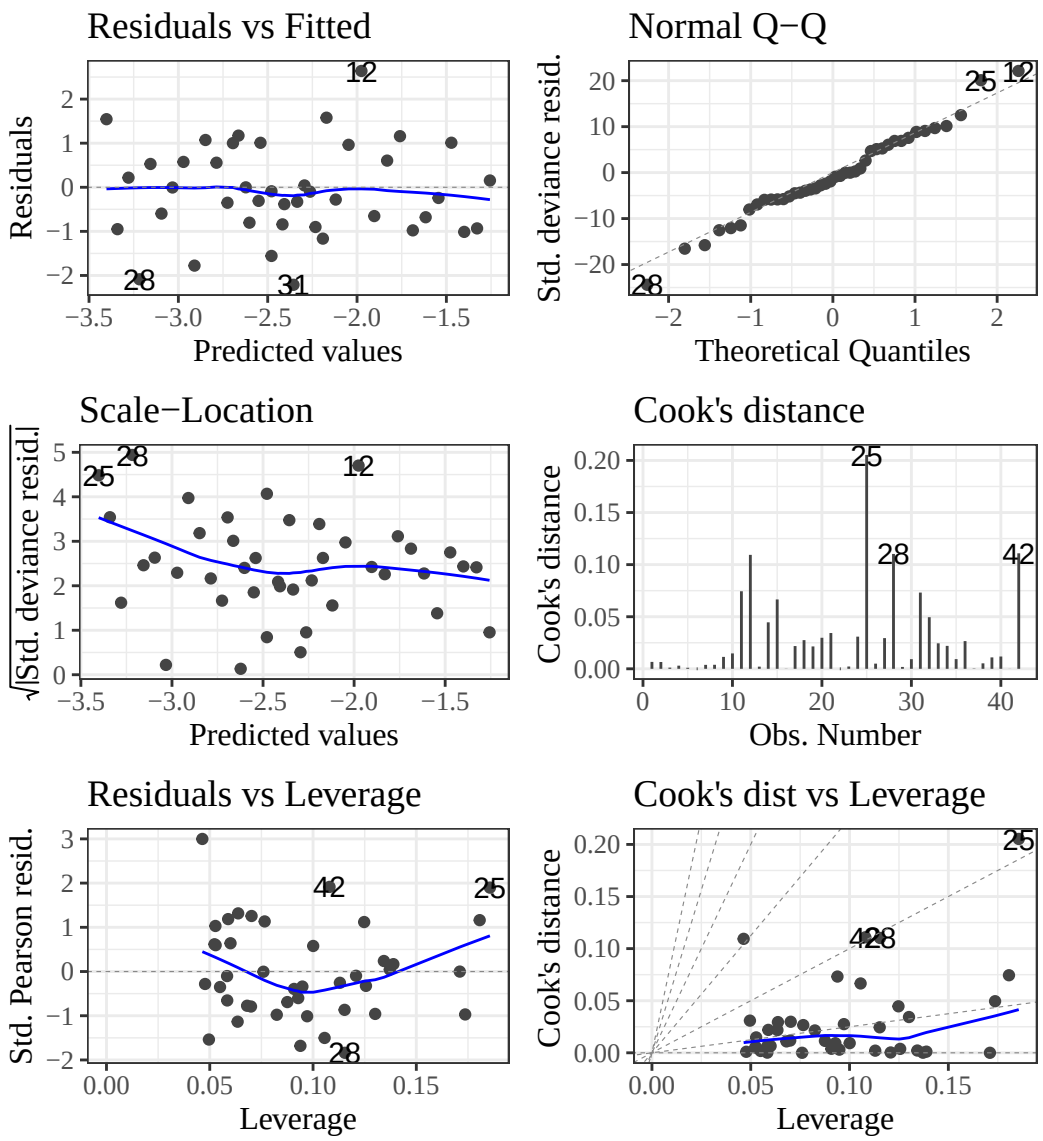


Figure 15: Diagnostics for CHD model

Things look pretty good here. The QQ plot is still usable: when the number of observations at each covariate pattern is large, the grouped residuals should be approximately Gaussian.

Exm

Example 9.5 (Diagnostics for the beetles model)**.** Let's look at the beetles model diagnostic plots for comparison:

```
beetles_glm_grouped |>
  autoplot(which = 1:6) |>
  print()
```

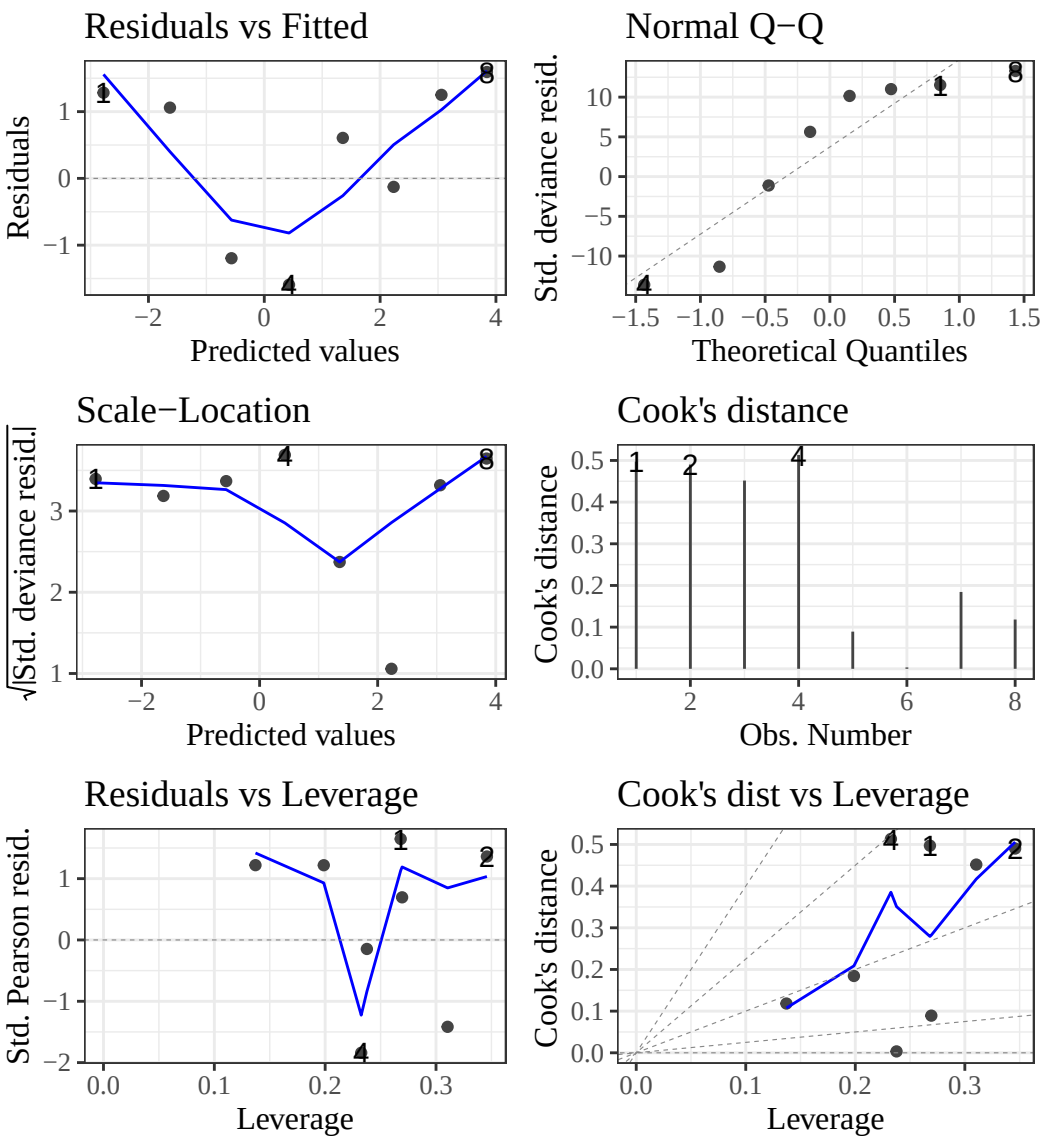


Figure 16: Diagnostics for logistic model of `BeetleMortality` data

Hard to tell much from so little data, but there might be some issues here.

10 Alternatives to reporting odds ratios

10.1 Objections to odds ratios

Some scholars have raised objections to the use of odds ratios as an effect measurement (Sackett et al. 1996; Norton et al. 2024).

As we saw in the figure comparing risk differences, risk ratios, and odds ratios[19], the odds ratio is not very closely correlated with the risk difference, and the risk difference is typically the metric that ultimately matters for policy decisions.

Another objection is that odds ratios (and risk ratios, and risk differences) depend on the set of covariates in a logistic regression model, even when those covariates are independent of the exposure of interest and do not interact with that exposure. For example, consider the following model:

[19] binary-outcome-associations.qmd#fig-rd-rr-or

$$\mathrm{P}(Y = y|X = x, C = c) = \pi(x, c)^y(1 - \pi(x, c))^{1-y}$$

$$\pi(x, c) = \text{expit}\{\eta_0 + \beta_X x + \beta_C c\} \tag{44}$$

Then:

$$\begin{aligned}
\mathrm{E}[Y|X = x] &= \mathrm{E}[\mathrm{E}[Y|X, C]|X = x] && \text{(law of total expectation)} \\
&= \mathrm{E}[\pi(X, C)|X = x] && \text{(Bernoulli conditional mean: } \mathrm{E}[Y|X, C] = \pi(X, C)) \\
&= \mathrm{E}[\text{expit}\{\eta_0 + \beta_X X + \beta_C C\}|X = x] && \text{(model definition of } \pi) \\
&= \mathrm{E}[\text{expit}\{\eta_0 + \beta_X x + \beta_C C\}|X = x] && \text{(conditioning on } X = x \text{ fixes } X \text{ at } x) \\
&= \int_c \text{expit}\{\eta_0 + \beta_X x + \beta_C c\}\ \mathrm{p}(C = c|X = x)\ \partial c && \text{(conditional expectation as an integral over } C\text{'s condition} \\
&= \int_c \pi(x, c)\ \mathrm{p}(C = c|X = x)\ \partial c && \text{(model definition of } \pi\text{, in reverse)}
\end{aligned}$$

Since the expit{} function is nonlinear, we can't change the order of the expectation and expit{} operators:

$$\mathrm{E}[\text{expit}\{\eta_0 + \beta_X X + \beta_C C\}|X] \neq \text{expit}\{\mathrm{E}[\eta_0 + \beta_X X + \beta_C C]|X\}$$

In contrast, consider a model with an identity link function:

$$\mathrm{P}(Y = y|X = x, C = c) = \pi(x, c)^y(1 - \pi(x, c))^{1-y}$$

$$\pi(x, c) = \eta_0 + \beta_X x + \beta_C c \tag{45}$$

Then:

$$\begin{aligned}
\mathrm{E}[Y|X = x] &= \mathrm{E}[\mathrm{E}[Y|X, C]|X = x] && \text{(law of total expectation)} \\
&= \mathrm{E}[\pi(X, C)|X = x] && \text{(Bernoulli conditional mean: } \mathrm{E}[Y|X, C] = \pi(X, C)) \\
&= \mathrm{E}[\eta_0 + \beta_X X + \beta_C C|X = x] && \text{(identity-link model definition of } \pi) \\
&= \mathrm{E}[\eta_0|X = x] + \mathrm{E}[\beta_X X|X = x] + \mathrm{E}[\beta_C C|X = x] && \text{(linearity of expectation)} \\
&= \eta_0 + \mathrm{E}[\beta_X X|X = x] + \mathrm{E}[\beta_C C|X = x] && \text{(expectation of a constant)} \\
&= \eta_0 + \beta_X\,\mathrm{E}[X|X = x] + \beta_C\,\mathrm{E}[C|X = x] && \text{(constant factors move outside expectations)} \\
&= \eta_0 + \beta_X x + \beta_C\,\mathrm{E}[C|X = x] && (\mathrm{E}[X|X = x] = x) \\
&= (\eta_0 + \beta_C\,\mathrm{E}[C|X = x]) + \beta_X x && \text{(commute and regroup the terms)}
\end{aligned}$$

If $C \perp\!\!\!\perp X$, then $\mathrm{E}[C|X = x] = \mathrm{E}[C]$, and we can simplify further:

$$\begin{aligned}
\mathrm{E}[Y|X = x] &= (\eta_0 + \beta_C\,\mathrm{E}[C|X = x]) + \beta_X x && \text{(marginal mean under the identity link)} \\
&= (\eta_0 + \beta_C\,\mathrm{E}[C]) + \beta_X x && (C \perp\!\!\!\perp X, \text{ so } \mathrm{E}[C|X = x] = \mathrm{E}[C]) \\
&= \eta_0^* + \beta_X x && (\text{definition: } \eta_0^* \stackrel{\text{def}}{=} \eta_0 + \beta_C\,\mathrm{E}[C])
\end{aligned}$$

Then:

$$\begin{aligned}
\frac{\partial}{\partial x}\,\mathrm{E}[Y|X=x] &= \frac{\partial}{\partial x}(\eta_0^* + \beta_X x) && \text{(marginal mean when } C \perp\!\!\!\perp X\text{)}\\
&= \frac{\partial}{\partial x}\eta_0^* + \frac{\partial}{\partial x}(\beta_X x) && \text{(linearity of differentiation)}\\
&= 0 + \frac{\partial}{\partial x}(\beta_X x) && (\eta_0^* \text{ does not depend on } x)\\
&= 0 + \beta_X && \text{(derivative of a linear term)}\\
&= \beta_X && \text{(additive identity)}
\end{aligned}$$

The conditional slope is the same:

$$\begin{aligned}
\frac{\partial}{\partial x}\,\mathrm{E}[Y|X=x, C=c] &= \frac{\partial}{\partial x}(\eta_0 + \beta_X x + \beta_C c) && \text{(identity-link model definition of } \pi\text{)}\\
&= 0 + \beta_X + 0 && \text{(term-by-term: only } \beta_X x \text{ depends on } x\text{)}\\
&= \beta_X && \text{(additive identity)}
\end{aligned}$$

Therefore:

$$\frac{\partial}{\partial x}\,\mathrm{E}[Y|X=x] = \beta_X = \frac{\partial}{\partial x}\,\mathrm{E}[Y|X=x, C=c]$$

In other words, for a model with an identity link function, if covariates X and C are independent, then the slope with respect to X doesn't depend on whether C is included in the model (and an analogous result holds if X is discrete or categorical).

Exercise 10.1. What are the expressions for $\mathrm{E}[Y|X=x]$ and $\frac{\partial}{\partial x}\,\mathrm{E}[Y|X=x]$ for the identity-link model in Equation 45, if $\mathrm{E}[C|X=x] = \gamma_0 + \gamma_x x$?

Solution

Solution 10.1.

$$\begin{aligned}
\mathrm{E}[Y|X=x] &= \mathrm{E}[\pi(x,C)|X=x] && \text{(law of total expectation, conditional on } X=x\text{)}\\
&= \mathrm{E}[\eta_0 + \beta_X x + \beta_C C|X=x] && \text{(identity-link model)}\\
&= \eta_0 + \beta_X x + \beta_C\,\mathrm{E}[C|X=x] && \text{(linearity of expectation)}\\
&= \eta_0 + \beta_X x + \beta_C(\gamma_0 + \gamma_x x) && \text{(substitute } \mathrm{E}[C|X=x] = \gamma_0 + \gamma_x x\text{)}\\
&= \eta_0 + \beta_X x + \beta_C\gamma_0 + \beta_C\gamma_x x && \text{(distribute)}\\
&= (\eta_0 + \beta_C\gamma_0) + (\beta_X + \beta_C\gamma_x)x && \text{(commute and regroup the terms)}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial x}\,\mathrm{E}[Y|X=x] &= \frac{\partial}{\partial x}\{(\eta_0 + \beta_C\gamma_0) + (\beta_X + \beta_C\gamma_x)x\} && \text{(marginal mean from the derivation of } \mathrm{E}[Y|X=x]\text{)}\\
&= \frac{\partial}{\partial x}(\eta_0 + \beta_C\gamma_0) + \frac{\partial}{\partial x}\{(\beta_X + \beta_C\gamma_x)x\} && \text{(linearity of differentiation)}\\
&= \beta_X + \beta_C\gamma_x && \text{(constants have derivative 0; derivative of a linear term)}
\end{aligned}$$

The marginal mean is still linear in x, but its slope $\beta_X + \beta_C\gamma_x$ differs from the conditional slope β_X unless $\beta_C = 0$ or $\gamma_x = 0$.

Exercise 10.2. What are the expressions for $\mathrm{E}[Y|X=x]$ and $\frac{\partial}{\partial x}\,\mathrm{E}[Y|X=x]$, if instead of the identity-link model in Equation 45,

$$\pi(x,c) = \eta_0 + \beta_X x + \beta_C c + \beta_{XC} xc$$

and $\mathrm{E}[C|X=x] = \gamma_0 + \gamma_x x$?

Solution

Solution 10.2.

$$
\begin{aligned}
\mathrm{E}[Y|X=x] &= \mathrm{E}[\pi(x,C)|X=x] && \text{(law of total expectation, conditional on } X=x\text{)} \\
&= \mathrm{E}[\eta_0 + \beta_X x + \beta_C C + \beta_{XC} x C | X=x] && \text{(interaction model)} \\
&= \eta_0 + \beta_X x + \beta_C \,\mathrm{E}[C|X=x] + \beta_{XC} x \,\mathrm{E}[C|X=x] && \text{(linearity of expectation)} \\
&= \eta_0 + \beta_X x + (\beta_C + \beta_{XC} x)\,\mathrm{E}[C|X=x] && \text{(factor out } \mathrm{E}[C|X=x]\text{)} \\
&= \eta_0 + \beta_X x + (\beta_C + \beta_{XC} x)(\gamma_0 + \gamma_x x) && \text{(substitute } \mathrm{E}[C|X=x] = \gamma_0 + \gamma_x x\text{)} \\
&= \eta_0 + \beta_X x + \beta_C \gamma_0 + \beta_C \gamma_x x + \beta_{XC} \gamma_0 x + \beta_{XC} \gamma_x x^2 && \text{(distribute)} \\
&= (\eta_0 + \beta_C \gamma_0) + (\beta_X + \beta_C \gamma_x + \beta_{XC} \gamma_0) x + \beta_{XC} \gamma_x x^2 && \text{(commute and regroup the terms)}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial x}\mathrm{E}[Y|X=x] &= \frac{\partial}{\partial x}\{(\eta_0 + \beta_C\gamma_0) + (\beta_X + \beta_C\gamma_x + \beta_{XC}\gamma_0)x + \beta_{XC}\gamma_x x^2\} && \text{(marginal mean from the derivat} \\
&= \frac{\partial}{\partial x}(\eta_0 + \beta_C\gamma_0) + \frac{\partial}{\partial x}\{(\beta_X + \beta_C\gamma_x + \beta_{XC}\gamma_0)x\} + \frac{\partial}{\partial x}(\beta_{XC}\gamma_x x^2) && \text{(linearity of differentiation)} \\
&= (\beta_X + \beta_C\gamma_x + \beta_{XC}\gamma_0) + 2\beta_{XC}\gamma_x x && \text{(constants have derivative 0; deri}
\end{aligned}
$$

So, answering the hint: yes — adding the interaction term changes the functional form of $\mathrm{E}[Y|X=x]$. The marginal mean is now quadratic in x (whenever $\beta_{XC} \neq 0$ and $\gamma_x \neq 0$), even though the conditional model is linear in x for each fixed c, and the marginal slope now varies with x.

10.2 Deriving risk ratios and risk differences from logistic regression models

If you want to report risk differences or risk ratios instead of odds ratios, you can obtain estimates from logistic regression models, as long as you didn't stratify sampling by the outcome; in other words, not in case-control studies (see Odds Ratios in Case-Control Studies[20]).

10.2.1 Computing marginal risk differences

Note

This section is **optional for 2026** and will not be tested on exams.

Conceptual approach

To compute risk ratios or risk differences from logistic regression models:

- Apply the expit function (definition[21]) to the linear predictor for each covariate pattern to compute the (estimated) risks,
- Then take the differences or ratios of the risks, as needed.

For example, suppose we have a logistic regression model with a binary exposure A and a covariate Z:

$$\mathrm{logit}\{\Pr(Y=1|A,Z)\} = \beta_0 + \beta_A A + \beta_Z Z \tag{46}$$

[20] binary-outcome-associations.qmd#sec-CC-studies
[21] binary-outcome-associations.qmd#def-expit

Definition 10.1 (Conditional predicted risk)**.** The **conditional predicted risk** (or **conditional risk**) for an individual with exposure $A = a$ and covariate $Z = z$ is the population probability

$$\pi(a, z) \stackrel{\text{def}}{=} \Pr(Y = 1 \mid A = a, Z = z).$$

Under the logistic model in Equation 46, this population risk equals

$$\pi(a, z) = \text{expit}\{\beta_0 + \beta_A a + \beta_Z z\}.$$

Definition 10.2 (Conditional risk difference)**.** The **conditional risk difference** comparing exposure levels $a = 1$ vs. $a = 0$ at a fixed covariate value z is, using the conditional predicted risk (Definition 10.1):

$$\text{RD}(z) \stackrel{\text{def}}{=} \pi(1, z) - \pi(0, z)$$

Definition 10.3 (Conditional risk ratio)**.** The **conditional risk ratio** comparing exposure levels $a = 1$ vs. $a = 0$ at a fixed covariate value z is, using the conditional predicted risk (Definition 10.1):

$$\text{RR}(z) \stackrel{\text{def}}{=} \frac{\pi(1, z)}{\pi(0, z)}$$

Exm

Example 10.1 (Numerical example: conditional RD and RR at a fixed covariate value)**.** Using the WCGS data with `dibpat` (behavior pattern) as the exposure and `chd69` (coronary heart disease by 1969) as the outcome, adjusted for `age`:

```
library(dplyr)
library(haven)

wcgs_clean <- rmb::wcgs |>
  mutate(
    across(where(is.labelled), haven::as_factor),
    dibpat = dibpat |> relevel(ref = "Type B"),
    chd69_binary = as.numeric(chd69 == "Yes")
  ) |>
  filter(!is.na(chd69_binary), !is.na(dibpat), !is.na(age))

fit <- glm(
  chd69_binary ~ dibpat + age,
  data = wcgs_clean,
  family = binomial(link = "logit")
)
```

At age $z = 50$ (within the WCGS sample's observed age range):

```r
z <- 50
dibpat_levels <- levels(wcgs_clean$dibpat)
wcgs_cf_a <- data.frame(dibpat = factor("Type A", levels = dibpat_levels), age = z)
wcgs_cf_b <- data.frame(dibpat = factor("Type B", levels = dibpat_levels), age = z)

pi_hat_1 <- predict(fit, newdata = wcgs_cf_a, type = "response")
pi_hat_0 <- predict(fit, newdata = wcgs_cf_b, type = "response")

rd_hat_z <- pi_hat_1 - pi_hat_0
rr_hat_z <- pi_hat_1 / pi_hat_0

tibble::tibble(
  quantity = c(
    "$\\hat\\pi(1, 50)$", "$\\hat\\pi(0, 50)$",
    "$\\widehat{\\text{RD}}(50)$", "$\\widehat{\\text{RR}}(50)$"
  ),
  value = c(pi_hat_1, pi_hat_0, rd_hat_z, rr_hat_z)
) |>
  knitr::kable(digits = 3)
```

quantity	value
$\hat{\pi}(1, 50)$	0.129
$\hat{\pi}(0, 50)$	0.063
$\widehat{\text{RD}}(50)$	0.067
$\widehat{\text{RR}}(50)$	2.066

These values are *conditional* estimates: both hold age fixed at 50 and compare only the two exposure levels at that one age.

Marginal vs. conditional effects

Note that these risk differences and risk ratios are **conditional** on the covariate value z. In many applications, we want to summarize the effect across the population, which requires computing a **marginal** risk difference or risk ratio.

Definition 10.4 (Observed marginal risk)**.** The **observed marginal risk** under exposure level $A = a$ is:

$$\pi(a) \stackrel{\text{def}}{=} \Pr(Y = 1 \mid A = a)$$

Exm

Example 10.2 (Illustration: observed marginal risk)**.** Continuing the WCGS example (Example 10.1):

```r
pi1_direct <- mean(wcgs_clean$chd69_binary[wcgs_clean$dibpat == "Type A"])
```

The observed marginal risk among the Type A men ($A = 1$) is simply the proportion of them who developed CHD by 1969:

$$\pi(1) = \Pr(Y = 1 \mid A = 1) = 0.112$$

This definition collapses the outcome over everything else in one step. When the model also conditions on covariates Z, we can unpack $\pi(a)$ as the covariate-specific risk $\pi(a, z)$ averaged over the distribution of Z among those with $A = a$:

Theorem 10.1 (Marginal risk as a conditional expectation)**.**

$$\pi(a) = \mathrm{E}[\pi(a, Z) \mid A = a]$$

where $\pi(a, z)$ is the conditional predicted risk (Definition 10.1).

Proof

Proof.

$$\begin{aligned}
\pi(a) &= \Pr(Y = 1 \mid A = a) && \text{[definition of observed marginal risk]} \\
&= \int \Pr(Y = 1 \mid Z = z, A = a)\, f_{Z|A}(z \mid a)\, dz && \text{[law of total probability]} \\
&= \int \pi(a, z)\, f_{Z|A}(z \mid a)\, dz && \text{[definition of conditional risk]} \\
&= \mathrm{E}[\pi(a, Z) \mid A = a] && \text{[definition of conditional expectation]}
\end{aligned}$$

□

Exm

Example 10.3 (Observed marginal risk by standardization)**.** Take Z to be an indicator for age 50 or older. Among the WCGS men with $A = 1$ (Type A):

```
wcgs_clean$age_ge50 <- as.numeric(wcgs_clean$age >= 50)
wcgs_exposed <- wcgs_clean[wcgs_clean$dibpat == "Type A", ]
pi1_z0 <- mean(wcgs_exposed$chd69_binary[wcgs_exposed$age_ge50 == 0])
pi1_z1 <- mean(wcgs_exposed$chd69_binary[wcgs_exposed$age_ge50 == 1])
p_z1_a1 <- mean(wcgs_exposed$age_ge50)
```

the age-specific risks are $\pi(1, 0) = 0.089$ (under 50) and $\pi(1, 1) = 0.159$ (50 or older), and $\Pr(Z = 1 \mid A = 1) = 0.329$ of the Type A men are 50 or older. Then Theorem 10.1 gives:

$$\begin{aligned}
\pi(1) &= \mathrm{E}[\pi(1, Z) \mid A = 1] && \text{[observed marginal risk theorem]} \\
&= (1 - 0.329)(0.089) + (0.329)(0.159) && \text{[substitute stratum risks and weights]} \\
&= 0.112 && \text{[arithmetic]}
\end{aligned}$$

matching the direct estimate $\pi(1) = 0.112$ from Example 10.2 (up to rounding).

Note

The rest of this section previews causal-inference concepts — potential outcomes[a], consistency[b], exchangeability[c], and positivity[d] — that are formally defined and developed fully in the causal inference chapter[e].

[a] causal-inference.qmd#def-potential-outcomes
[b] causal-inference.qmd#def-consistency
[c] causal-inference.qmd#def-exchangeability
[d] causal-inference.qmd#def-positivity
[e] causal-inference.qmd

In causal inference, potential outcomes[22] Y^a represent the outcome each unit would have under exposure level a. Population-level causal estimands summarize these potential outcomes:

Definition 10.5 (Potential mean)**.** The **potential mean** (or **mean potential outcome**) under treatment $A = a$ is the expected value of the potential outcome Y^a:

$$\mu_a \stackrel{\text{def}}{=} \mathrm{E}[Y^a]$$

[22] causal-inference.qmd#def-potential-outcomes

The potential mean is what the average outcome would be if the entire population were assigned treatment a.

Exm

Example 10.4 (Illustration: potential mean)**.** Suppose, hypothetically, that if every man in the WCGS cohort had exhibited Type A behavior pattern ($a = 1$), the fraction who would go on to develop CHD by 1969 were close to the fraction actually observed among the men who really did have that behavior pattern, $\pi(1) = 0.112$ (Example 10.2). Under that supposition, the potential mean is $\mu_1 = \mathrm{E}[Y^1] = 0.112$.

For a binary outcome, the potential mean is a probability, which we give its own name:

Definition 10.6 (Causal marginal risk)**.** For a binary outcome $Y \in \{0, 1\}$, the potential mean (Definition 10.5) is a probability, which we call the **causal marginal risk** (also: **potential probability**, **potential marginal risk**) under exposure level a:

$$\pi_a \stackrel{\text{def}}{=} \Pr(Y^a = 1)$$

We use the term **causal marginal risk** throughout.

Exm

Example 10.5 (Illustration: causal marginal risk)**.** Continuing Example 10.4, under that same supposition the causal marginal risk is $\pi_1 = \Pr(Y^1 = 1) = 0.112$ — the fraction who would develop CHD if every man in the cohort had Type A behavior pattern.

Theorem 10.2 (Causal marginal risk equals potential mean)**.** *The matching values in Example 10.4 and Example 10.5 are no coincidence: the causal marginal risk (Definition 10.6) is a special case of the potential mean (Definition 10.5):*

$$\pi_a = \mathrm{E}[Y^a]$$

Proof

Proof. Since Y^a is binary ($Y^a \in \{0, 1\}$):

$$\begin{aligned}
\mathrm{E}[Y^a] &= 1 \cdot \Pr(Y^a = 1) + 0 \cdot \Pr(Y^a = 0) && \text{[expectation of a binary variable]} \\
&= \Pr(Y^a = 1) && \text{[arithmetic]} \\
&= \pi_a && \text{[definition of causal marginal risk]}
\end{aligned}$$

□

Exm

Example 10.6 (Illustration: causal marginal risk equals potential mean)**.** With $\pi_1 = 0.112$ as in Example 10.5, reading the potential mean as an expectation gives the same value:

$$\mathrm{E}[Y^1] = 1 \cdot 0.112 + 0 \cdot 0.888 = 0.112 = \pi_1.$$

Under causal identification assumptions (see Consistency[23], Conditional Exchangeability[24], and Positivity[25]), the causal marginal risk π_a (Definition 10.6) can be expressed in terms of observed data:

[23] causal-inference.qmd#def-consistency
[24] causal-inference.qmd#def-exchangeability
[25] causal-inference.qmd#def-positivity

Theorem 10.3 (Causal marginal risk under identification assumptions)**.** *Under Consistency*[a]*, Conditional Exchangeability*[b]*, and Positivity*[c]*:*

$$\pi_a = \mathrm{E}[\pi(a, Z)]$$

where the expectation is over the ***marginal*** *distribution of Z (not the conditional distribution $f_{Z|A}(z \mid a)$).*

Proof

Proof.

$$\begin{aligned}
\pi_a &= \Pr(Y^a = 1) && \text{[definition of causal marginal risk]} \\
&= \int \Pr(Y^a = 1 \mid Z = z)\, f_Z(z)\, dz && \text{[law of total probability]} \\
&= \int \Pr(Y^a = 1 \mid A = a, Z = z)\, f_Z(z)\, dz && \text{[by conditional exchangeability]} \\
&= \int \Pr(Y = 1 \mid A = a, Z = z)\, f_Z(z)\, dz && \text{[by consistency]} \\
&= \int \pi(a, z)\, f_Z(z)\, dz && \text{[definition of conditional risk]} \\
&= \mathrm{E}[\pi(a, Z)] && \text{[definition of expectation]}
\end{aligned}$$

Conditional exchangeability licenses adding the conditioning on $A = a$; positivity ($\Pr(A = a \mid Z = z) > 0$ wherever $f_Z(z) > 0$) guarantees that conditional risk is defined in every stratum; and consistency replaces the potential outcome Y^a with the observed Y. □

[a] causal-inference.qmd#def-consistency
[b] causal-inference.qmd#def-exchangeability
[c] causal-inference.qmd#def-positivity

Exm

Example 10.7 (Causal marginal risk by standardization)**.** Continuing Example 10.3 with the same age-specific risks $\pi(1, 0) = 0.089$ and $\pi(1, 1) = 0.159$, now weight by the age distribution of the **whole** WCGS population rather than just the Type A men:

```
p_z1_pop <- mean(wcgs_clean$age_ge50)
```

$\Pr(Z = 1) = 0.287$ of the whole population is 50 or older — a smaller share than among the Type A men ($\Pr(Z = 1 \mid A = 1) = 0.329$), since behavior pattern and age are associated in this cohort. Then Theorem 10.3 gives:

$$\begin{aligned}
\pi_1 &= \mathrm{E}[\pi(1, Z)] && \text{[causal marginal risk theorem]} \\
&= (1 - 0.287)(0.089) + (0.287)(0.159) && \text{[substitute stratum risks and weights]} \\
&= 0.109 && \text{[arithmetic]}
\end{aligned}$$

This causal marginal risk differs (slightly) from the observed marginal risk $\pi(1) = 0.112$ because the Type A men skew somewhat older than the WCGS population as a whole.

We can now place the observed and causal marginal risks side by side:

Remark 10.1 (Observed vs. causal marginal risk). The two risks differ in what they define — one conditions on the **observed** exposure $A = a$, the other on a **potential outcome** Y^a — and, once unpacked through $\pi(a, z) = \Pr(Y = 1 \mid A = a, Z = z)$, in the distribution of Z they average over:

$$\begin{aligned}
\text{observed:} \quad \pi(a) &= \Pr(Y = 1 \mid A = a) && = \mathrm{E}[\pi(a, Z) \mid A = a] && = \int \pi(a, z)\, f_{Z|A}(z \mid a)\, dz \\
\text{causal:} \quad \pi_a &= \Pr(Y^a = 1) && = \mathrm{E}[\pi(a, Z)] && = \int \pi(a, z)\, f_Z(z)\, dz
\end{aligned}$$

The observed marginal risk $\pi(a)$ (Theorem 10.1) weights $\pi(a, z)$ by the covariate distribution $f_{Z|A}(z \mid a)$ **among those actually exposed to** $A = a$, so it inherits any association between A and Z (that is, any confounding).
The causal marginal risk π_a (Theorem 10.3) weights $\pi(a, z)$ by the **whole population's** covariate distribution $f_Z(z)$, which is the *same* for every exposure level a; that common weighting is what makes the causal contrast $\pi_1 - \pi_0$ an apples-to-apples comparison.
When A and Z are associated, the two risks diverge; they coincide in the absence of such association.

Corollary 10.1 (Observed and causal marginal risks coincide without confounding)**.** *Assume Consistency*[a]*, Conditional Exchangeability*[b]*, Positivity*[c]*, and* $Z \perp\!\!\!\perp A$*. Then the observed and causal marginal risks are equal:*

$$\pi(a) = \pi_a$$

Proof

Proof. When $Z \perp\!\!\!\perp A$ the conditional and marginal distributions of Z coincide, $f_{Z|A}(z \mid a) = f_Z(z)$, so Theorem 10.1 and Theorem 10.3 give:

$$\begin{aligned}
\pi(a) &= \mathrm{E}[\pi(a, Z) \mid A = a] && \text{[observed marginal risk theorem]} \\
&= \mathrm{E}[\pi(a, Z)] && \text{[no confounding]} \\
&= \pi_a && \text{[causal marginal risk theorem]}
\end{aligned}$$

□

[a] causal-inference.qmd#def-consistency
[b] causal-inference.qmd#def-exchangeability
[c] causal-inference.qmd#def-positivity

Under complete randomization of A, the design delivers $Z \perp\!\!\!\perp A$ together with the other two identification assumptions as parallel consequences — conditional exchangeability (randomized A is independent of the potential outcomes, hence also independent given Z) and positivity (every level has positive probability in each stratum) — so consistency and randomization alone suffice (Hernán and Robins 2020, chaps. 2–3). (Marginal independence $Z \perp\!\!\!\perp A$ on its own does not imply exchangeability; randomization is the stronger design assumption that yields all three.)

Exm

Example 10.8 (Illustration: coincidence under randomization)**.** Take the same age-specific risks $\pi(1, 0) = 0.089$ and $\pi(1, 1) = 0.159$ of Example 10.3, but now suppose A had been randomized, so the exposed would carry the population's age distribution instead of the actual, older-skewing one: $\Pr(Z = 1 \mid A = 1) = \Pr(Z = 1) = 0.287$. Then the observed marginal risk would be

$$\begin{aligned}
\pi(1) &= \mathrm{E}[\pi(1, Z) \mid A = 1] && \text{[observed marginal risk theorem]} \\
&= (1 - 0.287)(0.089) + (0.287)(0.159) && \text{[substitute randomized weights]} \\
&= 0.109 && \text{[arithmetic]}
\end{aligned}$$

equal to the causal marginal risk $\pi_1 = 0.109$ of Example 10.7, exactly as Corollary 10.1 predicts.

Two related procedures estimate these marginal risks from a fitted regression model:

Definition 10.7 (G-computation)**. G-computation** (also called the **g-formula** or **standardization**) is a procedure for estimating the causal marginal risk π_a (Definition 10.6) from observed data (Robins 1986; see also Hernán and Robins 2020, chap. 13):

1. Fit a regression model for $\hat{\pi}(a, z) = \widehat{\Pr}(Y = 1 \mid A = a, Z = z)$
2. For each observation i, predict the risk under exposure level a, writing $\hat{\pi}_{a,i} \stackrel{\text{def}}{=} \hat{\pi}(a, z_i)$ for short
3. Average these predicted risks over the observed covariate distribution:

$$\hat{\pi}_a \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \hat{\pi}_{a,i}$$

The following example carries out this procedure on the WCGS data.

Exm

Example 10.9 (Estimated causal marginal risks (WCGS))**.** Continuing with the same WCGS data and fitted model (`wcgs_clean` and `fit`) introduced in Example 10.1 — `dibpat` (behavior pattern) as the exposure and `chd69` (coronary heart disease by 1969) as the outcome, adjusted for `age`:
The **causal marginal risks** $\hat{\pi}_a = \frac{1}{n} \sum_{i=1}^{n} \hat{\pi}_{a,i}$ standardize the fitted risks $\hat{\pi}_{a,i} = \hat{\pi}(a, z_i)$ over the full covariate distribution (Definition 10.7):

```
dibpat_levels <- levels(wcgs_clean$dibpat)
wcgs_counterfactual_a <- wcgs_clean |>
  mutate(
    dibpat = factor("Type A", levels = dibpat_levels)
  )
wcgs_counterfactual_b <- wcgs_clean |>
  mutate(
    dibpat = factor("Type B", levels = dibpat_levels)
  )

pi1_hat <- mean(
  predict(fit, newdata = wcgs_counterfactual_a, type = "response")
)
pi0_hat <- mean(
  predict(fit, newdata = wcgs_counterfactual_b, type = "response")
)
```

$$\hat{\pi}_1 = 0.109, \quad \hat{\pi}_0 = 0.052$$

Under the identification assumptions of Theorem 10.3, this procedure consistently estimates the causal marginal risk:

Theorem 10.4 (Consistency of g-computation)**.** *Suppose:*

- *the data (Y_i, A_i, Z_i), $i = 1, \ldots, n$, are i.i.d. draws from the population;*
- *the outcome regression is correctly specified, so the fitted coefficients are consistent, $\hat{\beta} \xrightarrow{p} \beta^*$ (standard MLE consistency for a correctly specified GLM; see Agresti (2015), Chapter 4);*
- *the identification assumptions of Theorem 10.3 hold;*
- *the predictor vectors are uniformly bounded, $\|x_{a,i}\| \leq B < \infty$ (as holds when covariates are bounded).*

Then the g-computation estimator (Definition 10.7) is consistent for the causal marginal risk π_a (Definition 10.6):

$$\hat{\pi}_a \xrightarrow{p} \pi_a$$

Proof

Proof. Split the estimator into the average of the **true** conditional risks plus an estimation-error term:

$$\begin{aligned}
\hat{\pi}_a &= \frac{1}{n}\sum_{i=1}^{n}\hat{\pi}(a, z_i) && \text{[definition of the estimator]}\\
&= \frac{1}{n}\sum_{i=1}^{n}\pi(a, z_i) + \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\pi}(a, z_i) - \pi(a, z_i)\right) && \text{[add and subtract the true risks]}
\end{aligned}$$

The estimation-error term converges to zero in probability. By assumption, the fitted coefficients are consistent, $\hat{\beta} \xrightarrow{p} \beta^*$. Write the fitted and true risks as the logistic link of the linear predictor, $\hat{\pi}(a, z_i) \overset{\text{def}}{=} \operatorname{expit}\{x'_{a,i}\hat{\beta}\}$ and $\pi(a, z_i) \overset{\text{def}}{=} \operatorname{expit}\{x'_{a,i}\beta^*\}$, and recall that expit is Lipschitz with constant $\frac{1}{4}$ (its derivative $\operatorname{expit}\{u\}(1 - \operatorname{expit}\{u\})$ is at most $\frac{1}{4}$). Then

$$\begin{aligned}
|\hat{\pi}(a, z_i) - \pi(a, z_i)| &= \left|\operatorname{expit}\{x'_{a,i}\hat{\beta}\} - \operatorname{expit}\{x'_{a,i}\beta^*\}\right| && \text{[risk as logistic link of the linear predictor]}\\
&\le \tfrac{1}{4}\left|x'_{a,i}\hat{\beta} - x'_{a,i}\beta^*\right| && \text{[expit is } \tfrac{1}{4}\text{-Lipschitz, by the mean value theorem]}\\
&= \tfrac{1}{4}\left|x'_{a,i}(\hat{\beta} - \beta^*)\right| && \text{[factor out } x'_{a,i}\text{]}\\
&\le \tfrac{1}{4}\|x_{a,i}\|\,\|\hat{\beta} - \beta^*\| && \text{[Cauchy–Schwarz]}\\
&\le \tfrac{B}{4}\|\hat{\beta} - \beta^*\| && \text{[boundedness condition } \|x_{a,i}\| \le B\text{]}
\end{aligned}$$

uniformly in i, giving

$$\frac{1}{n}\sum_{i=1}^{n}|\hat{\pi}(a, z_i) - \pi(a, z_i)| \le \tfrac{B}{4}\|\hat{\beta} - \beta^*\| \xrightarrow{p} 0$$

(a uniform bound that pointwise convergence of the individual terms, correlated through the shared $\hat{\beta}$, would not provide on its own). The remaining term averages the true risks over an i.i.d. sample $z_1, \ldots, z_n$ from the marginal distribution f_Z, so the weak law of large numbers gives

$$\frac{1}{n}\sum_{i=1}^{n}\pi(a, z_i) \xrightarrow{p} \mathrm{E}[\pi(a, Z)]. \qquad \text{[weak law of large numbers]}$$

Finally, Theorem 10.3 identifies this expectation with the causal marginal risk:

$$\mathrm{E}[\pi(a, Z)] = \pi_a. \qquad \text{[causal marginal risk identification]}$$

Combining the three steps:

$$\begin{aligned}
\hat{\pi}_a &= \frac{1}{n}\sum_{i=1}^{n}\pi(a, z_i) + \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\pi}(a, z_i) - \pi(a, z_i)\right) && \text{[add-and-subtract decomposition]}\\
&\xrightarrow{p} \mathrm{E}[\pi(a, Z)] + 0 && \text{[Slutsky's theorem]}\\
&= \mathrm{E}[\pi(a, Z)] && \text{[additive identity]}\\
&= \pi_a. && \text{[causal marginal risk identification]}
\end{aligned}$$

□

The WCGS estimate $\hat{\pi}_1$ computed in Example 10.9 is a realization of the estimator this theorem covers: as the sample grows, that estimate is guaranteed to converge to the causal marginal risk π_1.

Definition 10.8 (Predictive margins)**.** A **predictive margin** is the fitted risk $\hat{\pi}(a, z)$ averaged over a chosen set of observations S (the term is due to Graubard and Korn 1999; Lane and Nelder 1982 develop the underlying standardization-as-prediction framework):

$$\widehat{\mathrm{PM}}(a \mid S) \stackrel{\text{def}}{=} \frac{1}{|S|} \sum_{i \in S} \hat{\pi}_{a,i}$$

using the shorthand $\hat{\pi}_{a,i} = \hat{\pi}(a, z_i)$ from Definition 10.7, where $|S|$ is the number of observations in S. The choice of the averaging set S determines which estimand the margin targets; the two standard choices each have their own name. The term "predictive margins" is commonly used in social science and in Stata's `margins` command.

Definition 10.9 (Full-sample predictive margin)**.** The **full-sample predictive margin** averages the fitted risks over the **whole sample** $S = \{1, \dots, n\}$:

$$\widehat{\mathrm{PM}}(a) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \hat{\pi}_{a,i} = \hat{\pi}_a$$

Exm

Example 10.10 (Illustration: full-sample predictive margin)**.** In the WCGS example (Example 10.9), the g-computation estimate $\hat{\pi}_1$ is simultaneously the full-sample predictive margin $\widehat{\mathrm{PM}}(1)$: both are the same average $\frac{1}{n}\sum_{i=1}^{n} \hat{\pi}_{1,i}$ of the fitted Type A risks over the whole sample.

Because the full-sample predictive margin is defined identically to the g-computation estimator (Example 10.10), it inherits the same consistency guarantee:

Corollary 10.2 (Full-sample predictive margin estimates the causal marginal risk)**.** *The full-sample predictive margin is consistent for the causal marginal risk π_a (Definition 10.6):*

$$\widehat{PM}(a) = \hat{\pi}_a \xrightarrow{p} \pi_a$$

Proof

Proof. The full-sample predictive margin $\frac{1}{n}\sum_{i=1}^{n} \hat{\pi}_{a,i}$ is, term for term, the g-computation estimator $\hat{\pi}_a$ of Definition 10.7, so its consistency for π_a is exactly Theorem 10.4. □

The WCGS predictive margin computed in Example 10.10 is therefore consistent for the causal marginal risk π_1 by this corollary.

Definition 10.10 (Exposed-subgroup predictive margin)**.** The **exposed-subgroup predictive margin** averages the fitted risks only over the observations actually exposed to $A = a$, $S = \{i : A_i = a\}$, and we give it its own symbol $\hat{\pi}_a(a)$:

$$\hat{\pi}_a(a) \stackrel{\text{def}}{=} \frac{1}{n_a} \sum_{i : A_i = a} \hat{\pi}_{a,i}$$

where n_a is the number of observations with $A_i = a$. The subscript records the treatment level the risks are predicted under, and the argument records the subgroup averaged over; the full-sample margin $\hat{\pi}_a$ averages over everyone instead.

Exm

Example 10.11 (Illustration: exposed-subgroup predictive margin)**.** Continuing Example 10.3, where the exposed ($A = 1$) have $\Pr(Z = 1 \mid A = 1) = 0.329$ and age-specific risks $\pi(1, 0) = 0.089$ and $\pi(1, 1) = 0.159$, averaging the fitted risks over the exposed subgroup gives

$$\hat{\pi}_1(1) = (1 - 0.329)(0.089) + (0.329)(0.159) = 0.112,$$

which coincides with the observed marginal risk $\pi(1) = 0.112$ (Example 10.3), not the causal marginal risk $\pi_1 = 0.109$ (Example 10.7); Theorem 10.5 shows this agreement is no coincidence.

Theorem 10.5 (Exposed-subgroup predictive margin estimates the observed marginal risk)**.** *With the data (Y_i, A_i, Z_i), $i = 1, \ldots, n$, drawn i.i.d. from the population, under correct model specification, and with the predictor vectors uniformly bounded ($\|x_{a,i}\| \leq B < \infty$, as in Theorem 10.4), the exposed-subgroup predictive margin is consistent for the observed marginal risk $\pi(a)$ (Definition 10.4):*

$$\hat{\pi}_a(a) \xrightarrow{p} \pi(a)$$

Proof

Proof. Split the estimator into the average of the **true** conditional risks over the exposed subgroup plus an estimation-error term:

$$\begin{aligned}
\hat{\pi}_a(a) &= \frac{1}{n_a} \sum_{i:A_i=a} \hat{\pi}(a, z_i) && \text{[definition of the estimator]} \\
&= \frac{1}{n_a} \sum_{i:A_i=a} \pi(a, z_i) + \frac{1}{n_a} \sum_{i:A_i=a} (\hat{\pi}(a, z_i) - \pi(a, z_i)) && \text{[add and subtract the true risks]}
\end{aligned}$$

The estimation-error term converges to zero in probability by the same bounded-predictor Lipschitz bound as in Theorem 10.4: $\frac{1}{n_a} \sum_{i:A_i=a} |\hat{\pi}(a, z_i) - \pi(a, z_i)| \leq \frac{B}{4} \|\hat{\beta} - \beta^*\| \xrightarrow{p} 0$. The observations with $A_i = a$ are an i.i.d. sample from the conditional distribution $f_{Z|A}(z \mid a)$, so the weak law of large numbers gives

$$\frac{1}{n_a} \sum_{i:A_i=a} \pi(a, z_i) \xrightarrow{p} \mathrm{E}[\pi(a, Z) \mid A = a]. \qquad \text{[weak law of large numbers]}$$

Finally, Theorem 10.1 identifies this conditional expectation with the observed marginal risk:

$$\mathrm{E}[\pi(a, Z) \mid A = a] = \pi(a). \qquad \text{[observed marginal risk theorem]}$$

Combining the three steps, $\hat{\pi}_a(a) \xrightarrow{p} \pi(a)$. □

The WCGS exposed-subgroup predictive margin from Example 10.11 is therefore consistent for the observed marginal risk $\pi(1)$ by this theorem.

Lemma 10.1 (Exposed-subgroup predictive margin reproduces the observed proportion)**.** *If the fitted model includes the exposure A as a categorical covariate, the exposed-subgroup predictive margin equals the observed event proportion in that subgroup **exactly**, in any finite sample:*

$$\hat{\pi}_a(a) = \bar{Y}_a \overset{def}{=} \frac{1}{n_a} \sum_{i:A_i=a} Y_i$$

Proof

Proof. The maximum-likelihood score equations for a logistic GLM set $\sum_i x_{ij}(Y_i - \hat{\mu}_i) = 0$ for each covariate column $x_{\cdot j}$, where $\hat{\mu}_i = \hat{\pi}(A_i, z_i)$ is observation i's fitted risk. Because A enters as a categorical covariate, the indicator column for level a contributes the score equation

$$\sum_{i:A_i=a} (Y_i - \hat{\mu}_i) = 0, \quad \text{i.e.} \quad \sum_{i:A_i=a} \hat{\mu}_i = \sum_{i:A_i=a} Y_i.$$

(For the reference level, which has no indicator column of its own, the same identity follows by subtracting the non-reference indicator score equations from the intercept score equation $\sum_{i=1}^{n}(Y_i - \hat{\mu}_i) = 0$.)
For an observation with $A_i = a$ we have $\hat{\mu}_i = \hat{\pi}(a, z_i)$, so dividing by n_a gives $\hat{\pi}_a(a) = \frac{1}{n_a}\sum_{i:A_i=a} \hat{\pi}(a, z_i) = \frac{1}{n_a}\sum_{i:A_i=a} Y_i = \bar{Y}_a$. □

The following example applies these predictive margins to the same WCGS data.

Exm

Example 10.12 (Estimated observed marginal risks (WCGS))**.** Continuing the WCGS example (Example 10.9), the **observed marginal risks** $\pi(a) = \Pr(Y = 1 \mid A = a)$ can be estimated in two equivalent ways.
Directly, as the sample proportion of events within each exposure group:

```
pi1_obs <- mean(wcgs_clean$chd69_binary[wcgs_clean$dibpat == "Type A"])
pi0_obs <- mean(wcgs_clean$chd69_binary[wcgs_clean$dibpat == "Type B"])
```

$$\hat{\pi}(1) = 0.112, \quad \hat{\pi}(0) = 0.05$$

Or via **predictive margins** (Definition 10.8): average the model's fitted risks over the **subgroup** actually exposed to $A = a$, keeping each man's own covariates and observed exposure:

```
fitted_risk <- predict(fit, type = "response")
pi1_obs_pm <- mean(fitted_risk[wcgs_clean$dibpat == "Type A"])
pi0_obs_pm <- mean(fitted_risk[wcgs_clean$dibpat == "Type B"])
```

$$\hat{\pi}_1(1) = 0.112, \quad \hat{\pi}_0(0) = 0.05$$

The two estimates agree exactly (to displayed precision), as Lemma 10.1 guarantees: because the model includes `dibpat` as a covariate, the exposed-subgroup predictive margin reproduces the raw event proportion in each `dibpat` group.
Note that $\hat{\pi}(1) \neq \hat{\pi}_1$ and $\hat{\pi}(0) \neq \hat{\pi}_0$ (Example 10.9): because age is associated with behavioral pattern (Type A and Type B men have different age distributions), the observed marginal and causal marginal risks differ. Standardization via g-computation adjusts for this difference.

Contrasting marginal risks across exposure levels gives the corresponding marginal risk differences:

Definition 10.11 (Observed marginal risk difference)**.** The **observed marginal risk difference** (**OMRD**) is:

$$\text{OMRD} \stackrel{\text{def}}{=} \pi(1) - \pi(0)$$

Exm

Example 10.13 (Estimated observed marginal risk difference)**.** Continuing from Example 10.12, the observed marginal risk difference estimate is:

```
rd_obs <- pi1_obs - pi0_obs
```

$$\begin{aligned}\widehat{\text{OMRD}} &= \hat{\pi}(1) - \hat{\pi}(0) && \text{[definition of OMRD]}\\ &= 0.112 - 0.05 && \text{[substitute estimates]}\\ &= 0.062 && \text{[arithmetic, computed from full-precision } \hat{\pi}(1) - \hat{\pi}(0)\text{]}\end{aligned}$$

Definition 10.12 (Causal marginal risk difference)**.** The **causal marginal risk difference** (**CMRD**) is:

$$\text{CMRD} \stackrel{\text{def}}{=} \pi_1 - \pi_0$$

Exm

Example 10.14 (Estimated causal marginal risk difference)**.** Continuing from Example 10.9, the causal marginal risk difference estimate is:

```
rd_hat <- pi1_hat - pi0_hat
```

$$\begin{aligned}\widehat{\text{CMRD}} &= \hat{\pi}_1 - \hat{\pi}_0 && \text{[definition of CMRD]}\\ &= 0.109 - 0.052 && \text{[substitute estimates]}\\ &= 0.056 && \text{[arithmetic, computed from full-precision } \hat{\pi}_1 - \hat{\pi}_0\text{]}\end{aligned}$$

Standard errors and confidence intervals

To quantify uncertainty for risk difference or risk ratio estimates derived from logistic regression models (e.g., to calculate SEs, CIs, and p-values), you will need to use the bootstrap, the multivariate delta method, or some other special technique.

The **bootstrap** is generally the most straightforward approach and does not require deriving complex formulas.

10.2.2 Bootstrap inference

Adapted from (Vittinghoff et al. 2012, sec. 3.6, p. 62), (Hastie et al. 2009, sec. 7.11), and (James et al. 2021, chap. 5, Section 5.2).

The bootstrap is a resampling method that allows us to estimate the sampling distribution of a statistic without making strong parametric assumptions. The method was introduced by Efron (1979); Efron and Tibshirani (1993) give a book-length treatment. For an introduction to the bootstrap, see Bootstrap Confidence Intervals[26].

Definition 10.13 (Bootstrap algorithm for a marginal risk difference)**.** To construct a bootstrap confidence interval for a marginal risk difference:

1. For $b = 1, \ldots, B$ (see Section 10.2.4 for guidance on choosing B):
   a. Draw a bootstrap sample of size n with replacement from the original data
   b. Fit the logistic regression model to the bootstrap sample
   c. Compute the marginal risk difference from the fitted model
2. The bootstrap distribution of the B risk difference estimates approximates the sampling distribution
3. Construct a confidence interval using the percentile method (e.g., the 2.5th and 97.5th percentiles for a 95% CI)

[26] basic-statistical-methods.qmd#sec-bootstrap-ci

Exm

Example 10.15 (Example: CHD risk and behavioral pattern)**.** We'll use the Western Collaborative Group Study (WCGS) data to illustrate computing marginal risk differences from a logistic regression model.
The research question is: what is the marginal effect of Type A behavioral pattern on CHD risk, adjusting for age?

Fit the model
We reuse `wcgs_clean` and `fit` from Example 10.9:

```
broom::tidy(fit, conf.int = TRUE) |>
  pander::pander()
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-6.142	0.554	-11.09	1.461e-28	-7.236	-5.063
dibpatType A	0.7994	0.1413	5.658	1.533e-08	0.5262	1.081
age	0.06868	0.01137	6.04	1.539e-09	0.0464	0.09101

Compute the point estimate
We define helper functions for g-computation — these helper functions are reused in the bootstrap loop of Section 10.15 — and extract the point estimate already computed in Example 10.14 for the contrast comparing Type A to Type B behavioral patterns:

```r
make_counterfactual_datasets <- function(data,
                                         exposure_var,
                                         exposed_level,
                                         unexposed_level) {
  if (!is.factor(data[[exposure_var]])) {
    stop("The exposure variable must be a factor.")
  }
  exposure_levels <- levels(data[[exposure_var]])
  if (!all(c(exposed_level, unexposed_level) %in% exposure_levels)) {
    stop(
      sprintf(
        paste0(
          "Specified exposure levels '%s' and '%s' ",
          "are not in factor levels: %s"
        ),
        exposed_level,
        unexposed_level,
        paste(exposure_levels, collapse = ", ")
      )
    )
  }

  data_exposed <- data
  data_exposed[[exposure_var]] <- factor(
    rep(exposed_level, nrow(data_exposed)),
    levels = exposure_levels
  )

  data_unexposed <- data
  data_unexposed[[exposure_var]] <- factor(
    rep(unexposed_level, nrow(data_unexposed)),
    levels = exposure_levels
  )

  list(
    data_exposed = data_exposed,
    data_unexposed = data_unexposed
  )
}

compute_marginal_rd <- function(model,
                                data,
                                exposure_var,
                                exposed_level,
                                unexposed_level) {
  counterfactual_data <- make_counterfactual_datasets(
    data = data,
    exposure_var = exposure_var,
    exposed_level = exposed_level,
    unexposed_level = unexposed_level
  )
  data_exposed <- counterfactual_data$data_exposed
  data_unexposed <- counterfactual_data$data_unexposed

  risk_exposed <- predict(
    model,
    newdata = data_exposed,
    type = "response"
  )
```



```r
  risk_unexposed <- predict(
    model,
```

Bootstrap confidence interval

Now we use the bootstrap to estimate the standard error and construct a 95% confidence interval. In the non-parametric bootstrap, each iteration applies both the model-fitting step and the covariate standardization to the same bootstrap resample. This joint refit-and-restandardize approach mirrors the original world in the bootstrap world and is standard practice for this type of estimator (Efron and Tibshirani 1993, chap. 6). (An alternative would fix the covariates at the observed values, but is less common for g-computation estimators.)

```
# Set seed for reproducibility
set.seed(20260512)

# Number of bootstrap iterations in this rendered example
# (illustrative only; use >= 2000 for final CIs)
B <- 300

boot_estimates <- numeric(B)

for (b in 1:B) {

  boot_indices <- sample(
    seq_len(nrow(wcgs_clean)),
    size = nrow(wcgs_clean),
    replace = TRUE
  )

  boot_data <- wcgs_clean[boot_indices, ]

  boot_fit <- glm(
    chd69_binary ~ dibpat + age,
    data = boot_data,
    family = binomial(link = "logit")
  )

  boot_estimates[b] <- compute_marginal_rd(
    model = boot_fit,
    data = boot_data,
    exposure_var = "dibpat",
    exposed_level = "Type A",
    unexposed_level = "Type B"
  )
}

boot_se <- sd(boot_estimates)

# percentile method: quantiles of the bootstrap distribution
ci_lower <- as.numeric(quantile(boot_estimates, 0.025))
ci_upper <- as.numeric(quantile(boot_estimates, 0.975))
```

```
tibble(
  Estimate = marginal_rd_est,
  `Std. Error` = boot_se,
  `95% CI Lower` = ci_lower,
  `95% CI Upper` = ci_upper
) |>
  knitr::kable(digits = 4)
```

Table 22: Causal marginal risk difference (CMRD) for Type A vs Type B behavioral pattern (adjusted for age) via g-computation, with bootstrap standard error and 95% confidence interval

Estimate	Std. Error	95% CI Lower	95% CI Upper
0.0562	0.0096	0.037	0.0732

Visualize the bootstrap distribution
We can visualize the bootstrap distribution to assess the sampling distribution of our estimate:

```
tibble(boot_rd = boot_estimates) |>
  ggplot() +
  aes(x = boot_rd) +
  geom_histogram(
    bins = 50,
    fill = "steelblue",
    alpha = 0.7,
    color = "black"
  ) +
  geom_vline(
    xintercept = marginal_rd_est,
    color = "red",
    linetype = "dashed",
    linewidth = 1
  ) +
  geom_vline(
    xintercept = ci_lower,
    color = "darkgreen",
    linetype = "dotted",
    linewidth = 1
  ) +
  geom_vline(
    xintercept = ci_upper,
    color = "darkgreen",
    linetype = "dotted",
    linewidth = 1
  ) +
  labs(
    x = "Bootstrap Marginal Risk Difference",
    y = "Frequency",
    title = "Bootstrap Distribution of Causal Marginal Risk Difference"
  ) +
  theme_minimal()
```

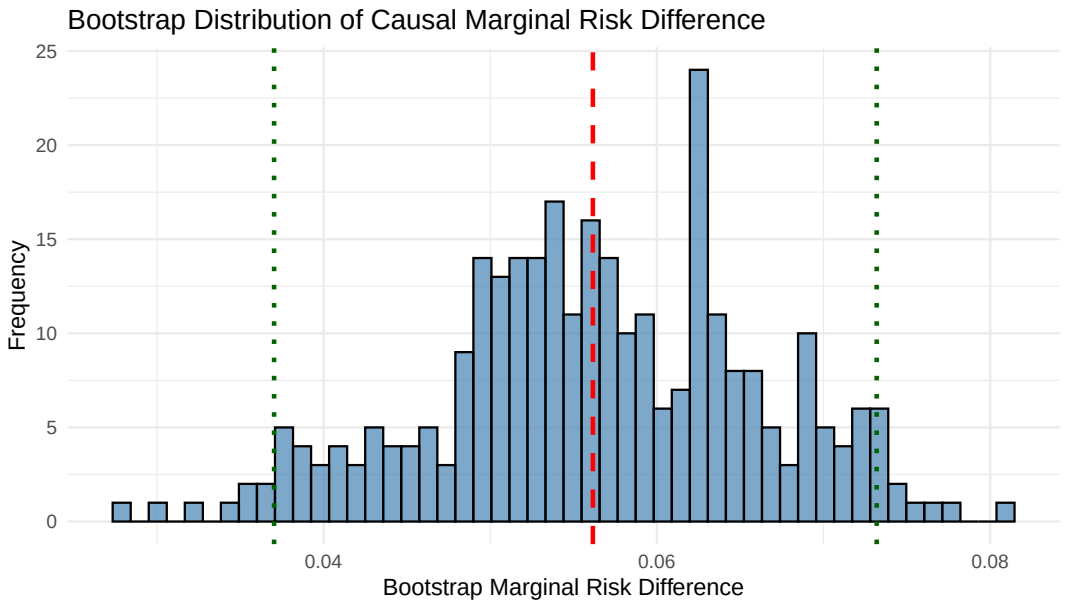


Figure 17: Bootstrap distribution of the marginal risk difference (Type A vs Type B behavioral pattern, adjusted for age). Red dashed line: point estimate. Green dotted lines: 95% CI bounds.

Interpretation
Under the identification assumptions (Consistency[a], Conditional Exchangeability[b], and Positivity[c]), we estimate that Type A behavioral pattern causally increases CHD risk by 5.62 percentage points compared to Type B behavioral pattern, after standardizing over the observed age distribution (95% CI: 3.7 to 7.32 percentage points).
Because this rendered example uses 300 bootstrap samples, the displayed standard error and confidence interval are illustrative. For final inferential reporting, rerun with at least 2000 bootstrap samples.
A 95% CI that excludes zero does not by itself imply practical or clinical importance; also interpret the estimated risk difference magnitude.

[a] causal-inference.qmd#def-consistency
[b] causal-inference.qmd#def-exchangeability
[c] causal-inference.qmd#def-positivity

Alternative: Using the boot package

The `boot`[27] package provides a more streamlined interface for bootstrap inference. Here's how to compute the same confidence interval using `boot::boot()`:

```
library(boot)

statistic_fn <- function(data, indices) {
  boot_data <- data[indices, ]

  boot_fit <- glm(
    chd69_binary ~ dibpat + age,
    data = boot_data,
    family = binomial(link = "logit")
  )

  compute_marginal_rd(
    model = boot_fit,
    data = boot_data,
    exposure_var = "dibpat",
    exposed_level = "Type A",
    unexposed_level = "Type B"
  )
}

set.seed(20260512)
boot_results <- boot(
  data = wcgs_clean,
  statistic = statistic_fn,
  # Keep the rendered example fast; increase to 2000+ for final analyses.
  R = 300
)

boot_results

boot.ci(boot_results, type = c("perc", "bca"))
```

The `boot`[28] package provides several types of confidence intervals, including the percentile method (`perc`) and the bias-corrected and accelerated (BCa) method (`bca`), which can provide better coverage in some situations. The BCa method is more demanding than the percentile method: its bias-correction and acceleration estimates are unstable at the illustrative `R = 300` used here (and

[27] https://cran.r-project.org/package=boot
[28] https://cran.r-project.org/package=boot

can warn or fail outright), so use a substantially larger `R` (at least 2000) before relying on `bca` intervals.

10.2.3 Computing marginal risk ratios

Contrasting marginal risks by ratio rather than by difference gives the corresponding marginal risk ratios:

Definition 10.14 (Observed marginal risk ratio)**.** The **observed marginal risk ratio** (**OMRR**) is:

$$\text{OMRR} \stackrel{\text{def}}{=} \frac{\pi(1)}{\pi(0)}$$

Exm

Example 10.16 (Estimated observed marginal risk ratio)**.** Continuing from Example 10.12, the observed marginal risk ratio estimate is:

```
rr_obs <- pi1_obs / pi0_obs
```

$$\begin{aligned}
\widehat{\text{OMRR}} &= \frac{\hat{\pi}(1)}{\hat{\pi}(0)} && \text{[definition of OMRR]} \\
&= \frac{0.112}{0.05} && \text{[substitute estimates]} \\
&= 2.22 && \text{[arithmetic, computed from full-precision } \hat{\pi}(1)/\hat{\pi}(0)\text{]}
\end{aligned}$$

Definition 10.15 (Causal marginal risk ratio)**.** The **causal marginal risk ratio** (**CMRR**) is:

$$\text{CMRR} \stackrel{\text{def}}{=} \frac{\pi_1}{\pi_0}$$

Exm

Example 10.17 (Estimated causal marginal risk ratio)**.** Continuing from Example 10.9, the causal marginal risk ratio estimate is:

```
rr_hat <- pi1_hat / pi0_hat
```

$$\begin{aligned}
\widehat{\text{CMRR}} &= \frac{\hat{\pi}_1}{\hat{\pi}_0} && \text{[definition of CMRR]} \\
&= \frac{0.109}{0.052} && \text{[substitute estimates]} \\
&= 2.07 && \text{[arithmetic, computed from full-precision } \hat{\pi}_1/\hat{\pi}_0\text{]}
\end{aligned}$$

The point estimates in Example 10.16 and Example 10.17 use the same fitted model as Example 10.14. To quantify their uncertainty, we bootstrap the marginal risk ratio, modifying the statistic function to compute ratios instead of differences:

```
library(boot)

compute_marginal_rr <- function(model,
                                data,
                                exposure_var,
                                exposed_level,
                                unexposed_level) {
```

```r
  counterfactual_data <- make_counterfactual_datasets(
    data = data,
    exposure_var = exposure_var,
    exposed_level = exposed_level,
    unexposed_level = unexposed_level
  )
  data_exposed <- counterfactual_data$data_exposed
  data_unexposed <- counterfactual_data$data_unexposed

  risk_exposed <- predict(
    model,
    newdata = data_exposed,
    type = "response"
  )

  risk_unexposed <- predict(
    model,
    newdata = data_unexposed,
    type = "response"
  )

  mean(risk_exposed) / mean(risk_unexposed)
}

statistic_fn_rr <- function(data, indices) {
  boot_data <- data[indices, ]
  boot_fit <- glm(
    chd69_binary ~ dibpat + age,
    data = boot_data,
    family = binomial(link = "logit")
  )
  compute_marginal_rr(
    model = boot_fit,
    data = boot_data,
    exposure_var = "dibpat",
    exposed_level = "Type A",
    unexposed_level = "Type B"
  )
}

set.seed(20260512)
boot_results_rr <- boot(
  data = wcgs_clean,
  statistic = statistic_fn_rr,
  # Keep the rendered example fast; increase to 2000+ for final analyses.
  R = 300
)

# BCa intervals can be unstable at this small R;
# increase to R = 2000+ before relying on bca results.
boot_ci_rr <- boot.ci(boot_results_rr, type = c("perc", "bca"))

cat("Marginal Risk Ratio (Type A vs Type B):",
    round(boot_results_rr$t0, 3), "\n")
cat("95% Percentile CI: (",
    round(boot_ci_rr$percent[4], 3), ",",
    round(boot_ci_rr$percent[5], 3), ")\n")
```

10.2.4 Notes and caveats

Definition 10.16 (Collapsibility)**.** An estimand θ is **collapsible** with respect to a covariate Z if the population-level estimand equals the marginal-distribution-weighted average of stratum-specific estimands:

$$\theta \overset{\text{def}}{=} \mathrm{E}[\theta(Z)]. \tag{47}$$

Collapsibility bias is the discrepancy that arises when $\theta \neq \mathrm{E}[\theta(Z)]$.

By LOTUS (the Law of the Unconscious Statistician[29]), Equation 47 expands into a weighted sum or integral over Z's distribution, depending on whether Z is discrete or continuous:

$$\mathrm{E}[\theta(Z)] = \begin{cases} \sum_z \theta(z)\ \mathrm{P}(Z = z), & Z \text{ discrete}, \\ \int \theta(z)\, f_Z(z)\, dz, & Z \text{ continuous}. \end{cases}$$

Theorem 10.6 (Which estimands are collapsible?)**.** *For **causal** estimands (Definition 10.6), and under the identification assumptions of Theorem 10.3, the collapsibility criterion (Definition 10.16) is evaluated with respect to the **marginal** distribution $f_Z(z)$. Additive (difference) measures are collapsible; ratio and log-odds measures are not collapsible in general.*
*For **observational** estimands (Definition 10.4), the marginal measure uses the conditional distribution $f_{Z|A}(z \mid a)$ as weights rather than the marginal $f_Z(z)$. Collapsibility with respect to f_Z therefore holds in particular when $Z \perp\!\!\!\perp A$ (no confounding).*

Proof

Proof. Additive measures are collapsible by linearity of expectation, applying the causal-marginal-risk identity (Theorem 10.3) to each potential probability:

$$\begin{aligned} \pi_1 - \pi_0 &= \mathrm{E}[\pi(1, Z)] - \mathrm{E}[\pi(0, Z)] && \text{[causal marginal risk theorem]} \\ &= \mathrm{E}[\pi(1, Z) - \pi(0, Z)]. && \text{[linearity of expectation]} \end{aligned}$$

Ratio and log-odds measures are not collapsible in general: Example 10.18 gives a counterexample in which the marginal risk ratio and odds ratio differ from the marginal-distribution-weighted average of their stratum-specific values, and Equation 49 derives the discrepancy algebraically for the risk-ratio case.
For the **observational** case, when $Z \perp\!\!\!\perp A$ the conditional weights coincide with the marginal ones, $f_{Z|A}(z \mid a) = f_Z(z)$, so by Theorem 10.1:

$$\begin{aligned} \pi(a) &= \mathrm{E}[\pi(a, Z) \mid A = a] && \text{[observed marginal risk theorem: law of total expectation, given } A = a\text{]} \\ &= \mathrm{E}[\pi(a, Z)], && [Z \perp\!\!\!\perp A\text{: conditioning on } A = a \text{ does not change } Z\text{'s distribution}] \end{aligned}$$

establishing that $\pi(a)$ equals the f_Z-weighted stratum average, which is collapsibility with respect to f_Z, requiring only $Z \perp\!\!\!\perp A$, not the full identification assumptions. □

Exm

Example 10.18 (Collapsibility: numerical illustration)**.** Continuing the WCGS example with `dibpat` as the exposure, `chd69` as the outcome, and age 50 or older as Z (Example 10.3): the causal marginal risk (Definition 10.6) standardizes **both** exposure groups' stratum-specific risks to the same population age distribution $f_Z(z)$ (Remark 10.1), so this standardization is exactly the common weighting Definition 10.16 requires – no simplifying assumptions needed:

[29] probability.qmd#thm-lotus

```r
wcgs_nonexposed <- wcgs_clean[wcgs_clean$dibpat == "Type B", ]
pi0_z0 <- mean(wcgs_nonexposed$chd69_binary[wcgs_nonexposed$age_ge50 == 0])
pi0_z1 <- mean(wcgs_nonexposed$chd69_binary[wcgs_nonexposed$age_ge50 == 1])

strata <- tibble::tibble(
  stratum = c("Z = 0 (under 50)", "Z = 1 (50 or older)"),
  weight  = c(1 - p_z1_pop, p_z1_pop),
  pi_0    = c(pi0_z0, pi0_z1),
  pi_1    = c(pi1_z0, pi1_z1)
) |>
  dplyr::mutate(
    RD = pi_1 - pi_0,
    RR = pi_1 / pi_0,
    OR = (pi_1 / (1 - pi_1)) / (pi_0 / (1 - pi_0))
  )

pi1_marg <- sum(strata$weight * strata$pi_1)
pi0_marg <- sum(strata$weight * strata$pi_0)

tibble::tibble(
  Measure = c("Risk difference", "Risk ratio", "Odds ratio"),
  Marginal = c(
    pi1_marg - pi0_marg,
    pi1_marg / pi0_marg,
    (pi1_marg / (1 - pi1_marg)) / (pi0_marg / (1 - pi0_marg))
  ),
  `Weighted avg. conditional` = c(
    sum(strata$weight * strata$RD),
    sum(strata$weight * strata$RR),
    sum(strata$weight * strata$OR)
  ),
  `Marginal = weighted avg. conditional?` = c("Yes", "No", "No")
) |>
  knitr::kable(digits = 4)
```

Measure	Marginal	Weighted avg. conditional	Marginal = weighted avg. conditional?
Risk difference	0.0572	0.0572	Yes
Risk ratio	2.1028	2.1034	No
Odds ratio	2.2378	2.2401	No

The marginal risk difference equals the weighted average of the conditional risk differences exactly, as Theorem 10.6 guarantees. The marginal risk ratio and marginal odds ratio both differ from the weighted average of their conditional counterparts – though the risk ratio's discrepancy only shows up in the fourth decimal place here, because the conditional risk ratios happen to be nearly homogeneous across the two age strata (2.105 vs. 2.100).
Non-collapsibility is distinct from effect-measure modification. The conditional odds ratios vary more across strata (2.213 vs. 2.308), and for the risk ratio, that near-homogeneity is exactly what keeps its discrepancy small here: if the conditional risk ratio were instead **constant** across strata, the marginal risk ratio would reproduce it exactly, just like the risk difference. The odds ratio's discrepancy is different in kind: even if the conditional odds ratio were held *constant* across strata, the marginal odds ratio would generally still differ from it (and lie closer to the null) (Hernán and Robins 2020, chap. 4; Greenland et al. 1999; Greenland and Robins 1986), whereas a constant conditional risk difference or risk ratio always reproduces the corresponding marginal measure exactly.

Important considerations when computing marginal effects

1. **Target population**: The marginal effect is specific to the covariate distribution in your sample. If your sample is not representative of your target population, you may want to use standardization or weighting.
2. **Collapsibility** (see Definition 10.16): Equation 48 and Equation 49 re-derive Theorem 10.6 concretely for the risk-difference and risk-ratio/odds-ratio cases. Risk differences are **collapsible**: the causal marginal RD $(\pi_1 - \pi_0)$ equals the average of stratum-specific causal RDs $(\pi(1,z) - \pi(0,z))$ weighted by the marginal distribution f_Z,

$$\begin{aligned} \pi_1 - \pi_0 &= \mathrm{E}[\pi(1,Z)] - \mathrm{E}[\pi(0,Z)] && \text{[causal marginal risk theorem]} \\ &= \mathrm{E}[\pi(1,Z) - \pi(0,Z)] && \text{[linearity of expectation]} \\ &= \mathrm{E}[\mathrm{RD}(Z)], && \text{[definition of conditional risk difference]} \end{aligned} \tag{48}$$

   so there is no collapsibility bias from averaging over a covariate Z. Risk ratios are **not** collapsible in general: when the conditional RR varies across strata, the marginal RR does not equal the marginal-distribution-weighted average of conditional RRs, $\mathrm{E}[\mathrm{RR}(Z)]$, because a ratio of expectations is not the same as an expectation of ratios:

$$\begin{aligned} \frac{\pi_1}{\pi_0} &= \frac{\mathrm{E}[\pi(1,Z)]}{\mathrm{E}[\pi(0,Z)]} && \text{[causal marginal risk theorem]} \\ &\neq \mathrm{E}\left[\frac{\pi(1,Z)}{\pi(0,Z)}\right] && \text{[ratio of expectations, not expectation of ratios]} \\ &= \mathrm{E}[\mathrm{RR}(Z)]. && \text{[definition of conditional risk ratio]} \end{aligned} \tag{49}$$

   Unlike the odds ratio, though, this discrepancy is entirely a consequence of the conditional RR *varying* across strata: when the conditional risk ratio is instead **constant**, $\mathrm{RR}(z) = r$ for every z, the marginal RR reproduces it exactly:

$$\begin{aligned} \pi_1 &= \mathrm{E}[\pi(1,Z)] && \text{[causal marginal risk theorem]} \\ &= \mathrm{E}[\mathrm{RR}(Z) \cdot \pi(0,Z)] && \text{[definition of conditional risk ratio]} \\ &= \mathrm{E}[r \cdot \pi(0,Z)] && \text{[constant conditional risk ratio]} \\ &= r \cdot \mathrm{E}[\pi(0,Z)] && \text{[linearity of expectation]} \\ &= r \cdot \pi_0 && \text{[causal marginal risk theorem]} \end{aligned}$$

   so $\pi_1/\pi_0 = r$: like the risk difference, the risk ratio is collapsible under a homogeneous conditional effect.

   Odds ratios are non-collapsible in a **stronger** sense than risk ratios: even when the conditional OR is the same in every stratum, the marginal OR is generally different from it (closer to the null) (Hernán and Robins 2020, chap. 4; Greenland et al. 1999; Greenland and Robins 1986), whereas a constant conditional RD or RR always reproduces the corresponding marginal RD or RR exactly (the marginal RD is the average $\mathrm{E}[\mathrm{RD}(Z)]$ of the stratum-specific RDs by Theorem 10.6, and the average of a constant is that constant).
3. **Bootstrap sample size**: A confidence interval needs more bootstrap resamples than a standard error, because it depends on the tails of the bootstrap distribution rather than just its spread (Efron and Tibshirani 1993, chap. 6; Davison and Hinkley 1997, chap. 2). As a conservative rule of thumb, use at least 1000 iterations for standard errors and at least 2000 for confidence intervals. More iterations give more stable estimates but take longer to compute.
4. **Alternative methods**: The multivariate delta method can also be used to compute standard errors analytically, but the formulas are complex. The `margins`[30] package in R provides implementations of the delta method for marginal effects.

[30] https://cran.r-project.org/package=margins

5. **Case-control studies**: These methods are **not valid** for case-control studies, where sampling is stratified by outcome. In case-control designs, only odds ratios can be estimated consistently (see Odds Ratios in Case-Control Studies[31]).

10.3 Other link functions for Bernoulli outcomes

Alternatively, if you want to estimate risk ratios more directly from the model, you can sometimes change the link function from logit{} to log{}; then you can obtain risk ratios by exponentiating coefficients [32], just like we did for odds ratios with the logit link:

```
data(anthers, package = "dobson")
anthers_sum <- aggregate(
  anthers[c("n", "y")],
  by = anthers[c("storage")], FUN = sum
)

anthers_glm_log <- glm(
  formula = cbind(y, n - y) ~ storage,
  data = anthers_sum,
  family = binomial(link = "log")
)

anthers_glm_log |>
  parameters() |>
  print_md()
```

Parameter	Log-Risk	SE	95% CI	z	p
(Intercept)	-0.80	0.12	(-1.04, -0.58)	-6.81	< .001
storage	0.17	0.07	(0.02, 0.31)	2.31	0.021

Now exp{β} gives us risk ratios instead of odds ratios:

```
anthers_glm_log |>
  parameters(exponentiate = TRUE) |>
  print_md()
```

Parameter	Risk Ratio	SE	95% CI	z	p
(Intercept)	0.45	0.05	(0.35, 0.56)	-6.81	< .001
storage	1.18	0.09	(1.03, 1.36)	2.31	0.021

Let's compare this model with a logistic model:

```
anthers_glm_logit <- glm(
  formula = cbind(y, n - y) ~ storage,
  data = anthers_sum,
  family = binomial(link = "logit")
)

anthers_glm_logit |>
  parameters(exponentiate = TRUE) |>
  print_md()
```

[31] binary-outcome-associations.qmd#sec-CC-studies

[32] or linear combinations of coefficients, depending on what covariate patterns you are contrasting

Parameter	Odds Ratio	SE	95% CI	z	p
(Intercept)	0.76	0.20	(0.45, 1.27)	-1.05	0.296
storage	1.49	0.26	(1.06, 2.10)	2.29	0.022

We can compare the two link functions by plotting their fitted germination probabilities across the range of the storage covariate, alongside the observed proportions:

```
library(tidyr)

anthers_fitted_grid <-
  data.frame(
    storage = seq(
      min(anthers_sum$storage),
      max(anthers_sum$storage),
      length.out = 100
    )
  ) |>
  mutate(
    `Log link` = predict(anthers_glm_log, pick(storage), type = "response"),
    `Logit link` = predict(anthers_glm_logit, pick(storage), type = "response")
  ) |>
  pivot_longer(
    cols = c(`Log link`, `Logit link`),
    names_to = "Link function",
    values_to = "fitted"
  )

anthers_observed <-
  anthers_sum |>
  mutate(proportion = y / n)

ggplot() +
  geom_line(
    data = anthers_fitted_grid,
    mapping = aes(x = storage, y = fitted, color = `Link function`)
  ) +
  geom_point(
    data = anthers_observed,
    mapping = aes(x = storage, y = proportion),
    size = 3
  ) +
  labs(
    x = "Storage condition",
    y = "Fitted probability of germination"
  ) +
  theme_bw()
```

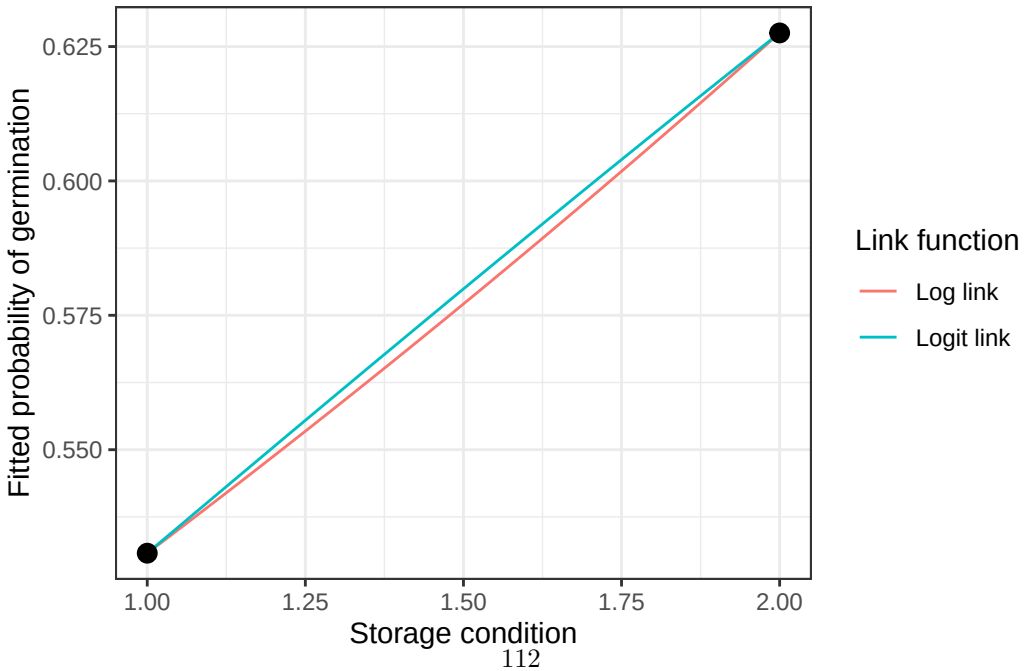


Figure 18: Fitted germination probabilities from the log-link and logit-link models for the `anthers` data, with the observed proportions as points. The two link functions nearly coincide here because

In practice, fitting a binomial model with `link = "log"` often fails to converge, producing errors about not finding good starting values for the estimation procedure. This convergence failure is likely because the model is producing fitted probabilities greater than 1.

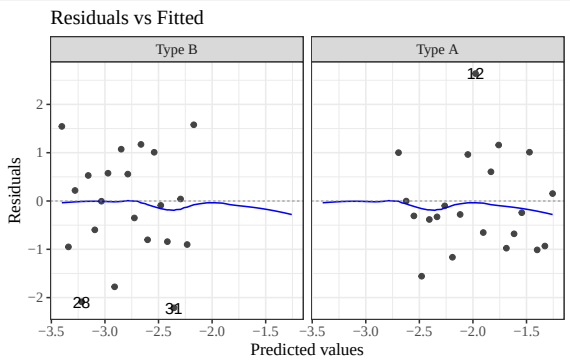
When this happens, you can instead switch the outcome distribution from binomial to Poisson and fit a Poisson regression model (we will see those soon!). Note that this substitution is a change in model *structure* — a different outcome distribution — not just a different way of fitting the same model. With a Poisson outcome distribution, you won't get warnings about fitted probabilities greater than 1, but that distribution isn't quite right for binary outcomes. Despite this mismatch, Poisson regression is the usual fallback when the log-binomial model fails to converge.

10.3.1 WCGS: link functions

```
wcgs_glm_logit_link <- chd_grouped_data |>
  mutate(type = relevel(dibpat, ref = "Type B")) |>
  glm(
    "formula" = cbind(x, `n - x`) ~ dibpat * age,
    "data" = _,
    "family" = binomial(link = "logit")
  )

wcgs_glm_identity_link <-
  chd_grouped_data |>
  mutate(type = relevel(dibpat, ref = "Type B")) |>
  glm(
    "formula" = cbind(x, `n - x`) ~ dibpat * age,
    "data" = _,
    "family" = binomial(link = "identity")
  )
wcgs_glm_identity_link |>
  coef() |>
  pander()
```

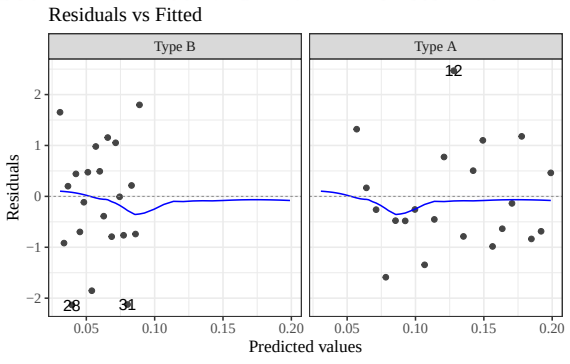
(Intercept)	dibpatType A	age	dibpatType A:age
-0.08257	-0.1374	0.002906	0.004194

```
library(ggfortify)
wcgs_glm_logit_link |> autoplot(which = c
```



(a) Logistic link

```
wcgs_glm_identity_link |> autoplot(which = c(1), ncol = 1)
```

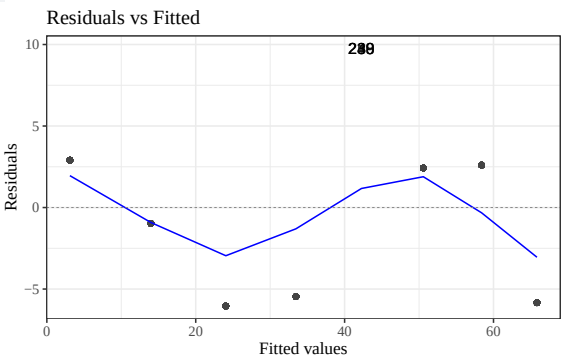


(b) Identity link

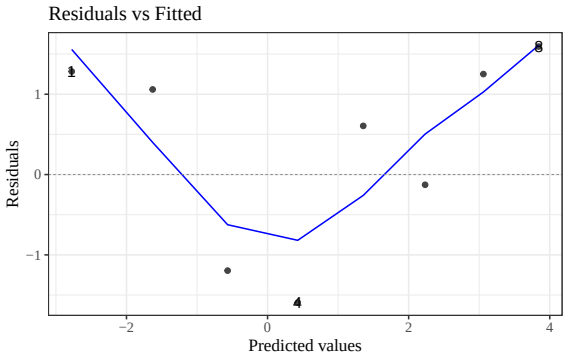
Figure 19: Residuals vs Fitted plot for `wcgs` models

```r
beetles_lm <-
  beetles_long |>
  lm(formula = died ~ dose)

beetles_glm_grouped <- beetles |>
  glm(formula = cbind(died, survived) ~ d


beetles <-
  beetles |> mutate(
    resid_logit = beetles_glm_grouped |>
  )
beetles_glm_grouped |> autoplot(which = c
```

```r
beetles_lm |> autoplot(which = c(1), ncol = 1)
```



(b) Identity link



(a) Logistic link

Figure 20: Residuals vs Fitted plot for `BeetleMortality` models

10.4 Quasibinomial

See Hua Zhou[33]'s lecture notes[34]

11 Further reading

- Hosmer et al. (2013) is a classic textbook on logistic regression

Exercises

Exercise 11.1 (Choosing a link function for binary outcomes)**.** (adapted from Dunn and Smyth (2018), Chapter 3)
Two researchers are modeling the probability that a patient experiences a side effect, as a function of dose. Define:

- x: the dose administered to the patient ($x > 0$)
- $\pi(x)$: the probability that a patient given dose x experiences the side effect
- β_0, β_1: the intercept and slope parameters of each model

The two models are:

- Researcher A uses the **logit** link: $\text{logit}(\pi) = \beta_0 + \beta_1 x$
- Researcher B uses the **log** link: $\log\{\pi\} = \beta_0 + \beta_1 x$

(a) For each model, write $\pi(x)$ as an explicit function of x.
(b) For the log-link model (Researcher B), what constraint on the parameters ensures $\pi(x) \leq 1$ for all $x > 0$?

[33] https://hua-zhou.github.io/
[34] https://ucla-biostat-200c-2020spring.github.io/slides/04-binomial/binomial.html#:~:text=0.05%20%27.%27%200.1%20%27%20%27%201-,Quasi%2Dbinomial,-Another%20way%20to

(c) Which link is more commonly used for binary outcomes? Give one practical advantage of that link.

Solution

Solution. **(a)**
Logit link (Researcher A):

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \text{expit}\{\beta_0 + \beta_1 x\}$$

This probability is always in $(0, 1)$, regardless of the values of β_0, β_1, or x.
Log link (Researcher B):

$$\pi(x) = e^{\beta_0+\beta_1 x}$$

This exponential expression is always positive, but may exceed 1.
(b)
For the log-link model, $\pi(x) \leq 1$ requires $e^{\beta_0+\beta_1 x} \leq 1$.
Since the exponential function is increasing, this requirement is equivalent to

$$\beta_0 + \beta_1 x \leq 0 \qquad \text{for all } x > 0.$$

Because this must hold for all positive x, a positive slope $\beta_1 > 0$ is not possible on this unbounded domain: for sufficiently large x, we would have $\beta_0 + \beta_1 x > 0$. Thus we need $\beta_1 \leq 0$. Also, to keep $\beta_0 + \beta_1 x \leq 0$ for x arbitrarily close to 0, we need $\beta_0 \leq 0$. So the parameter constraint is $\beta_0 \leq 0$ and $\beta_1 \leq 0$.
(c)
The **logit link** is more commonly used for binary outcomes.
One practical advantage: the logit link **automatically constrains** $\pi(x) \in (0, 1)$ for any real values of β_0, β_1, and x, eliminating the risk of predicted probabilities outside $[0, 1]$. Additionally, the logit link yields odds ratios as the natural measure of association, which are widely used in epidemiology (see Vittinghoff et al. (2012, chap. 5)).

Summary of definitions and results

Odds and log-odds

These association-measure results are covered in Measures of Association for Binary Outcomes[35].

Odds (def[36])

$$\omega \stackrel{\text{def}}{=} \frac{\Pr(A)}{\Pr(\neg A)}$$

Conditional odds (def[37])

$$\omega(A|B) \stackrel{\text{def}}{=} \frac{\Pr(A|B)}{\Pr(\neg A|B)}$$

Odds function (def[38])

$$\text{odds}\{\pi\} \stackrel{\text{def}}{=} \frac{\pi}{1-\pi}$$

[35] binary-outcome-associations.qmd
[36] binary-outcome-associations.qmd#def-odds
[37] binary-outcome-associations.qmd#def-c-odds
[38] binary-outcome-associations.qmd#def-odds-fn

Probability to odds (thm[39], cor[40])

$$\omega = \frac{\pi}{1-\pi} = \text{odds}\{\pi\}$$

Simplified odds expressions; odds of a non-event (thm[41], cor[42])

$$\text{odds}\{\pi\} = \frac{1}{\pi^{-1}-1} = (\pi^{-1}-1)^{-1}, \qquad \omega(\neg A) = \frac{1-\pi}{\pi} = \pi^{-1} - 1$$

Odds ratio (def[43])

$$\theta(\omega_1, \omega_2) \stackrel{\text{def}}{=} \frac{\omega_1}{\omega_2}$$

OR as ratio of probability ratios (thm[44])

$$\theta(\omega_1, \omega_2) = \frac{\omega_1}{\omega_2} = \frac{\left(\frac{\pi_1}{1-\pi_1}\right)}{\left(\frac{\pi_2}{1-\pi_2}\right)}$$

Odds ratios are reversible, marginally and conditionally (thm[45], thm[46])

$$\theta_\omega(A|B) = \theta_\omega(B|A), \qquad \theta_\omega(A|B,C) = \theta_\omega(B|A,C)$$

Inverse-odds and probability recovery

Odds to probability (thm[47])

$$\pi = \frac{\omega}{1+\omega}$$

Inverse-odds function (def[48]; recovers probability by cor[49])

$$\text{invodds}\{\omega\} \stackrel{\text{def}}{=} \frac{\omega}{1+\omega}, \qquad \pi = \text{invodds}\{\omega\}$$

Simplified inverse-odds (lem[50])

$$\text{invodds}\{\omega\} = \frac{1}{1+\omega^{-1}} = (1+\omega^{-1})^{-1}$$

One minus inverse-odds, and its complement (lem[51], cor[52])

$$1-\pi = \frac{1}{1+\omega}, \qquad 1+\omega = \frac{1}{1-\pi}$$

[39] binary-outcome-associations.qmd#thm-prob-to-odds
[40] binary-outcome-associations.qmd#cor-oddsf-to-odds
[41] binary-outcome-associations.qmd#thm-odds-simplified
[42] binary-outcome-associations.qmd#cor-odds-neg
[43] binary-outcome-associations.qmd#def-OR
[44] binary-outcome-associations.qmd#thm-or-ratio-ratio
[45] binary-outcome-associations.qmd#thm-or-swap
[46] binary-outcome-associations.qmd#thm-conditional-OR-swap
[47] binary-outcome-associations.qmd#thm-odds-to-prob
[48] binary-outcome-associations.qmd#def-inv-odds
[49] binary-outcome-associations.qmd#cor-invodds-pi
[50] binary-outcome-associations.qmd#lem-invodds-simplified
[51] binary-outcome-associations.qmd#lem-one-minus-odds-inv
[52] binary-outcome-associations.qmd#cor-inverse-odds-nonevent

Log-odds (logit) and expit

Log-odds, and log-odds from probability (def[53], thm[54])

$$\eta \overset{\text{def}}{=} \log\{\omega\}, \qquad \eta = \log\Big\{\frac{\pi}{1-\pi}\Big\}$$

Logit function, and expanded form (def[55], thm[56])

$$\text{logit}(\pi) \overset{\text{def}}{=} \log\{\text{odds}\{\pi\}\} = \log\Big\{\frac{\pi}{1-\pi}\Big\}$$

Log-odds equals logit (cor[57])

$$\eta = \text{logit}\{\pi\}$$

Odds and probability from log-odds (lem[58], thm[59])

$$\omega = \exp\{\eta\}, \qquad \pi = \frac{\exp\{\eta\}}{1+\exp\{\eta\}}$$

Expit / inverse-logit function, and equivalent expressions (def[60], thm[61])

$$\text{expit}(\eta) \overset{\text{def}}{=} \text{invodds}\{\exp\{\eta\}\} = \frac{\exp\{\eta\}}{1+\exp\{\eta\}} = (1+\exp\{-\eta\})^{-1}$$

Probability as expit (thm[62])

$$\pi = \text{expit}\{\eta\} \tag{50}$$

Logit and expit are inverses (thm[63])

$$\text{logit}\{\text{expit}\{\eta\}\} = \eta \qquad \text{expit}\{\text{logit}\{\pi\}\} = \pi$$

$$\Big[\pi \overset{\text{def}}{=} \Pr\big(Y=1|\tilde{X}=\tilde{x}\big)\Big] \overbrace{\underset{\frac{\omega}{1+\omega}}{\overset{\frac{\pi}{1-\pi}}{\rightleftarrows}} \Big[\omega \overset{\text{def}}{=} \text{odds}\Big\{Y=1|\tilde{X}=\tilde{x}\Big\}\Big] \underset{\exp\{\eta\}}{\overset{\log\{\omega\}}{\rightleftarrows}}}^{\text{expit}(\eta)} \Big[\eta(\tilde{x}) \overset{\text{def}}{=} \eta\big(Y=1|\tilde{X}=\tilde{x}\big)\Big]$$

Figure 21: Diagram of logistic regression link and inverse link functions

Rare events

Odds minus probability (thm[64])

$$\omega - \pi = \frac{\pi^2}{1-\pi}, \quad \text{where } \omega = \frac{\pi}{1-\pi}$$

[53] binary-outcome-associations.qmd#def-logodds
[54] binary-outcome-associations.qmd#thm-logodds-pi
[55] binary-outcome-associations.qmd#def-logit-fn
[56] binary-outcome-associations.qmd#thm-logit-function
[57] binary-outcome-associations.qmd#cor-logodds-logit
[58] binary-outcome-associations.qmd#lem-odds-from-logodds
[59] binary-outcome-associations.qmd#thm-prob-from-logodds
[60] binary-outcome-associations.qmd#def-expit
[61] binary-outcome-associations.qmd#thm-expit-expressions
[62] binary-outcome-associations.qmd#thm-expit-prob-logodds
[63] binary-outcome-associations.qmd#thm-logit-expit
[64] binary-outcome-associations.qmd#thm-odds-minus-probs

OR as RR times a correction factor (thm[65])

$$\theta(\omega_1, \omega_2) = \rho(\pi_1, \pi_2) \cdot \frac{1 - \pi_2}{1 - \pi_1}$$

When π_1 and π_2 are both small, the correction factor is close to 1, so the odds ratio is close to the risk ratio.

Derivatives

	π	ω	η	$\tilde{x}$	$\tilde{\beta}$
π	1	$(1+\omega)^2$	$\frac{(1+\omega)^2}{\omega}$	undef	undef
ω	$(1-\pi)^2$	1	$\frac{1}{\omega}$	undef	undef
η	$\pi(1-\pi)$	ω	1	undef	undef
$\tilde{x}$	$\underbrace{\tilde{\beta}\pi(1-\pi)}_{p\times 1}$	$\underbrace{\tilde{\beta}\omega}_{p\times 1}$	$\underbrace{\tilde{\beta}}_{p\times 1}$	$\underbrace{\mathbb{I}}_{p\times p}$	$\mathbf{0}_{p\times p}$
$\tilde{\beta}$	$\underbrace{\tilde{x}\pi(1-\pi)}_{p\times 1}$	$\underbrace{\tilde{x}\omega}_{p\times 1}$	$\underbrace{\tilde{x}}_{p\times 1}$	$\mathbf{0}_{p\times p}$	$\underbrace{\mathbb{I}}_{p\times p}$

Column labels indicate the numerators of the derivatives; row labels indicate the denominators (dimensions and details in tbl[66]).

Derivative of expit (cor[67])

$$\text{expit}'\{\eta\} = \text{expit}\{\eta\}(1 - \text{expit}\{\eta\})$$

Logistic regression model

Logistic regression (def[68])

$$Y_i|\tilde{X}_i \ \sim_{\perp\!\!\!\perp} \ \text{Ber}(\pi(\tilde{X}_i)), \qquad \text{logit}\{\pi(\tilde{x})\} = \tilde{x}^\top \tilde{\beta}$$

Throughout this summary, $\tilde{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$ is the $(p+1) \times 1$ coefficient vector (including the intercept), $\mathbf{X}$ is the $n \times (p+1)$ design matrix whose i^{th} row is $\tilde{x}_i^\top$, and $\tilde{\varepsilon}$ is the $n \times 1$ error vector with components $\varepsilon_i \overset{\text{def}}{=} y_i - \pi_i$.

Log-likelihood and score function

Log-likelihood component (lem[69])

$$\ell_i(\pi_i) = y_i \eta_i - \log\{1 + \omega_i\}$$

Score function as sum of components (lem[70], thm[71])

$$\tilde{\ell}'(\tilde{\beta}) = \sum_{i=1}^{n} \tilde{\ell}_i'(\tilde{\beta}), \qquad \ell_i'(\tilde{\beta}) = \tilde{x}_i \varepsilon_i$$

Score function (thm[72])

$$\tilde{\ell}'(\tilde{\beta}) = \sum_{i=1}^{n} \tilde{x}_i \varepsilon_i = \underbrace{\mathbf{X}^\top}_{(p+1)\times n} \underbrace{\tilde{\varepsilon}}_{n\times 1}$$

[65] binary-outcome-associations.qmd#thm-rare-disease-or-rr
[66] logistic-regression.qmd#tbl-logistic-deriv-matrix
[67] logistic-regression.qmd#cor-deriv-expit
[68] logistic-regression.qmd#def-logistic-regression
[69] logistic-regression.qmd#lem-logistic-loglik-component
[70] logistic-regression.qmd#lem-score-logistic
[71] logistic-regression.qmd#thm-logistic-score-comp
[72] logistic-regression.qmd#thm-logistic-score-fn

Maximum likelihood estimation

Estimating equation (eq[73]) The MLE $\hat{\tilde{\beta}}$ solves the score equation:

$$\underbrace{\tilde{\ell}'(\hat{\tilde{\beta}})}_{(p+1)\times 1} = \sum_{i=1}^{n} \underbrace{\tilde{x}_i}_{(p+1)\times 1} \underbrace{\left(y_i - \operatorname{expit}\left\{\tilde{x}_i^\top \hat{\tilde{\beta}}\right\}\right)}_{1\times 1} = \mathbf{0}_{(p+1)\times 1}$$

This estimating equation has no closed-form solution in general; $\hat{\tilde{\beta}}$ must be computed by numerical iteration.

Hessian (eq[74])

$$\underbrace{\ell''(\tilde{\beta})}_{(p+1)\times(p+1)} = -\underbrace{\mathbf{X}^\top}_{(p+1)\times n} \underbrace{\mathbf{D}}_{n\times n} \underbrace{\mathbf{X}}_{n\times(p+1)}, \qquad \mathbf{D} \stackrel{\text{def}}{=} \operatorname{diag}(\operatorname{Var}\left(Y_i | \tilde{X}_i = \tilde{x}_i\right)) = \operatorname{diag}(\pi_i(1-\pi_i))$$

Concavity (rem[75]): the Hessian is negative semi-definite for every $\tilde{\beta}$ (negative definite when $\mathbf{X}$ has full column rank), so the log-likelihood is concave and every solution of the score equation is a global maximum.

Newton-Raphson update (Newton-Raphson method[76])

$$\underbrace{\hat{\tilde{\beta}}^*}_{(p+1)\times 1} \leftarrow \underbrace{\hat{\tilde{\beta}}^*}_{(p+1)\times 1} - \underbrace{\left(\ell''\left(\tilde{y}; \hat{\tilde{\beta}}^*\right)\right)^{-1}}_{(p+1)\times(p+1)} \underbrace{\ell'\left(\tilde{y}; \hat{\tilde{\beta}}^*\right)}_{(p+1)\times 1}$$

One-sample MLE for odds (thm[77])

$$\hat{\omega} = \frac{x}{n-x}$$

Wald tests and confidence intervals for coefficients

Wald test statistic for $H_0 : \beta_k = \beta_{k,0}$ (CLT for MLEs[78])

$$z_k = \frac{\hat{\beta}_k - \beta_{k,0}}{\widehat{\operatorname{SE}}\left(\hat{\beta}_k\right)}, \qquad z_k \mathrel{\dot\sim} N(0,1) \text{ under } H_0$$

95% confidence intervals for β_k and e^{β_k} (MLE invariance)

$$\hat{\beta}_k \pm 1.96 \cdot \widehat{\operatorname{SE}}\left(\hat{\beta}_k\right), \qquad e^{\beta_k} \in \left(e^{\hat{\beta}_k - 1.96\cdot\widehat{\operatorname{SE}}(\hat{\beta}_k)},\ e^{\hat{\beta}_k + 1.96\cdot\widehat{\operatorname{SE}}(\hat{\beta}_k)}\right)$$

For $k \neq 0$, e^{β_k} is an odds ratio; for $k = 0$, e^{β_0} is the baseline odds.

Odds ratios in logistic regression

General OR formula (lem[79])

$$\theta_\omega(\tilde{x}, \tilde{x}^*) = \exp\{\eta(\tilde{x}) - \eta(\tilde{x}^*)\}$$

Difference in log-odds; OR in terms of $\Delta\eta$ (def[80], cor[81])

$$\Delta\eta \stackrel{\text{def}}{=} \eta(\tilde{x}) - \eta(\tilde{x}^*), \qquad \theta_\omega(\tilde{x}, \tilde{x}^*) = \exp\{\Delta\eta\}$$

[73] logistic-regression.qmd#eq-score-logistic
[74] logistic-regression.qmd#eq-logistic-hess
[75] logistic-regression.qmd#rem-hessian-nsd
[76] intro-MLEs.qmd#sec-newton-raphson
[77] binary-outcome-associations.qmd#thm-est-odds
[78] intro-MLEs.qmd#thm-dist-mle
[79] logistic-regression.qmd#lem-logistic-OR
[80] logistic-regression.qmd#def-difflogodds
[81] logistic-regression.qmd#cor-logistic-OR

$\Delta\eta$ from covariates (lem[82])

$$\Delta\eta = (\tilde{x} - \tilde{x}^*) \cdot \tilde{\beta}$$

Difference in covariate patterns; $\Delta\eta$ from $\Delta\tilde{x}$ (def[83], cor[84])

$$\Delta\tilde{x} \overset{\text{def}}{=} \tilde{x} - \tilde{x}^*, \qquad \Delta\eta = \Delta\tilde{x} \cdot \tilde{\beta}$$

OR in terms of $\Delta\tilde{x}$ (thm[85])

$$\theta_\omega(\tilde{x}, \tilde{x}^*) = \exp\Big\{(\Delta\tilde{x}) \cdot \tilde{\beta}\Big\}$$

Log OR equals $\Delta\eta$ (cor[86])

$$\log\{\theta_\omega(\tilde{x}, \tilde{x}^*)\} = \Delta\eta$$

Inference for log-odds and odds ratios

Estimated SE of log-odds (thm[87])

$$\widehat{\mathrm{Var}}(\hat{\eta}(\tilde{x})) = \underbrace{\tilde{x}^\top}_{1\times(p+1)} \underbrace{\hat{\Sigma}}_{(p+1)\times(p+1)} \underbrace{\tilde{x}}_{(p+1)\times 1}, \qquad \widehat{\mathrm{SE}}(\hat{\eta}(\tilde{x})) = \sqrt{\tilde{x}^\top \hat{\Sigma} \tilde{x}}$$

Estimated SE of $\Delta\hat{\eta}$ (thm[88])

$$\widehat{\mathrm{Var}}(\Delta\hat{\eta}) = \underbrace{\Delta\tilde{x}^\top}_{1\times(p+1)} \underbrace{\hat{\Sigma}}_{(p+1)\times(p+1)} \underbrace{(\Delta\tilde{x})}_{(p+1)\times 1}, \qquad \widehat{\mathrm{SE}}(\Delta\hat{\eta}) = \sqrt{\Delta\tilde{x}^\top \hat{\Sigma} (\Delta\tilde{x})}$$

CI for an odds ratio (eq[89]; details in sec[90]): exponentiate the endpoints of the CI for $\Delta\eta$:

$$\theta_\omega(\tilde{x}, \tilde{x}^*) \in (e^L, e^R), \qquad (L, R) = \widehat{\Delta\eta} \mp 1.96 \cdot \widehat{\mathrm{SE}}\Big(\widehat{\Delta\eta}\Big)$$

Odds ratios in study designs

Disease and exposure odds ratios (def[91], def[92])

$$\theta_\omega(D|E) \overset{\text{def}}{=} \frac{\omega(D|E)}{\omega(D|\neg E)}, \qquad \theta_\omega(E|D) \overset{\text{def}}{=} \frac{\omega(E|D)}{\omega(E|\neg D)}$$

Case-control studies (section[93]): outcome-stratified sampling breaks estimation of risks, risk ratios, and the components of the disease odds ratio (def[94]), but the exposure odds ratio (def[95]) is estimable, and it equals the disease odds ratio by reversibility (thm[96]).

2×2 table shortcut (table[97]; cell counts a, b = exposed events and non-events, c, d = unexposed events and non-events):

$$\hat{\omega}(Event|Exposed) = \frac{a}{b}, \qquad \hat{\omega}(Event|\neg Exposed) = \frac{c}{d}, \qquad \theta_\omega(Exposed, \neg Exposed) = \frac{ad}{bc}$$

[82] logistic-regression.qmd#lem-diff-logodds
[83] logistic-regression.qmd#def-diff-covs
[84] logistic-regression.qmd#cor-diff-logodds
[85] logistic-regression.qmd#thm-logistic-OR
[86] logistic-regression.qmd#cor-log-or
[87] logistic-regression.qmd#thm-est-se-logodds
[88] logistic-regression.qmd#thm-est-se-diff-logodds
[89] logistic-regression.qmd#eq-ci-OR-exp
[90] logistic-regression.qmd#sec-or-inference
[91] binary-outcome-associations.qmd#def-disease-OR
[92] binary-outcome-associations.qmd#def-exposure-OR
[93] binary-outcome-associations.qmd#sec-CC-studies
[94] binary-outcome-associations.qmd#def-disease-OR
[95] binary-outcome-associations.qmd#def-exposure-OR
[96] binary-outcome-associations.qmd#thm-or-swap
[97] binary-outcome-associations.qmd#tbl-2x2-generic

Model fit and comparison

Details in sec[98]; here q is the number of distinct covariate patterns, p is the number of β parameters, and ℓ_{full} is the saturated-model log-likelihood.

Deviance test (def[99])

$$D \overset{\text{def}}{=} 2(\ell_{\text{full}} - \ell(\hat{\beta}; \mathbf{x})), \qquad D \mathrel{\dot{\sim}} \chi^2(q-p)$$

Hosmer-Lemeshow test (def[100]; g groups by predicted probability)

$$X^2 \overset{\text{def}}{=} \sum_{c=1}^{g} \frac{(o_c - e_c)^2}{e_c} \mathrel{\dot{\sim}} \chi^2(g-2)$$

Pearson chi-squared GOF test (sum of squared Pearson residuals, def[101])

$$X^2 = \sum_{k=1}^{q} X_k^2 \mathrel{\dot{\sim}} \chi^2(q-p)$$

AIC and BIC [lower is better] (AIC[102] (Akaike 1974); BIC[103] (Schwarz 1978))

$$\text{AIC} = -2\ell(\hat{\theta}) + 2p, \qquad \text{BIC} = -2\ell(\hat{\theta}) + p\log\{n\}$$

Likelihood ratio test (def[104]; for nested models with $p_0 < p_1$ parameters — here ℓ_1 and ℓ_0 are the two fitted models' log-likelihoods, not the saturated model's)

$$\Lambda \overset{\text{def}}{=} 2(\ell_1(\hat{\theta}_1) - \ell_0(\hat{\theta}_0)), \qquad \Lambda \mathrel{\dot{\sim}} \chi^2(p_1 - p_0) \text{ under } H_0$$

by Wilks' theorem[105].

Residual-based diagnostics

Residuals are computed per covariate pattern k (grouped data); h_k denotes the leverage of pattern k (def[106]).

Response residual (def[107])

$$e_k \overset{\text{def}}{=} \bar{y}_k - \hat{\pi}(x_k)$$

Pearson residual (and standardized version) (def[108], def[109])

$$X_k \overset{\text{def}}{=} \frac{\bar{y}_k - \hat{\pi}_k}{\sqrt{\hat{\pi}_k(1-\hat{\pi}_k)/n_k}}, \qquad r_{P_k} \overset{\text{def}}{=} \frac{X_k}{\sqrt{1-h_k}}$$

Deviance residual (and standardized version) (def[110])

$$d_k \overset{\text{def}}{=} \text{sign}(y_k - n_k\hat{\pi}_k)\left\{\sqrt{2[\ell_{\text{full}}(x_k) - \ell(\hat{\beta}; x_k)]}\right\}, \qquad r_{D_k} \overset{\text{def}}{=} \frac{d_k}{\sqrt{1-h_k}}$$

[98] logistic-regression.qmd#sec-gof
[99] logistic-regression.qmd#def-deviance-test
[100] logistic-regression.qmd#def-hosmer-lemeshow
[101] logistic-regression.qmd#def-pearson-residual
[102] Linear-models-overview.qmd#def-aic
[103] Linear-models-overview.qmd#def-bic
[104] logistic-regression.qmd#def-lrt
[105] intro-MLEs.qmd#thm-wilks
[106] logistic-regression.qmd#def-leverage-glm
[107] logistic-regression.qmd#def-response-residual-logistic
[108] logistic-regression.qmd#def-pearson-residual
[109] logistic-regression.qmd#def-standardized-pearson-residual
[110] logistic-regression.qmd#def-deviance-residual-logistic

Cook's distance (def[111])

$$D_k \overset{\text{def}}{=} \frac{(r_{P_k})^2}{p} \cdot \frac{h_k}{1-h_k}$$

Leverage (def[112]): h_k is the k-th diagonal element of the weighted hat matrix (with $\mathbf{X}$ here the $q \times p$ design matrix of covariate patterns):

$$\underset{q\times q}{\mathbf{H}} \overset{\text{def}}{=} \underbrace{\mathbf{W}^{1/2}}_{q\times q}\,\underset{q\times p}{\mathbf{X}}\left(\underbrace{\mathbf{X}^\top\mathbf{W}\mathbf{X}}_{p\times p}\right)^{-1}\underset{p\times q}{\mathbf{X}^\top}\,\underbrace{\mathbf{W}^{1/2}}_{q\times q}, \qquad \underset{q\times q}{\mathbf{W}} \overset{\text{def}}{=} \operatorname{diag}(n_k\,\hat\pi_k\,(1-\hat\pi_k))$$

Comparing probabilities

Risk difference (def[113])

$$\delta(\pi_1,\pi_2) \overset{\text{def}}{=} \pi_1 - \pi_2$$

Risk ratio (def[114])

$$\rho(\pi_1,\pi_2) \overset{\text{def}}{=} \frac{\pi_1}{\pi_2}$$

Relative risk difference (def[115]; equals RR minus 1 by thm[116])

$$\xi(\pi_1,\pi_2) \overset{\text{def}}{=} \frac{\delta(\pi_1,\pi_2)}{\pi_2} = \rho(\pi_1,\pi_2) - 1$$

Marginal risks from logistic regression

Conditional predicted risk (def[117])

$$\pi(a,z) \overset{\text{def}}{=} \Pr(Y=1 \mid A=a, Z=z)$$

Conditional RD and RR (def[118], def[119])

$$\mathrm{RD}(z) \overset{\text{def}}{=} \pi(1,z) - \pi(0,z), \qquad \mathrm{RR}(z) \overset{\text{def}}{=} \frac{\pi(1,z)}{\pi(0,z)}$$

Observed marginal risk (def[120]; standardization form by thm[121])

$$\pi(a) \overset{\text{def}}{=} \Pr(Y=1 \mid A=a) = \mathrm{E}[\pi(a,Z) \mid A=a]$$

Potential mean and causal marginal risk (def[122], def[123]; identification form by thm[124])

$$\mu_a \overset{\text{def}}{=} \mathrm{E}[Y^a], \qquad \pi_a \overset{\text{def}}{=} \Pr(Y^a=1) = \mathrm{E}[\pi(a,Z)]$$

No confounding: observed = causal (cor[125]; the two estimands are contrasted in rem[126])

$$Z \perp\!\!\!\perp A \implies \pi(a) = \pi_a$$

[111] logistic-regression.qmd#def-cooks-distance-glm
[112] logistic-regression.qmd#def-leverage-glm
[113] binary-outcome-associations.qmd#def-RD
[114] binary-outcome-associations.qmd#def-RR
[115] binary-outcome-associations.qmd#def-RRD
[116] binary-outcome-associations.qmd#thm-rrd-vs-rr
[117] logistic-regression.qmd#def-cond-predicted-risk
[118] logistic-regression.qmd#def-cond-rd
[119] logistic-regression.qmd#def-cond-rrr
[120] logistic-regression.qmd#def-observed-marginal-risk
[121] logistic-regression.qmd#thm-observed-marginal-risk
[122] logistic-regression.qmd#def-potential-mean
[123] logistic-regression.qmd#def-causal-marginal-risk
[124] logistic-regression.qmd#thm-causal-marginal-risk
[125] logistic-regression.qmd#cor-observed-equals-causal
[126] logistic-regression.qmd#rem-observed-vs-causal-marginal-risk

G-computation estimator (def[127]; consistent for π_a by thm[128])

$$\hat{\pi}_a \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \hat{\pi}_{a,i}, \qquad \hat{\pi}_{a,i} \stackrel{\text{def}}{=} \hat{\pi}(a, z_i)$$

Predictive margins (def[129]; full-sample form def[130] is consistent for π_a by cor[131])

$$\widehat{\text{PM}}(a \mid S) \stackrel{\text{def}}{=} \frac{1}{|S|}\sum_{i \in S} \hat{\pi}_{a,i}, \qquad \widehat{\text{PM}}(a) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \hat{\pi}_{a,i} = \hat{\pi}_a$$

Exposed-subgroup predictive margin (def[132]; consistent for $\pi(a)$ by thm[133]; equals $\bar{Y}_a$ exactly by lem[134])

$$\hat{\pi}_a(a) \stackrel{\text{def}}{=} \frac{1}{n_a}\sum_{i:A_i=a} \hat{\pi}_{a,i} = \bar{Y}_a$$

Observed marginal RD and RR (def[135], def[136])

$$\text{OMRD} \stackrel{\text{def}}{=} \pi(1) - \pi(0), \qquad \text{OMRR} \stackrel{\text{def}}{=} \frac{\pi(1)}{\pi(0)}$$

Causal marginal RD and RR (def[137], def[138])

$$\text{CMRD} \stackrel{\text{def}}{=} \pi_1 - \pi_0, \qquad \text{CMRR} \stackrel{\text{def}}{=} \frac{\pi_1}{\pi_0}$$

Bootstrap SE and CI for marginal contrasts (def[139]; see Bootstrap Confidence Intervals[140] and exm[141]): resample the data with replacement B times; refit the logistic model and recompute the marginal contrast on each resample; the bootstrap SE is the standard deviation of the B estimates, and a 95% percentile CI uses their 2.5th and 97.5th percentiles.

Collapsibility (def[142]): an estimand θ is collapsible with respect to Z if

$$\theta \stackrel{\text{def}}{=} \text{E}[\theta(Z)]$$

Risk differences are collapsible; risk ratios and odds ratios are not collapsible in general (thm[143]; counterexample in exm[144]).

References

Agresti, Alan. 2015. *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons. https://www.wiley.com/en-us/Foundations+of+Linear+and+Generalized+Linear+Models-p-9781118730034.

[127] logistic-regression.qmd#def-g-computation
[128] logistic-regression.qmd#thm-g-computation-consistency
[129] logistic-regression.qmd#def-predictive-margins
[130] logistic-regression.qmd#def-full-sample-predictive-margin
[131] logistic-regression.qmd#cor-full-sample-pm
[132] logistic-regression.qmd#def-subgroup-predictive-margin
[133] logistic-regression.qmd#thm-subgroup-pm
[134] logistic-regression.qmd#lem-subgroup-pm-proportion
[135] logistic-regression.qmd#def-observed-marginal-rd
[136] logistic-regression.qmd#def-observed-marginal-rr
[137] logistic-regression.qmd#def-causal-marginal-rd
[138] logistic-regression.qmd#def-causal-marginal-rr
[139] logistic-regression.qmd#def-bootstrap-marginal-rd
[140] basic-statistical-methods.qmd#sec-bootstrap-ci
[141] logistic-regression.qmd#exm-wcgs-marginal-rd
[142] logistic-regression.qmd#def-collapsibility
[143] logistic-regression.qmd#thm-collapsibility
[144] logistic-regression.qmd#exm-collapsibility

Akaike, Hirotugu. 1974. "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control* 19 (6): 716–23. https://doi.org/10.1109/TAC.1974.1100705.

Bliss, C. I. 1935. "The Calculation of the Dosage-Mortality Curve." *Annals of Applied Biology* 22 (1): 134–67. https://doi.org/10.1111/j.1744-7348.1935.tb07713.x.

Boyd, Stephen, and Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press. https://web.stanford.edu/~boyd/cvxbook/.

Davison, Anthony C., and David V. Hinkley. 1997. *Bootstrap Methods and Their Application*. Cambridge Series in Statistical and Probabilistic Mathematics 1. Cambridge University Press. https://doi.org/10.1017/CBO9780511802843.

Dobson, Annette J, and Adrian G Barnett. 2018. *An Introduction to Generalized Linear Models*. 4th ed. CRC press. https://doi.org/10.1201/9781315182780.

Dunn, Peter K, and Gordon K Smyth. 2018. *Generalized Linear Models with Examples in R*. Vol. 53. Springer. https://link.springer.com/book/10.1007/978-1-4419-0118-7.

Efron, Bradley. 1979. "Bootstrap Methods: Another Look at the Jackknife." *The Annals of Statistics* 7 (1): 1–26. https://doi.org/10.1214/aos/1176344552.

Efron, Bradley, and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability 57. Chapman & Hall/CRC. https://doi.org/10.1201/9780429246593.

Graubard, Barry I., and Edward L. Korn. 1999. "Predictive Margins with Survey Data." *Biometrics* 55 (2): 652–59. https://doi.org/10.1111/j.0006-341X.1999.00652.x.

Greenland, Sander, and James M. Robins. 1986. "Identifiability, Exchangeability, and Epidemiological Confounding." *International Journal of Epidemiology* 15 (3): 413–19. https://doi.org/10.1093/ije/15.3.413.

Greenland, Sander, James M. Robins, and Judea Pearl. 1999. "Confounding and Collapsibility in Causal Inference." *Statistical Science* 14 (1): 29–46. https://doi.org/10.1214/ss/1009211805.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer. https://doi.org/10.1007/978-0-387-84858-7.

Hernán, Miguel A., and James M. Robins. 2020. *Causal Inference: What If*. Chapman & Hall/CRC. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

Hosmer, David W., Trina Hosmer, Saskia Le Cessie, and Stanley Lemeshow. 1997. "A Comparison of Goodness-of-Fit Tests for the Logistic Regression Model." *Statistics in Medicine* 16 (9): 965–80. https://doi.org/10.1002/(SICI)1097-0258(19970515)16:9%3C965::AID-SIM509%3E3.0.CO;2-O.

Hosmer, David W., and Stanley Lemeshow. 1980. "Goodness of Fit Tests for the Multiple Logistic Regression Model." *Communications in Statistics — Theory and Methods* 9 (10): 1043–69. https://doi.org/10.1080/03610928008827941.

Hosmer, David W, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied Logistic Regression*. John Wiley & Sons. https://onlinelibrary.wiley.com/doi/book/10.1002/9781118548387.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. *An Introduction to Statistical Learning: With Applications in R*. 2nd ed. Springer. https://doi.org/10.1007/978-1-0716-1418-1.

Lane, Peter W., and John A. Nelder. 1982. "Analysis of Covariance and Standardization as Instances of Prediction." *Biometrics* 38 (3): 613–21. https://doi.org/10.2307/2530043.

Nahhas, Ramzi W. 2024. *Introduction to Regression Methods for Public Health Using R*. CRC Press. https://www.bookdown.org/rwnahhas/RMPH/.

Norton, Edward C., Bryan E. Dowd, Melissa M. Garrido, and Matthew L. Maciejewski. 2024. "Requiem for Odds Ratios." *Health Services Research* 59 (4): e14337. https://doi.org/https://doi.org/10.1111/1475-6773.14337.

Pregibon, Daryl. 1981. "Logistic Regression Diagnostics." *The Annals of Statistics* 9 (4): 705–24. https://doi.org/10.1214/aos/1176345513.

Robins, James. 1986. "A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period—Application to Control of the Healthy Worker Survivor Effect." *Mathematical Modelling* 7 (9–12): 1393–512. https://doi.org/10.1016/0270-0255(86)90088-6.

Rosenman, Ray H, Richard J Brand, C David Jenkins, Meyer Friedman, Reuben Straus, and Moses Wurm. 1975. "Coronary Heart Disease in the Western Collaborative Group Study: Final Follow-up Experience of 8 1/2 Years." *JAMA* 233 (8): 872–77. https://doi.org/10.1001/jama.1975.03260080034016.

Sackett, David L, Jonathan J Deeks, and Doughs G Altman. 1996. "Down with Odds Ratios!" *BMJ Evidence-Based Medicine* 1 (6): 164.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics* 6 (2): 461–64. https://doi.org/10.1214/aos/1176344136.

Vittinghoff, Eric, David V Glidden, Stephen C Shiboski, and Charles E McCulloch. 2012. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*. 2nd ed. Springer. https://doi.org/10.1007/978-1-4614-1353-0.