Inference for Logistic Regression

Last modified: 2026-11-12 22:02:32 (PST)

Introduction

Logistic regression models a binary outcome, such as whether a patient recovers, from one or more predictors. Its fitted coefficients are maximum likelihood estimates, so the likelihood theory of Introduction to maximum likelihood gives them standard errors, confidence intervals and tests. This page specializes that theory to the logistic model and introduces the deviance, the measure of fit that the likelihood ratio test of a logistic model is built from [1].

Definition 0.1 (Logistic regression)**.** A **logistic regression** model for binary outcomes $y_1, \dots, y_n \in \{0, 1\}$ with predictor vectors $\underset{\sim}{x}_1, \dots, \underset{\sim}{x}_n$ says that the outcomes are independent and that

$$\Pr(Y_i = 1 \mid \underset{\sim}{x}_i) = \pi_i = \operatorname{expit}(\underset{\sim}{\beta} \cdot \underset{\sim}{x}_i), \qquad \operatorname{expit}(u) = \frac{1}{1 + e^{-u}}, \tag{1}$$

where $\underset{\sim}{x}_i = (1, x_{i1}, \dots, x_{ip})^\top$ and $\beta = (\beta_0, \beta_1, \dots, \beta_p)^\top$. The leading 1 in $\underset{\sim}{x}_i$ makes β_0 the intercept, and β_j for $j \geq 1$ is the slope for the jth predictor.
Source: Wikipedia contributors [2].

Example 0.1 (A six-patient trial)**.** Six patients are split into a control group, coded $x = -1$, and a treatment group, coded $x = +1$. The outcome is $y = 1$ if the patient recovered. Fitting the model of Equation 1 by maximum likelihood gives the coefficients below.

```r
# group coded -1 (control) and +1 (treatment); y = 1 means recovered
wf_x <- rep(c(-1, 1), each = 3)
wf_y <- c(0, 0, 1, 0, 1, 1)
# maximum likelihood fit of the logistic model
wf_fit <- glm(wf_y ~ wf_x, family = binomial)
# the fitted intercept and slope
round(coef(wf_fit), 3)
#> (Intercept)        wf_x
#>       0.000       0.693
```

The fitted slope is $\hat{\beta}_1 = 0.693$. The control group recovered at rate $1/3$ and the treatment group at $2/3$, so the model fits both group proportions exactly.

Odds and the coefficients

Lemma 0.1 (The log-odds in a logistic regression model)**.** *In the logistic regression model of Equation 1, the log-odds of $Y = 1$ at predictor vector $\underset{\sim}{x}$ are*

$$\eta(\underset{\sim}{x}) = \underset{\sim}{\beta} \cdot \underset{\sim}{x}.$$

Proof

Proof. Write $u \overset{\text{def}}{=} \underset{\sim}{\beta} \cdot \underset{\sim}{x}$ and $\pi \overset{\text{def}}{=} \operatorname{expit}(u)$, so that

$$\begin{aligned}
1 - \pi &= 1 - \frac{1}{1 + \mathrm{e}^{-u}} && \text{(the logistic model)} \\
&= \frac{1 + \mathrm{e}^{-u}}{1 + \mathrm{e}^{-u}} - \frac{1}{1 + \mathrm{e}^{-u}} && \text{(write 1 over the denominator } 1 + \mathrm{e}^{-u}\text{)} \\
&= \frac{(1 + \mathrm{e}^{-u}) - 1}{1 + \mathrm{e}^{-u}} && \text{(subtract fractions with the same denominator)} \\
&= \frac{\mathrm{e}^{-u}}{1 + \mathrm{e}^{-u}}. && ((1 + a) - 1 = a)
\end{aligned}$$

The first step substitutes the model of Equation 1. So the odds are

$$\begin{aligned}
\frac{\pi}{1 - \pi} &= \frac{1/(1 + \mathrm{e}^{-u})}{\mathrm{e}^{-u}/(1 + \mathrm{e}^{-u})} && \text{(the logistic model, and the display for } 1 - \pi\text{)} \\
&= \frac{1}{\mathrm{e}^{-u}} && \text{(divide fractions with the same denominator)} \\
&= \mathrm{e}^{u}, && (1/\mathrm{e}^{-u} = \mathrm{e}^{u})
\end{aligned}$$

and the log-odds are $\log\{e^u\} = u = \underset{\sim}{\beta} \cdot \underset{\sim}{x}$. □

Example 0.1 (The log-odds of recovery in the six-patient trial)**.**
In the six-patient trial, a treated patient has $\underset{\sim}{x} = (1, 1)^\top$. Under the maximum likelihood fit, the code computes the probability of recovery, its log-odds, and the dot product $\widehat{\underset{\sim}{\beta}} \cdot \underset{\sim}{x}$.

```
# fitted probability of recovery for a treated patient (x = +1)
wf_treated_data <- tibble::tibble(wf_x = 1)
wf_prob_treated <- predict(wf_fit, wf_treated_data, type = "response")
wf_prob_treated <- unname(wf_prob_treated)
wf_log_odds_treated <- log(wf_prob_treated / (1 - wf_prob_treated))
wf_coef <- coef(wf_fit)
wf_dot_treated <- sum(wf_coef)
c(prob = wf_prob_treated, log_odds = wf_log_odds_treated, dot = wf_dot_treated)
#>     prob log_odds      dot
#> 0.666667 0.693147 0.693147
```

That probability is 0.6667, and both the log-odds and the dot product are 0.6931, as Lemma 0.1 states.

Definition 0.1 (Odds multiplier)**.** Let:

- β be the coefficient vector of a logistic regression model.
- $\Delta\underset{\sim}{x}$ be a change in the predictor vector, whose first entry is 0, because every predictor vector has first entry 1.

The **odds multiplier** of $\Delta\underset{\sim}{x}$ is the exponential of the dot product of $\underset{\sim}{\beta}$ with $\Delta\underset{\sim}{x}$:

$$\theta_\omega(\Delta\underset{\sim}{x}) \stackrel{\text{def}}{=} e^{\underset{\sim}{\beta} \cdot \Delta\underset{\sim}{x}}.$$

Example 0.2 (The odds multiplier of moving from control to treatment)**.**
In the six-patient trial, moving a patient from the control group ($x = -1$) to the treatment group ($x = +1$) changes the predictor vector by $\Delta\underset{\sim}{x} = (0, 2)^\top$. The code computes the odds multiplier of that change under the maximum likelihood fit.

```
# change in the predictor vector from control to treatment
wf_change <- c(0, 2)
wf_change_terms <- coef(wf_fit) * wf_change
wf_change_dot <- sum(wf_change_terms)
wf_multiplier <- exp(wf_change_dot)
wf_multiplier
#> [1] 4
```

That odds multiplier is $e^{2\hat{\beta}_x} = 4$.

Example 0.3 (The risk ratio is not the odds multiplier)**.**
In the six-patient trial, the probability of recovery under the maximum likelihood fit is $2/3$ in the treatment group and $1/3$ in the control group. The code computes the risk ratio of treatment against control, and shows beside it the odds multiplier of the change $\Delta\underset{\sim}{x} = (0, 2)^\top$ from Example 0.2.

```
# fitted probabilities of recovery: control (x = -1), treatment (x = +1)
wf_group_codes <- c(-1, 1)
wf_groups <- tibble::tibble(wf_x = wf_group_codes)
wf_probs <- predict(wf_fit, wf_groups, type = "response")
wf_probs <- unname(wf_probs)
wf_prob_ratio <- wf_probs[2] / wf_probs[1]
c(risk_ratio = wf_prob_ratio, odds_multiplier = wf_multiplier)
#>      risk_ratio odds_multiplier
#>               2               4
```

The risk ratio is 2, but the odds multiplier is 4. The odds multiplier compares odds, 2 against $1/2$, while the risk ratio compares probabilities.

Theorem 0.1 (A change in the predictors multiplies the odds by its odds multiplier)**.**
Let:

- $\underset{\sim}{x}$ *be a predictor vector in the logistic regression model of Equation 1.*
- $\underset{\sim}{\Delta x}$ *be a change in the predictor vector, with first entry* 0*.*
- $\underset{\sim}{x}' \overset{\text{def}}{=} \underset{\sim}{x} + \Delta\underset{\sim}{x}$*.*

Then the odds of $Y = 1$ *at* $\underset{\sim}{x}'$ *are the odds at* $\underset{\sim}{x}$ *times the odds multiplier of* $\Delta\underset{\sim}{x}$*:*

$$\text{odds}\{\underset{\sim}{x}'\} = \text{odds}\{\underset{\sim}{x}\}\,\theta_\omega(\Delta\underset{\sim}{x}).$$

Proof

Proof. The log-odds are the natural logarithm of the odds, so the odds at a predictor vector are e raised to the log-odds there, which Lemma 0.1 gives.

$$\begin{aligned}
\operatorname{odds}\{\underset{\sim}{x}'\} &= \mathrm{e}^{\underset{\sim}{\beta}\cdot\underset{\sim}{x}'} && \text{(log-odds of the model at } \underset{\sim}{x}') \\
&= \mathrm{e}^{\underset{\sim}{\beta}\cdot(\underset{\sim}{x}+\Delta\underset{\sim}{x})} && \text{(definition of } \underset{\sim}{x}') \\
&= \mathrm{e}^{\underset{\sim}{\beta}\cdot\underset{\sim}{x}+\underset{\sim}{\beta}\cdot\Delta\underset{\sim}{x}} && \text{(the dot product is linear in each slot)} \\
&= \mathrm{e}^{\underset{\sim}{\beta}\cdot\underset{\sim}{x}}\,\mathrm{e}^{\underset{\sim}{\beta}\cdot\Delta\underset{\sim}{x}} && (\mathrm{e}^{a+b}=\mathrm{e}^{a}\mathrm{e}^{b}) \\
&= \operatorname{odds}\{\underset{\sim}{x}\}\,\mathrm{e}^{\underset{\sim}{\beta}\cdot\Delta\underset{\sim}{x}} && \text{(log-odds of the model at } \underset{\sim}{x}) \\
&= \operatorname{odds}\{\underset{\sim}{x}\}\,\theta_{\omega}(\Delta\underset{\sim}{x}). && \text{(definition of the odds multiplier)}
\end{aligned}$$

The first and fifth steps use Lemma 0.1, the third step uses the linearity of the dot product, and the last step uses Definition 0.1. □

Example 0.4 (The treated odds from the control odds)**.**
In the six-patient trial, take $\underset{\sim}{x} = (1, -1)^{\top}$, a control patient, and $\Delta\underset{\sim}{x} = (0, 2)^{\top}$, so that $\underset{\sim}{x}' = (1, 1)^{\top}$ is a treated patient. Under the maximum likelihood fit, the code computes the odds of recovery at both predictor vectors, and the control odds times the odds multiplier of $\Delta\underset{\sim}{x}$.

```
# fitted odds of recovery: control (x = -1), treatment (x = +1)
wf_odds <- wf_probs / (1 - wf_probs)
c(
  odds_control = wf_odds[1],
  odds_treated = wf_odds[2],
  control_times_multiplier = wf_odds[1] * wf_multiplier
)
#>             odds_control             odds_treated control_times_multiplier
#>                      0.5                      2.0                      2.0
```

The control odds are 0.5, and $0.5 \times 4 = 2$ equals the treated odds, as Theorem 0.1 states.

Corollary 0.1 (The odds ratio between two predictor vectors)**.** *Let $\underset{\sim}{x}$ and $\underset{\sim}{x}'$ be predictor vectors in the logistic regression model of Equation 1. The odds ratio of $\underset{\sim}{x}'$ compared with $\underset{\sim}{x}$ is the odds multiplier of their difference:*

$$\omega(\underset{\sim}{x}', \underset{\sim}{x}) = \theta_{\omega}(\underset{\sim}{x}' - \underset{\sim}{x}).$$

Proof

Proof. Let $\Delta \underset{\sim}{x} \overset{\text{def}}{=} \underset{\sim}{x}' - \underset{\sim}{x}$, so that $\underset{\sim}{x}' = \underset{\sim}{x} + \Delta \underset{\sim}{x}$, and the first entry of $\Delta \underset{\sim}{x}$ is $1 - 1 = 0$.

$$
\begin{aligned}
{}_{\omega}(\underset{\sim}{x}', \underset{\sim}{x}) &= \frac{\text{odds}\{\underset{\sim}{x}'\}}{\text{odds}\{\underset{\sim}{x}\}} && \text{(definition of the odds ratio)} \\
&= \frac{\text{odds}\{\underset{\sim}{x}\}\, \theta_{\omega}(\Delta \underset{\sim}{x})}{\text{odds}\{\underset{\sim}{x}\}} && \text{(odds multiplier theorem)} \\
&= \theta_{\omega}(\Delta \underset{\sim}{x}). && \text{(divide numerator and denominator by odds}\{\underset{\sim}{x}\} > 0)
\end{aligned}
$$

The first step is the definition of the odds ratio, and the second is Theorem 0.1. □

Example 0.5 (The odds ratio of treatment against control from the group counts)**.** In the six-patient trial, 2 of the 3 treated patients and 1 of the 3 control patients recovered. The code computes the odds ratio of treatment against control from those counts, and the odds multiplier under the maximum likelihood fit of $\underset{\sim}{x}' - \underset{\sim}{x} = (1, 1)^{\top} - (1, -1)^{\top}$.

```
# odds ratio from the counts: treated 2 of 3, control 1 of 3
count_or <- ((2 / 3) / (1 / 3)) / ((1 / 3) / (2 / 3))
wf_difference <- c(1, 1) - c(1, -1)
wf_difference_terms <- coef(wf_fit) * wf_difference
wf_difference_dot <- sum(wf_difference_terms)
wf_difference_multiplier <- exp(wf_difference_dot)
c(count_or = count_or, odds_multiplier = wf_difference_multiplier)
#>        count_or odds_multiplier
#>               4               4
```

Both are 4, as Corollary 0.1 states, because the model fits the two group proportions exactly.

Fitting by iteratively reweighted least squares

Remark 1: Source and license: iteratively reweighted least squares

This section is adapted from the generalized linear model part of the Newton's method lecture of Zhou [3], used under the MIT License. The license text is:


```
Copyright (c) 2025 ucla-biostat-257

Permission is hereby granted, free of charge, to any person obtaining a copy
of this software and associated documentation files (the "Software"), to deal
in the Software without restriction, including without limitation the rights
to use, copy, modify, merge, publish, distribute, sublicense, and/or sell
copies of the Software, and to permit persons to whom the Software is
furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all
copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR
IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY,
FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT.
IN NO EVENT SHALL THE
AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER
LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM,
OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE
SOFTWARE.
```


The source writes the logistic score and information, notes that Newton's method and Fisher scoring coincide, and rewrites the step as a weighted least squares fit. Here the step is stated as a theorem and proved, with the step length fixed at 1. The example, using the six-patient trial, is new. The source's treatment of general generalized linear models, Gauss-Newton and Levenberg-Marquardt is not part of this section. The general optimization methods behind IRLS are on the Optimization page of the math notes.

Definition 0.1 (Iteratively reweighted least squares (IRLS))**.** Suppose the outcomes follow the logistic regression model. Let X be the matrix whose rows are $\underset{\sim}{x}_i^\top$. From a current estimate $\underset{\sim}{\beta}^*$ with fitted probabilities π_i^* (collected in $\underset{\sim}{\pi}^*$), form

- the weights $w_i = \pi_i^*(1-\pi_i^*)$, collected in the diagonal matrix W;
- the working responses $z_i = \underset{\sim}{\beta}^* \cdot \underset{\sim}{x}_i + (y_i - \pi_i^*)/w_i$, collected in $\underset{\sim}{z}$.

Iteratively reweighted least squares replaces $\underset{\sim}{\beta}^*$ by the weighted least squares fit of

$\underset{\sim}{z}$ on X, the minimizer of $\sum_{i=1}^{n} w_i(z_i - \underset{\sim}{\beta} \cdot \underset{\sim}{x}_i)^2$, which is

$$((X^\top W X))^{-1} X^\top W \underset{\sim}{z},$$

and repeats until the estimate stops changing.

Exercise 0.1 (One IRLS step for the six-patient trial)**.** In the six-patient trial, start from $\underset{\sim}{\beta}^* = \underset{\sim}{0}$.

1. Find the fitted probabilities, the weights and the working responses.
2. Compute $X^\top W X$ and $X^\top W \underset{\sim}{z}$.
3. Compute the updated coefficients $((X^\top W X))^{-1} X^\top W \underset{\sim}{z}$.

Solution

Solution 0.1. Here the rows of X are $(1, -1)$ for the three patients in the group coded -1 and $(1, 1)$ for the three treated patients, and $y = (0, 0, 1, 0, 1, 1)^\top$.
1. At $\beta^* = 0$ every fitted probability is $\operatorname{expit}(0) = 1/2$, so every weight is $w_i = (1/2)(1/2) = 1/4$. The working responses are $z_i = 0 + (y_i - 1/2)/(1/4) = 4y_i - 2$:

$$\underset{\sim}{z} = (-2, -2, 2, -2, 2, 2)^\top.$$

2. With $W = \frac{1}{4}\mathbf{I}_6$,

$$X^\top W X = \tfrac{1}{4}\begin{bmatrix} 6 & 0 \\ 0 & 6 \end{bmatrix} = \begin{bmatrix} 3/2 & 0 \\ 0 & 3/2 \end{bmatrix},$$

because the entries of the first column of X sum to 6, the squares in the second column sum to 6, and the two columns multiply to a sum of $-3+3 = 0$. The first entry of $X^\top W \underset{\sim}{z}$ is $\frac{1}{4}(-2-2+2-2+2+2) = 0$, and the second is $\frac{1}{4}((-1)(-2-2+2) + (1)(-2+2+2)) = \frac{1}{4}(2+2) = 1$, so $X^\top W \underset{\sim}{z} = (0, 1)^\top$.
3. The updated coefficients are

$$\begin{bmatrix} 2/3 & 0 \\ 0 & 2/3 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 2/3 \end{bmatrix}.$$

The intercept stays at 0 and the slope moves from 0 to 2/3.

Theorem 0.1 (An IRLS step is a Newton-Raphson step and a Fisher scoring step)**.** *Suppose the outcomes follow the logistic regression model. One step of iteratively reweighted least squares from* $\underset{\sim}{\beta}^*$ *equals*

- *one step of the Newton-Raphson algorithm on the log-likelihood from* $\underset{\sim}{\beta}^*$*, and*

- *one step of Fisher scoring from* $\underset{\sim}{\beta}^*$*:*

$$\underset{\sim}{\beta}^* + ((X^\top W X))^{-1} X^\top (\underset{\sim}{y} - \underset{\sim}{\pi}^*). \tag{2}$$

Proof

Proof. The log-likelihood has score $X^\top(y - \pi)$ and Hessian $-X^\top W X$ at every β (with π and W computed from β), and the Hessian does not involve the outcomes (proof of Theorem 0.1). So the observed information $X^\top W X$ equals the expected information, and the Newton-Raphson and Fisher scoring steps both equal Equation 2.
For IRLS, the working responses satisfy $W\underset{\sim}{z} = WX\underset{\sim}{\beta}^* + (\underset{\sim}{y} - \underset{\sim}{\pi}^*)$, so

$$\begin{aligned} ((X^\top W X))^{-1} X^\top W \underset{\sim}{z} &= ((X^\top W X))^{-1} X^\top W X \underset{\sim}{\beta}^* + ((X^\top W X))^{-1} X^\top (\underset{\sim}{y} - \underset{\sim}{\pi}^*) && \text{(substitute } W\underset{\sim}{z}\text{)} \\ &= \underset{\sim}{\beta}^* + ((X^\top W X))^{-1} X^\top (\underset{\sim}{y} - \underset{\sim}{\pi}^*). && \text{(cancel } X^\top W X\text{)} \end{aligned}$$

□

Example 0.1 (IRLS on the six-patient trial)**.** The code runs IRLS from $\underset{\sim}{\beta}^* = \underset{\sim}{0}$ on the data of the six-patient trial and compares the result with the coefficients from `glm()`.

```
# design matrix: intercept, group (-1 control, +1 treatment); y = 1: recovered
irls_x <- cbind(1, rep(c(-1, 1), each = 3))
irls_y <- c(0, 0, 1, 0, 1, 1)
# start from zero coefficients and take six IRLS steps
irls_coef <- c(0, 0)
irls_path <- data.frame(step = 0, intercept = 0, slope = 0)
for (step in 1:6) {
  prob <- plogis(drop(irls_x %*% irls_coef))
  weight <- prob * (1 - prob)
  work_resp <- drop(irls_x %*% irls_coef) + (irls_y - prob) / weight
  irls_coef <- drop(solve(crossprod(irls_x, weight * irls_x),
                          crossprod(irls_x, weight * work_resp)))
  irls_path[step + 1, ] <- c(step, irls_coef)
}
irls_path
#>   step    intercept    slope
#> 1    0  0.00000e+00 0.000000
#> 2    1  0.00000e+00 0.666667
#> 3    2 -2.47644e-16 0.693033
#> 4    3  5.44066e-17 0.693147
#> 5    4 -1.12125e-16 0.693147
#> 6    5  2.20942e-16 0.693147
#> 7    6  5.77161e-17 0.693147
# the maximum likelihood fit from glm()
coef(glm(irls_y ~ irls_x[, 2], family = binomial))
#>  (Intercept)  irls_x[, 2]
#> -2.25619e-16  6.93147e-01
```

The first step moves the slope to 0.6667, which is 2/3 (the hand computation). After six steps the slope is 0.693147, and `glm()` reports 0.693147. The slope settles at $\ln 2 = 0.693147$, half the natural logarithm of the odds ratio between the treatment group (recovery odds 2) and the other group (recovery odds 1/2), since the groups are coded ± 1.

Deviance

Definition 0.1 (Deviance)**.** The **deviance** of a fitted logistic regression with labels $y_i \in \{0, 1\}$ is minus twice its maximized log-likelihood:

$$\text{dev} \stackrel{\text{def}}{=} -2\,\ell\big(\widehat{\underset{\sim}{\beta}}\big) = -2\sum_{i=1}^{n}\{y_i \ln \hat{\pi}_i + (1 - y_i)\ln(1 - \hat{\pi}_i)\}, \tag{3}$$

where $\hat{\pi}_i = \text{expit}\big(\widehat{\underset{\sim}{\beta}} \cdot \underset{\sim}{x}_i\big)$ is the fitted probability.
Source: Wikipedia contributors [4].

Source

Nelder and Wedderburn [5] introduced the deviance for the whole family of generalized linear models; logistic regression is the case of Bernoulli outcomes with the logit link. Hastie et al. [6, Sec. 4.4.2] defines the residual deviance of a fitted logistic regression as minus twice its log-likelihood.

Example 0.1 (The deviance of the six-patient trial)**.** For the fit of Example 0.1, R's `glm` reports the deviance. We check it against Equation 3 computed by hand.

```
# fitted probabilities of recovery
wf_p <- fitted(wf_fit)
# minus twice the log-likelihood, from the formula
wf_dev_hand <- -2 * sum(wf_y * log(wf_p) + (1 - wf_y) * log(1 - wf_p))
# the deviance that glm reports
wf_dev_glm <- deviance(wf_fit)
c(by_hand = wf_dev_hand, glm = wf_dev_glm)
#> by_hand     glm
#> 7.63817 7.63817
```

Both give dev $= 7.638$.

Remark (Why the deviance measures a shortfall)

Remark 0.1 (Why the deviance measures a shortfall). For labels $y_i \in \{0, 1\}$ the largest log-likelihood any model can reach is 0, by giving each label probability 1. So dev is twice the amount by which the maximized log-likelihood falls short of that largest value.

Definition 0.2 (Null deviance)**.** The **null deviance** is the deviance of the model with an intercept and no other predictors:

$$\text{dev}_0 = -2n\{\bar{y}\ln\bar{y} + (1-\bar{y})\ln(1-\bar{y})\}, \tag{4}$$

where $\bar{y}$ is the fraction of labels equal to 1.

Example 0.2 (The null deviance of the six-patient trial)**.** Three of the six patients recovered, so $\bar{y} = 1/2$ and Equation 4 gives $\text{dev}_0 = 2 \times 6 \ln 2$.

```
# null deviance from the formula with ybar = 1/2
wf_null_hand <- 2 * 6 * log(2)
# the null deviance that glm reports
wf_null_glm <- wf_fit$null.deviance
c(by_hand = wf_null_hand, glm = wf_null_glm)
#> by_hand     glm
#> 8.31777 8.31777
```

Both give $\text{dev}_0 = 8.318$.

Remark (Where the null deviance formula comes from)

Remark 0.2 (Where the null deviance formula comes from). The intercept-only model gives every observation the same fitted probability, and the maximizing value is the overall fraction $\bar{y}$ of labels equal to 1. Substituting $\bar{y}$ for each $\hat{\pi}_i$ in Equation 3 gives Equation 4.

Theorem 0.1 (The drop in deviance is the likelihood ratio statistic)**.** *Let* dev_0 *be the null deviance and* dev *the deviance of a logistic regression with* q *predictors besides the intercept. Then the likelihood ratio statistic of* H_0*: all* q *slopes are zero is*

$$\Lambda = \text{dev}_0 - \text{dev}\,.$$

Proof

Proof. In the notation of the likelihood ratio statistic, the unrestricted MLE $\hat{\theta}_{\text{ML}}$ is the fitted coefficient vector $\hat{\beta}$, and $\hat{\theta}_0$ is the fit of the intercept-only model that H_0 imposes. Then $\ell\big(\hat{\theta}_{\text{ML}}\big) = -\,\text{dev}\,/2$ and $\ell\big(\hat{\theta}_0\big) = -\,\text{dev}_0\,/2$. Substituting both into the definition gives $\Lambda = 2(-\,\text{dev}\,/2 + \text{dev}_0\,/2) = \text{dev}_0 - \text{dev}$. □

Example 0.3 (The likelihood ratio statistic of the treatment effect)**.** In the six-patient trial, $q = 1$.

```
# drop in deviance, the likelihood ratio statistic
wf_g2 <- wf_fit$null.deviance - deviance(wf_fit)
wf_g2
#> [1] 0.679596
```

The statistic is 0.68.

Corollary 0.1 (Distribution of the drop in deviance)**.** *Suppose the regularity conditions of Wilks' theorem hold for the logistic regression model. Then for a large sample the drop in deviance* $\text{dev}_0 - \text{dev}$ *has approximately a* χ^2_q *distribution when* H_0 *holds.*

Proof

Proof. By Theorem 0.1 the drop in deviance is the likelihood ratio statistic of q constraints, and Wilks' theorem gives its approximate distribution. □

Example 0.4 (Testing the treatment effect)**.** For the statistic of Example 0.3, the p-value comes from the chi-square distribution with $q = 1$ degree of freedom.

```
# upper tail of the chi-square distribution with 1 degree of freedom
wf_g2_p <- pchisq(wf_g2, df = 1, lower.tail = FALSE)
wf_g2_p
#> [1] 0.409726
```

The p-value is 0.41, so six patients give no evidence of a treatment effect.

Covariate patterns and the saturated model

Definition 0.1 (Covariate pattern)**.** A **covariate pattern** is a distinct value of the predictor vector among the observations $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$. The K distinct values are written $\underset{\sim}{x}^{(1)}, \ldots, \underset{\sim}{x}^{(K)}$, so $K \leq n$.

Definition 0.2 (Replicates)**.** The **replicates** of the covariate pattern $\underset{\sim}{x}^{(k)}$ are the observations i with $\underset{\sim}{x}_i = \underset{\sim}{x}^{(k)}$. Their number is written m_k, so $\sum_{k=1}^{K} m_k = n$.

Remark 2: Source: covariate patterns and replicates

Dobson and Barnett [7, Sec. 5.5] introduces both terms: observations that share a combination of factor levels and the same values of any continuous predictors have the same covariate pattern, and are called replicates.

Definition 0.3 (Events at a covariate pattern)**.** The number of **events** at the covariate pattern $\underset{\sim}{x}^{(k)}$ is the sum of the binary outcomes over its replicates:

$$s_k = \sum_{i:\, \underset{\sim}{x}_i = \underset{\sim}{x}^{(k)}} y_i,$$

so $0 \le s_k \le m_k$.

Example 0.1 (Covariate patterns in the low-birth-weight data)**.** The `birthwt` data record whether each of 189 babies had low birth weight (`low = 1`, under 2.5 kg), along with the mother's smoking status and race, among other predictors [8]. Take `smoke` (0 or 1) and `race` (1, 2 or 3) as the predictors. Counting the distinct pairs, and the babies and low-weight births at each, gives one row per covariate pattern.

```
bw <- read.csv("data/birthwt.csv")
bw$race <- factor(bw$race)
# one row per covariate pattern: replicates and events
bw_grouped <- bw |>
  dplyr::summarize(
    nreps = dplyr::n(),
    nevents = sum(low),
    .by = c(smoke, race)
  ) |>
  dplyr::arrange(smoke, race)
bw_grouped
#>   smoke race nreps nevents
#> 1     0    1    44       4
#> 2     0    2    16       5
#> 3     0    3    55      20
#> 4     1    1    52      19
#> 5     1    2    10       6
#> 6     1    3    12       5
```

There are $K = 6$ covariate patterns among $n = 189$ babies. The pattern with the most replicates has `smoke` $= 0$ and `race` $= 3$, and it has 55 of them.

Definition 0.4 (Saturated model (full model, maximal model))**.** The **saturated model** (also called the **full model** or **maximal model** [7, Sec. 5.5]) for binary outcomes gives each pattern its own free event probability:

$$\Pr\big(Y_i = 1 \mid \underset{\sim}{x}_i = \underset{\sim}{x}^{(k)}\big) = \pi_k, \qquad k = 1, \dots, K, \tag{5}$$

with the outcomes independent given the predictors.
Source: Wikipedia contributors [9].

Remark 3: Source: saturated model

Dobson and Barnett [7, Sec. 5.5] defines the saturated model as the model with the maximum number of parameters that can be estimated, one per distinct covariate pattern, and says it is also called a maximal or full model. In model selection, "full model" often means something else: the model with every candidate predictor [7, Sec. 6.3]. That model is generally smaller than the saturated model.

Theorem 0.1 (The saturated model fits each pattern's observed proportion)**.** *The maximum likelihood estimate of π_k in the saturated model is the observed proportion of events among the replicates of pattern k:*

$$\hat{\pi}_k = \frac{s_k}{m_k}.$$

Proof

Proof. Independence gives each pattern its own log-likelihood term, and the outcomes at pattern k contribute $s_k \ln \pi_k + (m_k - s_k)\ln(1 - \pi_k)$. The terms share no parameters, so each is maximized on its own. First suppose $0 < s_k < m_k$, so that both logarithms are defined for $0 < \pi_k < 1$. Setting the derivative with respect to π_k to zero gives $s_k/\pi_k - (m_k - s_k)/(1 - \pi_k) = 0$. Multiplying through by $\pi_k(1 - \pi_k)$ gives $s_k(1 - \pi_k) = (m_k - s_k)\pi_k$. Expanding gives $s_k - s_k\pi_k = m_k\pi_k - s_k\pi_k$. Adding $s_k\pi_k$ to both sides gives $s_k = m_k\pi_k$. Dividing both sides by m_k gives $\hat{\pi}_k = s_k/m_k$. The log-likelihood term increases to this point and decreases after it, so the point is a maximum. Now suppose $s_k = 0$. A count of 0 times a logarithm is taken as 0, so the term is $m_k\ln(1 - \pi_k)$, which decreases in π_k and so peaks at $\pi_k = 0$. Next suppose $s_k = m_k$. By the same convention, the term is $m_k \ln \pi_k$, which peaks at $\pi_k = 1$. Both agree with s_k/m_k. □

Definition 0.5 (Deviance relative to the saturated model)**.** The **deviance relative to the saturated model** of a fitted model is twice the amount by which its maximized

log-likelihood $\ell\big(\widehat{\underset{\sim}{\beta}}\big)$ falls short of the maximized log-likelihood ℓ_{sat} of the saturated model:

$$\text{dev}_{\text{sat}} = 2\big(\ell_{\text{sat}} - \ell\big(\widehat{\underset{\sim}{\beta}}\big)\big). \tag{6}$$

Theorem 0.2 (The deviance relative to the saturated model, from grouped counts)**.** *For a fitted logistic regression with fitted event probabilities $\hat{\pi}_k$ at the K covariate patterns,*

$$\text{dev}_{\text{sat}} = 2\sum_{k=1}^{K}\left\{ s_k \ln\left(\frac{s_k}{m_k\hat{\pi}_k}\right) + (m_k - s_k)\ln\left(\frac{m_k - s_k}{m_k(1-\hat{\pi}_k)}\right)\right\}, \tag{7}$$

where a term with $s_k = 0$ or $s_k = m_k$ contributes 0 in its zero-count part.

Proof

Proof. Grouping the outcomes by pattern, the fitted model's log-likelihood is $\ell\big(\widehat{\underset{\sim}{\beta}}\big) = \sum_{k=1}^{K}\{s_k \ln\hat{\pi}_k + (m_k - s_k)\ln(1-\hat{\pi}_k)\}$. By Theorem 0.1, the saturated log-likelihood is the same sum with $\hat{\pi}_k$ replaced by s_k/m_k. Subtracting the fitted from the saturated log-likelihood, term by term, replaces each $\ln(s_k/m_k) - \ln\hat{\pi}_k$ by $\ln(s_k/(m_k\hat{\pi}_k))$, and likewise for the non-events. Multiplying by 2, as Equation 6 does, gives Equation 7. A count of 0 times $\ln 0$ is taken as 0, its limit. □

Remark 4: Source: deviance from grouped counts

Dobson and Barnett [7, Sec. 5.6.1] derives the deviance of a binomial model relative to the saturated model, and Section 7.5 writes it in the form of Equation 7. Section 7.5 also notes that when every observation has its own covariate pattern, so each count is 0 or 1, this deviance is not a useful measure of fit.

Corollary 0.1 (The deviance is the saturated-model deviance shifted by a constant)**.** *The deviance of a fitted logistic regression and its deviance relative to the saturated model differ by a constant that depends on the data and not on the fitted model:*

$$\text{dev} = \text{dev}_{\text{sat}} - 2\ell_{\text{sat}}.$$

Proof

Proof. Definition 0.1 gives $\text{dev} = -2\ell\big(\widehat{\underset{\sim}{\beta}}\big)$. Distributing the 2 in Equation 6 gives $\text{dev}_{\text{sat}} = 2\ell_{\text{sat}} - 2\ell\big(\widehat{\underset{\sim}{\beta}}\big)$. Substituting $\text{dev} = -2\ell\big(\widehat{\underset{\sim}{\beta}}\big)$ into this expression gives $\text{dev}_{\text{sat}} = 2\ell_{\text{sat}} + \text{dev}$. Subtracting $2\ell_{\text{sat}}$ from both sides gives the claim. □

Remark (Why the shift does not change the drop in deviance)

Remark 0.1 (Why the shift does not change the drop in deviance). The constant $2\ell_{\text{sat}}$ in Corollary 0.1 is the same for every model fit to the same data, because the saturated model depends only on the data. So the difference between the deviances of two models, such as the drop in deviance of Theorem 0.1, is the same whether it is computed from dev or from dev_{sat}. The two deviances differ in their size, though: when every pattern has one replicate, the saturated model fits every label exactly, $\ell_{\text{sat}} = 0$, and $\text{dev}_{\text{sat}} = \text{dev}$.

Example 0.2 (The deviance of the low-birth-weight model, ungrouped and grouped)**.** Fit a logistic regression of low birth weight on smoking and race, once to the 189 individual babies and once to the 6 grouped counts of Example 0.1. R's `deviance()` is the deviance of Definition 0.1 for the first fit and the deviance relative to the saturated model of Definition 0.5 for the second. We check the second against Equation 7 computed by hand.

```r
# fit to the individual babies
bw_fit <- glm(low ~ smoke + race, family = binomial, data = bw)
# fit to the grouped counts: events and non-events at each pattern
bw_fit_grouped <- glm(
  cbind(nevents, nreps - nevents) ~ smoke + race,
  family = binomial,
  data = bw_grouped
)
# fitted probability at each pattern
bw_phat <- fitted(bw_fit_grouped)
# one term of the grouped formula; a zero count contributes 0
bw_term <- function(count, expected) {
  ifelse(count == 0, 0, count * log(count / expected))
}
bw_sat_dev_hand <- 2 * sum(
  bw_term(bw_grouped$nevents, bw_grouped$nreps * bw_phat) +
    bw_term(
      bw_grouped$nreps - bw_grouped$nevents,
      bw_grouped$nreps * (1 - bw_phat)
    )
)
c(
  by_hand = bw_sat_dev_hand,
  glm_grouped = deviance(bw_fit_grouped),
  glm_ungrouped = deviance(bw_fit)
)
#>       by_hand   glm_grouped glm_ungrouped
#>       3.15694       3.15694     219.97471
```

The deviance relative to the saturated model is $\text{dev}_{\text{sat}} = 3.157$ on the grouped counts, and the deviance is $\text{dev} = 219.975$ on the individual babies.

Example 0.3 (The shift between the two deviances)**.** For the two fits of Example 0.2, Corollary 0.1 says that $\text{dev} - \text{dev}_{\text{sat}} = -2\ell_{\text{sat}}$. The left side is 216.818. We compute the right side from the observed proportions of Theorem 0.1.

```r
# twice the saturated log-likelihood, from the observed proportions
bw_phat_sat <- bw_grouped$nevents / bw_grouped$nreps
bw_two_llik_sat <- 2 * sum(
  ifelse(bw_grouped$nevents == 0, 0, bw_grouped$nevents * log(bw_phat_sat)) +
    ifelse(
      bw_grouped$nreps == bw_grouped$nevents,
      0,
      (bw_grouped$nreps - bw_grouped$nevents) * log(1 - bw_phat_sat)
    )
)
c(
  difference = deviance(bw_fit) - deviance(bw_fit_grouped),
  minus_two_llik_sat = -bw_two_llik_sat
)
#>         difference minus_two_llik_sat
#>            216.818            216.818
```

Confidence intervals for the coefficients

Theorem 0.1 (Approximate distribution of the fitted coefficients)**.** *Suppose the outcomes follow the logistic regression model. Let X be the matrix whose rows are $\underset{\sim}{x}_i^\top$, and let W be the diagonal matrix with entries $\pi_i(1-\pi_i)$. Under regularity conditions, for a large sample the fitted coefficients are approximately Gaussian,*

$$\hat{\underset{\sim}{\beta}} \approx \mathrm{N}\big(\underset{\sim}{\beta}, (X^\top W X)^{-1}\big). \tag{8}$$

Proof

Proof. The log-likelihood has score $X^\top(y - \underset{\sim}{\pi})$, and its Hessian is $-X^\top W X$, which does not involve the outcomes. The observed information and expected information are therefore both $X^\top W X$. The claim is the Gaussian approximation of the central limit theorem for MLEs, with this information in place of the inverse covariance. The result is stated for logistic regression by Hastie et al. [6, Sec. 4.4.3]. For independent but non-identically distributed outcomes it holds under conditions on the predictors that Fahrmeir and Kaufmann [10] give. □

Corollary 0.1 (Wald interval for a logistic coefficient)**.** *Let $\hat{W}$ be W with each π_i replaced by its fitted value. The estimated standard error of $\hat{\beta}_j$ is the square root of the diagonal entry of $\left(X^\top \hat{W} X\right)^{-1}$ in the row and column of β_j; because β_0 comes first, that is the $(j+1)$th diagonal entry counting from 1, and a Wald confidence interval with level $1-\alpha$ for β_j is*

$$\hat{\beta}_j \pm z_{1-\alpha/2}\, \widehat{\mathrm{se}}\left(\hat{\beta}_j\right).$$

Example 0.1 (A Wald interval from six patients)**.** For the fit of Example 0.1, `vcov` returns $\left(X^\top \hat{W} X\right)^{-1}$.

```
# standard errors: square roots of the diagonal of (X'WX)^-1
wf_se <- sqrt(diag(vcov(wf_fit)))
wf_slope <- coef(wf_fit)[["wf_x"]]
# 95% Wald interval for the slope
wf_ci <- wf_slope + c(-1, 1) * qnorm(0.975) * wf_se[["wf_x"]]
wf_summary <- c(
  slope = wf_slope, se = wf_se[["wf_x"]], lower = wf_ci[1], upper = wf_ci[2]
)
round(wf_summary, 3)
#>  slope     se  lower  upper
#>  0.693  0.866 -1.004  2.391
```

The interval for β_1 is $(-1, 2.39)$. It contains 0, consistent with the test of Example 0.4.

Corollary 0.2 (Interval for the odds factor of a logistic coefficient)**.** *If (L, U) is a confidence interval for β_j from Corollary 0.1, then (e^L, e^U) is a confidence interval with the same level for e^{β_j}, the factor by which a one-unit increase in x_j, holding the other predictors fixed, multiplies the odds of $y = 1$.*

Proof

Proof. The exponential function is increasing, so $L \le \beta_j \le U$ holds exactly when $e^L \le e^{\beta_j} \le e^U$. □

Example 0.2 (How little six patients tell us)**.** The two groups of Example 0.1 sit two units apart on the x scale, so the odds ratio between them is $e^{2\beta_1}$.

```r
# fitted group odds ratio, and its interval: exponentiate twice the slope
wf_group_or_hat <- exp(2 * wf_slope)
wf_group_or <- exp(2 * wf_ci)
c(estimate = wf_group_or_hat, lower = wf_group_or[1], upper = wf_group_or[2])
#>   estimate      lower      upper
#>   4.000000   0.134195 119.229587
```

The fitted odds of recovery in the treatment group are 4 times those in the control group, with interval $(0.13, 119.2)$. The interval runs from a strong harm to a large benefit. Its width shows how little information six observations carry about the slope.

Further reading

- Hua Zhou's course page for "UCLA Biostat 257 - Statistical Computing" (2025 Spring) has lectures on fitting generalized linear models and on quasi-Newton optimization. Its source is on GitHub under the MIT License.
- David Rocke's EPI 204 lecture *Logistic Regression and Likelihood* (slides) fits logistic regression by maximum likelihood and compares models by deviance and AIC [11].
- Josh Starmer's StatQuest video *Logistic Regression Details Pt 2: Maximum Likelihood* shows how maximum likelihood estimation fits a logistic regression curve [12].

References

[1] Wikipedia contributors, "Likelihood-ratio test — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Likelihood-ratio_test

[2] Wikipedia contributors, "Logistic regression — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Logistic_regression

[3] H. Zhou, "Newton's method and Fisher scoring." Lecture notes for Biostat M257, Statistical Computing, University of California, Los Angeles, 2025. Accessed: Oct. 26, 2026. [Online]. Available: https://ucla-biostat-257.github.io/2025spring/slides/23-newton/newton.html

[4] Wikipedia contributors, "Deviance (statistics) — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Deviance_%28statistics%29

[5] J. A. Nelder and R. W. M. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society: Series A (General)*, vol. 135, no. 3, pp. 370–384, 1972, doi: 10.2307/2344614.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*, 2nd ed. Springer, 2009. doi: 10.1007/978-0-387-84858-7.

[7] A. J. Dobson and A. G. Barnett, *An introduction to generalized linear models*, 4th ed. CRC press, 2018. doi: 10.1201/9781315182780.

[8] D. W. Hosmer and S. Lemeshow, *Applied logistic regression*. New York: Wiley, 1989.

[9] Wikipedia contributors, "Saturated model — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Saturated_model

[10] L. Fahrmeir and H. Kaufmann, "Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models," *The Annals of Statistics*, vol. 13, no. 1, pp. 342–368, 1985, doi: 10.1214/aos/1176346597.

[11] D. M. Rocke, "EPI 204 lecture notes, Spring 2021." University of California, Davis, 2021. Available: https://dmrocke.ucdavis.edu/Class/EPI204-Spring-2021/EPI204-Spring-2021.html

[12] J. Starmer, "Logistic regression details, part 2: Maximum likelihood." StatQuest, video on YouTube, 2018. Available: https://www.youtube.com/watch?v=BfKanl1aSG0