Exam Formula Sheet

Epi 202: Probability

$$\begin{aligned}\operatorname{Var}\left(\tilde{a}\cdot\tilde{X}\right) &= \operatorname{Var}\left(\sum_{i=1}^{n} a_i X_i\right)\\ &= \tilde{a}^{\top}\operatorname{Var}\left(\tilde{X}\right)\tilde{a}\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \operatorname{Cov}\left(X_i, X_j\right)\end{aligned}$$

$$\mathrm{E}[Y] = \mathrm{E}[\mathrm{E}[Y \mid X]]$$

$$\mathrm{E}[Y \mid Z] = \mathrm{E}[\mathrm{E}[Y \mid X, Z] \mid Z]$$

$$\operatorname{Var}(Y) = \mathrm{E}[\operatorname{Var}(Y \mid X)] + \operatorname{Var}(\mathrm{E}[Y \mid X])$$

$$\operatorname{Cov}(Y, Z) = \mathrm{E}[\operatorname{Cov}(Y, Z \mid X)] + \operatorname{Cov}(\mathrm{E}[Y \mid X], \mathrm{E}[Z \mid X])$$

Epi 203: Statistical inference

$$\mathcal{L}(\theta) \stackrel{\text{def}}{=} \mathrm{p}(\tilde{X} = \tilde{x} | \Theta = \theta)$$

$$\ell \stackrel{\text{def}}{=} \log\{\mathcal{L}(\tilde{x}|\theta)\}$$

$$\ell' \stackrel{\text{def}}{=} \frac{\partial}{\partial\theta}\ell(\tilde{x}|\theta)$$

$$\ell'' \stackrel{\text{def}}{=} \frac{\partial}{\partial\tilde{\theta}}\frac{\partial}{\partial\tilde{\theta}^{\top}}\ell(\tilde{x}|\tilde{\theta})$$

$$\ell''_{ij} = \frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}\ell(\tilde{X} = \tilde{x}|\tilde{\theta})$$

$$\begin{aligned}I &\stackrel{\text{def}}{=} -\ell''(\tilde{x}|\tilde{\theta})\\ \mathcal{I} &\stackrel{\text{def}}{=} \mathrm{E}[I(\tilde{x}|\theta)]\end{aligned}$$

$$\hat{\theta}_{ML} \mathrel{\dot{\sim}} \mathrm{N}\left(\theta, \left[\mathcal{I}(\tilde{\theta})\right]^{-1}\right)$$

For one parameter θ_k:

$$\mathrm{CI}_{1-\alpha}(\theta_k) = \left[\hat{\theta}_k \pm z_{1-\frac{\alpha}{2}}\,\widehat{\mathrm{SE}}\left(\hat{\theta}_k\right)\right]$$

$$Z_k \stackrel{\text{def}}{=} \frac{\hat{\theta}_k - \theta_{k,0}}{\widehat{\mathrm{SE}}\left(\hat{\theta}_k\right)} \mathrel{\dot{\sim}} \mathrm{N}(0,1) \qquad \text{under } H_0 : \theta_k = \theta_{k,0} \qquad z_k = \text{observed } Z_k$$

$$\begin{aligned}p\text{-value} &= 2\Pr(|Z| \geq |z_k|), \quad Z \sim \mathrm{N}(0,1)\\ &= 2(1 - \Phi(|z_k|))\end{aligned}$$

Sta 108: Linear regression

$$t_k \stackrel{\text{def}}{=} \frac{\hat{\beta}_k - \beta_{k,0}}{\widehat{\text{SE}}\left(\hat{\beta}_k\right)} \dot{\sim} t_{n-p} \qquad \text{under } H_0 : \beta_k = \beta_{k,0} \qquad t_k^{\text{obs}} = \text{observed } t_k$$

$$\text{CI}_{1-\alpha}(\beta_k) = \left[\hat{\beta}_k \pm t_{n-p}\left(1 - \frac{\alpha}{2}\right) \widehat{\text{SE}}\left(\hat{\beta}_k\right)\right]$$

Let $T_{n-p} \sim t_{n-p}$.

$$\begin{aligned} p\text{-value} &= 2 \Pr\left(\left|T_{n-p}\right| \geq \left|t_k^{\text{obs}}\right|\right) \\ &= 2\left(1 - F_{t_{n-p}}\left(\left|t_k^{\text{obs}}\right|\right)\right) \end{aligned}$$

$$\text{CI}_{1-\alpha}(\mu(\tilde{x}^*)) = \left[\hat{\mu}(\tilde{x}^*) \pm t_{n-p}\left(1 - \frac{\alpha}{2}\right) \widehat{\text{SE}}(\hat{\mu}(\tilde{x}^*))\right]$$

Y^* denotes a new observation (not in the training data), with corresponding covariate pattern $\tilde{x}^*$. Let $\hat{Y}^* \stackrel{\text{def}}{=} \hat{\mu}(\tilde{x}^*)$.

$$\text{PI}_{1-\alpha}(Y^*|\tilde{x}^*) = \left[\hat{Y}^* \pm t_{n-p}\left(1 - \frac{\alpha}{2}\right) \widehat{\text{SE}}\left(Y^* - \hat{Y}^*\right)\right]$$

$$\widehat{\text{SE}}\left(Y^* - \hat{Y}^*\right) = \hat{\sigma}\sqrt{1 + (\tilde{x}^*)^\top (\mathbf{X}'\mathbf{X})^{-1}\tilde{x}^*}$$

$$\text{Var}\left(Y^* - \hat{Y}^*\right) = \sigma^2\left(1 + (\tilde{x}^*)^\top (\mathbf{X}'\mathbf{X})^{-1}\tilde{x}^*\right)$$

Let $\hat{\Sigma} \stackrel{\text{def}}{=} \widehat{\text{Var}}\left(\hat{\beta}\right) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$.

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

Let $\Delta\mu(\tilde{x}, \tilde{x}^*) = \mu(\tilde{x}) - \mu(\tilde{x}^*)$, and let $\Delta\tilde{x} = \tilde{x} - \tilde{x}^*$; then:

$$\widehat{\text{Var}}\left(\widehat{\Delta\mu}(\tilde{x}, \tilde{x}^*)\right) = \Delta\tilde{x}^\top \hat{\Sigma}\Delta\tilde{x}$$

Epi 204: Generalized linear models

Generalized linear models have three components:

1. The **outcome distribution** family: $\text{p}(Y|\mu(\tilde{x}))$
2. The **link function**: $g(\mu(\tilde{x})) = \eta(\tilde{x})$
3. The **linear component**: $\eta(\tilde{x}) = \tilde{x} \cdot \beta$

$$\theta_\omega(\tilde{x}, \tilde{x}^*) = \exp\left\{(\Delta\tilde{x}) \cdot \tilde{\beta}\right\}$$

Estimates of odds ratios from 2x2 contingency tables

$$\hat{\theta} = \frac{ad}{bc}$$

Summary of logistic regression definitions and results

Odds and log-odds

These association-measure results are covered in Measures of Association for Binary Outcomes[1].

Odds (def[2])

$$\omega \stackrel{\text{def}}{=} \frac{\Pr(A)}{\Pr(\neg A)}$$

Conditional odds (def[3])

$$\omega(A|B) \stackrel{\text{def}}{=} \frac{\Pr(A|B)}{\Pr(\neg A|B)}$$

Odds function (def[4])

$$\text{odds}\{\pi\} \stackrel{\text{def}}{=} \frac{\pi}{1 - \pi}$$

[1] binary-outcome-associations.qmd
[2] binary-outcome-associations.qmd#def-odds
[3] binary-outcome-associations.qmd#def-c-odds
[4] binary-outcome-associations.qmd#def-odds-fn

Probability to odds (thm[5], cor[6])

$$\omega = \frac{\pi}{1-\pi} = \text{odds}\{\pi\}$$

Simplified odds expressions; odds of a non-event (thm[7], cor[8])

$$\text{odds}\{\pi\} = \frac{1}{\pi^{-1}-1} = \left(\pi^{-1}-1\right)^{-1}, \qquad \omega(\neg A) = \frac{1-\pi}{\pi} = \pi^{-1}-1$$

Odds ratio (def[9])

$$\theta(\omega_1, \omega_2) \stackrel{\text{def}}{=} \frac{\omega_1}{\omega_2}$$

OR as ratio of probability ratios (thm[10])

$$\theta(\omega_1, \omega_2) = \frac{\omega_1}{\omega_2} = \frac{\left(\frac{\pi_1}{1-\pi_1}\right)}{\left(\frac{\pi_2}{1-\pi_2}\right)}$$

Odds ratios are reversible, marginally and conditionally (thm[11], thm[12])

$$\theta_\omega(A|B) = \theta_\omega(B|A), \qquad \theta_\omega(A|B,C) = \theta_\omega(B|A,C)$$

Inverse-odds and probability recovery
Odds to probability (thm[13])

$$\pi = \frac{\omega}{1+\omega}$$

Inverse-odds function (def[14]; recovers probability by cor[15])

$$\text{invodds}\{\omega\} \stackrel{\text{def}}{=} \frac{\omega}{1+\omega}, \qquad \pi = \text{invodds}\{\omega\}$$

Simplified inverse-odds (lem[16])

$$\text{invodds}\{\omega\} = \frac{1}{1+\omega^{-1}} = (1+\omega^{-1})^{-1}$$

One minus inverse-odds, and its complement (lem[17], cor[18])

$$1-\pi = \frac{1}{1+\omega}, \qquad 1+\omega = \frac{1}{1-\pi}$$

Log-odds (logit) and expit
Log-odds, and log-odds from probability (def[19], thm[20])

$$\eta \stackrel{\text{def}}{=} \log\{\omega\}, \qquad \eta = \log\left\{\frac{\pi}{1-\pi}\right\}$$

Logit function, and expanded form (def[21], thm[22])

$$\text{logit}(\pi) \stackrel{\text{def}}{=} \log\{\text{odds}\{\pi\}\} = \log\left\{\frac{\pi}{1-\pi}\right\}$$

[5] binary-outcome-associations.qmd#thm-prob-to-odds
[6] binary-outcome-associations.qmd#cor-oddsf-to-odds
[7] binary-outcome-associations.qmd#thm-odds-simplified
[8] binary-outcome-associations.qmd#cor-odds-neg
[9] binary-outcome-associations.qmd#def-OR
[10] binary-outcome-associations.qmd#thm-or-ratio-ratio
[11] binary-outcome-associations.qmd#thm-or-swap
[12] binary-outcome-associations.qmd#thm-conditional-OR-swap
[13] binary-outcome-associations.qmd#thm-odds-to-prob
[14] binary-outcome-associations.qmd#def-inv-odds
[15] binary-outcome-associations.qmd#cor-invodds-pi
[16] binary-outcome-associations.qmd#lem-invodds-simplified
[17] binary-outcome-associations.qmd#lem-one-minus-odds-inv
[18] binary-outcome-associations.qmd#cor-inverse-odds-nonevent
[19] binary-outcome-associations.qmd#def-logodds
[20] binary-outcome-associations.qmd#thm-logodds-pi
[21] binary-outcome-associations.qmd#def-logit-fn
[22] binary-outcome-associations.qmd#thm-logit-function

Log-odds equals logit (cor[23])

$$\eta = \text{logit}\{\pi\}$$

Odds and probability from log-odds (lem[24], thm[25])

$$\omega = \exp\{\eta\}, \qquad \pi = \frac{\exp\{\eta\}}{1+\exp\{\eta\}}$$

Expit / inverse-logit function, and equivalent expressions (def[26], thm[27])

$$\text{expit}(\eta) \stackrel{\text{def}}{=} \text{invodds}\{\exp\{\eta\}\} = \frac{\exp\{\eta\}}{1+\exp\{\eta\}} = (1+\exp\{-\eta\})^{-1}$$

Probability as expit (thm[28])

$$\pi = \text{expit}\{\eta\} \tag{1}$$

Logit and expit are inverses (thm[29])

$$\text{logit}\{\text{expit}\{\eta\}\} = \eta \qquad \text{expit}\{\text{logit}\{\pi\}\} = \pi$$

$$\left[\pi \stackrel{\text{def}}{=} \Pr\left(Y=1|\tilde{X}=\tilde{x}\right)\right] \overbrace{\underbrace{\underset{\frac{\omega}{1+\omega}}{\overset{\frac{\pi}{1-\pi}}{\rightleftarrows}} \left[\omega \stackrel{\text{def}}{=} \text{odds}\left\{Y=1|\tilde{X}=\tilde{x}\right\}\right] \underset{\exp\{\eta\}}{\overset{\log\{\omega\}}{\rightleftarrows}}}_{\text{expit}(\eta)}}^{\text{logit}(\pi)} \left[\eta(\tilde{x}) \stackrel{\text{def}}{=} \eta\left(Y=1|\tilde{X}=\tilde{x}\right)\right]$$

Figure 1: Diagram of logistic regression link and inverse link functions

Rare events

Odds minus probability (thm[30])

$$\omega - \pi = \frac{\pi^2}{1-\pi}, \quad \text{where } \omega = \frac{\pi}{1-\pi}$$

OR as RR times a correction factor (thm[31])

$$\theta(\omega_1, \omega_2) = \rho(\pi_1, \pi_2) \cdot \frac{1-\pi_2}{1-\pi_1}$$

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
$\tilde{x}$	$\underbrace{\tilde{\beta}\pi(1-\pi)}_{p\times 1}$	$\underbrace{\tilde{\beta}\omega}_{p\times 1}$	$\underbrace{\tilde{\beta}}_{p\times 1}$	$\underbrace{\mathbb{I}}_{p\times p}$	$\mathbf{0}_{p\times p}$
$\tilde{\beta}$	$\underbrace{\tilde{x}\pi(1-\pi)}_{p\times 1}$	$\underbrace{\tilde{x}\omega}_{p\times 1}$	$\underbrace{\tilde{x}}_{p\times 1}$	$\mathbf{0}_{p\times p}$	$\underbrace{\mathbb{I}}_{p\times p}$

[23] binary-outcome-associations.qmd#cor-logodds-logit
[24] binary-outcome-associations.qmd#lem-odds-from-logodds
[25] binary-outcome-associations.qmd#thm-prob-from-logodds
[26] binary-outcome-associations.qmd#def-expit
[27] binary-outcome-associations.qmd#thm-expit-expressions
[28] binary-outcome-associations.qmd#thm-expit-prob-logodds
[29] binary-outcome-associations.qmd#thm-logit-expit
[30] binary-outcome-associations.qmd#thm-odds-minus-probs
[31] binary-outcome-associations.qmd#thm-rare-disease-or-rr

Column labels indicate the numerators of the derivatives; row labels indicate the denominators (dimensions and details in tbl[32]).

Derivative of expit (cor[33])

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

Logistic regression model

Logistic regression (def[34])

$$Y_i|\tilde{X}_i \ \sim_{\perp\!\!\!\perp} \ \text{Ber}(\pi(\tilde{X}_i)), \qquad \text{logit}\{\pi(\tilde{x})\} = \tilde{x}^\top \tilde{\beta}$$

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 $\varepsilon_i \stackrel{\text{def}}{=} y_i - \pi_i$.

Log-likelihood and score function

Log-likelihood component (lem[35])

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

Score function as sum of components (lem[36], thm[37])

$$\tilde{\ell}'(\tilde{\beta}) = \sum_{i=1}^{n} \tilde{\ell}_i'(\tilde{\beta}), \qquad \ell_i'(\tilde{\beta}) = \tilde{x}_i \varepsilon_i$$

Score function (thm[38])

$$\tilde{\ell}'(\tilde{\beta}) = \sum_{i=1}^{n} \tilde{x}_i \varepsilon_i = \underbrace{\mathbf{X}^\top}_{(p+1)\times n} \underbrace{\tilde{\varepsilon}}_{n \times 1}$$

Maximum likelihood estimation

Estimating equation (eq[39]) The MLE $\hat{\tilde{\beta}}$ solves the score equation:

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

This estimating equation has no closed-form solution in general; $\hat{\tilde{\beta}}$ must be computed by numerical iteration.

Hessian (eq[40])

$$\underbrace{\ell''(\tilde{\beta})}_{(p+1)\times(p+1)} = -\underbrace{\mathbf{X}^\top}_{(p+1)\times n} \underbrace{\mathbf{D}}_{n \times n} \underbrace{\mathbf{X}}_{n \times (p+1)}, \qquad \mathbf{D} \stackrel{\text{def}}{=} \text{diag}(\text{Var}\left(Y_i|\tilde{X}_i = \tilde{x}_i\right)) = \text{diag}(\pi_i(1-\pi_i))$$

Concavity (rem[41]): 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[42])

$$\underbrace{\hat{\tilde{\beta}}^*}_{(p+1)\times 1} \leftarrow \underbrace{\hat{\tilde{\beta}}^*}_{(p+1)\times 1} - \underbrace{\left(\ell''\left(\tilde{y}; \hat{\tilde{\beta}}^*\right)\right)^{-1}}_{(p+1)\times(p+1)} \underbrace{\ell'\left(\tilde{y}; \hat{\tilde{\beta}}^*\right)}_{(p+1)\times 1}$$

One-sample MLE for odds (thm[43])

$$\hat{\omega} = \frac{x}{n - x}$$

[32] logistic-regression.qmd#tbl-logistic-deriv-matrix
[33] logistic-regression.qmd#cor-deriv-expit
[34] logistic-regression.qmd#def-logistic-regression
[35] logistic-regression.qmd#lem-logistic-loglik-component
[36] logistic-regression.qmd#lem-score-logistic
[37] logistic-regression.qmd#thm-logistic-score-comp
[38] logistic-regression.qmd#thm-logistic-score-fn
[39] logistic-regression.qmd#eq-score-logistic
[40] logistic-regression.qmd#eq-logistic-hess
[41] logistic-regression.qmd#rem-hessian-nsd
[42] intro-MLEs.qmd#sec-newton-raphson
[43] binary-outcome-associations.qmd#thm-est-odds

Wald tests and confidence intervals for coefficients
Wald test statistic for $H_0 : \beta_k = \beta_{k,0}$ (CLT for MLEs[44])

$$z_k = \frac{\hat{\beta}_k - \beta_{k,0}}{\widehat{\text{SE}}\left(\hat{\beta}_k\right)}, \qquad z_k \mathrel{\dot\sim} N(0,1) \text{ under } H_0$$

95% confidence intervals for β_k and e^{β_k} (MLE invariance)

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

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[45])

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

Difference in log-odds; OR in terms of $\Delta\eta$ (def[46], cor[47])

$$\Delta\eta \stackrel{\text{def}}{=} \eta(\tilde{x}) - \eta(\tilde{x}^*), \qquad \theta_\omega(\tilde{x}, \tilde{x}^*) = \exp\{\Delta\eta\}$$

$\Delta\eta$ from covariates (lem[48])

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

Difference in covariate patterns; $\Delta\eta$ from $\Delta\tilde{x}$ (def[49], cor[50])

$$\Delta\tilde{x} \stackrel{\text{def}}{=} \tilde{x} - \tilde{x}^*, \qquad \Delta\eta = \Delta\tilde{x} \cdot \tilde{\beta}$$

OR in terms of $\Delta\tilde{x}$ (thm[51])

$$\theta_\omega(\tilde{x}, \tilde{x}^*) = \exp\left\{(\Delta\tilde{x}) \cdot \tilde{\beta}\right\}$$

Log OR equals $\Delta\eta$ (cor[52])

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

Inference for log-odds and odds ratios
Estimated SE of log-odds (thm[53])

$$\widehat{\text{Var}}(\hat{\eta}(\tilde{x})) = \underbrace{\tilde{x}^\top}_{1\times(p+1)} \underbrace{\hat{\Sigma}}_{(p+1)\times(p+1)} \underbrace{\tilde{x}}_{(p+1)\times 1}, \qquad \widehat{\text{SE}}(\hat{\eta}(\tilde{x})) = \sqrt{\tilde{x}^\top \hat{\Sigma} \tilde{x}}$$

Estimated SE of $\Delta\hat{\eta}$ (thm[54])

$$\widehat{\text{Var}}(\Delta\hat{\eta}) = \underbrace{\Delta\tilde{x}^\top}_{1\times(p+1)} \underbrace{\hat{\Sigma}}_{(p+1)\times(p+1)} \underbrace{(\Delta\tilde{x})}_{(p+1)\times 1}, \qquad \widehat{\text{SE}}(\Delta\hat{\eta}) = \sqrt{\Delta\tilde{x}^\top \hat{\Sigma} (\Delta\tilde{x})}$$

CI for an odds ratio (eq[55]; details in sec[56]): exponentiate the endpoints of the CI for $\Delta\eta$:

$$\theta_\omega(\tilde{x}, \tilde{x}^*) \in (e^L, e^R), \qquad (L, R) = \widehat{\Delta\eta} \mp 1.96 \cdot \widehat{\text{SE}}\left(\widehat{\Delta\eta}\right)$$

[44] intro-MLEs.qmd#thm-dist-mle
[45] logistic-regression.qmd#lem-logistic-OR
[46] logistic-regression.qmd#def-difflogodds
[47] logistic-regression.qmd#cor-logistic-OR
[48] logistic-regression.qmd#lem-diff-logodds
[49] logistic-regression.qmd#def-diff-covs
[50] logistic-regression.qmd#cor-diff-logodds
[51] logistic-regression.qmd#thm-logistic-OR
[52] logistic-regression.qmd#cor-log-or
[53] logistic-regression.qmd#thm-est-se-logodds
[54] logistic-regression.qmd#thm-est-se-diff-logodds
[55] logistic-regression.qmd#eq-ci-OR-exp
[56] logistic-regression.qmd#sec-or-inference

Odds ratios in study designs

Disease and exposure odds ratios (def[57], def[58])

$$\theta_\omega(D|E) \overset{\text{def}}{=} \frac{\omega(D|E)}{\omega(D|\neg E)}, \qquad \theta_\omega(E|D) \overset{\text{def}}{=} \frac{\omega(E|D)}{\omega(E|\neg D)}$$

Case-control studies (section[59]): outcome-stratified sampling breaks estimation of risks, risk ratios, and the components of the disease odds ratio (def[60]), but the exposure odds ratio (def[61]) is estimable, and it equals the disease odds ratio by reversibility (thm[62]).

2×2 table shortcut (table[63]; cell counts a, b = exposed events and non-events, c, d = unexposed events and non-events):

$$\hat{\omega}(Event|Exposed) = \frac{a}{b}, \qquad \hat{\omega}(Event|\neg Exposed) = \frac{c}{d}, \qquad \theta_\omega(Exposed, \neg Exposed) = \frac{ad}{bc}$$

Model fit and comparison

Details in sec[64]; 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[65])

$$D \overset{\text{def}}{=} 2(\ell_{\text{full}} - \ell(\hat{\beta}; \mathbf{x})), \qquad D \mathrel{\dot\sim} \chi^2(q-p)$$

Hosmer-Lemeshow test (def[66]; 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[67])

$$X^2 = \sum_{k=1}^{q} X_k^2 \mathrel{\dot\sim} \chi^2(q-p)$$

AIC and BIC [lower is better] (AIC[68] (Akaike 1974); BIC[69] (Schwarz 1978))

$$\text{AIC} = -2\ell(\hat{\theta}) + 2p, \qquad \text{BIC} = -2\ell(\hat{\theta}) + p\log\{n\}$$

Likelihood ratio test (def[70]; 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[71].

Residual-based diagnostics

Residuals are computed per covariate pattern k (grouped data); h_k denotes the leverage of pattern k (def[72]).

Response residual (def[73])

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

Pearson residual (and standardized version) (def[74], def[75])

$$X_k \overset{\text{def}}{=} \frac{\bar{y}_k - \hat{\pi}_k}{\sqrt{\hat{\pi}_k(1-\hat{\pi}_k)/n_k}}, \qquad r_{P_k} \overset{\text{def}}{=} \frac{X_k}{\sqrt{1-h_k}}$$

[57] binary-outcome-associations.qmd#def-disease-OR
[58] binary-outcome-associations.qmd#def-exposure-OR
[59] binary-outcome-associations.qmd#sec-CC-studies
[60] binary-outcome-associations.qmd#def-disease-OR
[61] binary-outcome-associations.qmd#def-exposure-OR
[62] binary-outcome-associations.qmd#thm-or-swap
[63] binary-outcome-associations.qmd#tbl-2x2-generic
[64] logistic-regression.qmd#sec-gof
[65] logistic-regression.qmd#def-deviance-test
[66] logistic-regression.qmd#def-hosmer-lemeshow
[67] logistic-regression.qmd#def-pearson-residual
[68] Linear-models-overview.qmd#def-aic
[69] Linear-models-overview.qmd#def-bic
[70] logistic-regression.qmd#def-lrt
[71] intro-MLEs.qmd#thm-wilks
[72] logistic-regression.qmd#def-leverage-glm
[73] logistic-regression.qmd#def-response-residual-logistic
[74] logistic-regression.qmd#def-pearson-residual
[75] logistic-regression.qmd#def-standardized-pearson-residual

Deviance residual (and standardized version) (def[76])

$$d_k \overset{\text{def}}{=} \operatorname{sign}(y_k - n_k\hat{\pi}_k)\left\{\sqrt{2[\ell_{\text{full}}(x_k) - \ell(\hat{\beta}; x_k)]}\right\}, \qquad r_{D_k} \overset{\text{def}}{=} \frac{d_k}{\sqrt{1-h_k}}$$

Cook's distance (def[77])

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

Leverage (def[78]): 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):

$$\underbrace{\mathbf{H}}_{q\times q} \overset{\text{def}}{=} \underbrace{\mathbf{W}^{1/2}}_{q\times q} \underbrace{\mathbf{X}}_{q\times p} \Bigg(\underbrace{\mathbf{X}^\top \mathbf{W} \mathbf{X}}_{p\times p}\Bigg)^{-1} \underbrace{\mathbf{X}^\top}_{p\times q} \underbrace{\mathbf{W}^{1/2}}_{q\times q}, \qquad \underbrace{\mathbf{W}}_{q\times q} \overset{\text{def}}{=} \operatorname{diag}(n_k\, \hat{\pi}_k\, (1-\hat{\pi}_k))$$

Comparing probabilities
Risk difference (def[79])

$$\delta(\pi_1, \pi_2) \overset{\text{def}}{=} \pi_1 - \pi_2$$

Risk ratio (def[80])

$$\rho(\pi_1, \pi_2) \overset{\text{def}}{=} \frac{\pi_1}{\pi_2}$$

Relative risk difference (def[81]; equals RR minus 1 by thm[82])

$$\xi(\pi_1, \pi_2) \overset{\text{def}}{=} \frac{\delta(\pi_1, \pi_2)}{\pi_2} = \rho(\pi_1, \pi_2) - 1$$

Marginal risks from logistic regression
Conditional predicted risk (def[83])

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

Conditional RD and RR (def[84], def[85])

$$\operatorname{RD}(z) \overset{\text{def}}{=} \pi(1, z) - \pi(0, z), \qquad \operatorname{RR}(z) \overset{\text{def}}{=} \frac{\pi(1, z)}{\pi(0, z)}$$

Observed marginal risk (def[86]; standardization form by thm[87])

$$\pi(a) \overset{\text{def}}{=} \Pr(Y = 1 \mid A = a) = \operatorname{E}[\pi(a, Z) \mid A = a]$$

Potential mean and causal marginal risk (def[88], def[89]; identification form by thm[90])

$$\mu_a \overset{\text{def}}{=} \operatorname{E}[Y^a], \qquad \pi_a \overset{\text{def}}{=} \Pr(Y^a = 1) = \operatorname{E}[\pi(a, Z)]$$

No confounding: observed = causal (cor[91]; the two estimands are contrasted in rem[92])

$$Z \perp\!\!\!\perp A \implies \pi(a) = \pi_a$$

[76] logistic-regression.qmd#def-deviance-residual-logistic
[77] logistic-regression.qmd#def-cooks-distance-glm
[78] logistic-regression.qmd#def-leverage-glm
[79] binary-outcome-associations.qmd#def-RD
[80] binary-outcome-associations.qmd#def-RR
[81] binary-outcome-associations.qmd#def-RRD
[82] binary-outcome-associations.qmd#thm-rrd-vs-rr
[83] logistic-regression.qmd#def-cond-predicted-risk
[84] logistic-regression.qmd#def-cond-rd
[85] logistic-regression.qmd#def-cond-rrr
[86] logistic-regression.qmd#def-observed-marginal-risk
[87] logistic-regression.qmd#thm-observed-marginal-risk
[88] logistic-regression.qmd#def-potential-mean
[89] logistic-regression.qmd#def-causal-marginal-risk
[90] logistic-regression.qmd#thm-causal-marginal-risk
[91] logistic-regression.qmd#cor-observed-equals-causal
[92] logistic-regression.qmd#rem-observed-vs-causal-marginal-risk

G-computation estimator (def[93]; consistent for π_a by thm[94])

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

Predictive margins (def[95]; full-sample form def[96] is consistent for π_a by cor[97])

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

Exposed-subgroup predictive margin (def[98]; consistent for $\pi(a)$ by thm[99]; equals $\bar{Y}_a$ exactly by lem[100])

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

Observed marginal RD and RR (def[101], def[102])

$$\text{OMRD} \stackrel{\text{def}}{=} \pi(1) - \pi(0), \qquad \text{OMRR} \stackrel{\text{def}}{=} \frac{\pi(1)}{\pi(0)}$$

Causal marginal RD and RR (def[103], def[104])

$$\text{CMRD} \stackrel{\text{def}}{=} \pi_1 - \pi_0, \qquad \text{CMRR} \stackrel{\text{def}}{=} \frac{\pi_1}{\pi_0}$$

Bootstrap SE and CI for marginal contrasts (def[105]; see Bootstrap Confidence Intervals[106] and exm[107]): 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[108]): 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[109]; counterexample in exm[110]).

Survival analysis

Probability distribution functions

Table 1: Probability distribution functions

Name	Symbols	Definition
Probability density function (PDF)	$\text{f}(t), \text{p}(t)$	$\text{p}(T = t)$
Cumulative distribution function (CDF)	$\text{F}(t), \text{P}(t)$	$\text{P}(T \le t)$
Survival function	$\text{S}(t), \bar{\text{F}}(t)$	$\text{P}(T > t)$
Hazard function	$\lambda(t), \text{h}(t)$	$\text{p}(T = t \mid T \ge t)$
Cumulative hazard function	$\Lambda(t), \text{H}(t)$	$\int_{u=-\infty}^{t} \lambda(u)du$
Log-hazard function	$\eta(t)$	$\log\{\lambda(t)\}$

[93] logistic-regression.qmd#def-g-computation
[94] logistic-regression.qmd#thm-g-computation-consistency
[95] logistic-regression.qmd#def-predictive-margins
[96] logistic-regression.qmd#def-full-sample-predictive-margin
[97] logistic-regression.qmd#cor-full-sample-pm
[98] logistic-regression.qmd#def-subgroup-predictive-margin
[99] logistic-regression.qmd#thm-subgroup-pm
[100] logistic-regression.qmd#lem-subgroup-pm-proportion
[101] logistic-regression.qmd#def-observed-marginal-rd
[102] logistic-regression.qmd#def-observed-marginal-rr
[103] logistic-regression.qmd#def-causal-marginal-rd
[104] logistic-regression.qmd#def-causal-marginal-rr
[105] logistic-regression.qmd#def-bootstrap-marginal-rd
[106] basic-statistical-methods.qmd#sec-bootstrap-ci
[107] logistic-regression.qmd#exm-wcgs-marginal-rd
[108] logistic-regression.qmd#def-collapsibility
[109] logistic-regression.qmd#thm-collapsibility
[110] logistic-regression.qmd#exm-collapsibility

Diagram of survival distribution function relationships

$$\mathrm{f}(t) \xleftarrow[\mathrm{S}(t)\lambda(t)]{-S'(t)} \mathrm{S}(t) \xleftarrow{\exp\{-\Lambda(t)\}} \Lambda(t) \xleftarrow{\int_{u=0}^{t}\lambda(u)du} \lambda(t) \xleftarrow{\exp\{\eta(t)\}} \eta(t)$$

$$\mathrm{f}(t) \xrightarrow[\int_{u=t}^{\infty}\mathrm{f}(u)du]{\mathrm{f}(t)/\lambda(t)} \mathrm{S}(t) \xrightarrow[-\log \mathrm{S}(t)]{} \Lambda(t) \xrightarrow[\Lambda'(t)]{} \lambda(t) \xrightarrow[\log\{\lambda(t)\}]{} \eta(t)$$

Survival likelihood contributions, assuming non-informative censoring

$$\begin{aligned}\mathrm{p}(Y=y,D=d) &= [\mathrm{f}_T(y)]^d[\mathrm{S}_T(y)]^{1-d}\\ &= [\lambda_T(y)]^d[\mathrm{S}_T(y)]\end{aligned}$$

Nonparametric time-to-event distribution estimators

$$\hat{\lambda}_i = \frac{d_i}{r_i}$$

$$\hat{\kappa}_i = 1 - \hat{\lambda}_i = \frac{r_i - d_i}{r_i}$$

$$\hat{\mathrm{S}}_{\mathrm{KM}}(t) \stackrel{\mathrm{def}}{=} \prod_{\{i:\ t_i \le t\}} \hat{\kappa}_i$$

$$\hat{\Lambda}_{NA}(t) \stackrel{\mathrm{def}}{=} \sum_{\{i:\ t_i \le t\}} \hat{\lambda}_i$$

Proportional hazards model structure

Formula	Description
$\lambda(t \mid \tilde{x}) = \lambda_0(t) \cdot \theta_\lambda(\tilde{x})$	Proportional hazards assumption
$\Lambda(t \mid \tilde{x}) = \Lambda_0(t) \cdot \theta_\lambda(\tilde{x})$	Cumulative hazard factorization
$\eta(t \mid \tilde{x}) = \eta_0(t) + \Delta\eta(\tilde{x})$	Log-hazard decomposition
$\Delta\eta(\tilde{x}) = \tilde{x} \cdot \tilde{\beta} = \beta_1 x_1 + \cdots + \beta_p x_p$	Linear predictor
$\theta_\lambda(\tilde{x}) = \exp\{\Delta\eta(\tilde{x})\}$	Hazard multiplier
$\theta_\lambda(t\|\tilde{x} : \tilde{x}^*) \stackrel{\mathrm{def}}{=} \frac{\lambda(t\|\tilde{x})}{\lambda(t\|\tilde{x}^*)}$	Hazard ratio definition
$\theta_\lambda(t \mid \tilde{x} : \tilde{x}^*) = \exp\{\Delta\eta(\tilde{x}) - \Delta\eta(\tilde{x}^*)\} = \exp\left\{(\tilde{x} - \tilde{x}^*) \cdot \tilde{\beta}\right\}$	Hazard ratio formula

Proportional hazards model partial likelihood formula:

Definition 0.1 (Cox PH partial likelihood)**.**

$$\mathcal{L}_i^*(\tilde{\beta}) \stackrel{\mathrm{def}}{=} \frac{\theta_\lambda(\tilde{x}_{(i)})}{\sum_{k \in R(t_i)} \theta_\lambda(\tilde{x}_k)}$$
$$\mathcal{L}^*(\tilde{\beta}) \stackrel{\mathrm{def}}{=} \prod_{\{i:\ d_i = 1\}} \mathcal{L}_i^*(\tilde{\beta})$$

where, for the distinct event times $t_1 < \cdots < t_K$ of the data set:

- (i) is the subject who fails at t_i (unique under the no-ties assumption);
- $R(t_i) \stackrel{\mathrm{def}}{=} \left\{j : Y_j \ge t_i\right\}$ is the risk set at t_i (where $Y_j = \min(T_j, C_j)$ is subject j's observed follow-up time) — the subjects still under observation just before t_i;
- d_i is the number of events at t_i (so $d_i = 1$ when there are no ties);
- $\theta_\lambda(\tilde{x}) \stackrel{\mathrm{def}}{=} \exp\left\{\tilde{x}^\top \tilde{\beta}\right\}$ is the hazard ratio (risk score) for covariate pattern $\tilde{x}$.

Proportional hazards model baseline cumulative hazard estimator:

Definition 0.2 (Breslow estimator of the baseline cumulative hazard)**.**

$$\hat{\Lambda}_0(t) \stackrel{\mathrm{def}}{=} \sum_{t_i \le t} \frac{d_i}{\sum_{k \in R(t_i)} \theta_\lambda(\tilde{x}_k)}$$

where, for the distinct event times $t_1 < \cdots < t_K$ of the data set:

- d_i is the number of events at t_i;
- $R(t_i) \stackrel{\text{def}}{=} \{j : Y_j \geq t_i\}$ is the risk set at t_i (where $Y_j = \min(T_j, C_j)$ is subject j's observed follow-up time) — the subjects still under observation just before t_i;
- $\theta_\lambda(\tilde{x}) \stackrel{\text{def}}{=} \exp\left\{\tilde{x}^\top \tilde{\beta}\right\}$ is the hazard ratio (risk score) for covariate pattern $\tilde{x}$.

Exm

Example 0.1 (Breslow estimator with two event times)**.** Suppose there are two distinct event times $t_1 = 2$ and $t_2 = 4$ (no ties, $d_1 = d_2 = 1$), with risk-set-weighted totals $\sum_{k \in R(2)} \theta_\lambda(\tilde{x}_k) = 10$ and $\sum_{k \in R(4)} \theta_\lambda(\tilde{x}_k) = 6$. Then

$$\hat{\Lambda}_0(4) = \frac{1}{10} + \frac{1}{6} \approx 0.267.$$

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.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics* 6 (2): 461–64. https://doi.org/10.1214/aos/1176344136.