Chapter 4: Introduction to Statistical Models

In this chapter, we will explore how the methods of statistical inference that we developed for the setting of n IID observations $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$ may be applied to other types of data and statistical models. We will introduce five statistical models using different motivating examples, and then use our tools from previous chapters to solve several inference questions in this example[1]:

- Bradley Terry Model
- Linear Regression Model
- Logistic Regression Model
- Poisson Regression
- Cox Proportional Hazards Model

A parametric model for a data vector $\mathbf{Y}$ (not necessarily consisting of IID coordinates) is a specification of the joint distribution of $\mathbf{Y}$ in terms of a small number of parameters θ. The likelihood $\operatorname{lik}(\theta) = f(\mathbf{Y} \mid \theta)$ is the joint PMF or PDF of $\mathbf{Y}$ viewed as a function of θ. The log-likelihood is $l(\theta) = \log \operatorname{lik}(\theta)$, and the MLE $\hat{\theta}$ is the value of θ that maximizes $\operatorname{lik}(\theta)$. To extend the theory of maximum likelihood and Fisher information to the non-IID setting, note that for IID data $X_1, \ldots, X_n \overset{IID}{\sim} f(x \mid \theta)$, we may introduce the notation

$$I_{\mathbf{X}}(\theta) := nI(\theta) = \sum_{i=1}^{n} -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f\left(X_i \mid \theta\right) \right] = -\mathbb{E}_\theta \left[l''(\theta) \right],$$

which represents the total Fisher information of all n observations $\mathbf{X} = (X_1, \ldots, X_n)$. Our main theorem regarding the MLE $\hat{\theta}$ states that it is approximately distributed as $\mathcal{N}\left(\theta_0, \frac{1}{n} I\left(\theta_0\right)^{-1}\right) = \mathcal{N}\left(\theta_0, I_{\mathbf{X}}\left(\theta_0\right)^{-1}\right)$ for large n if the parametric model is correct and the true parameter is θ_0. For non-IID data and the general log-likelihood $l(\theta) = \log f(\mathbf{Y} \mid \theta)$, let us define

$$I_{\mathbf{Y}}(\theta) = -\mathbb{E}_\theta \left[l''(\theta) \right]$$

in the single-parameter case $\theta \in \mathbb{R}$ and

[1] Agresti, A. (2003). Categorical data analysis. John Wiley & Sons.

$$I_{\mathbf{Y}}(\theta) = -\mathbb{E}_\theta\left[\nabla^2 l(\theta)\right]$$

in the multi-parameter case $\theta \in \mathbb{R}^k$, where

$$\nabla^2 l(\theta) = \left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} l(\theta)\right)_{1\leq i,j\leq k}$$

is the second-derivative (Hessian) matrix for $l(\theta)$. In all of the non-IID settings we will consider, under appropriate asymptotic conditions, the approximate sampling distribution of $\hat{\theta}$ is still given by the (multivariate) normal distribution $\mathcal{N}\left(\theta_0, I_{\mathbf{Y}}(\theta_0)^{-1}\right)$ when the total sample size is large. We will appeal to this approximation without proof in our examples.

1 The Bradley-Terry model

Example 1.0.1. *There are 30 basketball teams in the NBA, each playing 82 games in the regular season (so there are 1230 total games). We observe, at the end of the regular season, which two teams (i, j) played in each game, and whether team i or team j won. How can we rank the teams and/or determine the strength of each team?*

The simplest strategy might be to compare the number of games won by each team. However, the NBA season is structured so that every team plays every other team a different number of times (between 2 and 4). So the teams have different "strengths of schedule", meaning that some teams play stronger opponents more frequently than do other teams. These teams might have worse win-loss records, but in fact, be better than other teams that won more games against weaker opponents.

A model-based approach to address this problem is the following: Let $\beta_i \in \mathbb{R}$ represent the "strength" of team i, and let the outcome of a game between teams (i, j) be determined by $\beta_i - \beta_j$. The ***Bradley-Terry model*** *treats this outcome as an independent Bernoulli random variable with distribution* Bernoulli (p_{ij}), *where the log-odds corresponding to the probability p_{ij} that team i beats team j is modeled as*

$$\log \frac{p_{ij}}{1 - p_{ij}} = \beta_i - \beta_j.$$

Equivalently, solving for p_{ij} yields

$$p_{ij} = \frac{e^{\beta_i - \beta_j}}{1 + e^{\beta_i - \beta_j}} = \frac{e^{\beta_i}}{e^{\beta_i} + e^{\beta_j}}.$$

This model is over-parametrized in the sense that it is exactly the same if we add a fixed constant c to all values β_i, because the differences $\beta_i - \beta_j$ remain unchanged. We *may fix this problem by setting $\beta_i \equiv 0$ for a particular team, for example, $\beta_{Warriors} \equiv 0$. Then for every other team $j, \beta_j = \beta_j - 0$ represents the log-odds that team j beats the Warriors.*

If we always order each pair (i, j) so that team i is the home team and j is the away team, then we may incorporate a home-court advantage by including an intercept term α:

$$\log \frac{p_{ij}}{1-p_{ij}} = \alpha + \beta_i - \beta_j,$$

or equivalently

$$p_{ij} = \frac{e^{\alpha+\beta_i-\beta_j}}{1+e^{\alpha+\beta_i-\beta_j}}. \tag{1}$$

This increases the log-odds of the home team winning in every game by a constant value α.

More generally, the Bradley-Terry model assigns scores to a fixed set of items based on pairwise comparisons of these items, where the log-odds of item i "beating" item j is given by the difference of their scores. An intercept term may be included to account for a systematic difference between the first and second item of each comparison.

1.1 Statistical inference

Let $k = 30$ be the number of NBA teams, and denote the Warriors as team 1. We might be interested in the following inferential tasks:

- Estimate the home-court advantage α and the team strengths $\beta_1, \ldots, \beta_k$ (constraining, say, $\beta_1 = \beta_{\text{Warriors}} \equiv 0$)
- Test the null hypothesis of no home-court effect, $\alpha = 0$
- Obtain a confidence interval for $\beta_i - \beta_j$ for two particular teams (i, j)

Suppose we observe n total games $(i_1, j_1), \ldots, (i_n, j_n)$ between these k teams, where each (i, j) is a pair of distinct teams in $\{1, \ldots, k\}$ and the home team is team i. Let $Y_1, \ldots, Y_n \in \{0, 1\}$ be such that $Y_m = 1$ if i_m beat j_m in the m^{th} game and $Y_m = 0$ otherwise. The likelihood for the parameters $\theta = (\alpha, \beta_2, \ldots, \beta_k)$ is then given by

$$\text{lik}(\alpha, \beta_2, \ldots, \beta_k) = \prod_{m=1}^{n} p_{i_m j_m}^{Y_m} (1 - p_{i_m j_m})^{1-Y_m} = \prod_{m=1}^{n} (1 - p_{i_m j_m}) \left(\frac{p_{i_m j_m}}{1 - p_{i_m j_m}} \right)^{Y_m},$$

where p_{ij} is given as a function of α, β_i, and β_j by equation (1) and we set $\beta_1 \equiv 0$. The log-likelihood is

$$\begin{aligned} l\left(\alpha, \beta_2, \ldots, \beta_k\right) &= \sum_{m=1}^{n} Y_m \log\left(\frac{p_{i_m j_m}}{1-p_{i_m j_m}}\right) + \log\left(1-p_{i_m j_m}\right) \\ &= \sum_{m=1}^{n} Y_m\left(\alpha + \beta_{i_m} - \beta_{j_m}\right) - \log\left(1+e^{\alpha+\beta_{i_m}-\beta_{j_m}}\right). \end{aligned} \tag{2}$$

To estimate the parameters $\theta = (\alpha, \beta_2, \ldots, \beta_k)$ using the MLE, we set the partial derivative with respect to each parameter $\alpha, \beta_2, \ldots, \beta_k$ equal to 0 :

$$0 = \frac{\partial l}{\partial \alpha} = \sum_{m=1}^{n} Y_m - \frac{e^{\alpha+\beta_{i_m}-\beta_{j_m}}}{1+e^{\alpha+\beta_{i_m}-\beta_{j_m}}} \tag{3}$$

$$0 = \frac{\partial l}{\partial \beta_i} = \sum_{m:i_m=i}\left(Y_m - \frac{e^{\alpha+\beta_{i_m}-\beta_{j_m}}}{1+e^{\alpha+\beta_{i_m}-\beta_{j_m}}}\right) + \sum_{m:j_m=i}\left(-Y_m + \frac{e^{\alpha+\beta_{i_m}-\beta_{j_m}}}{1+e^{\alpha+\beta_{i_m}-\beta_{j_m}}}\right). \tag{4}$$

This yields a system of k equations in the k unknowns $\alpha, \beta_2, \ldots, \beta_k$, which may be solved numerically using the Newton-Raphson algorithm. The solution is the MLE $\hat{\theta} = \left(\hat{\alpha}^2, \hat{\beta}_2, \ldots, \hat{\beta}_k\right)$.

To test the null hypothesis $H_0 : \alpha = 0$, we may use the generalized likelihood ratio test (GLRT): Under the sub-model where $\alpha = 0$, the log-likelihood function is

$$l\left(\beta_2, \ldots, \beta_k\right) = \sum_{m=1}^{n} Y_m\left(\beta_{i_m} - \beta_{j_m}\right) - \log\left(1+e^{\beta_{i_m}-\beta_{j_m}}\right),$$

and the system of score equations satisfied by the sub-model MLE is

$$0 = \frac{\partial l}{\partial \beta_i} = \sum_{m:i_m=i}\left(Y_m - \frac{e^{\beta_{i_m}-\beta_{j_m}}}{1+e^{\beta_{i_m}-\beta_{j_m}}}\right) + \sum_{m:j_m=i}\left(-Y_m + \frac{e^{\beta_{i_m}-\beta_{j_m}}}{1+e^{\beta_{i_m}-\beta_{j_m}}}\right)$$

for $i = 2, \ldots, k$. We may solve these equations using Newton-Raphson to obtain the submodel MLEs $\hat{\beta}_{2,0}, \ldots, \hat{\beta}_{k,0}$. The GLRT of $\alpha = 0$ is based on the test statistic

$$-2\log\Lambda = -2\log\frac{\operatorname{lik}\left(0, \hat{\beta}_{2,0}, \ldots, \hat{\beta}_{k,0}\right)}{\operatorname{lik}\left(\hat{\alpha}, \hat{\beta}_2, \ldots, \hat{\beta}_k\right)},$$

and an approximate level-0.05 test rejects H_0 when $-2\log\Lambda > \chi_1^2(0.05)$. (The number of degrees of freedom is 1 because the full model has one more parameter, α, than the submodel.)

We may obtain a confidence interval for $\beta_i - \beta_j$ by centering it around $\hat{\beta}_i - \hat{\beta}_j$, and estimating the standard error of $\hat{\beta}_i - \hat{\beta}_j$. Let us first consider the sampling distribution of the entire vector of MLE estimates $\hat{\theta} = \left(\hat{\alpha}, \hat{\beta}_2, \ldots, \hat{\beta}_k\right)$. When the number of total games n is large, this is approximately $\mathcal{N}\left(\theta, I_{\mathbf{Y}}(\theta)^{-1}\right)$, where $I_{\mathbf{Y}}(\theta) = -\mathbb{E}_\theta\left[\nabla^2 l(\theta)\right]$. The Hessian matrix $\nabla^2 l(\theta)$ may be computed by differentiating the right sides of the score equations (3) and (4) a second time with respect to the variables $\alpha, \beta_2, \ldots, \beta_k$. (We will do this explicitly for the more general logistic regression model.) It is easy to see that $\nabla^2 l(\theta)$ is a constant quantity that does not involve $Y_1, \ldots, Y_n$, so $I_{\mathbf{Y}}(\theta) = -\nabla^2 l(\theta)$.

Finally, since $\hat{\beta}_i - \hat{\beta}_j$ is a linear combination of the coordinates of $\hat{\theta}$, it is approximately normal when $\hat{\theta}$ is approximately multivariate normal. Its mean is $\mathbb{E}\left[\hat{\beta}_i - \hat{\beta}_j\right] \approx \beta_i - \beta_j$, and its variance is

$$\begin{aligned}\operatorname{Var}\left[\hat{\beta}_i - \hat{\beta}_j\right] &= \operatorname{Cov}\left[\hat{\beta}_i - \hat{\beta}_j, \hat{\beta}_i - \hat{\beta}_j\right] \\ &= \operatorname{Var}\left[\hat{\beta}_i\right] + \operatorname{Var}\left[\hat{\beta}_j\right] - 2\operatorname{Cov}\left[\hat{\beta}_i, \hat{\beta}_j\right] \\ &\approx \left(I_{\mathbf{Y}}^{-1}(\theta)\right)_{ii} + \left(I_{\mathbf{Y}}^{-1}(\theta)\right)_{jj} - 2\left(I_{\mathbf{Y}}^{-1}(\theta)\right)_{ij}.\end{aligned}$$

We may estimate the standard error of $\hat{\beta}_i - \hat{\beta}_j$ by the plug-in estimate

$$\hat{\mathbf{se}}_{ij} = \sqrt{\left(I_{\mathbf{Y}}^{-1}(\hat{\theta})\right)_{ii} + \left(I_{\mathbf{Y}}^{-1}(\hat{\theta})\right)_{jj} - 2\left(I_{\mathbf{Y}}^{-1}(\hat{\theta})\right)_{ij}}.$$

A 95% confidence interval for $\beta_i - \beta_j$, assuming correctness of the Bradley-Terry model, is then given by $\hat{\beta}_i - \hat{\beta}_j \pm z(0.025)\hat{\text{se}}_{ij}$.

2 The Linear Model

Example 2.0.1. *When a string instrument sustains a note at a particular pitch, the resulting sound wave is periodic with some fixed frequency f (say 440 Hz). For a "pure" tone at this pitch, the sound wave is a perfect sinusoidal wave with frequency f, but the sound produced by any real string instrument is not a pure tone. Instead, it is a superposition of sinusoidal waves with frequencies $f, 2f, 3f, 4f$, etc., corresponding to different vibrating modes of the string. These frequencies are called resonance harmonics, and the relative volumes, or amplitudes, of the resonance harmonics, determine the timbre or "color" of the sound.*

A recording device measures the sound wave produced by an instrument (sustaining a single note) at n points in time $t_1, \ldots, t_n$, where the measurements are contaminated by white noise. We consider the problem of estimating the amplitudes of the resonance harmonics for this instrument. Let $Y_1, \ldots, Y_n \in \mathbb{R}$ be measurements of the sound wave at these time points. Suppose, for simplicity, we have scaled our units so that the sine and cosine curves corresponding to the fundamental frequency f are $\sin(t)$ and $\cos(t)$. Then, assuming the existence of resonance harmonics up to frequency kf, we may model each measurement Y_i as

$$Y_i = \beta_1 \sin(t_i) + \beta_2 \cos(t_i) + \beta_3 \sin(2t_i) + \beta_4 \sin(2t_i) + \ldots + \beta_{2k-1} \sin(kt_i) + \beta_{2k} \cos(kt_i) + \varepsilon_i, \tag{5}$$

for some coefficients $\beta_1, \ldots, \beta_{2k} \in \mathbb{R}$, where the errors $\varepsilon_i \overset{IID}{\sim} \mathcal{N}\left(0, \sigma_0^2\right)$ correspond to the white noise and the variance σ_0^2 signifies the noise level. If we construct the matrix

$$X = \begin{pmatrix} \sin(t_1) & \cos(t_1) & \cdots & \sin(kt_1) & \cos(kt_1) \\ \sin(t_2) & \cos(t_2) & \cdots & \sin(kt_2) & \cos(kt_2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sin(t_n) & \cos(t_n) & \cdots & \sin(kt_n) & \cos(kt_n) \end{pmatrix}$$

and denote its entries as x_{ij}, then we may write the above as

$$Y_i = \sum_{j=1}^{2k} \beta_j x_{ij} + \varepsilon_i,$$

or more succinctly in matrix notation, for all $i = 1, \ldots, n$, as

$$Y = X\beta + \varepsilon.$$

*Here, Y denotes the column vector $(Y_1, \ldots, Y_n)$, β denotes the column vector $(\beta_1, \ldots, \beta_{2k})$, and ε denotes the column vector $(\varepsilon_1, \ldots, \varepsilon_n)$. This is called a **linear model**.*

More generally, given a vector of **responses** $Y_1, \ldots, Y_n$, the linear model models each Y_i as a certain linear combination $\beta_1 x_{i1} + \ldots + \beta_p x_{ip}$ of corresponding **covariates** $x_{i1}, \ldots, x_{ip}$, plus IID Gaussian errors. (The coefficients $\beta_1, \ldots, \beta_p$ are the same for all n responses $Y_1, \ldots, Y_n$.) We will treat the covariates as fixed and known constants. The values of the Gaussian errors are not directly observed; for simplicity, however, we'll assume in this section that their variance σ_0^2 is known. The parameters of the model are the regression coefficients $\beta_1, \ldots, \beta_p$. (Much of our analysis in this section may be extended to the more realistic setting where σ_0^2 is unknown, in which case it would also be a parameter of the model.)

2.1 Statistical inference

In the model of equation (5), the amplitudes of the k resonance harmonics are defined as $A_1 = \sqrt{\beta_1^2 + \beta_2^2}, A_2 = \sqrt{\beta_3^2 + \beta_4^2}, \ldots, A_k = \sqrt{\beta_{2k-1}^2 + \beta_{2k}^2}$. We will discuss the following inferential tasks:

- Estimate the amplitudes $A_1, \ldots, A_k$
- Provide confidence intervals corresponding to these estimates

Let $p = 2k$. To write down the likelihood for the linear model, note that $Y_1, \ldots, Y_n$ are independent and distributed as $Y_i \sim \mathcal{N}\left(\sum_j \beta_j x_{ij}, \sigma_0^2\right)$. Then

$$\operatorname{lik}(\beta_1, \ldots, \beta_p) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2\sigma_0^2}\left(Y_i - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2\right),$$

and the log-likelihood is

$$l(\beta_1, \ldots, \beta_p) = -\frac{n}{2}\log\left(2\pi\sigma_0^2\right) - \frac{1}{2\sigma_0^2}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2.$$

For any $\sigma_0^2 > 0$, this log-likelihood is maximized when $\beta_1, \ldots, \beta_k$ are the **least-squares estimators** minimizing the total squared error

$$\text{err} = \sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2.$$

So the MLEs $\hat{\beta}_1, \ldots, \hat{\beta}_k$ are equal to the least-squares estimators. To compute these MLEs, we solve the system of p equations for $m = 1, \ldots, p$

$$0 = \frac{\partial l}{\partial \beta_m} = \frac{1}{\sigma_0^2}\sum_{i=1}^{n} x_{im}\left(Y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right).$$

Letting X_m denote the m th column of X, these equations may be written as

$$0 = \frac{1}{\sigma_0^2} X_m^T (Y - X\beta),$$

or even more succinctly for all $m = 1, \ldots, p$ as $0 = \frac{1}{\sigma_0^2} X^T (Y - X\beta)$, where both sides of this equation are vectors of length p. Solving for β yields the MLEs/least-squares estimates

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y.$$

To estimate an amplitude of a resonance harmonic, say $A_1 = \sqrt{\beta_1^2 + \beta_2^2}$, we may use the plugin estimate $\hat{A}_1 = \sqrt{\hat{\beta}_1^2 + \hat{\beta}_2^2}$.

To obtain a confidence interval for A_1, we will derive an approximate standard error for $\hat{A}_1$, by first deriving the sampling distribution of $\hat{\beta}$ and then applying the delta method. We compute the Fisher information $I_{\mathbf{Y}}(\beta) = -\mathbb{E}_\beta\left[\nabla^2 l(\beta)\right]$, by computing the second-order derivatives of l:

$$\frac{\partial^2 l}{\partial \beta_m \partial \beta_l} = -\frac{1}{\sigma_0^2} \sum_{i=1}^{n} x_{im} x_{il} = -\frac{1}{\sigma_0^2} X_m^T X_l$$

Then the Hessian matrix is $\nabla^2 l(\beta) = -\frac{1}{\sigma_0^2} X^T X$, so $I_{\mathbf{Y}}(\beta) = \frac{1}{\sigma_0^2} X^T X$. The distribution of $\hat{\beta}$ is then approximately $\mathcal{N}\left(\beta, \sigma_0^2 \left(X^T X\right)^{-1}\right)$ for large n.

In fact, the distribution of $\hat{\beta}$ is exactly this multivariate normal distribution even for small n, because $Y = X\beta + \varepsilon \sim \mathcal{N}\left(X\beta, \sigma_0^2 I\right)$ so that $\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \sim \mathcal{N}\left(\left(X^T X\right)^{-1} X^T X \beta, \sigma_0^2 \left(X^T X\right)^{-1} X^T X \left(X^T X\right)^{-1}\right) = \mathcal{N}\left(\beta, \sigma_0^2 \left(X^T X\right)^{-1}\right)$.

We now apply the **delta method:** Defining $g(x, y) = \sqrt{x^2 + y^2}$, a Taylor expansion yields

$$\begin{aligned}
\hat{A}_1 - A_1 = g\left(\hat{\beta}_1, \hat{\beta}_2\right) - g\left(\beta_1, \beta_2\right) &\approx \frac{\partial g}{\partial x}\left(\beta_1, \beta_2\right) \times \left(\hat{\beta}_1 - \beta_1\right) + \frac{\partial g}{\partial y}\left(\beta_1, \beta_2\right) \times \left(\hat{\beta}_2 - \beta_2\right) \\
&= \frac{\beta_1}{\sqrt{\beta_1^2 + \beta_2^2}}\left(\hat{\beta}_1 - \beta_1\right) + \frac{\beta_2}{\sqrt{\beta_1^2 + \beta_2^2}}\left(\hat{\beta}_2 - \beta_2\right).
\end{aligned}$$

Letting $c_1 = \beta_1 / \sqrt{\beta_1^2 + \beta_2^2}, c_2 = \beta_2 / \sqrt{\beta_1^2 + \beta_2^2}, Z_1 = \hat{\beta}_1 - \beta_1$, and $Z_2 = \hat{\beta}_2 - \beta_2$, the above sampling distribution for $\hat{\beta}$ implies that (Z_1, Z_2) is approximately bivariate normal with mean 0 and covariance given by the upper-left 2×2 block of $\sigma_0^2 \left(X^T X\right)^{-1}$. So $\hat{A}_1 - A_1 \approx c_1 Z_1 + c_2 Z_2$ is approximately normal with mean 0 and variance

$$\begin{aligned}
\operatorname{Var}\left[c_1 Z_1 + c_2 Z_2\right] &= \operatorname{Cov}\left[c_1 Z_1 + c_2 Z_2, c_1 Z_1 + c_2 Z_2\right] \\
&= c_1^2 \operatorname{Var}\left[Z_1\right] + c_2^2 \operatorname{Var}\left[Z_2\right] + 2 c_1 c_2 \operatorname{Cov}\left[Z_1, Z_2\right] \\
&= c_1^2 \sigma_0^2 \left(\left(X^T X\right)^{-1}\right)_{11} + c_2^2 \sigma_0^2 \left(\left(X^T X\right)^{-1}\right)_{22} + 2 c_1 c_2 \sigma_0^2 \left(\left(X^T X\right)^{-1}\right)_{12}.
\end{aligned}$$

Letting $\hat{c}_1 = \hat{\beta}_1/\sqrt{\hat{\beta}_1^2 + \hat{\beta}_2^2}$ and $\hat{c}_2 = \hat{\beta}_2/\sqrt{\hat{\beta}_1^2 + \hat{\beta}_2^2}$, we may estimate the standard error of $\hat{A}_1$ by

$$\hat{\text{se}} = \sqrt{\hat{c}_1^2\sigma_0^2\left((X^TX)^{-1}\right)_{11} + \hat{c}_2^2\sigma_0^2\left((X^TX)^{-1}\right)_{22} + 2\hat{c}_1\hat{c}_2\sigma_0^2\left((X^TX)^{-1}\right)_{12}},$$

and construct a 95% confidence interval for A_1 as $\hat{A}_1 \pm z(0.025)\hat{\text{se}}$.

3 Logistic Regression

3.1 The logistic regression model

Example 3.1.1. *An internet company would like to understand what factors influence whether a visitor to a webpage clicks on an advertisement. Suppose it has available historical data of* n *ad impressions, each impression corresponding to a single ad being shown to a single visitor. For the* i^{th} *impression, let* $Y_i \in \{0, 1\}$ *be such that* $Y_i = 1$ *if the visitor clicked on the ad, and* $Y_i = 0$ *otherwise. The internet company also has available various attributes for each impression, such as the position and size of the ad on the webpage, the company or product being advertised, the age and gender of the visitor, the time of day, the month of the year, etc. For each* i^{th} *impression, suppose that all of these attributes are encoded numerically as* p *covariates* $x_{i1}, \ldots, x_{ip} \in \mathbb{R}$.

The **logistic regression model** assumes each response Y_i is an independent random variable with distribution Bernoulli (p_i), where the log-odds corresponding to p_i is modeled as a linear combination of the covariates plus a possible intercept term:

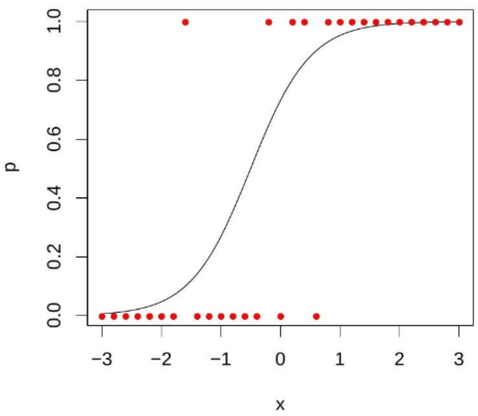
$$\log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}.$$

The intercept β_0 represents the "baseline" log-odds of the visitor clicking on the ad, if all of the covariates take value 0. Each coefficient β_j represents the amount of increase or decrease in the log-odds, if the value of the covariate x_{ij} is increased by 1 unit. The above may be equivalently written as

$$\mathbb{P}[Y_i = 1] = p_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}}}. \tag{6}$$

As in the case of the linear model, we will treat the covariates as fixed and known quantities. The unknown parameters are the regression coefficients $\beta = (\beta_0, \ldots, \beta_p)$.

When there is only one covariate, $p = 1$, we simply write $x_1 = x_{11}, \ldots, x_n = x_{n1}$. The picture below illustrates the logistic regression model, where the red points correspond to the data values $(x_1, Y_1), \ldots, (x_n, Y_n)$ of the covariate and response, and the black curve shows the probability function $p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$:



3.2 Statistical inference

We will explore the following inferential questions:

- Estimate the regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$
- Estimate the "conversion" probability that a new impression, with covariate values $(\tilde{x}_1, \ldots, \tilde{x}_p)$, will lead to click on the ad
- Test whether $\beta_j = 0$ for a particular covariate j, say the age of the visitor and provide a confidence interval for β_j

Since the responses $Y_1, \ldots, Y_n$ are independent Bernoulli random variables, the likelihood for the logistic regression model is given by

$$\operatorname{lik}(\beta_0, \ldots, \beta_p) = \prod_{i=1}^{n} p_i^{Y_i} (1-p_i)^{1-Y_i} = \prod_{i=1}^{n} (1-p_i) \left(\frac{p_i}{1-p_i} \right)^{Y_i},$$

where p_i is defined as a function of $\beta_0, \ldots, \beta_p$ and the covariates $x_{i1}, \ldots, x_{ip}$ by equation (6). Then, introducing for convenience a covariate $x_{i0} \equiv 1$ for all i that captures the intercept term, the log-likelihood is

$$l(\beta_0, \ldots, \beta_p) = \sum_{i=1}^{n} Y_i \log \frac{p_i}{1-p_i} + \log(1-p_i) = \sum_{i=1}^{n} \left(Y_i \sum_{j=0}^{p} \beta_j x_{ij} - \log\left(1 + e^{\sum_{j=0}^{p} \beta_j x_{ij}}\right) \right).$$

To estimate the parameters $\beta_0, \ldots, \beta_p$, we may compute the MLE. For the function $f(x) = \log(1+e^x), f'(x) = \frac{e^x}{1+e^x} = 1 - \frac{1}{1+e^x}$. Then setting the partial derivatives of the log-likelihood equal to 0 and applying the chain rule yields the equations (for $m=0, \ldots, p$)

$$0=\frac{\partial l}{\partial \beta_m}=\sum_{i=1}^n x_{i m}\left(Y_i-\frac{e^{\sum_{j=0}^p \beta_j x_{i j}}}{1+e^{\sum_{j=0}^p \beta_j x_{i j}}}\right) . \tag{7}$$

These equations may be solved numerically (e.g. by Newton-Raphson) to obtain the MLEs $\hat{\beta}_0, \ldots, \hat{\beta}_p$. To estimate the conversion probability for a new impression with covariates $\tilde{x}_1, \ldots, \tilde{x}_p$, we may use the plugin estimate

$$\hat{p}=\frac{e^{\hat{\beta}_0+\hat{\beta}_1 \tilde{x}_1+\ldots+\hat{\beta}_p \tilde{x}_p}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 \tilde{x}_1+\ldots+\hat{\beta}_p \tilde{x}_p}} .$$

To test if a particular coefficient is 0 , say $H_0: \beta_p=0$, one method is using the generalized likelihood ratio test. This null hypothesis corresponds to a sub-model with one fewer free parameter. We may calculate the sub-model MLEs $\hat{\beta}_{0,0}, \ldots, \hat{\beta}_{0, p-1}$ from the same score equations as (7) except with the $p^{t h}$ covariate removed, and use the generalized likelihood ratio statistic

$$-2 \log \Lambda=-2 \log \frac{\operatorname{lik}\left(\hat{\beta}_{0,0}, \ldots, \hat{\beta}_{0, p-1}, 0\right)}{\operatorname{lik}\left(\hat{\beta}_0, \ldots, \hat{\beta}_p\right)} .$$

When the number of impressions n is large, we may perform an approximate level-α test of H_0 by rejecting H_0 when $D>\chi_1^2(\alpha)$, since the difference between model dimensionalities here is 1.

We may obtain a confidence interval for β_j from the MLE estimate $\hat{\beta}_j$ and an estimate of its standard error: We compute the Fisher information $I_{\mathbf{Y}}(\beta)=-\mathbb{E}_\beta\left[\nabla^2 l(\beta)\right]$ by calculating the second partial derivatives of l: For $f(x)=\log \left(1+e^x\right), f^{\prime \prime}(x)=\frac{e^x}{\left(1+e^x\right)^2}$. Then

$$\frac{\partial^2 l}{\partial \beta_m \partial \beta_l}=-\sum_{i=1}^n x_{i m} x_{i l} \frac{e^{\sum_{j=0}^p \beta_j x_{i j}}}{\left(1+e^{\sum_{j=0}^p \beta_j x_{i j}}\right)^2}=-X_m^T W X_l,$$

where we have set $X_j=\left(x_{1 j}, \ldots, x_{n j}\right)$ as the j th column of the matrix of covariates as in Section 2, and defined the $n \times n$ diagonal matrix

$$W:=W(\beta)=\operatorname{diag}\left(\frac{e^{\sum_{j=0}^p \beta_j x_{1 j}}}{\left(1+e^{\sum_{j=0}^p \beta_j x_{1 j}}\right)^2}, \cdots, \frac{e^{\sum_{j=0}^p \beta_j x_{n j}}}{\left(1+e^{\sum_{j=0}^p \beta_j x_{n j}}\right)^2}\right) .$$

So $\nabla^2 l(\beta) = -X^T W X, I_{\mathbf{Y}}(\beta) = X^T W X$, and the approximate sampling distribution of $\hat{\beta}$ for large n is $\mathcal{N}\left(\beta, \left(X^T W X\right)^{-1}\right)$. Letting $\hat{W} = W(\hat{\beta})$ be the plugin estimate of the diagonal matrix W, we may estimate the standard error of $\hat{\beta}_j$ by $\hat{\text{se}}_j = \sqrt{\left(\left(X^T \hat{W} X\right)^{-1}\right)_{jj}}$, and obtain a 95% confidence interval for β_j as $\hat{\beta}_j \pm z(0.025)\hat{\text{se}}_j$.

Remark 3.1. *A word of caution regarding model misspecification: The above standard error estimates $\hat{\text{se}}_j$ (which are the standard errors reported by most logistic regression software) are only expected to be accurate when the logistic regression model is correctly specified that is, when the Y_i's are truly independent random variables with distribution* Bernoulli (p_i), *where the log-odds for each p_i is the same linear combination of the covariates. This is because, as in the case of n IID observations, the covariance of $\hat{\beta}$ is given by the inverse Fisher information only in a correctly specified model.*

Logistic regression is still oftentimes used as a tool for binary classification problems even if the model does not yield an extremely accurate fit to the data, as long as the model has good classification accuracy. In such settings, the MLE $\hat{\beta}$ represents the "closest" logistic regression model (in the given covariates) to the true distribution of $Y_1, \ldots, Y_n$, in the sense of KL-divergence as in Section 3 of Chapter 3. The standard error for $\hat{\beta}_j$ may be robustly estimated using either a sandwich estimator or the nonparametric bootstrap. For the logistic regression model, the sandwich estimate of the covariance matrix of $\hat{\beta}$ is given by[2]

$$\left(X^T \hat{W} X\right)^{-1} \left(X^T \tilde{W} X\right) \left(X^T \hat{W} X\right)^{-1},$$

where $\tilde{W} = \operatorname{diag}\left((Y_1 - \hat{p}_1)^2, \ldots, (Y_n - \hat{p}_n)^2\right)$ and $\hat{p}_i$ is the fit probability for the i^{th} observation, defined by the right side of equation (6) with $\hat{\beta}$ in place of β. The (j, j) element of this matrix gives a sandwich estimate for the variance of $\hat{\beta}_j$.

*Alternatively, one may use the **pairs bootstrap**, which pairs the covariates and response for each i^{th} observation into a single data vector $(x_{i1}, \ldots, x_{ip}, Y_i)$, and then draws bootstrap samples by randomly selecting, with replacement, n of these vectors. The logistic regression model is fit to each such bootstrap sample to yield an MLE $\hat{\beta}^*$, and the standard error of $\hat{\beta}_j$ is estimated by the empirical standard deviation of $\hat{\beta}^*_j$ across bootstrap samples.*

[2] See Liang and Zeger, "Longitudinal data analysis using generalized linear models" or Agresti, "Categorical Data Analysis" Section 12.3.3.

4 Poisson Regression

4.1 The Poisson log-linear model

Example 4.1.1. *Neurons in the central nervous system transmit signals via a series of action potentials, or "spikes". The spiking of a single neuron may be measured by a microelectrode, and its sequence of spikes over time is called a spike train. A simple and commonly-used statistical model for a spike train is an inhomogeneous Poisson point process, which has the following property: For n time windows of length Δ, letting Y_i denote the number of spikes generated by the neuron in the i^{th} time window, the random variables $Y_1, \ldots, Y_n$ are independent and distributed as $Y_i \sim \text{Poisson}\,(\lambda_i \Delta)$, where the parameter λ_i controls the spiking rate in the i^{th} time window. For simplicity, we will assume $\Delta = 1$.*

The spiking rate λ_i of a neuron may be influenced by external sensory stimuli present in this i^{th} window of time, for example, the intensity and pattern of light visible to the eye or the texture of an object presented to the touch. To understand the effects of these sensory stimuli on the spiking rate of a particular neuron, we may perform an experiment that applies different stimuli in different windows of time and records the neural response. Encoding the stimuli applied in the i^{th} window of time by a set of p covariates $x_{i1}, \ldots, x_{ip}$, a simple model for the Poisson rate parameter λ_i is given by

$$\log \lambda_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \tag{8}$$

or equivalently,

$$\lambda_i = e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}}.$$

Together with the distributional assumption $Y_i \sim \text{Poisson}\,(\lambda_i)$, this is called the **Poisson log-linear model**, or the Poisson regression model. It is a special case of what is known in neuroscience as the linear-nonlinear Poisson cascade model.

More generally, the Poisson log-linear model is a model for n responses $Y_1, \ldots, Y_n$ that take integer count values. Each Y_i is modeled as an independent $\text{Poisson}\,(\lambda_i)$ random variable, where $\log \lambda_i$ is a linear combination of the covariates corresponding to the i^{th} observation. As in the cases of linear and logistic regression, we treat the covariates as fixed constants, and the model parameters to be inferred are the regression coefficients $\beta = (\beta_0, \ldots, \beta_p)$.

4.2 Statistical inference

We will describe the procedure for maximum-likelihood estimation of the regression coefficients and Fisher-information based estimation of their standard errors, and discuss some issues concerning model misspecification and robust standard error estimates.

Since $Y_1, \ldots, Y_n$ are independent Poisson random variables, the likelihood function is given by

$$\operatorname{lik}\left(\beta_0, \ldots, \beta_p\right) = \prod_{i=1}^{n} \frac{\lambda_i^{Y_i} e^{-\lambda_i}}{Y_i!},$$

where λ_i is defined in terms of $\beta_0, \ldots, \beta_p$ and the covariates $x_{i1}, \ldots, x_{ip}$ via equation (8). Setting $x_{i0} \equiv 1$ for all i, the log-likelihood is then

$$\begin{aligned} l\left(\beta_0, \ldots, \beta_p\right) &= \sum_{i=1}^{n} Y_i \log \lambda_i - \lambda_i - \log Y_i! \\ &= \sum_{i=1}^{n} Y_i \left(\sum_{j=0}^{p} \beta_j x_{ij}\right) - e^{\sum_{j=0}^{p} \beta_j x_{ij}} - \log Y_i! \end{aligned}$$

and the MLEs are the solutions to the system of score equations, for $m = 0, \ldots, p$,

$$0 = \frac{\partial l}{\partial \beta_m} = \sum_{i=1}^{n} x_{im} \left(Y_i - e^{\sum_{j=0}^{p} \beta_j x_{ij}}\right).$$

These equations may be solved numerically using the Newton-Raphson method.

The Fisher information matrix $I_{\mathbf{Y}}(\beta) = -\mathbb{E}_\beta\left[\nabla^2 l(\beta)\right]$ may be obtained by computing the second-order partial derivatives of l:

$$\frac{\partial^2 l}{\partial \beta_m \partial \beta_l} = -\sum_{i=1}^{n} x_{im} x_{il} e^{\sum_{j=0}^{p} \beta_j x_{ij}}.$$

Writing $X_j = (x_{1j}, \ldots, x_{nj})$ as the j^{th} column of the covariate matrix X and defining the diagonal matrix

$$W = W(\beta) := \operatorname{diag}\left(e^{\sum_{j=0}^{p} \beta_j x_{1j}}, \ldots, e^{\sum_{j=0}^{p} \beta_j x_{nj}}\right),$$

the above may be written as $\frac{\partial^2 l}{\partial \beta_m \partial \beta_l} = -X_m^T W X_l$, so $\nabla^2 l(\beta) = -X^T W X$ and $I_{\mathbf{Y}}(\beta) = X^T W X$. For large n, if the Poisson log-linear model is correct, then the MLE vector $\hat{\beta}$ is approximately distributed as $\mathcal{N}\left(\beta, \left(X^T W X\right)^{-1}\right)$. We may then estimate the standard error of $\hat{\beta}_j$ by

$$\hat{\operatorname{se}}_j = \sqrt{\left(\left(X^T \hat{W} X\right)^{-1}\right)_{jj}},$$

where $\hat{W} = W(\hat{\beta})$ is the plugin estimate for W. These formulas are the same as for the case of logistic regression in Section 3, except with a different form of the diagonal matrix W.

The modeling assumption of a Poisson distribution for Y_i is rather restrictive, as it implies that the variance of Y_i must be equal to its mean. This is rarely true in practice, and it is frequently the case that the observed variance of Y_i is larger than its mean – this problem is known as **overdispersion**. Nonetheless, the Poisson regression model is oftentimes used in overdispersed settings: As long as $Y_1, \ldots, Y_n$ are independent and

$$\log \mathbb{E}\left[Y_i\right] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

for each i (so the model for the means of the Y_i' 's is correct), then it may be shown that the MLE $\hat{\beta}$ in the Poisson regression model is unbiased for β, even if the distribution of Y_i is not Poisson and the variance of Y_i exceeds its mean. The above standard error estimate $\hat{se}_j$ and the associated confidence interval for β_j, though, would not correct in the overdispersed setting. One may use instead the robust sandwich estimate of the covariance of $\hat{\beta}$, given by

$$\left(X^T \hat{W} X\right)^{-1} \left(X^T \tilde{W} X\right) \left(X^T \hat{W} X\right)^{-1},$$

where

$$\tilde{W} = \operatorname{diag}\left(\left(Y_1 - \hat{\lambda}_1\right)^2, \ldots, \left(Y_n - \hat{\lambda}_n\right)^2\right)$$

and $\hat{\lambda}_i = e^{\sum_{j=0}^{p} \hat{\beta}_j x_{ij}}$ is the fitted value of λ for the i^{th} observation. Alternatively, one may use the pairs bootstrap procedure as described in Section 3.

Remark 4.1. *The linear model, logistic regression model, and Poisson regression model are all examples of the* ***generalized linear model (GLM)****. In a generalized linear model, $Y_1, \ldots, Y_n$ are modeled as independent observations with distributions $Y_i \sim f\left(y \mid \theta_i\right)$ for some one-parameter family $f(y \mid \theta)$. The parameter θ_i is modeled as*

$$g\left(\theta_i\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

for some one-to-one transformation $g : \mathbb{R} \to \mathbb{R}$ called the ***link function****, where $x_{i1}, \ldots, x_{ip}$ are covariates corresponding to Y_i. In the linear model considered in Section 2, the parameter was $\theta \equiv \mu$ where $f(y \mid \mu)$ was the* PDF *of the $\mathcal{N}\left(\mu, \sigma_0^2\right)$ distribution (for a known variance σ_0^2), and $g(\mu) = \mu$. In logistic regression, the parameter was $\theta \equiv p$ where $f(y \mid p)$ was the PMF of the Bernoulli (p) distribution, and $g(p) = \log \frac{p}{1-p}$. In Poisson regression, the parameter was $\theta \equiv \lambda$ where $f(y \mid \lambda)$ was the PMF of the*

Poisson (λ) *distribution, and* $g(\lambda) = \log \lambda$.

The choice of the link function g is an important modeling decision, as it determines which transform of the model parameter should be modeled as linear in the observed covariates. In each of the three examples discussed, we used what is called the **natural link**, which is motivated by considering a change-of-variable for the parameter, $\theta \mapsto \eta(\theta)$, so that the PDF/PMF$f(y \mid \eta)$ in terms of the new parameter η has the form

$$f(y \mid \eta) = e^{\eta y - A(\eta)} h(y)$$

for some functions A and h. For example, the Bernoulli PMF is

$$f(y) = p^y (1-p)^{1-y} = (1-p) \left(\frac{p}{1-p} \right)^y = e^{\left(\log \frac{p}{1-p} \right) y + \log(1-p)},$$

so we may set $\eta = \log \frac{p}{1-p}$, $A(\eta) = -\log(1-p) = \log\left(1 + e^{\eta}\right)$, and $h(y) = 1$. This is called the **exponential family** form of the PDF/PMF, and η is called the **natural parameter**. In each example, the natural link simply sets $g(\theta) = \eta(\theta)$ (or equivalently, $g(\theta) = c\eta(\theta)$ for a constant c).

Use of the natural link leads to some nice mathematical properties for likelihood-based inference - for instance since η is modeled as linear in β, the second-order partial derivatives of

$$\log f(Y \mid \eta) = \eta Y - A(\eta) + \log h(Y)$$

with respect to β do not depend on Y, so the Fisher information is always given by $-\nabla^2 l(\beta)$ without needing to take an expectation. (We sometimes say in this case that the "observed and expected Fisher information matrices" are the same.) On the other hand, from the modeling perspective, there is usually no intrinsic reason to believe that the natural link $g(\theta) = \eta(\theta)$ is the correct transformation of θ that is well-modeled as a linear combination of the covariates, and other link functions are also commonly used, especially if they lead to a better fit for the data.

5 The Proportional Hazards Model

Example 5.0.1. *A clinical trial is performed to study the effect of a drug for maintaining/prolonging remission induced by chemotherapy in the treatment of acute leukemia. (Remission is the disappearance of leukemic cells and other symptoms of the disease.) For each i^{th} patient in the trial, let T_i denote the length of the remission (or equivalently, the time until recurrence of cancer), which we wish to model in terms of patient-specific covariates $x_{i1}, \ldots, x_{ip}$. The first covariate x_{i1} may be a $0-1$ variable indicating whether patient i received the drug or a placebo, and the remaining covariates are other factors, such as the age of the patient, that may affect the remission length.*

Modeling T_i as a continuous, positive-valued random variable with CDF *$F_i(t)$ and* PDF *$f_i(t) = F_i'(t)$, it is useful to think about the distribution of T_i in terms of its **hazard function** $\lambda_i(t)$, which represents the "instantaneous risk" of recurrence at time t:*

$$\lambda_i(t) := \lim_{\delta \to 0} \frac{1}{\delta} \mathbb{P}\left[T_i \leq t + \delta \mid T_i \geq t\right].$$

In other words, for small δ, the probability that a recurrence of cancer occurs in the time window $[t, t+\delta]$, conditional on it not having occurred up to time t, is approximately $\delta\lambda_i(t)$. The hazard function may be expressed in terms of the CDF $F_i(t)$ and PDF *$f_i(t)$ as*

$$\lambda_i(t) = \lim_{\delta \to 0} \frac{\mathbb{P}\left[t \leq T_i \leq t + \delta\right]}{\delta \mathbb{P}\left[T_i \geq t\right]} = \lim_{\delta \to 0} \frac{F_i(t+\delta) - F_i(t)}{\delta\left(1 - F_i(t)\right)} = \frac{f_i(t)}{1 - F_i(t)}.$$

To develop some intuition for the hazard function, consider a simple example where $T_i \sim \text{Exponential}(\theta)$. Then the PDF *is $f_i(t) = \theta e^{-\theta t}$, the* CDF *is $F_i(t) = 1 - e^{-\theta t}$, so the hazard function is*

$$\lambda_i(t) = \frac{\theta e^{-\theta t}}{1 - (1 - e^{-\theta t})} = \theta.$$

In this case, the hazard function is constant in time (which is a special property of the exponential distribution). Intuitively, this means that assuming the remission has lasted until time t, the probability of the recurrence occurring in the next instant of time is the same for every t and is determined only by θ. The parameter θ governs how quickly the exponential distribution decays - the larger the value of θ, the faster the rate of decay, and the more likely it is that the recurrence of cancer will occur at any next instant of time.

Cox's **proportional hazards model** does not assume that the distribution of T_i is exponential, or that it follows any particular parametric form. Instead, it models the hazard function for T_i as

$$\lambda_i(t) = \lambda_0(t) \exp\left(\beta_1 x_{i1} + \ldots + \beta_p x_{ip}\right)$$

The regression coefficients $\beta_1, \ldots, \beta_p$ are unknown parameters determining the effects of the covariates on the remission length T_i, and $\lambda_0(t)$ is a completely unknown **baseline hazard function**. In this model, $\lambda_0(t)$ controls the shape of the hazard function over time for all patients, and the factor $\exp\left(\beta_1 x_{i1} + \ldots + \beta_p x_{ip}\right)$ controls the scale of the hazard function for each patient i. Thus the model asserts that for any two patients i and j, their hazard functions have the same shape and differ only in scale so that the ratio of their hazard functions $\lambda_i(t)/\lambda_j(t)$ is constant over time (hence the name "proportional" hazards). The model posits that this scale ratio is determined by a linear combination of the differences of the patients' covariates.

In the clinical trial, the remission for a patient i may last longer than the duration for which the patient participates in the trial. In this case, we do not observe their true remission length T_i (which can take the value ∞ if the cancer never returns), but instead, we only observe that $T_i > l_i$ where l_i is the length of time for which the patient is in the trial. This type of observation is called **right-censored**. The method of inference developed below for the proportional hazards model will naturally handle data in which some of the observations are right-censored. We treat l_i as a fixed and known constant for every patient, so that we either observe a value of T_i that is at most l_i, or we observe that $T_i > l_i$. We will make an important assumption that T_i (the true remission length) does not depend on l_i (the right-censoring time).

The proportional hazards model may be used to model the time-to-onset of any event pertaining to an individual in terms of observed covariates for that individual; example applications include medical trials as above, as well as industrial reliability experiments that model the time-to-failures of devices. According to a 2014 list in the scientific journal *Nature*, the 1972 paper by David Cox that introduced this model is the 2^{nd} most cited paper in statistics and the 24^{th} most cited paper in all of science.

5.1 Statistical inference

In many applications, the regression coefficients $\beta_1, \ldots, \beta_p$ are of greater interest than the baseline hazard function $\lambda_0(t)$. If the first covariate x_{i1} corresponds to an indicator variable representing assignment to the treatment group (drug) or the control group (placebo), then the coefficient β_1 represents the log-hazards-ratio between the two groups after controlling for the other covariates $x_{i2}, \ldots, x_{ip}$. We will discuss inference procedures for the following tasks:

- Estimate $\beta_1, \ldots, \beta_p$.

- Test whether $\beta_1 = 0$.

Perhaps surprisingly, it is possible to perform these inference tasks without any knowledge of, and without any assumptions regarding, the baseline hazard function $\lambda_0(t)$.

In previous models, we performed inference by writing down the likelihood of the model parameters. Inference in the proportional hazards model will be slightly different from these previous examples, because the baseline hazard function $\lambda_0(t)$ is completely unknown, and the likelihood function and MLEs for $\beta_1, \ldots, \beta_p$ would depend on $\lambda_0(t)$. If $\lambda_0(t)$ were modeled parametrically using a small number of additional parameters, then we may include these as parameters of the model and fit the entire model by computing the joint MLEs of these additional parameters and $\beta_1, \ldots, \beta_p$. However, without parametric modeling assumptions on $\lambda_0(t)$, in this course, we have not discussed procedures for how to estimate an entire unknown function $\lambda_0(t)$.

We will circumvent this problem by conditioning on the set of all distinct observed recurrence times $t_{(1)} < t_{(2)} < \ldots < t_{(m)}$ across all patients. (This idea is quite similar to how we conditioned on the set of all distinct observed values in permutation two-sample tests, to address the problem of an unknown common distribution function $F = G$ under the null hypothesis.) Since we are modeling T_i as continuous random variables, we may assume that each observed recurrence time $t_{(k)}$ corresponds to only one patient (i.e. there are no ties in recurrence times), so m is just the total number of non-right-censored observations. For each $t_{(k)}$, the *risk set* $\mathcal{R}_{(k)}$ immediately before time $t_{(k)}$ is the set of patients who have not yet left the study (been right-censored) and are still in remission-this represents the candidate set of patients for which we may have observed a recurrence at time $t_{(k)}$. Conditional on the fact that some patient in this risk set $\mathcal{R}_{(k)}$ has a recurrence at time $t_{(k)}$, the probability that it is a particular patient $I_k \in \mathcal{R}_{(k)}$ is

$$\frac{\lambda_{I_k}\left(t_{(k)}\right)}{\sum_{i \in \mathcal{R}_{(k)}} \lambda_i\left(t_{(k)}\right)}$$

(the ratio of the "instantaneous rate" of recurrence for patient I_k to the sum of the rates for all candidate patients). Under the proportional hazards model, this is

$$\frac{\lambda_0\left(t_{(k)}\right) \exp\left(\beta_1 x_{I_k 1} + \ldots + \beta_p x_{I_k p}\right)}{\sum_{i \in \mathcal{R}_{(k)}} \lambda_0\left(t_{(k)}\right) \exp\left(\beta_1 x_{i1} + \ldots + \beta_p x_{ip}\right)}.$$

Importantly, the factor $\lambda_0\left(t_{(k)}\right)$ cancels from the numerator and denominator of this expression, yielding a quantity that does not depend on the baseline hazard function $\lambda_0(t)$. Taking a product of the above expression overall observed recurrence times yield

$$\operatorname{plik}(\beta_1, \ldots, \beta_p) = \prod_{k=1}^{m} \frac{\exp(\beta_1 x_{I_k 1} + \ldots + \beta_p x_{I_k p})}{\sum_{i \in \mathcal{R}_{(k)}} \exp(\beta_1 x_{i1} + \ldots + \beta_p x_{ip})}$$

This quantity is called the **partial likelihood function** of $\beta_1, \ldots, \beta_p$. Intuitively, it captures all of the information contained by the observations that at each time $t_{(k)}$, the particular recurrence was for patient I_k as opposed to the other patients for which we could have observed a recurrence at that time. We may perform likelihood-based inference using this partial likelihood in place of the usual likelihood function.

We may estimate $\beta_1, \ldots, \beta_p$ by maximizing the partial likelihood over these parameters. As with usual MLE calculations, it is computationally convenient first to take a logarithm, so we consider the log-partial likelihood

$$l(\beta_1, \ldots, \beta_p) = \sum_{k=1}^{m} \left(\beta_1 x_{I_k 1} + \ldots + \beta_p x_{I_k p} - \log \sum_{i \in \mathcal{R}_{(k)}} \exp(\beta_1 x_{i1} + \ldots + \beta_p x_{ip}) \right).$$

We may maximize this quantity by setting its derivative with respect to each $\beta_1, \ldots, \beta_p$ equal to 0:

$$0 = \frac{\partial l}{\partial \beta_j} = \sum_{k=1}^{m} \left(x_{I_k j} - \frac{\sum_{i \in \mathcal{R}_{(k)}} x_{ij} \exp(\beta_1 x_{i1} + \ldots + \beta_p x_{ip})}{\sum_{i \in \mathcal{R}_{(k)}} \exp(\beta_1 x_{i1} + \ldots + \beta_p x_{ip})} \right).$$

Solving numerically this system of p equations in p unknowns $\beta_1, \ldots, \beta_p$ yields the maximum partial-likelihood estimates $\hat{\beta}_1, \ldots, \hat{\beta}_p$.

The asymptotic theory for the maximum partial-likelihood estimate is analogous to that of the MLE in usual parametric models (although the mathematical derivation of this theory requires more advanced probabilistic tools that we did not cover in this course). In particular, the usual generalized likelihood ratio test applies: To test $H_0 : \beta_1 = 0$, we may compute the maximum partial likelihood estimates $\hat{\beta}_{2,0}, \ldots, \hat{\beta}_{p,0}$ in this sub-model, using the same procedure as above with the first covariate removed. The test statistic

$$-2 \log \Lambda = -2 \log \frac{\operatorname{plik}\left(0, \hat{\beta}_{2,0}, \ldots, \hat{\beta}_{p,0}\right)}{\operatorname{plik}\left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)}$$

is, under mild regularity conditions, distributed as χ_1^2 in the limit of large n, and an asymptotic level-α test would reject H_0 when $-2 \log \Lambda$ exceeds the upper-α point $\chi_1^2(\alpha)$.