Regression Cheatsheet

Contents

Binary Logistic Regression

Data Modelling Framework

The Binary Logistic regression model has a response variable in the form:

$$Y_i = \begin{cases} 1 & \text{if the } i\text{th observation is a success,} \\ 0 & \text{otherwise.} \end{cases}$$

As the response variable can only take the values 0 or 1, the key parameter becomes the probability that Y_i takes on the value of 1, i.e. the probability of success, denoted as p_i. Hence:

$$Y_i \sim \text{Bernoulli}(p_i).$$

The Binary Logistic regression approach models the probability of success, p_i, of the binary response Y_i. To re-express p_i **on an unrestricted scale**, the modelling is done in terms of the logit function (the link function in this model).

Specifically, for a training set of size n, p_i ($i = 1, 2, \dots, n$) will depend on the values of the k regressors $X_{i,1}, X_{i,2}, \dots, X_{i,k}$ in the form:

$$h(p_i) = \text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 X_{i,1} + \beta_1 X_{i,2} + \ldots + \beta_k X_{i,k},$$

or equivalently

$$p_i = \frac{\exp\left[\text{logit}(p_i)\right]}{1 + \exp\left[\text{logit}(p_i)\right]}.$$

Note that the $\log(\cdot)$ notation in the model above refers to the **natural logarithm**, i.e., **logarithm base** e. The equation above for p_i shows that this Binary Logistic regression model will result in values of the probability of success p_i that are always between 0 and 1. Furthermore, the response in this GLM is called the log-odds, the logarithm of the odds

$$\frac{p_i}{1 - p_i},$$

the ratio of the probability of the event to the probability of the non-event.

Estimation

Under a general framework with k regressors, the **regression parameters** $\beta_0, \beta_1, \dots, \beta_k$ in this model are also unknown. In order to fit the model, we can use the function `glm()` and its argument `family = binomial` (required to specify the binary nature of the response), which obtains the estimates $\hat{\beta}_0, \hat{\beta}_1, \dots \hat{\beta}_k$ (note the hat notation).

The estimates are obtained through **maximum likelihood** where we assume a **joint probability mass function of the** n **responses** Y_i.

Inference

We can determine **whether a regressor is statistically associated with the logarithm of the response's odds** through **hypothesis testing** for the parameters β_j. We will need information about the estimated regression coefficient $\hat{\beta}_j$ and its corresponding variability which is reflected in the **standard error** of the estimate, $\text{se}\left(\hat{\beta}_j\right)$.

To determine the **statistical significance** of $\hat{\beta}_j$, you can use the **Wald statistic** z_j

$$z_j = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta}_j)}$$

to test the hypotheses

$$\begin{gathered}H_0 : \beta_j = 0 \\ H_a : \beta_j \neq 0.\end{gathered}$$

A statistic like z_j is analogous to the t-value in Ordinary Least-squares (OLS) regression. However, in Binary Logistic regression, provided the sample size n is large enough, z_j has an **approximately Standard Normal distribution** under H_0 rather than a t-distribution.

R provides the corresponding p**-value** for each β_j. The smaller the p-value, the stronger the evidence against the null hypothesis H_0. Hence, a small enough p-value (less than the significance level α) indicates that the data provides evidence in favour of **association** (**or causation in the case of an experimental study!**) between the log-dds and the jth regressor. Furthermore, given a specified level of confidence, we can construct approximate $(1 - \alpha) \times 100\%$ **confidence intervals** (CIs) for the corresponding true value of β_j:

$$\hat{\beta}_j \pm z_{\alpha/2} \text{se}(\hat{\beta}_j),$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the **Standard Normal distribution**.

Model Selection

Analysis of Deviance

The **deviance** (D_k) criterion can be used to compare a given model with k regressors with that of a **baseline model**. The usual baseline model is the **saturated** or **full model**, which perfectly fits the data because it allows a distinct probability of success p_i for the ith observation in the training dataset ($i = 1, \dots, n$), **unrelated to the k regressors**.

Given the definition of the saturated or full model under this context, we can view it as an **overfitted model**. Thus, we aim to avoid this type of model.

The **maximized likelihood** of this full model is denoted as $\hat{\ell}_f$. Now, let $\hat{\ell}_k$ be the value of the maximized likelihood computed from our dataset of n observation with k regressors.

We can compare the fits provided by these two models by the deviance D_k given by

$$D_k = -2 \log \left(\frac{\hat{\ell}_k}{\hat{\ell}_f} \right) = -2 \left[\log \left(\hat{\ell}_k \right) - \log \left(\hat{\ell}_f \right) \right].$$

Note that D_k expresses **how much our given model deviates from the full model on log-likelihood scale**. This metric is interpreted as follows:

- **Large values** of D_k arise when $\hat{\ell}_k$ is small relative to $\hat{\ell}_f$, indicating that **our given model fits the data poorly compared to the baseline model**.
- **Small values** of D_k arise when $\hat{\ell}_k$ is similar to $\hat{\ell}_f$, indicating that **our given model provides a good fit to the data compared to the baseline model**.

For the specific case of the Binary Logistic regression, it can be shown that D_k is represented by the following equation:

$$D_k = -2 \sum_{i=1}^{n} \left[\hat{p}_i \text{logit}(\hat{p}_i) + \log(1 - \hat{p}_i) \right], \tag{29}$$

where $\hat{p}_i$ is the estimated probability of success for the ith observation for $i = 1, \dots, n$ in our training set **with our fitted model of k regressors**. Equation (29) above comes from

maximum likelihood estimation (MLE).

For the specific case of Binary Logistic regression, deviance D_k (29) cannot be used as a standalone metric of **goodness of fit** because of **data sparsity**; i.e., each ith observation has a different set of observed values for the k regressors if at least one of them is of **continuous-type**.

This data sparsity puts D_k just in function of the fitted probabilities $\hat{p}_i$ and not on the observed values y_i (which tells us nothing about the agreement of our model with k regressors to the observed data!).

Still, for the case of Binary Logistic regression, we can use the analysis of deviance to perform model selection **between two models where one is nested in the other**.

Suppose **Model 1** is nested in **Model 2**. Hence, this specific model selection will involve a hypothesis testing. The hypotheses are:

$$\begin{gathered}H_0 : \textbf{Model 1} \text{ fits the data better than } \textbf{Model 2} \\ H_a : \textbf{Model 2} \text{ fits the data better than } \textbf{Model 1}.\end{gathered}$$

Let D_2 be the deviance for **Model 2** and D_1 the deviance for **Model 1**. The test statistic Δ_D for the analysis of deviance is given by:

$$\Delta_D = D_1 - D_2 \sim \chi^2_d,$$

which **assymptotically** (i.e., $n \to \infty$) is **Chi-squared distributed** with d degrees of freedom under H_0. The d degrees of freedom are the **regression parameters of difference between both models**.

Formally, this nested hypothesis testing is called the **likelihood-ratio test**.

Akaike Information Criterion

One of the drawbacks of the analysis of deviance is that it only allows to test **nested** regression models when we have sparse data (i.e., each response is associated with a different set of values in the regressors).

Fortunately, we have alternatives for model selection. **The Akaike Information Criterion (AIC) makes it possible to compare models that are either nested or not.** For a model with k regressors and a deviance D_k is defined as:

$$\mathrm{AIC}_k = D_k + 2k.$$

Models with **smaller** values of AIC_k are preferred. That said, AIC_k favours models with small values of D_k.

However, AIC_k penalizes for including more regressors in the model. Hence, it discourages overfitting, which is key in model selection. This is why we select that model with the smallest AIC_k.

Bayesian Information Criterion

An alternative to AIC is the Bayesian Information Criterion (BIC). **The BIC also makes it possible to compare models that are either nested or not.** For a model with k regressors, n observations used for training, and a deviance D_k; it is defined as:

$$\mathrm{BIC}_k = D_k + k \log(n).$$

Models with **smaller** values of BIC_k are preferred. That said, BIC_k also favours models with small values of D_k.

The differences between AIC and BIC will be more pronounced in datasets with large sample sizes n. As the BIC penalty of $k \log(n)$ will always be larger than the AIC penalty of $2k$ when $n > 7$, **BIC tends to select models with fewer regressors than AIC**.

Cox Proportional Hazards Model

It is a widely popular **semiparametric** regression model (namely, **Cox regression**). **In Statistics**, we call a regression model semiparametric when we only define the systematic component (i.e., regressors and regression estimates). Yet, we **do not assume any specific distribution** for our response of interest.

Data Modelling Framework

The Cox Proportional Hazards model is a commonly used survival model that allows us to interpret how regressors influence a censored response.

Using a training set of size n, the idea is to model the **hazard function** $\lambda_i(t)$ directly for the i th observation ($i = 1, \dots, n$) subject to k regressors $X_{i,j}$ ($j = 1, \dots, k$):

$$\lambda_i \left(t | X_{i,1}, \dots, X_{i,k} \right) = \lambda_0(t) \exp \left(\sum_{j=1}^{k} \beta_j X_{i,j} \right).$$

This model has certain particularities, which will be highlighted below:

- We model the ith individual hazard function $\lambda_i \left(t | X_{i,1}, \dots, X_{i,k} \right)$ along with a baseline hazard $\lambda_0(t)$, which is equal for all the n observations, multiplied by $\exp \left(\sum_{j=1}^{k} \beta_j X_{i,j} \right)$.
- The model does not have an intercept β_0. Nevertheless, the baseline hazard $\lambda_0(t)$ acts as the intercept in this survival framework.
- The baseline hazard $\lambda_0(t)$ is the reference **for all subjects** when all covariates are equal to zero. Note this baseline hazard depends on time t **unlike the regressors** $X_{i,j}$. Cox regression does not assume any distribution on this baseline hazard, which is the non-parametric part of the model.
- The parametric part of the model is reflected in $\exp \left(\sum_{j=1}^{k} \beta_j X_{i,j} \right)$.
- Therefore, given the combinations of both situations, Cox regression is overall defined as semiparametric.

To exemplify the **proportional hazards** assumption, let us consider a simple Cox regression model with a continuous regressor $X_{i,1}$. Moreover, assume these two subjects $i = 1, 2$ have the following regressor values:

$$X_{2,1} = X_{1,1} + 1 \qquad \text{Subject 2's regressor value is one unit larger.}$$

Their responses can be modelled as:

$$\lambda_1 \left(t | X_{1,1} \right) = \lambda_0(t) \exp \left(\beta_1 X_{1,1} \right) \qquad \text{Subject 1}$$

$$\begin{aligned}\lambda_2\left(t \mid X_{2,1}\right) &= \lambda_0(t) \exp\left(\beta_1 X_{2,1}\right) \\ &= \lambda_0(t) \exp\left[\beta_1\left(X_{1,1}+1\right)\right] \\ &= \lambda_0(t) \exp\left(\beta_1 X_{1,1}\right) \exp\left(\beta_1\right) \qquad \text{Subject 2}\end{aligned}$$

Then, the hazard ratio comes into play as:

$$\frac{\lambda_2\left(t \mid X_{2,1}\right)}{\lambda_1\left(t \mid X_{1,1}\right)} = \frac{\lambda_0(t) \exp\left(\beta_1 X_{1,1}\right) \exp\left(\beta_1\right)}{\lambda_0(t) \exp\left(\beta_1 X_{1,1}\right)} = \exp\left(\beta_1\right) \quad \Rightarrow \quad \lambda_2(t) = \lambda_1(t) \exp\left(\beta_1\right)$$

In plain words and generally, the proportional hazards assumption assumes that the hazard for any subject is proportional to the hazard of any other subject **via the exponentiated regression coefficients**.

Estimation

Parameter estimation in Cox regression is done through another special maximum likelihood technique using a **partial likelihood**. A partial likelihood is a specific class of **quasi-likelihood**, which does not require assuming any specific PDF for the continuous survival times Y_i (thus, the baseline hazard is not required to take any distributional form).

However, we use the parametric part

$$\exp\left(\sum_{j=1}^{k} \beta_j X_{i,j}\right)$$

to perform our corresponding estimation.

Inference

We can determine **whether a regressor is statistically associated with the response's hazard function** through **hypothesis testing** for β_j. We will need the estimated regression coefficient $\hat{\beta}_j$ and its **standard error**, $\mathrm{se}(\hat{\beta}_j)$.

You can test the below hypotheses via the **Wald statistic** $z_j = \frac{\hat{\beta}_j}{\mathrm{se}(\hat{\beta}_j)}$:

$$
\begin{gathered}
H_0 : \beta_j = 0 \\
H_a : \beta_j \neq 0.
\end{gathered}
$$

Given a large enough training size n, z_j has an **approximately Standard Normal distribution** under H_0.

R provides the corresponding p**-value** for each β_j. The smaller the p-value, the stronger the evidence against the null hypothesis H_0. Hence, a small enough p-value (less than the significance level α) indicates that the data provides evidence in favour of **association** (**or causation in the case of an experimental study!**) between the hazard function and the jth regressor. Furthermore, given a specified level of confidence, we can construct approximate $(1-\alpha) \times 100\%$ CIs for the corresponding true value of β_j:

$$
\hat{\beta}_j \pm z_{\alpha/2} \text{se}(\hat{\beta}_j),
$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the **Standard Normal distribution**.

Prediction

Even though Cox regression models the hazard function $\lambda_i(t)$, it is possible to obtain **a given estimated survival function** via the following equation:

$$
S\left(t | X_1, \dots X_k\right) = S_0(t)^{\exp\left(\sum_{j=1}^{k} \beta_j X_{i,j}\right)},
$$

where

$$
S_0(t) = \exp\left[-\Lambda_0(t)\right]
$$

is the **baseline survival function**.

Note that $\Lambda_0(t)$ is the cumulative baseline hazard function:

$$
\Lambda_0(t) = \int_0^t \lambda_0(u) du.
$$

We use the baseline cumulative hazard $\Lambda_0(t)$ to obtain $S_0(t)$.

Cumulative Hazard Function

Let us consider a **continuous** random variable

$$Y = \text{Time until an event occurs.}$$

The cumulative hazard function $\Lambda(t)$ depict the accumulated amount of hazard up to time t as follows:

$$\Lambda(t) = \int_0^t \lambda(u) du.$$

Furthermore, the cumulative hazard function $\Lambda(t)$ and survival function $S_Y(t)$ are mathematically related as follows:

$$\Lambda(t) = -\log S_Y(t).$$

Hazard Function

Let us consider a **continuous** random variable

$$Y = \text{Time until an event occurs.}$$

Their survival and PDF functions are $S_Y(t)$ and $f_Y(t)$. That said, the hazard function $\lambda(t)$, the **instantaneous rate of event occurrence per unit of time**, is given by

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq Y < t + \Delta t | Y \geq t)}{\Delta t} = \frac{f_Y(t)}{S_Y(t)}.$$

One can interpret $\lambda(t)\Delta t$ as the approximate probability of the event occurring immediately, given that the event has not occurred up until time t.

Link Functions in Generalized Linear Models

(GLMs)

A GLM has the components of the conceptual regression model in a training set of n elements as:

- **Random component.** Each *response* $Y_1, \dots, Y_n$ is a random variable with its respective mean μ_i.
- **Systematic component.** How the k regressors come into the model denoted as a **linear combination**:

$$\eta_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_k X_{i,k} \quad \text{for } i = 1, \ldots, n.$$

- **Link function.** The element that connects the **random component** with the **systematic component** η_i. The connection is made through $h(\mu_i)$:

$$h(\mu_i) = \eta_i.$$

Note the following:

- The link function needs to be monotonic so we can allow putting the systematic component η_i in terms of the corresponding mean μ_i, i.e.:

$$\mu_i = h^{-1}(\eta_i).$$

- Furthermore, it needs to be differentiable **since we rely on maximum likelihood estimation** to obtain $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$.

Multinomial Logistic Regression

Data Modelling Framework

Let us suppose that a given **discrete nominal response** Y_i **has categories** $1, 2, \ldots, m$. Categories $1, 2, \ldots, m$ **are merely labels here**. Thus, they **do not** implicate an ordinal scale.

This regression approach assumes a **Multinomial distribution** where $p_{i,1}, p_{i,2}, \ldots, p_{i,m}$ are the probabilities that Y_i will belong to categories $1, 2, \ldots, m$ respectively; i.e.,

$$P(Y_i = 1) = p_{i,1} \quad P(Y_i = 2) = p_{i,2} \quad \dots \quad P(Y_i = m) = p_{i,m},$$

where

$$\sum_{j=1}^{m} p_{i,j} = p_{i,1} + p_{i,2} + \cdots + p_{i,m} = 1.$$

The Multinomial Logistic regression **also models the logarithm of the odds**. However, only one logarithm of the odds (or **logit**) will not be enough anymore. Recall we can capture the odds between two categories with a single logit function. **What about adding some other ones?**

Here is what we can do:

1. Pick one of the categories to be the **baseline**. For example, the category "1".
2. For each of the **other** categories, we model the logarithm of the odds to the baseline category.

What is the math for the general case with m response categories and k regressors? For the ith observation, we end up with a system of $m-1$ link functions in the Multinomial Logistic regression model as follows:

$$\begin{gathered}
\eta_i^{(2,1)} = \log\left[\frac{P(Y_i = 2 \mid X_{i,1}, \dots, X_{i,k})}{P(Y_i = 1 \mid X_{i,1}, \dots, X_{i,k})}\right] = \beta_0^{(2,1)} + \beta_1^{(2,1)} X_{i,1} + \beta_2^{(2,1)} X_{i,2} + . . \\
\eta_i^{(3,1)} = \log\left[\frac{P(Y_i = 3 \mid X_{i,1}, \dots, X_{i,k})}{P(Y_i = 1 \mid X_{i,1}, \dots, X_{i,k})}\right] = \beta_0^{(3,1)} + \beta_1^{(3,1)} X_{i,1} + \beta_2^{(3,1)} X_{i,2} + . . \\
\vdots \\
\eta_i^{(m,1)} = \log\left[\frac{P(Y_i = m \mid X_{i,1}, \dots, X_{i,k})}{P(Y_i = 1 \mid X_{i,1}, \dots, X_{i,k})}\right] = \beta_0^{(m,1)} + \beta_1^{(m,1)} X_{i,1} + \beta_2^{(m,1)} X_{i,2} + .
\end{gathered} \tag{30}$$

Note that the superscript $(j, 1)$ in (30) indicates that the equation is on level j (for $j = 2, \dots, m$) with respect to level 1. Furthermore, **the regression intercept and coefficients are different for each link function**.

With some algebraic manipulation, we can show that the probabilities $p_{i,1}, p_{i,2}, \dots, p_{i,m}$ of Y_i belonging to categories $1, 2, \dots, m$ are:

$$
\begin{aligned}
p_{i,1} = P(Y_i = 1 \mid X_{i,1}, \ldots, X_{i,k}) &= \frac{1}{1 + \sum_{j=2}^{m} \exp\left(\eta_i^{(j,1)}\right)} \\
p_{i,2} = P(Y_i = 2 \mid X_{i,1}, \ldots, X_{i,k}) &= \frac{\exp\left(\eta_i^{(2,1)}\right)}{1 + \sum_{j=2}^{m} \exp\left(\eta_i^{(j,1)}\right)} \\
&\vdots \\
p_{i,m} = P(Y_i = m \mid X_{i,1}, \ldots, X_{i,k}) &= \frac{\exp\left(\eta_i^{(m,1)}\right)}{1 + \sum_{j=2}^{m} \exp\left(\eta_i^{(j,1)}\right)}.
\end{aligned}
\tag{31}
$$

If we sum all m probabilities in (31), the sum will be equal to 1 for the ith observation. **This is particularly important when we want to use this model for making predictions in classification matters**. In a Multinomial Logistic regression model, each link function has its own intercept and regression coefficients.

Estimation

All parameters in the Multinomial Logistic regression model are also unknown. To fit the model with the package `nnet`, we use the function `multinom()`, which obtains the corresponding estimates. The estimates are obtained through **maximum likelihood**, where we assume a **Multinomial joint probability mass function** of the n responses Y_i. You can find more information on this matter **here**.

Inference

We can determine whether a regressor is statistically associated with the logarithm of the odds through hypothesis testing for the parameters $\beta_j^{(u,v)}$ **by link function**. We also use the **Wald statistic** $z_j^{(u,v)}$:

$$
z_j^{(u,v)} = \frac{\hat{\beta}_j^{(u,v)}}{\text{se}\left(\hat{\beta}_j^{(u,v)}\right)}
$$

to test the hypotheses

$$
\begin{gathered}
H_0: \beta_j^{(u,v)} = 0 \\
H_a: \beta_j^{(u,v)} \neq 0.
\end{gathered}
$$

Provided the sample size n is large enough, z_j has an **approximately Standard Normal distribution** under H_0.

R provides the corresponding p**-values** for each $\beta_j^{(u,v)}$. The smaller the p-value, the stronger the evidence against the null hypothesis H_0. As in the previous regression models, we would set a predetermined significance level α (usually taken to be 0.05) to infer if the p-value is small enough. If the p-value is smaller than the predetermined level α, then you could claim that there is evidence to reject the null hypothesis. Hence, p-values that are small enough indicate that the data provides evidence in favour of **association** (**or causation in the case of an experimental study!**) between the response variable and the jth regressor.

Furthermore, given a specified level of confidence where α is the significance level, we can construct approximate $(1-\alpha) \times 100\%$ **confidence intervals** for the corresponding true value of $\beta_j^{(u,v)}$:

$$
\hat{\beta}_j^{(u,v)} \pm z_{\alpha/2} \text{se}\left(\hat{\beta}_j^{(u,v)}\right),
$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the **Standard Normal distribution**.

Model Selection

To perform model selection, we can use the same techniques from the Binary Logistic regression model (check Model Selection).

Negative Binomial Regression

Let

$$
Y_i \sim \text{Negative Binomial}(m, p_i) \quad \text{for} \quad i = 1, \dots, n.
$$

From **DSCI 551**, recall that a Negative Binomial distribution has the following probability mass function (PMF):

$$P(Y_i = y_i \mid m, p_i) = \binom{m-1+y_i}{y_i} p_i^m (1-p_i)^{y_i} \quad \text{for} \quad y_i = 0, 1, \dots \tag{32}$$

A Negative Binomial random variable depicts **the number of y_i failed independent Bernoulli trials before experiencing m successes** with a probability of success p_i.

This distribution has the following mean and variance:

$$\begin{aligned} \mathbb{E}(Y_i) &= \frac{m(1-p_i)}{p_i} \\ \text{Var}(Y_i) &= \frac{m(1-p_i)}{p_i^2}. \end{aligned}$$

Reparametrization

Under the following parametrization:

$$\lambda_i = \frac{m(1-p_i)}{p_i} \qquad \Rightarrow \qquad p_i = \frac{m}{m+\lambda_i}, \tag{33}$$

the mean and variance of a Negative Binomial random variable can be reexpressed as

$$\begin{aligned} \mathbb{E}(Y_i) &= \lambda_i \\ \text{Var}(Y_i) &= \lambda_i \left(1 + \frac{\lambda_i}{m}\right). \end{aligned} \tag{34}$$

This reparametrized variance indicates that a Negative Binomial random variable allows for overdispersion through factor $\left(1 + \frac{\lambda_i}{m}\right)$.

Finally, by applying parametrization (33) in PMF (32), we have the following:

(35)

$$
\begin{aligned}
P(Y_i = y_i \mid m, p_i) &= \binom{m-1+y_i}{y_i} p_i^m (1-p_i)^{y_i} \\
&= \frac{(m+y_i-1)!}{y_i!(m-1)!} \left(\frac{m}{m+\lambda}\right)^m \left(1 - \frac{m}{m+\lambda}\right)^{y_i} \\
&= \frac{\Gamma(y_i+m)}{\Gamma(y_i+1)\Gamma(m)} \left(\frac{m}{m+\lambda}\right)^m \left(1 - \frac{m}{m+\lambda}\right)^{y_i},
\end{aligned}
$$

where $\Gamma(\cdot)$ is the **Gamma function**. We actually use the property

$$\Gamma(a) = (a-1)!,$$

where $a \geq 1$ is an integer.

Data Modelling Framework

As in the case of Poisson regression with k regressors, the Negative Binomial case is a GLM with the following link function:

$$h(\lambda_i) = \log(\lambda_i) = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k}. \tag{36}$$

Lastly, note the following:

- From (34), let $\theta = \frac{1}{m}$. Then, Negative Binomial regression will assume the following variance:

$$
\begin{aligned}
\text{Var}(Y_i) &= \lambda_i \left(1 + \frac{\lambda_i}{m}\right) \\
&= \lambda_i + \frac{\lambda_i^2}{m} \\
&= \lambda_i + \theta\lambda_i^2.
\end{aligned}
$$

Therefore, the model has even more flexibility to deal with overdispersion compared to Poisson regression.

Estimation

Via a training set of size n whose responses are **independent counts** Y_i $(i = 1, \dots, n)$, we use the reparametrized PMF (35) along with the link function (36) via **maximum likelihood estimation** to obtain $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k, \hat{\theta}$.

To fit a Negative Binomial regression via `R`, we can use the function `glm.nb()` from package `MASS`.

Inference, Coefficient Interpretation, Prediction, Goodness of Fit, and Model Selection

Since the link function in Negative Binomial regression is the same as in Poisson regression; **inference, coefficient interpretation, and prediction are performed similarly** (even with the same `broom` functions!). Regarding **model selection**, since we use a regular maximum likelihood approach to estimate the regression parameters, we can use analysis of deviance, AIC, and BIC to perform model selection and/or goodness of fit testing.

Ordinal Logistic Regression

Data Modelling Framework

Let us suppose that a given **discrete ordinal response** Y_i (for $i = 1, \dots, n$) has categories $1, 2, \dots, m$ in a training set of size n.

Categories $1, 2, \dots, m$ **implicate an ordinal scale here**, i.e., $1 < 2 < \dots < m$.

Also, note there is more than one class of Ordinal Logistic regression. We will review the **proportional odds** model (a **cumulative logit model**).

We have to point out that Ordinal Logistic regression will indicate how each one of the k regressors $X_{i,1}, \dots, X_{i,k}$ affects the **cumulative logarithm of the odds** in the ordinal response for the following $m - 1$ situations:

$$\begin{gathered}\text{Level } m-1 \text{ or any lesser degree versus level } m\\ \text{Level } m-2 \text{ or any lesser degree versus level } m-1 \text{ or any higher degree}\\ \vdots\\ \text{Level 2 or any lesser degree versus level 3 or any higher degree}\\ \text{Level 1 versus level 2 or any higher degree}\end{gathered}$$

These $m-1$ situations are translated into cumulative probabilities using the logarithms of the odds on the left-hand side ($m-1$ link functions) subject to the linear combination of the k regressors $X_{i,j}$ (for $j=1,\ldots,k$):

$$\begin{aligned}\eta_i^{(m-1)} &= \log\left[\frac{P(Y_i \leq m-1 \mid X_{i,1},\ldots,X_{i,k})}{P(Y_i = m \mid X_{i,1},\ldots,X_{i,k})}\right] = \beta_0^{(m-1)} - \beta_1 X_{i,1} - \beta_2 X_{i,2} - \ldots -\\ \eta_i^{(m-2)} &= \log\left[\frac{P(Y_i \leq m-2 \mid X_{i,1},\ldots,X_{i,k})}{P(Y_i > m-2 \mid X_{i,1},\ldots,X_{i,k})}\right] = \beta_0^{(m-2)} - \beta_1 X_{i,1} - \beta_2 X_{i,2} - \ldots -\\ &\vdots\\ \eta_i^{(2)} &= \log\left[\frac{P(Y_i \leq 2 \mid X_{i,1},\ldots,X_{i,k})}{P(Y_i > 2 \mid X_{i,1},\ldots,X_{i,k})}\right] = \beta_0^{(2)} - \beta_1 X_{i,1} - \beta_2 X_{i,2} - \ldots - \beta_k X_{i,k}\\ \eta_i^{(1)} &= \log\left[\frac{P(Y_i = 1 \mid X_{i,1},\ldots,X_{i,k})}{P(Y_i > 1 \mid X_{i,1},\ldots,X_{i,k})}\right] = \beta_0^{(1)} - \beta_1 X_{i,1} - \beta_2 X_{i,2} - \ldots - \beta_k X_{i,k}\end{aligned}$$

Note that the system above has $m-1$ intercepts but **only k regression coefficients**. In general, the previous $m-1$ equations can be generalized for levels $j=m-1,\ldots,1$ as follows:

$$\begin{gathered}\eta_i^{(j)} = \log\left[\frac{P(Y_i \leq j \mid X_{i,1},\ldots,X_{i,k})}{P(Y_i > j \mid X_{i,1},\ldots,X_{i,k})}\right] = \beta_0^{(j)} - \beta_1 X_{i,1} - \beta_2 X_{i,2} - \ldots - \beta_k X_{i,k}\\ \Rightarrow \; P(Y_i \leq j \mid X_{i,1},\ldots,X_{i,k}) = \frac{\exp\left(\beta_0^{(j)} - \beta_1 X_{i,1} - \beta_2 X_{i,2} - \ldots - \beta_k X_{i,k}\right)}{1+\exp\left(\beta_0^{(j)} - \beta_1 X_{i,1} - \beta_2 X_{i,2} - \ldots - \beta_k X_{i,k}\right.}\end{gathered}$$

The probability that Y_i will fall in the category j can be computed as follows:

$$p_{i,j} = P(Y_i = j \mid X_{i,1},\ldots,X_{i,k}) = P(Y_i \leq j \mid X_{i,1},\ldots,X_{i,k}) - P(Y_i \leq j-1 \mid X_{i,1},\ldots$$

which leads to

$$P(Y_i = 1) = p_{i,1} \quad P(Y_i = 2) = p_{i,2} \quad \dots \quad P(Y_i = m) = p_{i,m}$$

where

$$\sum_{j=1}^{m} p_{i,j} = p_{i,1} + p_{i,2} + \cdots + p_{i,m} = 1.$$

Estimation

All parameters in the Ordinal Logistic regression model are also unknown. Therefore, model estimates are obtained through **maximum likelihood**, where we also assume a **Multinomial joint probability mass function** of the n responses Y_i. Moreover, this Multinomial assumption plays around with cumulative probabilities in the joint likelihood function **as discussed in Agresti (2013) in Chapter 8 (Section 8.2.2)**.

To fit the model with the package `MASS`, we use the function `polr()`, which obtains the corresponding estimates. The argument `Hess = TRUE` is required to compute the Hessian matrix of the **log-likelihood function**, which is used to obtain the standard errors of the estimates.

Inference

We can determine whether a regressor is statistically associated with the logarithm of the cumulative odds through hypothesis testing for the parameters β_j. We also use the **Wald statistic** z_j:

$$z_j = \frac{\hat{\beta}_j}{\mathrm{se}(\hat{\beta}_j)}$$

to test the hypotheses

$$\begin{gathered} H_0 : \beta_j = 0 \\ H_a : \beta_j \neq 0. \end{gathered}$$

The **null hypothesis** H_0 indicates that the jth regressor associated to β_j does not affect the response variable in the model, and the **alternative hypothesis** H_a otherwise. Moreover, provided the sample size n is large enough, z_j has an **approximately Standard Normal distribution** under H_0.

R provides the corresponding p**-values** for each β_j. The smaller the p-value, the stronger the evidence against the null hypothesis H_0. As in the previous regression models, we would set a predetermined significance level α (usually taken to be 0.05) to infer if the p-value is small enough. If the p-value is smaller than the predetermined level α, then you could claim that there is evidence to reject the null hypothesis. Hence, p-values that are small enough indicate that the data provides evidence in favour of **association** (**or causation in the case of an experimental study!**) between the response variable and the jth regressor.

Furthermore, given a specified level of confidence where α is the significance level, we can construct approximate $(1-\alpha) \times 100\%$ **confidence intervals** for the corresponding true value of β_j:

$$\hat{\beta}_j \pm z_{\alpha/2} \text{se}(\hat{\beta}_j),$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the **Standard Normal distribution**.

Model Selection

To perform model selection, we can use the same techniques from the Binary Logistic regression model (check Model Selection).

Non-proportional Odds Model

We might wonder

> What happens if we do not fulfil the proportional odds assumption in our Ordinal Logistic regression model?

The Brant-Wald Test

It is essential to remember that the Ordinal Logistic model under the proportional odds assumption is **the first step** when performing Regression Analysis on an ordinal response.

Once this model has been fitted, **it is possible to assess whether it fulfils this strong assumption statistically**.

We can do it via the Brant-Wald test:

- This tool statistically assesses whether our model globally fulfils this assumption. It has the following hypotheses:

$$H_0 : \text{Our Ordinal Logistic regression model globally fulfils the proportional odds as}$$
$$H_a : \text{Otherwise.}$$

- Moreover, it also performs further hypothesis testing on each regressor. With k regressors for $j = 1, \dots, k$; we have the following hypotheses:

$$H_0 : \text{The } j\text{th regressor in our Ordinal Logistic regression model fulfils the proportio}$$
$$H_a : \text{Otherwise.}$$

Function `brant()` from package `brant` implements this tool, which can be used in our `polr()` object.

The Brant Wald test essentially compares this basic Ordinal Logistic regression model of $k - 1$ cumulative logit functions versus a combination of $k - 1$ correlated Binary Logistic regressions.

Ordinary Least-squares Regression

Data Modelling Framework

Conceptually, the OLS regression model can be expressed as:

$$\text{Response} = \text{Systematic Component} + \text{Random Component}. \tag{37}$$

For the ith observation in our **random sample** or **training data** ($i = 1, \dots, n$), the conceptual model (37) is mathematically represented as:

$$\underbrace{Y_i}_{\text{Response}} = \underbrace{\beta_0 + \beta_1 g_1(X_{i,1}) + \ldots + \beta_k g_k(X_{i,k})}_{\text{Systematic Component}} + \underbrace{\varepsilon_i.}_{\text{Random Component}} \tag{38}$$

Note the following:

- The response Y_i is equal to the sum of $k + 2$ terms on the right-hand side.
- The systematic component is the sum of:
  - An **unknown intercept** β_0 and
  - k **regressor functions** $g_j(X_{i,j})$ multiplied by their respective **unknown regression coefficient** β_j ($j = 1, \dots, k$).
- ε_i is the **random noise**.

Modelling Assumptions

For the **random component** in (38), we assume the following:

$$\begin{gathered}\mathbb{E}(\varepsilon_i) = 0 \\ \text{Var}(\varepsilon_i) = \sigma^2 \\ \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \\ \varepsilon_i \perp\!\!\!\perp \varepsilon_k \quad \text{for } i \neq k \quad \text{(independence)}.\end{gathered}$$

Hence, **each Y_i is also assumed to be independent and normally distributed**:

$$Y_i \mid X_{i,j} \sim \mathcal{N}\big(\beta_0 + \beta_1 g_1(X_{i,1}) + \ldots + \beta_k g_k(X_{i,k}), \sigma^2\big).$$

Definition of Linearity

The classical OLS model, from **DSCI 561**, implicates the identity function $g_j(X_{i,j}) = X_{i,j}$ in equation (38). This leads to:

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 g_1(X_{i,1}) + \ldots + \beta_k g_k(X_{i,k}) + \varepsilon_i \\
&= \beta_0 + \beta_1 X_{i,1} + \ldots + \beta_k X_{i,k} + \varepsilon_i.
\end{aligned}
$$

Note the model is “linear” on the parameters (i.e., regression terms), not the regressors.

Inference

In terms of inference, we use the fitted model to identify the relationship between the response and regressors. We will need the jth estimated regression coefficient $\hat{\beta}_j$ and its corresponding variability which is reflected in the **standard error** of the estimate, $\text{se}\left(\hat{\beta}_j\right)$. To determine the statistical significance of $\hat{\beta}_j$, we use the **test statistic**

$$
t_j = \frac{\hat{\beta}_j}{\text{se}\left(\hat{\beta}_j\right)}
$$

to test the hypotheses

$$
\begin{gathered}
H_0 : \beta_j = 0 \\
H_a : \beta_j \neq 0.
\end{gathered}
$$

A statistic like t_j is referred to as a t-value. It has a t-distribution **under the null hypothesis** H_0 with $n - k - 1$ degrees of freedom.

We can obtain the corresponding p-values for each β_j associated to the t-values under the null hypothesis H_0. **The smaller the p-value, the stronger the evidence against the null hypothesis H_0 in our sample**. Hence, small p-values (less than the significance level α) indicate that the data provides evidence in favour of association (or **causation** if that is the case) between the response variable and the jth regressor.

Similarly, given a specified $(1 - \alpha) \times 100\%$ level of confidence, we can construct **confidence intervals** for the corresponding true value of β_j:

$$
\hat{\beta}_j \pm t_{\alpha/2, n-k-1} \text{se}\left(\hat{\beta}_j\right),
$$

where $t_{\alpha/2,n-k-1}$ is the upper $\alpha/2$ quantile of the t-distribution with $n-k-1$ degrees of freedom.

Poisson Regression

Data Modelling Framework

Besides OLS and Binary Logistic regressions, another alternative is count data modelling, as in **Poisson regression**. Unlike Binary Logistic regression, **we use counts as a response variable**. Hence, we have to modify the modelling framework to consider this fact. Poisson regression would be the primary resource when it comes to modelling counts. Note this model also fits into the GLM class.

> What is the distributional key difference between the Poisson and the OLS regression models in terms of the response?

First of all, we have to specify what a Poisson random variable is. Recall **DSCI 551**, a Poisson random variable refers to discrete data with non-negative integer values that count something. **These counts could happen during a given timeframe or even a space such as a geographic unit!**

The Poisson regression model assumes a random sample of n count observations Y_is, hence **independent** (**but not identically distributed!**), which have the following distribution:

$$Y_i \sim \text{Poisson}(\lambda_i).$$

Each ith observation has its own $\mathbb{E}(Y_i) = \lambda_i > 0$, which also implicates $\text{Var}(Y_i) = \lambda_i > 0$.

The equality of the expected value and variance in a random variable is called **equidispersion**. Parameter λ_i is the risk of an event occurrence, coming from the definition of the Poisson random variable, **in a given timeframe or even a space**. Furthermore, we have to highlight another particularity in the Poisson distribution: λ_i **is a continuous distributional parameter!**

Let us generalize the Poisson regression model with k regressors as:

$$h(\lambda_i) = \log(\lambda_i) = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k}. \tag{39}$$

In model (39), each one of the k regression coefficients $\beta_1, \dots, \beta_k$ represents **the expected change in the natural logarithm of the mean λ_i per unit change in their respective regressors $X_{i,1}, \dots, X_{i,k}$**. Nonetheless, we could make more sense in the interpretation by exponentiating (39):

$$\lambda_i = \exp\left(\beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k}\right),$$

where an increase in one unit in any of the k regressors (**while keeping the rest of them constant**) **multiplies the mean λ_i by a factor $\exp\left(\beta_j\right)$, for all $j = 1, \dots, k$.**

As a side note, we have to clarify that the **systematic component** in the Poisson regression model is explicitly depicted by the regressors and their coefficients as in multiple linear regression. The **random component** is implicitly contained in each random variable

$$Y_i \sim \text{Poisson}(\lambda_i).$$

Estimation

Under a general framework with k regressors, the **regression parameters** $\beta_0, \beta_1, \dots, \beta_k$ in the model are also unknown. In order to estimate them, we will use function `glm()` and its argument `family = poisson` (required to specify the Poisson nature of the response), which obtains the estimates $\hat{\beta}_0, \hat{\beta}_1, \dots \hat{\beta}_k$.

The estimates are obtained through **maximum likelihood** where we assume a **Poisson joint probability mass function** of the n responses Y_i.

For the sake of coding clarity, you could also use `family = poisson(link = "log")`. Nevertheless, `link = "log"` is a default in `glm()` for Poisson regression. Thus, `family = poisson` suffices when using the logarithmic link function.

Inference

The fitted regression model will be used to identify the relationship between the logarithm of the response's mean and regressors. To determine the **statistical**

significance of β_j in this model, we also use the **Wald statistic**:

$$z_j = \frac{\hat{\beta}_j}{\mathrm{se}\left(\hat{\beta}_j\right)}$$

to test the hypotheses

$$\begin{gathered} H_0 : \beta_j = 0 \\ H_a : \beta_j \neq 0; \end{gathered}$$

where the **null hypothesis** H_0 indicates that the jth regressor corresponding to β_j does not have any association on the response variable in the model, and the **alternative hypothesis** H_a otherwise. Provided the sample size n is large enough, z_j has an **approximately Standard Normal distribution** under H_0.

R provides the corresponding p**-values** for each β_j. The smaller the p-value, the stronger the evidence against the null hypothesis H_0. As in the previous regression models, we would set a predetermined significance level α to infer if the p-value is small enough. If the p-value is smaller than the predetermined level α, then we could claim that there is evidence to reject the null hypothesis. Hence, p-values that are small enough indicate that the data provides evidence in favour of **association** (**or causation in the case of an experimental study!**) between the response variable and the jth regressor.

Furthermore, given a specified level of confidence where α is the significance level, we can construct approximate $(1 - \alpha) \times 100\%$ **confidence intervals** for the corresponding true value of β_j:

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

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the **Standard Normal distribution**.

Model Selection

Goodness of Fit Test

The **deviance** (D_k) criterion can be used to compare a given model with k regressors with that of a **baseline model**. The usual baseline model is the **saturated** or **full model**, which perfectly fits the data because it allows a distinct Poisson mean λ_i for the ith observation in the training dataset ($i = 1, \dots, n$), **unrelated to the k regressors**.

The **maximized likelihood** of this full model is denoted as $\hat{\ell}_f$. Now, let $\hat{\ell}_k$ be the value of the maximized likelihood computed from our dataset of n observation with k regressors.

We can compare the fits provided by these two models by the deviance D_k given by

$$D_k = -2 \log \left(\frac{\hat{\ell}_k}{\hat{\ell}_f} \right) = -2 \left[\log \left(\hat{\ell}_k \right) - \log \left(\hat{\ell}_f \right) \right]. \tag{40}$$

Note that D_k expresses **how much our given model deviates from the full model on log-likelihood scale**. This metric is interpreted as follows:

- **Large values** of D_k arise when $\hat{\ell}_k$ is small relative to $\hat{\ell}_f$, indicating that **our given model fits the data poorly compared to the baseline model**.
- **Small values** of D_k arise when $\hat{\ell}_k$ is similar to $\hat{\ell}_f$, indicating that **our given model provides a good fit to the data compared to the baseline model**.

Specifically for Poisson regression with k regressors, it can be shown that D_k (40) is defined as follows:

$$\begin{gathered} \hat{\lambda}_i = \exp \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \dots + \hat{\beta}_k x_{i,k} \right) \\ D_k = 2 \sum_{i=1}^{n} \left[y_i \log \left(\frac{y_i}{\hat{\lambda}_i} \right) - \left(y_i - \hat{\lambda}_i \right) \right] \end{gathered} \tag{41}$$

where y_i is the ith observed response in the training set of size n. Note that when $y_i = 0$ counts, then $\log \left(\frac{y_i}{\hat{\lambda}_i} \right)$ is assumed as 0.

Equation (41) depicts the agreement of our model with k regressors to the observed data. Hence, we can use (41) to test the goodness of fit; i.e., **whether our fitted model fits the data better than the saturated model, which makes it correctly specified (with a level of significance α!)**.

The hypothesis are the following:

$$\begin{gather*} H_0 : \text{Our } \textbf{Model with } k \textbf{ regressors} \text{ fits the data better than the } \textbf{Saturated Mod} \\ H_a : \text{otherwise.} \end{gather*}$$

We use the **residual deviance Δ_k** as a **test statistic**. **Asymptotically**, we have the following **null distribution**:

$$\Delta_k \sim \chi^2_{n-(k+1)}.$$

Analysis of Deviance for Nested Models

Suppose **Model 1** is nested in **Model 2**. Hence, this specific model selection will involve a hypothesis testing. The hypotheses are:

$$\begin{gather*} H_0 : \textbf{Model 1} \text{ fits the data better than } \textbf{Model 2} \\ H_a : \textbf{Model 2} \text{ fits the data better than } \textbf{Model 1}. \end{gather*}$$

Let D_2 be the deviance for **Model 2** and D_1 the deviance for **Model 1**. The test statistic Δ_D for the analysis of deviance is given by:

$$\Delta_D = D_1 - D_2 \sim \chi^2_d,$$

which **assymptotically** (i.e., $n \to \infty$) is **Chi-squared distributed** with d degrees of freedom under H_0. The d degrees of freedom are the **regression parameters of difference between both models**.

Formally, this nested hypothesis testing is called the **likelihood-ratio test**.

Akaike Information Criterion

One of the drawbacks of the analysis of deviance is that it only allows to test **nested** regression models when we have sparse data (i.e., each response is associated with a different set of values in the regressors).

Fortunately, we have alternatives for model selection. **The Akaike Information Criterion (AIC) makes it possible to compare models that are either nested or not.** For a model with k regressors and a deviance D_k is defined as:

$$\text{AIC}_k = D_k + 2k.$$

Models with **smaller** values of AIC_k are preferred. That said, AIC_k favours models with small values of D_k.

However, AIC_k penalizes for including more regressors in the model. Hence, it discourages overfitting, which is key in model selection. This is why we select that model with the smallest AIC_k.

Bayesian Information Criterion

An alternative to AIC is the Bayesian Information Criterion (BIC). **The BIC also makes it possible to compare models that are either nested or not.** For a model with k regressors, n observations used for training, and a deviance D_k; it is defined as:

$$\text{BIC}_k = D_k + k \log(n).$$

Models with **smaller** values of BIC_k are preferred. That said, BIC_k also favours models with small values of D_k.

The differences between AIC and BIC will be more pronounced in datasets with large sample sizes n. As the BIC penalty of $k \log(n)$ will always be larger than the AIC penalty of $2k$ when $n > 7$, **BIC tends to select models with fewer regressors than AIC**.

Survival Function

For this function, let us consider a **continuous** random variable

$$Y = \text{Time until an event occurs.}$$

We know that the cumulative distribution function (CDF) of Y tells us the probability that the event of interest occurs **before a certain point in time** t:

$$F_Y(t) = P(Y \leq t).$$

However, as the name suggests, in Survival Analysis, we are more interested in the probability that the event **WILL NOT** occur before a certain point in time t (i.e., **the subject survives at least at point** t):

$$S_Y(t) = P(Y > t) = 1 - F_Y(t).$$

$S_Y(t)$ is called the **survival function**.