

1 Binary Logistic Regression

1.1 Data Modelling Framework

Models a binary response variable Y_i .

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

The key parameter is the probability of success, $p_i = P(Y_i = 1)$. The response is assumed to follow a Bernoulli distribution, $Y_i \sim \text{Bernoulli}(p_i)$.

1.2 Link Function (Logit)

The logit function links the probability p_i to a linear combination of predictors. It transforms a probability $[0, 1]$ to an unrestricted scale $(-\infty, \infty)$.

$$h(p_i) = \text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_{i,1} + \dots + \beta_k X_{i,k}$$

This is the log-odds of success. The inverse is:

$$p_i = \frac{\exp(\text{logit}(p_i))}{1 + \exp(\text{logit}(p_i))}$$

1.3 Estimation

Parameters β_0, \dots, β_k are estimated using maximum likelihood estimation (MLE). In R, use `glm()` with `family = binomial`.

1.4 Inference

To test the significance of a coefficient β_j , we use the Wald statistic:

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

For large n , z_j follows a Standard Normal distribution under $H_0 : \beta_j = 0$. Confidence intervals are calculated as:

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

1.5 Model Selection

Deviance (D_k): Compares the fit of a model with k predictors to a saturated (full) model which perfectly fits the data. Smaller is better.

$$D_k = -2[\log(L_k) - \log(L_f)]$$

Likelihood-ratio test: For nested models, compare deviances: $\Delta D = D_1 - D_2 \sim \chi^2_d$, where d is the difference in the number of parameters. **AIC/BIC:** For non-nested models.

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

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

Lower AIC/BIC indicates a better model. BIC penalizes complexity more heavily.

2 Cox Proportional Hazards Model

2.1 Data Modelling Framework

A semi-parametric survival model. It models the hazard function $\lambda_i(t)$ directly for a censored response.

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

- $\lambda_0(t)$: baseline hazard function (non-parametric part).
- $\exp(\dots)$: parametric part.
- No intercept β_0 .

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. The ratio of hazards is constant over time.

$$\frac{\lambda_2(t)}{\lambda_1(t)} = \exp(\beta_1(X_{2,1} - X_{1,1}))$$

2.2 Estimation

Parameters are estimated using partial likelihood, which does not require specifying the baseline hazard $\lambda_0(t)$.

2.3 Inference And Prediction

Similar to logistic regression, Wald statistics are used to test coefficient significance ($H_0 : \beta_j = 0$).

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

Prediction involves estimating the survival function:

$$S(t|\mathbf{X}) = S_0(t)^{\exp(\sum \beta_j X_j)}$$

where $S_0(t) = \exp(-\int_0^t \lambda_0(u)du)$ is the baseline survival function.

3 Multinomial Logistic Regression

3.1 Data Modelling Framework

For a nominal response Y_i with m categories. It models the log-odds of a category relative to a baseline category (e.g., category "1"). This results in $m - 1$ link functions:

$$\log\left(\frac{P(Y_i = j)}{P(Y_i = 1)}\right) = \beta_0^{(j,1)} + \sum_{k=1}^K \beta_k^{(j,1)} X_{i,k}$$

for $j = 2, \dots, m$. Each equation has its own set of coefficients.

3.2 Estimation And Inference

Parameters are estimated via MLE. In R, use `multinom()` from the `nnet` package. Inference is done using Wald statistics for each coefficient.

4 Poisson Regression

4.1 Data Modelling Framework

For count data response variables, $Y_i \sim \text{Poisson}(\lambda_i)$. A key assumption is that the mean equals the variance: $E[Y_i] = \text{Var}(Y_i) = \lambda_i$. The link function is the log:

$$\log(\lambda_i) = \beta_0 + \sum_{j=1}^k \beta_j X_{i,j}$$

Coefficients β_j represent the change in the log of the mean count for a one-unit change in X_j .

5 Negative Binomial Regression

5.1 Data Modelling Framework

An extension of Poisson regression for overdispersed count data (when variance > mean). It introduces a dispersion parameter θ such that:

$$\text{Var}(Y_i) = \lambda_i + \theta \lambda_i^2$$

As $\theta \rightarrow 0$, it converges to Poisson regression. The link function is also the log. In R, use `glm.nb()` from the `MASS` package.

6 Ordinal Logistic Regression

6.1 Data Modelling Framework

For an ordinal response Y_i with m ordered categories. This models the cumulative log-odds. The Proportional Odds Model assumes that the effect of predictors is constant across all cumulative splits.

$$\text{logit}(P(Y_i \leq j)) = \beta_j^{(0)} - \sum_{k=1}^K \beta_k X_{i,k}$$

for $j = 1, \dots, m - 1$. Note there is a different intercept $\beta_j^{(0)}$ for each split, but the regression coefficients β_k are the same.

6.2 Brant-Wald Test

Used to check the proportional odds assumption. If the test is significant, it suggests the assumption is violated, and a more complex model (like non-proportional odds model) might be needed.

7 OLS Regression

7.1 Modelling Assumptions

For response Y_i and random error ϵ_i :

- Linearity: $Y_i = \beta_0 + \sum \beta_j X_{ij} + \epsilon_i$
- Zero Mean Error: $E[\epsilon_i] = 0$
- Constant Variance (Homoscedasticity): $\text{Var}(\epsilon_i) = \sigma^2$
- Normality: $\epsilon_i \sim N(0, \sigma^2)$
- Independence: Errors are uncorrelated.

7.2 Inference

To test $H_0 : \beta_j = 0$, the t-statistic is used:

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

which follows a t-distribution with $n - k - 1$ degrees of freedom.