

1 Binary Logistic Regression
1.1 Data Modelling Framework

Models a binary response variable Yi.

Yi = { 1 if the i^th observation is a success
0 otherwise.

The key parameter is the probability of success, pi = P(Yi = 1). The response is assumed to follow a Bernoulli distribution, Yi ~ Bernoulli(pi).

1.2 Link Function (Logit)

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

h(pi) = logit(pi) = log(p\_i / (1 - p\_i))

logit(pi) = beta\_0 + beta\_1 Xi,1 + ... + beta\_k Xi,k

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

pi = exp(logit(pi)) / (1 + exp(logit(pi)))

1.3 Estimation

Parameters beta\_0, ..., beta\_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 beta\_j, we use the Wald statistic:

zj = beta\_hat\_j / se(beta\_hat\_j)

For large n, zj follows a Standard Normal distribution under H0 : beta\_j = 0. Confidence intervals are calculated as:

beta\_hat\_j +/- z\_alpha/2 se(beta\_hat\_j)

1.5 Model Selection

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

Dk = -2[log(Lk) - log(Lf)]

Likelihood-ratio test: For nested models, compare deviances: Delta D = D1 - D2 ~ chi^2\_d, where d is the difference in the number of parameters. AIC/BIC: For non-nested models.

AIC\_k = D\_k + 2k

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|Xi) = lambda\_0(t) exp(sum\_{j=1}^k beta\_j Xi,j)

- lambda\_0(t): baseline hazard function (non-parametric part).
exp(...): parametric part.
No intercept beta\_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.

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 (H0 : beta\_j = 0).

zj = beta\_hat\_j / se(beta\_hat\_j)

Prediction involves estimating the survival function:

S(t|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 Yi 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(P(Yi = j) / P(Yi = 1)) = beta\_0^{(j,1)} + sum\_{k=1}^K beta\_k^{(j,1)} Xi,k

for j = 2, ..., 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, Yi ~ Poisson(lambda\_i). A key assumption is that the mean equals the variance: E[Yi] = Var(Yi) = lambda\_i. The link function is the log:

log(lambda\_i) = beta\_0 + sum\_{j=1}^k beta\_j Xi,j

Coefficients beta\_j represent the change in the log of the mean count for a one-unit change in Xj.

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 theta such that:

Var(Yi) = lambda\_i + theta lambda\_i^2

As theta -> 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 Yi 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.

logit(P(Yi <= j)) = beta\_j^{(0)} - sum\_{k=1}^K beta\_k Xi,k

for j = 1, ..., m - 1. Note there is a different intercept beta\_j^{(0)} for each split, but the regression coefficients beta\_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 Yi and random error epsilon\_i:

- Linearity: Yi = beta\_0 + sum beta\_j Xi,j + epsilon\_i
Zero Mean Error: E[epsilon\_i] = 0
Constant Variance (Homoscedasticity): Var(epsilon\_i) = sigma^2
Normality: epsilon\_i ~ N(0, sigma^2)
Independence: Errors are uncorrelated.

7.2 Inference

To test H0 : beta\_j = 0, the t-statistic is used:

tj = beta\_hat\_j / se(beta\_hat\_j)

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