

Statistical Analysis of Corpus Data with R

A short introduction to regression and linear models

Designed by Marco Baroni¹ and Stefan Evert²

¹Center for Mind/Brain Sciences (CIMeC)
University of Trento

²Institute of Cognitive Science (IKW)
University of Osnabrück

Linear regression

- Can random variable Y be predicted from r. v. X ?
 ☞ focus on linear relationship between variables

- Linear predictor:

$$Y \approx \beta_0 + \beta_1 \cdot X$$

- β_0 = intercept of regression line
- β_1 = slope of regression line

- Least-squares regression minimizes prediction error

$$Q = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2$$

for data points $(x_1, y_1), \dots, (x_n, y_n)$

Outline

- Regression
 - Simple linear regression
 - General linear regression
- Linear statistical models
 - A statistical model of linear regression
 - Statistical inference
- Generalised linear models

Simple linear regression

- Coefficients of least-squares line

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x}_n \bar{y}_n}{\sum_{i=1}^n x_i^2 - n \bar{x}_n^2}$$

$$\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n$$

- Mathematical derivation of regression coefficients
 - minimum of $Q(\beta_0, \beta_1)$ satisfies $\partial Q / \partial \beta_0 = \partial Q / \partial \beta_1 = 0$
 - leads to normal equations (system of 2 linear equations)

$$-2 \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)] = 0 \rightarrow \beta_0 n + \beta_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i$$

$$-2 \sum_{i=1}^n x_i [y_i - (\beta_0 + \beta_1 x_i)] = 0 \rightarrow \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$$

- regression coefficients = unique solution $\hat{\beta}_0, \hat{\beta}_1$

The Pearson correlation coefficient

- Measuring the “goodness of fit” of the linear prediction
 - variation among observed values of Y = sum of squares S_y^2
 - closely related to (sample estimate for) variance of Y

$$S_y^2 = \sum_{i=1}^n (y_i - \bar{y}_n)^2$$

- residual variation wrt. linear prediction: $S_{\text{resid}}^2 = Q$

- Pearson correlation = amount of variation “explained” by X

$$R^2 = 1 - \frac{S_{\text{resid}}^2}{S_y^2} = 1 - \frac{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_n)^2}$$

- correlation vs. slope of regression line

$$R^2 = \hat{\beta}_1(y \sim x) \cdot \hat{\beta}_1(x \sim y)$$

Multiple linear regression

- Linear regression with multiple predictor variables

$$Y \approx \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$$

minimises

$$Q = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})]^2$$

for data points $(x_{11}, \dots, x_{1k}, y_1), \dots, (x_{n1}, \dots, x_{nk}, y_n)$

- Multiple linear regression fits n -dimensional hyperplane instead of regression line

Multiple linear regression: The design matrix

- Matrix notation of linear regression problem

$$\mathbf{y} \approx \mathbf{Z}\beta$$

- “Design matrix” \mathbf{Z} of the regression data

$$\mathbf{Z} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$$

$$\mathbf{y} = [y_1 \ y_2 \ \dots \ y_n]'$$

$$\beta = [\beta_0 \ \beta_1 \ \beta_2 \ \dots \ \beta_k]'$$

- \mathbf{A}' denotes transpose of a matrix; \mathbf{y}, β are column vectors

General linear regression

- Matrix notation of linear regression problem

$$\mathbf{y} \approx \mathbf{Z}\beta$$

- Residual error

$$Q = (\mathbf{y} - \mathbf{Z}\beta)'(\mathbf{y} - \mathbf{Z}\beta)$$

- System of normal equations satisfying $\nabla_{\beta} Q = 0$:

$$\mathbf{Z}'\mathbf{Z}\beta = \mathbf{Z}'\mathbf{y}$$

- Leads to regression coefficients

$$\hat{\beta} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y}$$

General linear regression

- Predictor variables can also be functions of the observed variables \rightarrow regression only has to be linear in coefficients β
- E.g. polynomial regression with design matrix

$$\mathbf{Z} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^k \\ 1 & x_2 & x_2^2 & \cdots & x_2^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^k \end{bmatrix}$$

corresponding to regression model

$$Y \approx \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_k X^k$$

Linear statistical models

- Linear statistical model ($\epsilon =$ random error)

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$

$$\epsilon \sim N(0, \sigma^2)$$

- \Leftrightarrow x_1, \dots, x_k are not treated as random variables!
 $\rightarrow \sim$ "is distributed as"; $N(\mu, \sigma^2) =$ normal distribution

- Mathematical notation:

$$Y | x_1, \dots, x_k \sim N(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \sigma^2)$$

- Assumptions

- error terms ϵ_j are i.i.d. (independent, same distribution)
- error terms follow normal (Gaussian) distributions
- equal (but unknown) variance $\sigma^2 =$ homoscedasticity

Linear statistical models

- Probability density function for simple linear model

$$\Pr(\mathbf{y} | \mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \cdot \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2\right]$$

- $\mathbf{y} = (y_1, \dots, y_n)$ = observed values of Y (sample size n)
- $\mathbf{x} = (x_1, \dots, x_n)$ = observed values of X
- Log-likelihood has a familiar form:

$$\log \Pr(\mathbf{y} | \mathbf{x}) = C - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \propto Q$$

- \rightarrow MLE parameter estimates $\hat{\beta}_0, \hat{\beta}_1$ from linear regression

Statistical inference for linear models

- Model comparison with ANOVA techniques

- Is variance reduced significantly by taking a specific explanatory factor into account?
- intuitive: proportion of variance explained (like R^2)
- mathematical: F statistic $\rightarrow p$ -value

- Parameter estimates $\hat{\beta}_0, \hat{\beta}_1, \dots$ are random variables

- t -tests ($H_0: \beta_j = 0$) and confidence intervals for β_j
- confidence intervals for new predictions

- Categorical factors: dummy-coding with binary variables

- e.g. factor x with levels *low*, *med*, *high* is represented by three binary dummy variables $x_{low}, x_{med}, x_{high}$
- one parameter for each factor level: $\beta_{low}, \beta_{med}, \beta_{high}$
- NB: β_{low} is "absorbed" into intercept β_0
- model parameters are usually $\beta_{med} - \beta_{low}$ and $\beta_{high} - \beta_{low}$
- \Leftrightarrow mathematical basis for standard ANOVA

Interaction terms

- Standard linear models assume independent, additive contribution from each predictor variable x_j ($j = 1, \dots, k$)
- Joint effects of variables can be modelled by adding interaction terms to the design matrix (+ parameters)
- Interaction of numerical variables (interval scale)
 - interaction term for variables x_i and $x_j =$ product $x_i \cdot x_j$
 - e.g. in multivariate polynomial regression:
 $Y = p(x_1, \dots, x_k) + \epsilon$ with polynomial p over k variables
- Interaction of categorical factor variables (nominal scale)
 - interaction of x_i and x_j coded by one dummy variable for each combination of a level of x_i with a level of x_j
 - alternative codings e.g. to have separate parameters for independent additive effects of x_i and x_j
- Interaction of categorical factor with numerical variable

Generalised linear models

- Linear models are flexible analysis tool, but they ...
 - only work for a numerical response variable (interval scale)
 - assume independent (i.i.d.) Gaussian error terms
 - assume equal variance of errors (homoscedasticity)
 - cannot limit the range of predicted values
 - Linguistic frequency data problematic in all four respects
 - each data point $y_i =$ frequency f_i in one text sample
 - f_i are discrete variables with binomial distribution (or more complex distribution if there are non-randomness effects)
 - linear model uses relative frequencies $p_i = f_i/n_i$
 - Gaussian approximation not valid for small text size n_i
 - sampling variance depends on text size n_i and "success probability" π_i (= relative frequency predicted by model)
 - model predictions must be restricted to range $0 \leq p_i \leq 1$
- ➔ Generalised linear models (GLM)

Generalised linear model for corpus frequency data

- Sampling family (binomial)

$$f_i \sim B(n_i, \pi_i)$$

- Link function (success probability $\pi \leftrightarrow$ odds θ)

$$\pi_i = \frac{1}{1 + e^{-\theta_i}}$$

- Linear predictor

$$\theta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}$$

- ➔ Estimation and ANOVA based on likelihood ratios

⊗ iterative methods needed for parameter estimation