Chapter 6: Simple and Multiple Linear Regression

An Online Course

Sponsored by The Georgia R School

Presented by Geoffrey S. Hubona

How Old Is The Universe?

- Using Hubble Space Telescope, Freedman et al. (2001) gave relative velocity and distance of 24 galaxies.
    - Velocities are assessed by measuring the Doppler red shift in the spectrum of light observed from the galaxies concerned.
- Obtain scatterplot of velocity and distance. Fit a regression model to the data:

$$\text{velocity} = \beta_1 \text{distance} + \varepsilon$$

- Known as ***Hubble's Law***, $\boldsymbol{\beta_1}$ known as ***Hubble's constant***
    - $\boldsymbol{\beta_1^{-1}}$ gives an approximate age of the universe.

How Old Is The Universe?

- 24 Galaxies (observations), 3 Data Variables:
    - **Galaxy:** Identification number;
    - **Velocity:** Velocity of galaxy;
    - **Distance:** Distance of galaxy.

Cloud Seeding

- Weather modification, or ***cloud seeding***, is the treatment of individual clouds or storm systems with various inorganic and organic materials in the hope of achieving an increase in rainfall.
- Data collected in summer of 1975 in an experiment using massive quantities of silver iodide.
    - 24 days deemed suitable for seeding when *S-Ne* was not less than 1.5. (large ‘Seedability’; small rainfall)
- Question of interest: ***How is rainfall related to the explanatory variables? How effective is seeding?***
    - Multiple linear regression.

Cloud Seeding

- 24 Observations, 2 Data Variables:
  - **Seeding:** a factor indicating whether seeding occurred;
  - **Time:** number of days after the first day of the experiment.
  - **Cloudcover:** the percentage cloud cover in the experimental area, measured using radat.
  - **Prewetness:** the total rainfall in the target area one hour before seeding.
  - **Echomotion:** factor showing whether the radar echo was moving or stationary.
  - **Rainfall:** the amount of rain in cubic metres.
  - **Sne:** Suitability criterion.

Multiple Linear Regression

Assume y_i represents the value of the response variable on the ith individual, and that $x_{i1}, x_{i2}, \ldots, x_{iq}$ represents the individual's values on q explanatory variables, with $i = 1, \ldots, n$.

The multiple linear regression model is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq} + \varepsilon_i .$$

The residual or error terms ε_i, $i = 1, \ldots, n$, are assumed to be independent random variables having a normal distribution with mean zero and constant variance σ^2.

Multiple Linear Regression

Consequently, the distribution of the random response variable, y, is also normal with expected value given by the linear combination of the explanatory variables

$$E(y|x_1, \ldots, x_q) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$$

and with variance σ^2.

The parameters of the model β_k, $k = 1, \ldots, q$, are known as regression coefficients with β_0 corresponding to the overall mean.

The multiple linear regression model can be written most conveniently for all n individuals by using matrices and vectors as

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

Inference

$\hat{y}_i$ is the predicted value of the response variable for the ith individual $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_q x_{q1}$ and $\bar{y} = \sum_{i=1}^{n} y_i / n$ is the mean of the response variable.

The mean square ratio

$$F = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 / q}{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 / (n - q - 1)} \sim F_{q,n-q-1}$$

provides an F-test of the general hypothesis

$$H_0 : \beta_1 = \cdots = \beta_q = 0.$$

Variance Estimation

An estimate of the variance σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-q-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Individual regression coefficients can be assessed by using the ratio t-statistics $t_j = \hat{\beta}_j / \sqrt{\mathrm{Var}(\hat{\beta})_{jj}}$, although these ratios should only be used as rough guides to the 'significance' of the coefficients.