Chapter 11

- Simple linear regression

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

x is regressor (predictor, independent variable, explanatory variable), Y is response (dependent variable), ϵ is error (random variable), $E\epsilon = 0$; β_0 (intercept) and β_1 (slope) are parameters of interest.

- Data: $(x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n)$,

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $\epsilon_1, \epsilon_2, ..., \epsilon_n$ are independent normally distributed, $E\epsilon_i = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$.

If $\hat{\beta}_0$ and $\hat{\beta}_1$ are some estimators of β_0 and β_1, then

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

are fitted values, and the differences

$$e_i = Y_i - \hat{Y}_i$$

are residuals.

A visual analysis of the residuals gives useful preliminary information about data and model. On the basis of this analysis, the model can be corrected.

- Least squares estimators: minimization of

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2.$$

The estimators are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}.$$

- Properties of the estimators. $\hat{\beta}_1$ and $\hat{\beta}_0$ are normally distributed.

$$E\hat{\beta}_1 = \beta_1, \ E\hat{\beta}_0 = \beta_0,$$

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n}(x_i - \bar{x})^2}\sigma^2.$$

- Unbiased estimator of σ^2 is

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

S^2 is independent on $\bar{Y}$, $\hat{\beta}_0$ and $\hat{\beta}_1$. Distribution of $(n-2)S^2/\sigma^2$ is χ^2-distribution with $n-2$ degrees of freedom.

- Inference on the slope and intercept.

$(1-\alpha)$ confidence interval for β_1:

$$\left[\hat{\beta}_1 - t_{\alpha/2,n-2}\frac{S}{\sqrt{\sum(x_i-\bar{x})^2}}, \hat{\beta}_1 + t_{\alpha/2,n-2}\frac{S}{\sqrt{\sum(x_i-\bar{x})^2}}\right].$$

$(1-\alpha)$ confidence interval for β_0:

$$\left[\hat{\beta}_0 - t_{\alpha/2,n-2}\frac{S\sqrt{\sum x_i^2}}{\sqrt{n\sum(x_i-\bar{x})^2}}, \hat{\beta}_0 + t_{\alpha/2,n-2}\frac{S\sqrt{\sum x_i^2}}{\sqrt{n\sum(x_i-\bar{x})^2}}\right].$$

Testing $H_0 : \beta_1 = \beta_{10}$. Alternatives a) $H_1 : \beta_1 > \beta_{10}$, b) $H_1 : \beta_1 < \beta_{10}$, c) $H_1 : \beta_1 \neq \beta_{10}$. Significance level α.

Test statistic

$$T = \frac{\hat{\beta}_1 - \beta_{10}}{S/\sqrt{S_{xx}}}$$

where $S_{xx} = \sum(x_i-\bar{x})^2$. Under H_0, T has t-distribution with $n-2$ degrees of freedom. Critical region: a) $T \geq t_{\alpha,n-2}$, b) $T \leq -t_{\alpha,n-2}$, c) $|T| \geq t_{\alpha/2,n-2}$.

Testing $H_0 : \beta_0 = \beta_{00}$. Alternatives a) $H_1 : \beta_0 > \beta_{00}$, b) $H_1 : \beta_0 < \beta_{00}$, c) $H_1 : \beta_0 \neq \beta_{00}$. Significance level α.

Test statistic

$$T = \frac{\hat{\beta}_0 - \beta_{00}}{S\sqrt{\sum x_i^2/nS_{xx}}}$$

where $S_{xx} = \sum(x_i-\bar{x})^2$. Under H_0, T has t-distribution with $n-2$ degrees of freedom. Critical region: a) $T \geq t_{\alpha,n-2}$, b) $T \leq -t_{\alpha,n-2}$, c) $|T| \geq t_{\alpha/2,n-2}$.