Exercise class 2 formula sheet

András Bálint, andras.balint@chalmers.se

January 29, 2020

Simple linear regression

Formula for simple linear regression:
$Y = \beta_0 + \beta_1 X + \varepsilon$

Observed data in form of (predictor, response) pairs:
$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

i-th predicted response and residual:
$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \quad e_i = y_i - \hat{y}_i$

Residual squared at observation i:
$e_i^2 = (y_i - \hat{y}_i)^2$

Residual sum of squares:
$\mathrm{RSS} = e_1^2 + e_2^2 + \ldots + e_n^2$

Least squares coefficient estimates:
$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$
$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$, $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$

Estimating standard error of the random error term by residual standard error:
$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^n (y_i - \hat{y}_i)^2}$

Estimating the standard error of coefficients:
$\mathrm{SE}(\hat{\beta}_0) = \mathrm{RSE} \times \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$

$\mathrm{SE}(\hat{\beta}_1) = \mathrm{RSE} \times \sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}}$

Confidence intervals for the coefficients:
$\hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0)$
$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1)$
Note: for the precise confidence intervals, 2 should be replaced by the 97.5% quantile of the t distribution with $n - 2$ degrees of freedom.

Proportion of variability in the response that is explained by the predictor:
$R^2 = \frac{\text{TSS}-\text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$

Total sum of squares:
$\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

Correlation of X and Y:
$r = \text{Cor}(X,Y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$

Correlation and R^2: $\quad R^2 = r^2$

Multiple linear regression

Formula for multivariate linear regression:
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$

Observed data of n observations each containing values for the p predictors and the response:
$(x_{1,1}, x_{1,2}, \ldots, x_{1,p}, y_1), (x_{2,1}, x_{2,2}, \ldots, x_{2,p}, y_2), \ldots, (x_{n,1}, x_{n,2}, \ldots, x_{n,p}, y_n)$

i-th predicted response:
$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \ldots + \hat{\beta}_p x_{i,p}$

i-th residual:
$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \hat{\beta}_2 x_{i,2} - \ldots - \hat{\beta}_p x_{i,p}$

Residual sum of squares:
$\text{RSS} = e_1^2 + e_2^2 + \ldots + e_n^2$

Residual standard error:
$\text{RSE} = \sqrt{\frac{1}{n-p-1}\text{RSS}}$

Proportion of variability in the response that is explained by the predictor:
$R^2 = \frac{\text{TSS}-\text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$

Total sum of squares:
$\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

Test the null hypothesis that all coefficients are zero:
$H_0 : \beta_1 = \beta_2 = \ldots \beta_p = 0$
$H_a :$ at least one β_j is not zero
Compute F-statistic: $\quad F = \frac{(\text{TSS}-\text{RSS})/p}{\text{RSS}/(n-p-1)}$

Test relationship of variable X_j with the response Y in the presence of all other predictors:
$H_0 : \beta_j = 0$
$H_a : \beta_j \neq 0$
Compute t-statistic: $t = \frac{\hat{\beta}_j}{\text{SE}(\hat{\beta}_j)}$