11.4 - 11.5

With the notation $S_{xx} = \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2$, we get

$$Var(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^{n} x_{1i}^2}{n S_{xx}} \quad \text{and} \quad Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

An estimator for σ^2 is given by

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

Also $T_{\beta_0} = \frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\frac{\sum_{i=1}^{n} x_{1i}^2}{n S_{xx}}}}$ and $T_{\beta_1} = \frac{\hat{\beta}_1 - \beta_1}{\frac{S}{\sqrt{S_{xx}}}}$ are t_{n-2} under the assumption of normally distributed data.

These can be used for hypothesis testing of the coefficients.

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$$

The p-value is $2P(T_{\beta_1} \geq |t_{obs}| \mid H_0)$ and can be compared to the significance level α.

For $\quad H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 > 0$

then p-value is $P(T_{\beta_1} > t_{obs} \mid H_0)$

Similarly for tests on β_0

From $P\left(-t_{\frac{\alpha}{2}, n-2} \leq \frac{\hat{\beta}_0 - \beta_0}{S\sqrt{\frac{\sum_{i=1}^{n} x_{1i}^2}{n S_{xx}}}} \leq t_{\frac{\alpha}{2}, n-2}\right) = 1 - \alpha$

we get ~~a~~ $100(1-\alpha)\%$ confidence interval for β_0

$$\left(b_0 - t_{\frac{\alpha}{2}, n-2}\, S\sqrt{\frac{\sum_{i=1}^{n} x_{1i}^2}{n S_{xx}}},\ b_0 + t_{\frac{\alpha}{2}, n-2}\, S\sqrt{\frac{\sum_{i=1}^{n} x_{1i}^2}{n S_{xx}}} \right)$$

Similarly for β_1

Confidence and Prediction intervals 11.6

A confidence interval for $E[Y|X_1 = x_0] = \overset{\beta_0+\beta_1 x_0 =}{\mu_{Y|x_0}}$ can be obtained as follows.

Let $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ be the estimator

$$\hat{Y}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_1 x_0 = \bar{Y} + \hat{\beta}_1(x_0 - \bar{x}_1)$$

Assuming normally distributed data, $\bar{Y}$ and $\hat{\beta}$ are independent.

Hence $Var(\hat{Y}_0) = \frac{\sigma^2}{n} + \frac{(x_0 - \bar{x}_1)^2 \cdot \sigma^2}{S_{xx}}$

With unknown variance $\frac{\hat{Y}_0 - \mu_{Y|x_0}}{s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x}_1)^2}{S_{xx}}}} \sim t_{n-2}$

and a $100(1-\alpha)\%$ confidence interval is given by

$$\left(\hat{y}_0 - t_{\frac{\alpha}{2}, n-2}\, s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x}_1)^2}{S_{xx}}},\ \hat{y}_0 + t_{\frac{\alpha}{2}, n-2}\, s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x}_1)^2}{S_{xx}}} \right)$$

For a new observation, Y_0, when $X_1 = x_0$ we may construct a prediction interval as follows.

$E[\hat{Y}_0 - Y_0] = \beta_0 + \beta_1 x_0 - (\beta_0 + \beta_1 x_0) = 0$

and $Var(\hat{Y}_0 - Y_0) = Var(\hat{Y}_0) + Var(Y_0) = \frac{\sigma^2}{n} + \frac{(x_0 - \bar{x}_1)^2 \cdot \sigma^2}{S_{xx}} + \sigma^2$

Hence $\frac{\hat{Y}_0 - Y_0}{s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x}_1)^2}{S_{xx}}}} \sim t_{n-2}$ and a $100(1-\alpha)\%$ prediction interval for Y_0 is

$$\left(\hat{y}_0 - t_{\frac{\alpha}{2}, n-2}\, s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x}_1)^2}{S_{xx}}},\ \hat{y}_0 + t_{\frac{\alpha}{2}, n-2}\, s\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x}_1)^2}{S_{xx}}} \right)$$

Notice that the 1. normal equation gives

$$\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$$ The sum of residuals are zero

Further we get from the normal equations in general that

$$\underline{X}'(\underline{y}-\hat{\underline{y}}) = \underline{0} \Rightarrow \underline{b}'\underline{X}'(\underline{y}-\hat{\underline{y}}) = \underline{0} \Leftrightarrow \hat{\underline{y}}'(\underline{y}-\hat{\underline{y}}) = \underline{0}$$

$$\Leftrightarrow \sum_{i=1}^{n} \hat{y}_i (y_i - \hat{y}_i) = 0$$

<u>Decomposing the variation</u>

$$y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y}$$

such that $$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)\hat{y}_i}_{=0} - \underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)\bar{y}}_{=0}$$

such that $$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SS_T = SS_E + SS_R$$

Total sum of squares = error sum of squares + sum of squares for regression.

<u>Teorem 12.1</u>

In the linear regressionmodel

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{\varepsilon}$$

an unbiased estimator for σ^2 is given by

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n - (k+1)}.$$ The estimate is $\frac{SS_E}{n-(k+1)} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)}{n-(k+1)}$

The sum of squares, SS_T, SS_E and SS_R are normally gathered in an analysis of variance (ANOVA) table.

Source	SS	d.f.	F
Regression	$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	k	$\frac{SS_R}{k} \Big/ \frac{SS_E}{n-k-1}$
Error	$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$n-k-1$	
Total	$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$n-1$	

Let $\varepsilon_i \sim N(0, \sigma^2)$ and independent, $i = 1, 2, \ldots, n$

The F statistics is used to test the hypothesis

$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ $\qquad$ H_1: at least one β_i is different from 0

i.e. of the regression is significant

Reject H_0 if $F = \dfrac{\frac{SS_R}{k}}{\frac{SS_E}{n-k-1}} \geq f_{\alpha, k, n-k-1}$

eventually if. $P(F \geq F_{obs}) \leq \alpha$.

12.5 Inference in multippel linear regression

Test for the significant impact on the response given that the other variables are in the model

$H_0: \beta_j = 0$ $\qquad$ $H_1: \beta_j \neq 0$

More general $H_0: \beta_j = \beta_{j0}$ $\qquad$ $H_1: \beta_j \gtreqless \beta_{j0}$