Review - Week 12

Read: Chapters 22-28

Review: Inference for Regression

The least-squares line computed from a sample can be viewed as an estimate of the true regression line for the population.

The population regression line is given by \( \mu_y = \beta_0 + \beta_1 x \). It describes how the mean response changes as \( x \) changes.

The observed response \( y_i \) when \( x = x_i \) is \( y_i = \beta_0 + \beta_1 x_i + \epsilon_i \). The deviations \( \epsilon_i \) are assumed to be independent and follow the \( N(0, \sigma) \) model.

The parameters of the model are \( \beta_0 \), \( \beta_1 \) and \( \sigma \). They are estimated as follows:

\[
\begin{align*}
  b_0 &= \bar{y} - b_1 \bar{x} \\
  b_1 &= r \frac{s_y}{s_x} \\
  s_e &= \sqrt{\frac{\sum e_i^2}{n-2}}
\end{align*}
\]

The predicted response is \( \hat{y}_i = b_0 + b_1 x_i \) and the residuals are \( e_i = y_i - \hat{y}_i \).

A level C confidence interval for \( \beta_1 \) is \( b_1 \pm t_{n-2}^\ast \cdot SE(b_1) \), where the critical value comes from a \( t \) model with \( n-2 \) degrees of freedom.

To test the hypothesis, \( H_0 : \beta_1 = 0 \) vs. \( H_a : \beta_1 \neq 0 \) (No linear relationship between \( x \) and \( y \)), we use the test statistic \( t = \frac{b_1}{SE(b_1)} \).

P-values for this test can be calculated using the \( t \) distribution with \( n-2 \) degrees of freedom.

The mean predicted value for the subpopulation corresponding to the value \( x^* \) of the explanatory variable is \( \hat{\mu}_y = b_0 + b_1 x^* \).

A level C confidence interval for the mean response \( \mu_y \) when \( x \) takes the value \( x^* \) is

\[
\hat{\mu}_y \pm t^\ast \cdot SE(\hat{\mu})
\]

where \( t^\ast \) is calculated using the \( t \)-model with (n-2) degrees of freedom and

\[
SE(\hat{\mu}_y) = \sqrt{\left( (x^* - \bar{x})SE(b_1) \right)^2 + \frac{s_e^2}{n}}
\]
The predicted value of $y$ for a future observation from the subpopulation corresponding to the value $x^*$ of the explanatory variable is $\hat{y} = b_0 + b_1 x^*$. 

A level C confidence interval for the estimated response is

$$\hat{y} \pm t^* SE(\hat{y})$$

where $t^*$ is calculated from the t-model with (n-2) degrees of freedom and

$$SE(\hat{y}) = \sqrt{\left(\bar{x} - x^*\right)^2 SE(b_1)} + \frac{s^2}{n} + s_e^2$$

When performing regression analysis it is very important to check that the assumptions you made are correct. Always look at the residual plot for any patterns and make a normal quantile plot of the residuals, to check for a straight line.

**Exercise 1:** A grass seed company conducts a study to determine the relationship between the density of seeds planted (in pounds per 500 sq ft) and the quality of the resulting lawn. Eight similar plots of land are selected and each is planted with a particular density of seed. One month later the quality of each lawn is rated on a scale of 0 to 100. The regression analysis is given below:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>SE(coefficient)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>33.148</td>
<td>7.511</td>
</tr>
<tr>
<td>Seed Density</td>
<td>4.537</td>
<td>2.469</td>
</tr>
</tbody>
</table>

Dependent variable is: Lawn Quality
R squared = 36.0%
$s = 9.073602$ with 8 - 2 = 6 degrees of freedom

At the 5% level of significance, is there evidence of an association between seed density and lawn quality?