1. (a) Model: \( Y_i = \alpha + \beta x_i + \epsilon_i \), or an equivalent form.

Assumptions: Response, \( Y_i \), is normally distributed; equal variance in the response variable, \( Y_i \), for each value of the explanatory variable, \( x_i \). [Or, \( E_i \sim N(0,1) \).] All observations are independent.

Check normality with a normal probability plot of the residuals. This should be, roughly, a straight line.

Check constant \( \sigma \) with a plot of the residuals versus the fitted values. The plot should be a random scatter about 0.

(b) \( H_0 : \beta = 0 \) vs \( H_1 : \beta \neq 0 \). Either of the following:

- \( T = 5.21 \Rightarrow P - \text{value} = 0.000 \).
- \( F = 27.12 \Rightarrow P - \text{value} = 0.000 \).

So there is evidence to reject the null hypothesis and we conclude that there is a linear relationship between stress level and distance travelled.

(c) For zero distance travelled the stress level is 2.8 (this may well be so but in that case the stress level has little or nothing to do with commuting distance). In the data set the distance travelled is in the range 4.5 to 18.4 kilometres. Zero is outside this range and thus the fitted model may not apply. It is very risky to extrapolate.

(d) \[ 16.027 \times 100 \approx 73.1\% \]

(e) Adjusted \( R^2 = 1 - \frac{MSE_{\text{model}}}{MSE_{\text{total}}} = 1 - \frac{0.591}{21.937/11} \approx 70.4\% \).

(f) Estimate of \( \sigma \), the standard deviation of the observations not explained by the model. Calculated using \( \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n-p}} \), where \( n \) is the number of observations and \( p \) is the number of parameters estimated from the data (in this case, 2).

\[ \text{se}(\hat{\mu}(15)) = \sqrt{\frac{0.7687^2}{12} + (15 - 13.73)^2 \times 0.05875^2} = 0.234. \]

95% confidence interval: \( 7.37 \pm 1.975 \times 0.234 = (6.85, 7.89) \).

2. (a) Analysis of Variance for yield

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variety</td>
<td>4</td>
<td>1096.9</td>
<td>274.23</td>
<td>3.73</td>
<td>0.01 &lt; P &lt; 0.025</td>
</tr>
<tr>
<td>Error</td>
<td>30</td>
<td>2205.3</td>
<td>73.51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>3302.2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( H_0 : \mu_1 = \mu_2 = \ldots = \mu_5 \) vs \( H_1 : \text{not} \ H_0 \).

The \( P \)-value is < 0.05, so there is evidence to reject the null hypothesis. We conclude that at least one pair of mean yields is significantly different.

(b) \( \hat{\mu}_A = 23.714, \hat{\mu}_B = 34.857 \).

95% confidence interval: \( 23.714 - 34.857 \pm 1.975 \sqrt{MSE \left( \frac{1}{7} + \frac{1}{7} \right)} = (-20.5, -1.8) \).

(c) • The spread of the groups varies. The ANOVA model assumes that the observations for each group have the same variance.

- The distribution for the observations in troup 1, 2 and 4 are noticeably asymmetric. ANOVA assumes that the observations for each group are normally distributed.

- The means for the groups appear different, particularly \( \bar{x}_1 \) and \( \bar{x}_3 \). This supports the conclusion drawn from the hypothesis test in (b).
\(\psi = \frac{1}{4}(\mu_1 + \mu_2 + \mu_4 + \mu_5) - \mu_3.\)

\(H_0 : \psi = 0 \quad \text{vs} \quad H_1 : \psi \neq 0.\)

\(\hat{\psi} = -9.9285 \quad \text{and} \quad \text{se}(\hat{\psi}) = 3.623.\)

\(T = -2.74 \frac{d}{t_{30}} \Rightarrow 0.01 < P\text{-value} < 0.02.\)

Therefore there is evidence to reject the null hypothesis and we conclude that the combined mean yield of the Californian varieties differs from the mean yield of the Florida variety.

3. See solutions to 270 ASSIGNMENT 4 Question 1.
4. See solutions to 270 ASSIGNMENT 4 Question 2.
5. See solutions to 270 ASSIGNMENT 4 Question 3.
7. (a) This is an additive 2-way mixed effects model: \(Y_{ijk} = \mu + \alpha_i + B_j + E_{ijk},\) where

- \(E_{ijk} \sim N(0, \sigma^2)\)
- \(\sum \alpha_i = 0: \) \(i\) is the \(i\)th fixed level of fertilizer
- \(B_j \sim N(0, \sigma^2_B): \) \(j\) is the \(j\)th random level of area

(b) Error: \(\hat{\sigma}^2 = 0.001007.\)

Area: \(\hat{\sigma}^2_B = \frac{0.008167 - 0.001007}{3} = 0.0024.\)

(c) Fertilizer: \(H_0 : \alpha_i = 0, \) for all \(i.\)

Area: \(H_0 : \hat{\sigma}^2_B = 0.\)

(d) \(\hat{\sigma}^2_Y = \hat{\sigma}^2_B + \hat{\sigma}^2 = 0.0034, \quad \Rightarrow \quad \text{se}(\hat{Y}) = 0.058.\)

(e) There is only one value, the mean, reported per treatment combination (cell).

(f) Fertilizer: \(F = 8.94. \) Now \(F_{2,8}(0.09) = 8.649.\) Therefore fertilizer has a significant effect on yield at the 0.01% level.

Area: \(F = 8.11. \) Now \(F_{4,8}(0.09) = 7.006.\) Therefore area has a significant effect on yield at the 0.01% level.

The celery grower would conclude that not all fertilizers result in the same mean growth and not all areas give the same mean growth.

(g) A randomised block design with the areas as the blocks.

(h) Yes, it was a good design. Area has a significant effect on yield and, within each area, fertilizers have a significant effect on yield. Unless area was used as a blocking factor the effect of the different fertilizers on yield would be confounded by the effect of area.

8. (a) \(\Pr(\bar{X} \geq -0.538) = \Pr(\bar{X}_{S} \geq 1.957) = 1 - 0.9748 = 0.0252.\)

Therefore, if the milk is not watered, the observed sample mean is likely to occur approximately 2.5 times per 100. This is less than 5%...ie. it is good evidence that the producer is adding water to the milk.

(b) \(H_0 : \mu = -0.545 \quad \text{vs} \quad H_1 : \mu > -0.545.\)

Under \(H_0: c_{0.95} = -0.545 + 1.645 \times \frac{0.538}{\sqrt{5}} = -0.539.\)

\(\Pr(\bar{X} > -0.539|\mu = -0.53) = 0.9941.\)

Yes...99.4% of the time a mean freezing point this high will be detected.