Some formulae:

\[ b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \]
\[ b_0 = \bar{y} - b_1 \bar{x} \]
\[ \text{Var}(b_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \]
\[ \text{Var}(b_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right) \]
\[ \text{Cov}(b_0, b_1) = -\frac{\sigma^2 \bar{x}}{\sum (x_i - \bar{x})^2} \]
\[ \text{SSE} = \sum (y_i - \hat{y}_i)^2 \]
\[ \text{SSTO} = \sum (y_i - \bar{y})^2 \]
\[ \text{SSR} = b_1^2 \sum (x_i - \bar{x})^2 = \sum (\hat{y}_i - \bar{y})^2 \]
\[ \sigma^2 \{ \hat{y}_h \} = \text{Var}(\hat{y}_h) = \sigma^2 \left( \frac{1}{n} + \frac{(X_h - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) \]
\[ \sigma^2 \{ \text{pred} \} = \text{Var}(y_h - \hat{y}_h) = \sigma^2 \left( 1 + \frac{1}{n} + \frac{(X_h - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) \]

Working-Hotelling coefficient: \[ W = \sqrt{\frac{2F_{2,n-2;1-\alpha}}{n}} \]

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3 abcd</th>
<th>3 efg</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1. (a) (2 points) Consider the simple linear regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where the $\epsilon_i$'s are independent and identically distributed with the $N(0, \sigma^2)$ distribution. Assume the $X_i$'s are fixed. What is the distribution of $Y_i$ when $X_i$ is 10?

$Y_i \sim N(\beta_0 + 10\beta_1, \sigma^2)$

(b) (3 points) The least squares estimate of the $Y$ intercept for the model in (a) is $b_0$ as given on the first page. Show that $b_0$ is an unbiased estimate of the intercept in the model. You may take as known any results that were proved in lecture.

$$E(b_0) = E(Y - b_1X)$$

$$= \beta_0 + \beta_1\bar{X} - E(b_1)\bar{X}$$

$$= \beta_0 \text{ since } E(b_1) = \beta_1$$

(c) (4 points) Show that the sum of the residuals is 0 for the least squares fit for the model above. What assumptions about the model did you use for this calculation?

$$\sum e_i = \sum (Y_i - b_0 - b_1 X_i)$$

$$= \sum (Y_i - \bar{Y} + b_1\bar{X} - b_1 X_i)$$

$$= n\bar{Y} - n\bar{Y} + nb_1\bar{X} - b_1 n\bar{X}$$

$$= 0$$

*Don’t need any statistical assumptions – just the form of the model.*
2. (6 points (2 each)) For each of the following statements, say whether it is true or false. Give a brief justification of your answer.

(a) A value of $R^2$ close to 1 indicates that the linear regression model is a good fit to the data.

*False. Points could be closely scattered around in a line in a curvilinear fashion, or could be caused by one overly influential point.*

(b) The estimate of the error variance, $s^2$, is a random variable.

*True. $s^2$ is a function of the $Y$’s which are random variables.*

(c) $\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = 0$

*False. This is what is minimized in least squares.*
3. Two species of predatory birds, collard flycatchers and tits, compete for nest holes during breeding season. Frequently, dead flycatchers are found in nest boxes occupied by tits. A field study examined whether the risk of mortality to flycatchers is related to the degree of competition between the two bird species for nest sites. At each of 14 locations, the following data were collected: the number of flycatchers killed (the response variable labelled $fc_{killed}$) and the nest box occupancy measured as a percentage (the predictor variable labelled $tit_{occ}$). 

The data and some SAS output are given below. Some numbers from the SAS output have been purposely deleted.

<table>
<thead>
<tr>
<th>Location</th>
<th>1 2 3 4 5 6 7 8 9 10 11 12 13 14</th>
</tr>
</thead>
<tbody>
<tr>
<td>$fc_{killed}$</td>
<td>0 0 0 0 0 1 1 1 1 2 2 3 4 5</td>
</tr>
<tr>
<td>$tit_{occ}$</td>
<td>24 33 34 43 50 35 35 38 40 31 43 55 57 64</td>
</tr>
</tbody>
</table>

The REG Procedure

Descriptive Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Sum</th>
<th>Mean</th>
<th>SS</th>
<th>Variance</th>
<th>Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>14.00000</td>
<td>1.00000</td>
<td>14.00000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$tit_{occ}$</td>
<td>582.00000</td>
<td>41.57143</td>
<td>25844</td>
<td>126.87912</td>
<td>11.26406</td>
</tr>
<tr>
<td>$fc_{killed}$</td>
<td>20.00000</td>
<td>1.42857</td>
<td>62.00000</td>
<td>2.57143</td>
<td>1.60357</td>
</tr>
</tbody>
</table>

The REG Procedure

Model: MODEL1

Dependent Variable: $fc_{killed}$

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>(A)</td>
<td>19.11669</td>
<td>19.11669</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>(B)</td>
<td>14.31188</td>
<td>(C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>13</td>
<td>33.42857</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE | 1.09209 | R-Square | (D) |
Dependent Mean | 1.42857 | Adj R-Sq | 0.5362 |
Coeff Var | 76.44618 | |

Parameter Estimates

| Variable | DF | Estimate | Error | t Value | Pr > |t| |
|----------|----|----------|-------|---------|------|
| Intercept | 1 | -3.04686 | 1.15533 | -2.64 | 0.0217 |
| $tit_{occ}$ | 1 | 0.10766 | 0.02689 | (E) | 0.0018 |

Questions pertaining to this output are on the next two pages.
(a) (5 points) What are the values of the 5 missing numbers (A through E) in the SAS output?

\[
\begin{align*}
A &= 1 \\
B &= 12 \\
C &= \frac{14.31188}{12} = (1.09209)^2 = 1.1927 \\
D &= \frac{19.11669}{33.42857} = 0.5719 \\
E &= \frac{.10766}{.02689} = 4.0037
\end{align*}
\]

(b) (5 points) For the analysis of variance $F$ test, state the null and alternative hypotheses, the value of the test statistic, the distribution of the test statistic under the null hypothesis, the $p$-value as accurately as possible, and an appropriate conclusion.

\[
H_0 : \beta_1 = 0 \text{ versus } H_a : \beta_1 \neq 0
\]

Test statistic: $F_{obs} = E^2 = \frac{19.11669}{C} = 16.028$

Under $H_0$, $F_{obs} \sim F_{1,12}$

$p$-value = .0018

Strong evidence that the slope is not zero.

(c) (2 points) Estimate the mean change in the number of flycatchers killed when the nest box tit occupancy increases by 10%.

\[
0.10766(10) = 1.0766
\]

(d) (3 points) Give a 95% confidence interval for the slope of the line.

\[
t_{12,.025} = 2.179 \\
.10766 \pm 2.179(.02689) = (.0491, .1662)
\]
(e) (5 points) Suppose an additional location was later found to have a nest box tit occupancy of 30%. Give a 90% prediction interval for this new value.

\[ \hat{Y} = -3.04686 + 0.10766(30) = .1829 \]

\[ t_{12.05} = 1.782 \]

\[ \sum (X_i - \bar{X})^2 = 25844 - 14(41.57143)^2 = 13(126.87912) = 1649.4 \]

\textit{Prediction Interval:} \[ .1829 \pm 1.782(1.09209) \sqrt{1 + \frac{1}{14} + \frac{(30-41.57143)^2}{1649.4}} = (-1.906, 2.272) \]

(f) (3 marks) Would a 90% confidence interval for the mean number of flycatchers killed when the tit occupancy is 30% be wider or narrower than your interval in part (e). Explain why the width of the intervals differ. An answer that only points out the differences in the formulae will receive no marks.

\textit{Narrower. The s.e. of the estimate of} \( E(Y) \) \textit{only takes into account the variance in the estimation of the line while the s.e. of the estimate of the value of} \( Y \) \textit{has this source of variance plus the model error variance (how points are scattered about the line).}

(g) (2 marks) Basing your answer only on the information you have from the data and SAS output that was given, do you have any concerns about the validity of the prediction interval you found in the part (e)? Explain.

\textit{Yes. Looking at the values of the response variable, they clearly aren’t samples from a normal distribution.}