1. In placing a weekly order, a concessionaire that provides services at a baseball stadium must know what size crowd is expected during the coming week in order to know how much food, etc., to order. Since advance ticket sales give an indication of expected attendance, food needs might be predicted on the basis of the advanced sales. The table below gives data from 7 previous games.

<table>
<thead>
<tr>
<th>GAME</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>X (in thousands)</td>
<td>50.4</td>
<td>41.8</td>
<td>28.8</td>
<td>64.2</td>
<td>64.4</td>
<td>55.9</td>
<td>43.2</td>
</tr>
<tr>
<td>Y (in thousands)</td>
<td>29.1</td>
<td>25.9</td>
<td>10.8</td>
<td>32.4</td>
<td>36.0</td>
<td>30.7</td>
<td>19.9</td>
</tr>
</tbody>
</table>

(a) Fit a simple linear regression (least squares) line to the data,
given $\bar{Y} = 26.4$, $\bar{X} = 49.814285$
SSy = 439.8, SSx = 1006.6485, and SSxy = 628.91376

$$b_1 = \frac{SSXY}{SSX} = \frac{628.9136}{1006.6485} = .62476$$

$$b_0 = \bar{Y} - b_1 \bar{X} = 26.4 - .62476(49.814285) = -4.71297$$

Estimated regression function:

$$\hat{Y} = -4.72197 + .624756 X$$

(b) Sketch a scatterplot of Y vs X. Draw your fitted line on the same plot. Comment on the fit.

The line seems to fit the data fairly well.
(c) Fill in the following ANOVA table:

<table>
<thead>
<tr>
<th>SOURCE</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F*</th>
</tr>
</thead>
<tbody>
<tr>
<td>regress</td>
<td>1</td>
<td>392.92</td>
<td>392.92</td>
<td></td>
</tr>
<tr>
<td>error</td>
<td>5</td>
<td>46.88</td>
<td>9.38</td>
<td>41.90</td>
</tr>
<tr>
<td>total</td>
<td>6</td>
<td>439.80</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(d) Calculate the coefficient of determination $r^2$ and interpret its value.

$$r^2 = \frac{SSR}{SSTO} = \frac{392.92}{439.80} = .8934$$

About 89% of the variation in hot dog sales can be explained by the variation in advanced ticket sales.

(e) Conduct a test to decide whether or not there is a linear association between number of advanced ticket sales (in thousands), $X$, and the number of hot dogs sold (in thousands), $Y$. Use a level of significance $\alpha = .05$. State the alternatives, test statistic, decision rule, and conclusion. Do both a t-test and an F-test. Calculate the p-value for the F-test. Demonstrate numerically the equivalency of the t-test and the F-test.

i) t-test for linear association --

hypotheses: $H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

test statistic: $t^* = \frac{b_1}{s(b_1)} = \frac{.62476}{.09651} = 6.47$

decision rule: Reject $H_0$ if $|t^*| \geq t_{5,.975,*}$

conclusion: Since $6.47 \geq 2.57058$, we reject $H_0$ and conclude that there is a statistically significant linear relationship between the number of hot dogs sold and the number of advanced ticket sales.

ii) F-test for linear association --

hypotheses: $H_0: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

test statistic: $F^* = \frac{MSR}{MSE} = \frac{392.92}{9.38} = 41.9$

decision rule: Reject $H_0$ if $F^* \geq F_{1,5,.95,*}$

p-value = .001

conclusion: Since $41.9 \geq 6.608$, we reject $H_0$ and conclude that there is a statistically significant linear relationship between the number of hot dogs sold and the number of advanced ticket sales.

iii) The t-test and F-test are numerically equivalent here since

$$(t^*)^2 = F^* \text{ and } (t_{5,.975})^2 = F_{1,5,.95,*}.$$
Find a 95% confidence interval for the mean number of hot dogs sold when the advanced ticket sales equal 57,300 (i.e., when $X = 57.3$).

$$\hat{Y} \pm t_{s\{\hat{Y}_h\}} \Rightarrow (27500, 34580)$$

If the advanced ticket sales this week equal 57,300, a 95% prediction interval for the number of hot dogs that will be purchased this week at the game is 22,460 to 39,700. Compare this interval to the interval in (f). How are they alike? How are they different?

Both intervals are approximately centered about $\hat{Y} = 31077$, but the prediction interval is wider than the confidence interval because $s\{Y_{h_{\text{new}}\}} > s\{\hat{Y}_h\}$.

2. Fill in the blank(s) with an appropriate answer or circle the most appropriate answer below.

(a) For the simple linear model with normal errors $\epsilon_i$, the errors are assumed to have $E(\epsilon_i) = 0$ and constant variance for all $i = 1,...,n$. Furthermore, the errors $\epsilon_i$ are assumed to be independent.

(b) In hypothesis testing, the probability of rejecting the null hypothesis when it is actually false is called the power of the test.

(c) In the general linear test, the null hypothesis specifies the reduced model and the test statistic is

$$F^* = \frac{\text{SSE}(R) - \text{SSE}(F)}{d_{\text{R}} - d_{\text{F}}} \div \frac{\text{SSE}(F)}{d_{\text{F}}}$$

(d) The least square estimator $b_1$ is “best” in the sense that, among all linear unbiased estimators of $\beta_1$, $b_1$ has minimum variance. This result is part of the Gauss-Markov Theorem.

(e) The method of least squares regression involves finding a regression line that has the least error sum of squares.

(f) For the simple normal error regression model, $\frac{\text{SSE}}{\sigma^2}$ has a $\chi^2$ distribution with $n - 2$ degrees of freedom.

If $\beta_1 = 0$, then $\frac{\text{SSR}}{\sigma^2}$ also has a $\chi^2$ distribution with 1 df, due to Cochran's Theorem. Hence, under the null hypothesis that $\beta_1 = 0$, $\frac{\text{MSR}}{\text{MSE}}$ has an $F_{1,n-2}$ distribution.

3. Prove that, for a simple linear regression model, the least squares estimator $b_1$ is unbiased.

$$E\{b_1\} = E\{ \frac{\text{SSXY}}{\text{SSX}} \} = \frac{\sum(X - \bar{X})E\{Y\}}{\text{SSX}}$$

$$= \frac{\sum(X - \bar{X}) (\beta_0 + \beta_1 X)}{\text{SSX}} = \beta_0 \frac{\sum(X - \bar{X})}{\text{SSX}} + \beta_1 \frac{\sum(X - \bar{X})X}{\text{SSX}}$$

$$= \beta_1$$