Homework 4/28

Chap 13: 66(a-l)

Chap 14: 4(not f), 23b, 29, 63 (for i only do Confidence interval, not prediction interval).

Note: On problem 29b, 63j use Seq. SS method for r-sq shown in class

On 63, do NOT do part g.

Point Distribution 66 (20 points); 4 (13 pts), 23b (3pts), 29(4 pts); 63 (20pts); Extra
#1(20 points), Extra #2 (20 pts)

Additional Problems:

1) Given the following information, compute the appropriate values and tests at
alpha=.05 (note, be sure to state your hypotheses and critical values of the test statistic).

Covariance Matrix of X and Y

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>3.85606</td>
<td></td>
</tr>
<tr>
<td>X</td>
<td>11.81061</td>
<td>96.99242</td>
</tr>
</tbody>
</table>

Variable    N  N*    Mean  SE Mean  StDev
Y         12   0  22.617    0.567  1.964
X         12   0  123.42     2.84   9.85

Sum of Squares Regression=15.82
Sum of Squares Error=26.597
Sample size n=12
Standard error of slope (Sb1)= .05

a) Find the correlation between X and Y
b) Test whether the correlation indicates a significant relationship
c) Find the regression line equation.
d) Interpret generally the coefficients b0 and b1.
e) Test using ANOVA whether there is a relationship
f) Find r^2
g) Test whether the slope is zero

2) A school district wanted to predict the number of sick days used as a function of total
number of teachers in a school. Use the following partial output to answer questions:

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1.843</td>
<td>2.468</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teachers</td>
<td>0.50054</td>
<td>0.04906</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

S = 10.3780  R-Sq =
### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td></td>
<td>11214</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual Error</td>
<td></td>
<td>4954</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### New

<table>
<thead>
<tr>
<th>Obs</th>
<th>Fit</th>
<th>SE Fit</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.67</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a) Write the estimated regression equation and interpret the coefficients.
b) Test whether teacher is a significant variable.
c) Test whether there is an overall relationship.
d) Find the value of r-squared and test if it is significant.
e) For a school with 25 teachers, find the predicted number of sick days and a 95% confidence interval.
a) Scatterplot

Scatterplot of Delivery Time vs Number of cases

b) B0 = 24.8; b1 = 0.140

c) The regression equation is

\[ \text{Delivery Time} = 24.8 + 0.140 \times \text{Number of cases} \]

d) The intercept is 24.8, this is the delivery time for 0 cases. The slope (b1) is 0.140, this means that for every additional case, delivery time increases by 0.140 minutes.

e) Using the equation above, yhat for x = 150 is 45.838, see below for output

Predicted Values for New Observations

<table>
<thead>
<tr>
<th>New Obs</th>
<th>Fit</th>
<th>SE Fit</th>
<th>95% CI</th>
<th>95% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>45.838</td>
<td>(44.876, 46.801)</td>
<td>(41.555, 50.121)</td>
</tr>
</tbody>
</table>

f) No, it would not be appropriate to predict for x = 500, since as shown below, its range in the sample is only 52-298.

Descriptive Statistics: Number of cases

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>N*</th>
<th>Mean</th>
<th>SE Mean</th>
<th>StDev</th>
<th>Minimum</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of cases</td>
<td>20</td>
<td>0</td>
<td>169.9</td>
<td>18.1</td>
<td>81.0</td>
<td>52.0</td>
<td>97.0</td>
<td>159.0</td>
<td>251.3</td>
</tr>
</tbody>
</table>

Variable Maximum

Number of cases 298.0
g) \( r^{2} = 97.2\% \)
\[ S = 1.98650 \quad R-Sq = 97.2\% \quad R-Sq(adj) = 97.0\% \]

h) Correlation = \( r = \sqrt{0.972} = 0.9859 \)

i) Standard error of the estimate is \( s = 1.98650 \)

j. Residual analysis

The residuals do NOT appear very normal. The probability plot shows they do not follow the normal pattern as does the histogram. There are too many in the tails and not enough nearer the mean.

k) To test for a linear relationship, we use a t-test of the slope. (note they may also use ANOVA and the F-test of \( H_{0}: \) there is a relationship) For t-test, we use critical \( t = \frac{\alpha}{2}, n-k-1 = 0.025,18 = 2.093 \) to test \( H_{0}: \) slope = 0

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>24.835</td>
<td>1.054</td>
<td>23.56</td>
<td>0.000</td>
</tr>
<tr>
<td>Number of cases</td>
<td>0.140026</td>
<td>0.005627</td>
<td>24.88</td>
<td>0.000</td>
</tr>
</tbody>
</table>

This shows that since 24.88 > 2.093 we reject the \( H_{0} \), there is a relationship. (also can use p-value)

For anova test (alternative) we test \( H_{0} \): no relationship by testing \( F(\alpha, 1,18) = 4.41 \).

Analysis of Variance

<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>Regression</td>
<td>1</td>
<td>2443.5</td>
<td>2443.5</td>
<td>619.20</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>18</td>
<td>71.0</td>
<td>3.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>19</td>
<td>2514.5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Since $619.20 > 4.41$, we reject $H_0$. Also can use p-value

1) 95% CI for prediction for $x=150 = \hat{y} + t*se(\hat{y})$  \[ \hat{y} = 45.838, \text{se}(\hat{y}) = 0.458, \]
    $t(\alpha/2, 18) = 2.1009 = 45.838 + 2.1009*0.458 = P(44.875 < y < 46.800) = .95$, or can show output below.

<table>
<thead>
<tr>
<th>New Obs</th>
<th>Fit</th>
<th>SE Fit</th>
<th>95% CI</th>
<th>95% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>45.838</td>
<td>0.458</td>
<td>(44.876, 46.801)</td>
<td>(41.555, 50.121)</td>
</tr>
</tbody>
</table>

Chapter 14.4

a) Regression equation

**Regression Analysis: DistCost versus Sales, Orders**

The regression equation is

\[ \text{DistCost} = -2.73 + 0.0471 \text{Sales} + 0.0119 \text{Orders} \]

b) $B_0 =$ intercept = -2.73 thousand dollars or -$2,730 if sales and orders=0
   $B_1 =$ slope of sales=.0471, as sales increase by one unit (thousand of dollars)
   costs increase by .0471 thousand dollar or $47.1
   $B_2 =$ slope of orders as orders increase by one unit, costs increase by .0119 thousand dollars or $11.9

c) The intercept ($b_0$)= -2.73, this would mean that if sales and orders were zero, costs would be -2.73 thousand dollars (-$2,730). This would be impossible, so it has no practical interpretation

d) Predict for sales=400,000 (enter as 400) and order=4,500

<table>
<thead>
<tr>
<th>New Obs</th>
<th>Fit</th>
<th>SE Fit</th>
<th>95% CI</th>
<th>95% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>69.878</td>
<td>1.663</td>
<td>(66.420, 73.337)</td>
<td>(59.381, 80.376)</td>
</tr>
</tbody>
</table>

e) Confidence interval = fit +/- $t*se(\hat{y}) = 66.420$ to $73.337$ given above

f) Skip this

g) R-squared = SSR/SST or from minitab

$R^2 = 87.6\%$  \hspace{1cm} $R^2(\text{adj}) = 86.4\%$

h) $R^2(\text{adjusted})$ given above or found by

\[
rsq(\text{adj}) = \left[ 1 - (1 - rsq) \cdot \frac{n-1}{n-k-1} \right] = \left[ 1 - (1-.876) \cdot \frac{23}{21} \right] = .864
\]
14.23b Test slopes b1, b2 using t-test for Ho: slope (b1 or b2) = 0 indicating no relationship

For b1 t=b1/sb1 = .04711/.02033 = 2.32 test against t(alpha/2, 21)=2.0796, therefore reject Ho, sales has significant slope and is related.
For b2, t=5.31, tcritical=2.0796, therefore, we reject, orders has a significant slope and is related.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-2.728</td>
<td>6.158</td>
<td>-0.44</td>
<td>0.662</td>
</tr>
<tr>
<td>Sales</td>
<td>0.04711</td>
<td>0.02033</td>
<td>2.32</td>
<td>0.031</td>
</tr>
<tr>
<td>Orders</td>
<td>0.011947</td>
<td>0.002249</td>
<td>5.31</td>
<td>0.000</td>
</tr>
</tbody>
</table>

29.
A) Test individual variables. See 23b above. Both orders and sales are significantly related.
b) find partial r-squared using Sequential SS method described in class as r-sq(partial) = SeqSS/SST

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Seq SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales</td>
<td>1</td>
<td>2726.8</td>
</tr>
<tr>
<td>Orders</td>
<td>1</td>
<td>641.3</td>
</tr>
</tbody>
</table>

SST=3845.1
63. Don’t do part g.

a) Regression equation

**Regression Analysis: Height versus breast height di, bark thickness**

The regression equation is

$$\text{Height} = 62.1 + 2.06 \text{ breast height diameter} + 15.6 \text{ bark thickness}$$

b) Interpret coefficients:

- $B_0=62.1$ is intercept, value at other variables=0
- $B_1=2.06$, for every 1 inch diameter, height increases by 2.06 feet
- $B_2=15.6$, for every 1 inch thicker, height increases by 15.6 feet

c) $\hat{y}$ for diameter=25, thickness=2 is 144.84 feet (see below)

**Predicted Values for New Observations**

<table>
<thead>
<tr>
<th>New Obs</th>
<th>Fit</th>
<th>SE Fit</th>
<th>95% CI</th>
<th>95% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>144.84</td>
<td>5.16</td>
<td>(134.01, 155.68)</td>
<td>(96.15, 193.54)</td>
</tr>
</tbody>
</table>

d) $r^2= 78.6\%$. This means that the model explains 78.6% of the total variation in height.

e) Residual analysis

Looking at the residuals, they seem broadly normal, but they are not smoothly distributed. There are more negative residuals (left end) than expected.

f) Test overall relationship using Anova and F-test

**Analysis of Variance**

<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>Regression</td>
<td>2</td>
<td>33718</td>
<td>16859</td>
<td>33.01</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Test Ho: No relationship by comparing F to Fcritical with (alpha, 2, 18) = 3.55 Since 33.01>3.55 we reject Ho, there is a relationship.

g) Don’t do this part

h) Use t-test for each slope to test Ho: b=0, or no relationship

Predictor      Coef  SE Coef     T      P
Constant      62.14    13.50  4.60  0.000
breast height diameter  2.0567   0.4428  4.64  0.000
bark thickness  15.642    7.148  2.19  0.042

Critical t(\alpha/2, 18) = 2.1009
For Diameter:  t=4.64>2.1009, we reject the Ho, there is a relationship
For thickness, t=2.19>2.1009, we reject the Ho, there is a relationship

i) Confidence interval for predicted y-hat

By hand or from computer

New Obs     Fit  SE Fit      95% CI            95% PI
1  144.84    5.16  (134.01, 155.68)  (96.15, 193.54)

j) Partial r-squared, using Seq SS method used in class r-sq(partial)=SeqSS/SST

Source                 DF  Seq SS           r-sq(partial) =SeqSS/SST
breast height diameter  1   31273  =31273/42911=.72878 or 72.88%
bark thickness        1    2445  = 2445/42911 = .0569 or 5.69%
SST =                42911

This means that diameter predicts 72.88% of the variation and thickness explains only 5.69%. Clearly diameter is a more important variable.
Extra Problem 1:

l) Correlation = \( \frac{S_{xy}}{(S_x)(S_y)} = \frac{11.81}{9.848 \times 1.96} = 0.6118 \)

m) \( T = \frac{r}{\sqrt{1-r^2}} = \frac{0.6118}{\sqrt{1-0.3744}} = 2.446 \)

Test Ho: \( r = 0 \), reject if \( T > t(\alpha/2, n-k-1) = 2.2281 \) (\( \alpha = 0.05 \)). Since 2.446 > 2.22, we reject Ho and conclude it is a significant relationship.

n) \( b_1 = \frac{S_{xy}}{S_x^2} = \frac{11.81}{96.99} = 0.1218 \)

\( b_0 = \bar{y} - b_1 \bar{x} = 22.617 - 0.1218 \times 123.42 = 7.68 \)

Equation \( y_{\hat{}} = 7.68 + 0.1218 \times x \)

The \( b_0 \) coefficient is the y-intercept, at a value of \( x = 0 \), \( y = 7.68 \);

The \( b_1 \) coefficient is the slope, for every 1 unit increase in \( x \), \( y \) will increase by 0.1218. It is a positive relationship.

p) Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>15.82</td>
<td>15.82</td>
<td>5.94</td>
</tr>
<tr>
<td>Residual Error</td>
<td>10</td>
<td>26.597</td>
<td>2.6597</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>42.414</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Test Ho: There is no relationship or explanatory power, reject if \( F > F(\alpha, 1, 10) = 4.96 \). Therefore, since 5.94 > 4.96, reject Ho, the model has power.

f) \( r^2 = \frac{SSR}{SST} = \frac{15.82}{42.414} = 0.3729 \)

or \( r^2 = \text{correlation found in part (a)} \) squared = \( 0.6118^2 = 0.3742 \)

Note error due to rounding

g) Test if slope = 0 using \( T \), reject if \( T > t(\alpha/2, n-k-1) = 2.22 \).

\( T = \frac{b_1}{s_{b_1}} = \frac{0.1281}{0.05} = 2.562 \), therefore we reject Ho, the slope is not zero, therefore, the variable is significant and there is a relationship.

Extra Problem #2

a) Write the estimated regression equation and interpret the coefficients.

From output: \( y = 1.803 + 0.50054 \times \text{Teachers} \)

b) Test whether teacher is a significant variable.

Use t_statistic, Ho: Slope (beta) = 0. Reject if \( t > t(\alpha/2, n-k-1) = t(0.025, 89) = 1.987 \)

\( T = \frac{b_1}{s_{b_1}} = \frac{0.50054}{0.04906} = 10.20 \), therefore reject, there is a relationship.

c) Test whether there is an overall relationship.

Use Analysis of variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>11214</td>
<td>11214</td>
<td>201.57</td>
</tr>
<tr>
<td>Residual Error</td>
<td>89</td>
<td>4954</td>
<td>55.66</td>
<td></td>
</tr>
</tbody>
</table>
Ho: model has no overall relationship or explanatory power  
Reject if $F>F(\alpha, 1,89)= 3.92$ (approx) therefore, we reject Ho  
d) Find the value of r-squared and test if it is significant.  
\[ r^2 = \frac{SSR}{SST} = .6936 \]  
Ho: $r=0$, test using $t$  
\[ T = \frac{r}{\sqrt{1-r^2}} = \frac{.6936}{\sqrt{1-.4811}} = 9.08 \]  
Critical $t= t(\alpha/2, 89) = 1.987$, since 9.08>1.987, we reject the Ho and conclude there is a significant relationship.  
e) For a school with 25 teachers, find the predicted number of sick days and a 95% confidence interval.  
\[ y-hat = 1.803 + .50054*Teachers \]  
\[ = 1.803 + .50054*25 = 14.31 \]  
\[ CI = y-hat +/- t*se(fit) \]  
\[ = 14.31 +/- 1.987*1.67 = 14.31 +/- 3.31 \]  
\[ P(11 < y < 17.62) = .95 \]