MATERIAL COVERED ON FINAL EXAM (the final exam is NOT cumulative):

<table>
<thead>
<tr>
<th>Coursepack sections</th>
<th>Corresponding textbook sections 3rd ed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture Unit 5: all</td>
<td>Chapter 9: §9.2-9.4; Chapter 10: §10.1-10.6; Chapter 11: §11.4-§11.7</td>
</tr>
<tr>
<td>Lecture Unit 6: sections §6.1 and §6.2 (skip section §6.3)</td>
<td>Chapter 13: §13.4; Chapter 14: §14.5</td>
</tr>
<tr>
<td>Lecture Unit 7: all,</td>
<td>Chapter 4: §4.1-§4.6, §4.9</td>
</tr>
<tr>
<td>Lecture Unit 8: §8.1 - §8.4</td>
<td>Chapter 15: §15.1-§15.4; Chapter 17: §17.1-§17.4, Chapter 18: §18.1-§18.2</td>
</tr>
</tbody>
</table>

The above material is covered on webassign homework 9, 10, 11 and 12.

NOTE: tables of the normal distribution and t-distributions will be supplied with the test. WARNING! these sample problems may not cover all topics for which you are responsible on the final exam

1. (linear regression) A copy machine dealer has data on the number of copy machines and the number of service calls in a month at each location. Summary calculations give \( \bar{x} = 8.4 \), \( s_x = 2.1 \), \( \bar{y} = 14.2 \), \( s_y = 3.8 \), and \( r = .86 \). What is the slope of the least squares regression line of number of service calls on number of copiers?
   a. 0.86  b. 1.56  c. 0.48  d. none of these  e. can't tell from information given

2. (linear regression) In the setting of the previous problem, about what percent of the variation in number of service calls is explained by the linear relation between number of service calls and number of machines?
   a. 86%  b. 93%  c. 74%  d. none of these  e. can't tell from information given

3. (correlation) Outdoor temperature influences natural gas consumption for the purpose of heating a house. The usual measure of the need for heating is heating degree days. The number of heating degree days for a particular day is the number of degrees the average temperature for that day is below 65°F, where the average temperature for a day is the mean of the high and low temperatures for that day. An average temperature of 20°F, for example, corresponds to 45 heating degree days. A homeowner interested in switching to solar heating panels collects the following data on her natural gas use for the months October through June, where \( x \) is heating degree days per day for the month and \( y \) is gas consumption per day in hundreds of cubic feet.

<table>
<thead>
<tr>
<th>Month</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>June</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x )</td>
<td>15.6</td>
<td>26.8</td>
<td>37.8</td>
<td>36.4</td>
<td>35.5</td>
<td>18.6</td>
<td>15.3</td>
<td>7.9</td>
<td>0</td>
</tr>
<tr>
<td>( y )</td>
<td>5.2</td>
<td>6.1</td>
<td>8.7</td>
<td>8.5</td>
<td>8.8</td>
<td>4.9</td>
<td>4.5</td>
<td>2.5</td>
<td>1.1</td>
</tr>
</tbody>
</table>

If \( \sum_{i=1}^{9} (x_i - \bar{x})(y_i - \bar{y}) = 291.31 \), \( s_x = 13.42 \), \( s_y = 2.74 \), calculate the correlation coefficient \( r \) and interpret its value; draw a scatterplot of the data.

4. (correlation) Each of the following statements contains a blunder. In each case explain what is wrong.
   a. “There is a high correlation between the sex of American workers and their income.”
   b. “We found a high correlation \( r = 1.09 \) between students' ratings of faculty teaching and ratings made by other faculty members.”
   c. “The correlation between planting rate and yield of corn was found to be \( r = .23 \) bushel.”

5. (linear regression) A study of 1,000 families gave the following results:
   average height of husband = \( \bar{x} = 68 \) inches, \( s_x = 2.7 \) in.;
   average height of wife = \( \bar{y} = 63 \) inches, \( s_y = 2.5 \) in.; \( r = .25 \).
   Estimate the height of a wife when her husband is 72 inches tall.
   a. 63 inches  b. 72 inches  c. 64 inches  d. none of these  e. need more information
6. (linear regression) Outdoor temperature influences natural gas consumption for the purpose of heating a house. The usual measure of the need for heating is *heating degree days*. The number of heating degree days for a particular day is the number of degrees the average temperature for that day is below 65°F, where the average temperature for a day is the mean of the high and low temperatures for that day. An average temperature of 20°F, for example, corresponds to 45 heating degree days. A homeowner interested in switching to solar heating panels collects the following data on her natural gas use for the months October through June, where the explanatory variable \( x \) is heating degree days per day for the month and the response variable \( y \) is gas consumption per day in hundreds of cubic feet.

<table>
<thead>
<tr>
<th>Month</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>June</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x )</td>
<td>15.6</td>
<td>26.8</td>
<td>37.8</td>
<td>36.4</td>
<td>35.5</td>
<td>18.6</td>
<td>15.3</td>
<td>7.9</td>
<td>0</td>
</tr>
<tr>
<td>( y )</td>
<td>5.2</td>
<td>6.1</td>
<td>8.7</td>
<td>8.5</td>
<td>8.8</td>
<td>4.9</td>
<td>4.5</td>
<td>2.5</td>
<td>1.1</td>
</tr>
</tbody>
</table>

Summary statistics are as follows: \( \bar{x} = 21.54, \bar{y} = 5.59, s_x = 13.42, s_y = 2.74, r = .990 \). Calculate the least squares regression line \( y = b_0 + b_1x \) of gas consumption \( y \) on heating degree days \( x \). Draw the regression line on a scatterplot of the data.

7. (confidence interval for mean) As the sample size \( n \) increases, the width of the confidence interval at a fixed confidence level for the population mean tends to:
   a. increase  
   b. decrease  
   c. stay the same

8. (confidence interval for mean) In a study to establish the absolute threshold of hearing, 71 male college freshmen are asked to participate. Each subject is seated in a soundproof room and a 150 Hz tone is presented at a large number of stimulus levels in a randomized order. The subject is instructed to press a button if he detects the tone. The mean for the group was 21.6 db with \( s = 2.1 \). Estimate the mean absolute threshold of all men 19-21 years of age with a 99% confidence interval.

9. (confidence interval for mean) A 95% confidence interval for the mean mileage of the 2005 Mazda 626s is (23.8, 29.6). Then
   a. At the 95% confidence level, we can conclude that the mean mileage for 1990 Mazda 626 exceeds 23.8 miles per gallon.
   b. If the same data were used to construct a 99% confidence interval, the resulting interval would have length shorter than the one given above.
   c. If we repeatedly took samples of the same size, and each time computed a 95% confidence interval, approximately 95% of the resulting intervals would include the true mean mileage for 1990 Mazda 626
   d. None of the above

10. (confidence interval for mean) A random sample of size 49 selected from a population had a sample standard deviation \( s = 8.726 \). Suppose the confidence interval formed for \( \mu \) was \( 63 < \mu < 69 \).
    a. What is the sample mean, \( \bar{x} \)?
    b. What is the estimated standard deviation of \( \bar{x} \)?
    c. What was the confidence coefficient used to form the confidence interval?

11. (confidence interval) A confidence coefficient of 0.99 can correctly be interpreted to mean that
    a. 99% of the time in repeated sampling, intervals using an appropriate formula will contain the sample value.
    b. 99% of the time in repeated sampling, intervals using an appropriate formula will contain the relevant population parameter.
    c. 99% of the time in repeated sampling, intervals using an appropriate formula will contain the sample value as the midpoint of the interval.
    d. 99% of the time in repeated sampling, intervals using an appropriate formula will contain the sample mean as their midpoints.
12. (correlation) If a pair of variables have a strong curvilinear (that is, nonlinear) relationship, which of the following is true?

a. The correlation coefficient will be able to indicate that a nonlinear relationship is present
b. A scatter plot will not be needed to indicate that a nonlinear relationship is present
c. The correlation coefficient will not be able to indicate the relationship is nonlinear
d. The correlation coefficient will be exactly equal to zero

13. (confidence interval for mean) The minimum sample size needed to estimate a population mean within \( \pm 5 \) at a 95% confidence level when the standard deviation is 40 is:
   a. 62   b. 44   c. 1,537   d. 175   e. 246

14. (linear regression) If a residual plot exhibits a curved pattern in the residuals, this means that:
   a. the \( x \) and \( y \) variable should be reversed so that a least squares line is appropriate for this data
   b. a least squares line is not appropriate because there is a nonlinear relation between \( x \) and \( y \)
   c. there is no significant relation between \( x \) and \( y \)
   d. there is a problem with the accuracy of the data

Use the following Excel output to answer questions 15, 16, and 17 below.

```
Regression Statistics
Multiple R 0.8851
R Square 0.7835
Adjusted R Square 0.7474
Standard Error 5.4006
Observations 8

<table>
<thead>
<tr>
<th>ANOVA</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>633.242</td>
<td>633.242</td>
<td>21.711</td>
</tr>
<tr>
<td>Residual</td>
<td>6</td>
<td>175.000</td>
<td>29.167</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>808.242</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>5.93118</td>
<td>1.41989</td>
<td>0.20545</td>
</tr>
<tr>
<td>x</td>
<td>-2.71551</td>
<td>-4.65952</td>
<td>0.00347</td>
</tr>
</tbody>
</table>
```

15. (linear regression) Which of the following is true?
   a. The correlation between \( x \) and \( y \) must be approximately 0.8851
   b. The correlation between \( x \) and \( y \) must be approximately -0.8851
   c. The correlation between \( x \) and \( y \) must be approximately 0.7835
   d. The correlation between \( x \) and \( y \) must be approximately -0.7835

16. (linear regression) Which of the following is true?
   a. \( x \) explains about 88.5% of the variation in \( y \)
   b. \( y \) explains about 88.5% of the variation in \( x \)
   c. \( x \) explains about 78.4% of the variation in \( y \)
   d. \( y \) explains about 78.4% of the variation in \( x \)

17. (linear regression) The predicted value of \( y \) when \( x = 15 \) is approximately
   a. -34.80   b. 86.25   c. 46.66   d. 20.93
Use the following Excel output to answer questions 18, 19, and 20 below.

<table>
<thead>
<tr>
<th>Regression Statistics</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
<td>0.754525991</td>
<td></td>
</tr>
<tr>
<td>Adjusted R Square</td>
<td>0.507782253</td>
<td></td>
</tr>
<tr>
<td>Standard Error</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ANOVA</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>27.3727758</td>
<td>3.910397</td>
</tr>
<tr>
<td>Residual</td>
<td>7</td>
<td>63.56555556</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>8</td>
<td>63.56555556</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.822953737</td>
</tr>
<tr>
<td>X</td>
<td>0.053825623</td>
</tr>
</tbody>
</table>

18. (linear regression) Given this information, what percent of the variation in the y variable is explained by the variation of the independent variable?
   a. About 75 percent
   b. Approximately 57 percent
   c. Can't be determined without having the actual data available
   d. About 25 percent

19. (linear regression) Given this information, what was the sample size used in the study?
   a. 8   b. 18   c. 9   d. 16

20. (linear regression) If the x variable increased by 2 units, then y would change by approximately
   a. 9.64   b. 1.509   c. −0.1076   d. −9.64   e. 0.1076
21. (linear regression) If a least squares line were determined for the data in each scatterplot, which would have the smallest sum of the squares of the residuals?


22. (multiple regression) For all students at Walden University, the prediction equation for \( y = \text{college GPA} \) (range 0-4.0) and \( x_1 = \text{high school GPA} \) (range 0-4.0) and \( x_2 = \text{college board score} \) (range 200-800) is

\[
\hat{y} = 0.20 + 0.50x_1 + 0.002x_2
\]

a. Find the predicted college GPA for students having (i) high school GPA = 4.0 and college board score =800; (ii) \( x_1 = 2.0 \) and \( x_2 = 200 \).

b. A high school student trying to gain admission to Walden University re-takes the college board exam during the summer after his senior year in high school. If he increases his college board score by 100 points, what will be the change in his predicted college GPA? Note that his high school GPA does not change since he has already graduated from high school.

c. For all students with \( x_2 = 500 \), what is the predicted college GPA when high school GPA is the explanatory variable?
23. (multiple regression) For a study of crime in the United States, data for each of the 50 states and Washington, DC was collected on the violent crime rate (per 100,000 citizens), poverty rate (percent of the population), single parent households (percent of all state households), and urbanization (percent of state population living in urban areas). The multiple regression output is shown below where \( y = \text{violent crime rate}, x_1 = \text{poverty rate}, x_2 = \text{single parent households}, \) and \( x_3 = \text{urbanization} \).

### Regression Statistics

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
<td>0.845428508</td>
</tr>
<tr>
<td>R Square</td>
<td>0.714749363</td>
</tr>
<tr>
<td>Adjusted R Square</td>
<td>0.696541875</td>
</tr>
<tr>
<td>Standard Error</td>
<td>132.9791841</td>
</tr>
<tr>
<td>Observations</td>
<td>51</td>
</tr>
</tbody>
</table>

### ANOVA

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>3</td>
<td>2082535.142</td>
<td>694178.4</td>
<td>39.25579</td>
<td>7.4E-05</td>
</tr>
<tr>
<td>Residual</td>
<td>47</td>
<td>831122.7794</td>
<td>17683.46</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>2913657.922</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Coefficients

<table>
<thead>
<tr>
<th></th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-786.7533445</td>
<td>-6.7577</td>
<td>1.91E-08</td>
</tr>
<tr>
<td>poverty</td>
<td>13.40434162</td>
<td>1.762847</td>
<td>0.084428</td>
</tr>
<tr>
<td>single parent</td>
<td>33.02182927</td>
<td>5.979317</td>
<td>2.89E-07</td>
</tr>
<tr>
<td>urbanization</td>
<td>4.401587623</td>
<td>4.449291</td>
<td>5.26E-05</td>
</tr>
</tbody>
</table>

a. What is the least squares prediction equation for the violent crime rate?

b. What is the sum of the squares of the residuals SSE for this multiple regression model?

c. What proportion of the variation in the violent crime rate is explained by poverty rate, single family households, and urbanization?

d. If the poverty rate increased by 1 percent, with single parent households and urbanization unchanged, how would the violent crime rate change?

e. Is this model useful for predicting the violent crime rate?
24. (linear regression) In 1998 sociologists were of the opinion that since 1975 there had been a decrease in the difference in ages at first marriage of husbands and wives. We want to examine data to determine if this decrease is significant. The following data summary and regression results were obtained, where the $x$ variable is year and the $y$ variable is the age difference (husband age – wife age) at first marriage.

<table>
<thead>
<tr>
<th>Variable (x)</th>
<th>Count</th>
<th>Mean</th>
<th>StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year</td>
<td>24</td>
<td>1986.5</td>
<td>7.071</td>
</tr>
<tr>
<td>husband-wife age (y)</td>
<td>24</td>
<td>2.3125</td>
<td>0.249</td>
</tr>
</tbody>
</table>

Regression Statistics

| Multiple R  | 0.680259874 |
| Adj. R Square | 0.462753496 |
| Standard Err | 0.18662649  |
| Observations | 24          |

ANOVA

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Signif F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>0.660002174</td>
<td>0.6600062</td>
<td>10.94965</td>
<td>0.000254659</td>
</tr>
<tr>
<td>Residual</td>
<td>22</td>
<td>0.765247026</td>
<td>0.034829</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>1.42525</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Regression Coefficients

<table>
<thead>
<tr>
<th>Intercept</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
<th>Lower 95%</th>
<th>Upper 95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>49.90213043</td>
<td>10.9324015</td>
<td>4.564608</td>
<td>0.000152</td>
<td>27.72971749</td>
<td>72.75454337</td>
</tr>
<tr>
<td>Year</td>
<td>-0.023956522</td>
<td>-0.005593315</td>
<td>-3.5311</td>
<td>0.000255</td>
<td>-0.035367</td>
</tr>
</tbody>
</table>

a. Interpret the value of the least squares slope $b_1$.
b. What is the value of the test statistic for testing $H_0 : \beta_1 = 0$?
c. For the hypothesis test $H_0 : \beta_1 = 0 \text{ vs } H_a : \beta_1 < 0$, select the choice below that gives the correct P-value and correct conclusion.
   i. The P-value is 0.68; do not reject $H_0 : \beta_1 = 0$; there is no linear relationship since 1975 between year and age difference between husband and wife at first marriage.
   ii. The P-value is 0.000152; reject $H_0 : \beta_1 = 0$; there is evidence that since 1975 the age difference (husband age – wife age) has increased.
   iii The P-value is 0.0001275; reject $H_0 : \beta_1 = 0$; there is evidence that since 1975 the age difference (husband age – wife age) has decreased.
   iv The P-value is 0.0001275; do not reject $H_0 : \beta_1 = 0$; there is no linear relationship since 1975 between year and age difference between husband and wife at first marriage.
d. What is a 95% confidence interval for the slope?
e. Find a 95% confidence interval for the mean difference (husband age – wife age) at first marriage in 1998.
f. Find a 95% prediction interval for the difference (husband age – wife age) for a particular couple getting married for the first time in 1998.
25. (linear regression) Over 6 decades the Gallup Organization has periodically asked the following question:

*If your party nominated a generally well-qualified person for president who happened to be a woman, would you vote for that person?*

Below is a table showing the percentage answering “yes” and the year of the century (37 = 1937).

<table>
<thead>
<tr>
<th>Year</th>
<th>% Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>99</td>
<td>92</td>
</tr>
<tr>
<td>87</td>
<td>82</td>
</tr>
<tr>
<td>84</td>
<td>78</td>
</tr>
<tr>
<td>83</td>
<td>76</td>
</tr>
<tr>
<td>78</td>
<td>73</td>
</tr>
<tr>
<td>75</td>
<td>66</td>
</tr>
<tr>
<td>71</td>
<td>63</td>
</tr>
<tr>
<td>69</td>
<td>57</td>
</tr>
<tr>
<td>67</td>
<td>55</td>
</tr>
<tr>
<td>63</td>
<td>57</td>
</tr>
<tr>
<td>59</td>
<td>54</td>
</tr>
<tr>
<td>58</td>
<td>48</td>
</tr>
<tr>
<td>55</td>
<td>33</td>
</tr>
<tr>
<td>49</td>
<td>33</td>
</tr>
</tbody>
</table>

**summary statistics:**

| \( \bar{x} \) = 67.44 | \( s_x = 16.7 \) | \( \bar{y} = 61.81 \) | \( s_y = 17.19 \) | \( r = .971 \) |

- **a.** Determine the estimates \( b_0 \) and \( b_1 \) of the parameters \( \beta_0 \) and \( \beta_1 \) in the linear model \( y = \beta_0 + \beta_1 x + \epsilon \), where \( x \) is year and \( y \) is the percentage who respond “yes”.
- **b.** Use the least squares line to estimate the percentage of respondents that would say “yes” in 1997.
- **c.** Determine the estimate \( s_e \) of the standard deviation \( \sigma \) of the error component \( \epsilon \) (note that the sum of squares of residuals \( \sum_{i=1}^{10} (y_i - \hat{y}_i)^2 = 255.748 \)).
- **d.** Calculate a 95% confidence interval for the slope \( \beta_1 \). (Note that \( SE(b_1) = \frac{s_e}{\sqrt{n-2} s_x} \))
- **e.** Conduct an appropriate hypothesis test (use \( \alpha = .05 \)) to determine if the year of the century is useful for predicting the percentage of respondents that would answer “yes” to the above question. State the hypotheses, find the value of the test statistic, and state your conclusion based on the P-value or the rejection region.

26. (confidence interval) A random sample of 100 individuals was taken to determine the true percentage of people who smoke in a region of the eastern United States. Forty-six of them said “yes” when they were asked if they smoked. A 95% confidence interval for the true proportion of nonsmokers is:

- **a.** (.46, .54)
- **b.** (.36, .56)
- **c.** (.95, 1.0)
- **d.** (.44, .64)
- **e.** (.00, .95)

27. (sample size) The credit manager of a department store would like to know what proportion of the credit-card customers take advantage of the store's deferred payment plan each year. She would like to estimate this proportion within \( \pm .10 \) at a 90% confidence level, but has no good idea about what this proportion might be. How many customers should she sample?

- **a.** 271
- **b.** 41
- **c.** 17
- **d.** 68
- **e.** cannot be determined

28. (hypothesis test) To test the hypotheses

\[ H_0 : p = .06 \]
\[ H_a : p < .06 \]

a random sample of size 1000 revealed that the number of successes was 40. Compute the P-value and explain how to use it to test the hypothesis.

29. (hypothesis test) A professor claims that 70% of College of Business graduates earn more than $45,000 per year. In a random sample of 300 graduates, 195 earn more than $45,000. Perform a 2-tailed hypothesis test to test the professor's claim; compute the P-value for the test.

30. (hypothesis test) A candy company claims that in a large bag of St. Patrick's Day candy half the candies are green and half are white. You select candies at random from a bag and discover that of the first 50 you eat, only 15 are green. This makes you suspicious of the company's claim that half are green and half are white. Perform a hypothesis test to test the company's claim.
31. (hypothesis test) In the 1980's it was generally believed that congenital abnormalities affected about 5% of the nation's children. Some people believe that the increase in the number of chemicals in the environment has led to an increase in the incidence of abnormalities. A recent study examined 408 children and found that 48 of them showed signs of an abnormality. Is this evidence that the risk has increased?

i) Let \( p \) denote the proportion of children with genetic abnormalities. Choose the correct null and alternative hypotheses.

A. \( H_0 : p = 0.1176 \) vs \( H_A : p > 0.1176 \);  
B. \( H_0 : \hat{p} = .05 \) vs \( H_A : \hat{p} > .05 \);  
C. \( H_0 : p = .05 \) vs \( H_A : p \neq .05 \);  
D. \( H_0 : p = .05 \) vs \( H_A : p > .05 \);  
E. \( H_0 : p = .05 \) vs \( H_A : p < .05 \);  
F. \( H_0 : \hat{p} = 0.1176 \) vs \( H_A : \hat{p} \neq 0.1176 \).

ii) What is the value of the test statistic?

iii) What is the \( P \)-value?

iv) What is your conclusion?

32. (confidence interval) The minimum sample size needed to estimate a population mean within \( \pm 5 \) at a 95% confidence level when the standard deviation is 40 is:

   a. 62    b. 44    c. 1,537    d. 175    e. 246

33. (multiple regression, global F-test, hypothesis test for individual coefficients) Suppose you fit the quadratic model \( y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon \) to a set of \( n = 20 \) data points and found \( R^2 = 0.90 \), \( \sum_{i=1}^{20} (y_i - \bar{y})^2 = 37.14 \), and \( SSE = 3.87 \).

33a. (global F-test) Perform a global F-test to test if there sufficient evidence to indicate that the model contributes information for predicting \( y \) by answering questions i) - iv) below.

i) The null hypothesis is \( H_0 \):
   a) \( \beta_0 = \beta_1 = \beta_2 = 0 \)   
   b) \( \beta_0 = \beta_1 = \beta_2 \neq 0 \)   
   c) \( \beta_1 = \beta_2 = 0 \)   
   d) \( \beta_1 = \beta_2 \neq 0 \)   
   e) \( \beta_0 = \beta_1 = \beta_2 > 0 \)

ii) The alternative hypothesis is \( H_A \):
   a) at least one of \( \beta_0, \beta_1, \beta_2 > 0 \)   
   b) at least one of \( \beta_0, \beta_1, \beta_2 \neq 0 \)   
   c) at least one of \( \beta_1, \beta_2 > 0 \)  
   d) at least one of \( \beta_1, \beta_2 \neq 0 \)   
   e) at least one of \( \beta_1, \beta_2 < 0 \)

iii) What is the value of the test statistic?
   a) 65.13    b) 126.4    c) 73.06    d) 48.88    e) 83.23

iv) If the \( P \)-value = 0.005, what is the conclusion for this test?
   a) Do not reject the null hypothesis. Conclude that there is not sufficient evidence that the model is useful for predicting \( y \).
   b) Reject the null hypothesis. Conclude that there is not sufficient evidence that the model is useful for predicting \( y \).
   c) Reject the null hypothesis. Conclude that at least one of \( \beta_1, \beta_2 \) is nonzero and that the model is useful for predicting \( y \).
   d) Reject the null hypothesis. Conclude that both \( \beta_1 \) and \( \beta_2 \) are greater than 0 and that the model is useful for predicting \( y \).
   e) Do not reject the null hypothesis. Conclude that since \( \beta_1 = 0 \) and \( \beta_2 = 0 \), then \( \beta_0 \) must be significantly different from 0.
33b. (t-test for individual coefficient) What null and alternative hypotheses would you test to determine whether upward curvature exists?
   a)  \( H_0 : \beta_2 = 0, \ H_a : \beta_2 \neq 0 \)  
   b)  \( H_0 : \beta_2 = 0, \ H_a : \beta_2 < 0 \)  
   c)  \( H_0 : \beta_2 = 0, \ H_a : \beta_2 > 0 \)  
   d)  \( H_0 : \beta_2 > 0, \ H_a : \beta_2 \neq 0 \)  
   e)  \( H_0 : \beta_2 > 0, \ H_a : \beta_2 < 0 \)  

33c. (t-test for individual coefficient) What null and alternative hypotheses would you test to determine whether downward curvature exists?
   a)  \( H_0 : \beta_2 = 0, \ H_a : \beta_2 \neq 0 \)  
   b)  \( H_0 : \beta_2 = 0, \ H_a : \beta_2 < 0 \)  
   c)  \( H_0 : \beta_2 = 0, \ H_a : \beta_2 > 0 \)  
   d)  \( H_0 : \beta_2 > 0, \ H_a : \beta_2 \neq 0 \)  
   e)  \( H_0 : \beta_2 > 0, \ H_a : \beta_2 < 0 \)  

34. (multiple regression with interaction) Consider the following prediction equation with an interaction term:
\[ \hat{y} = 3.6 + 1.2x_1 + 2.4x_2 + 0.2x_1x_2. \]  
When \( x_2 \) is held fixed at 3, how much does the predicted value of \( y \) change when \( x_1 \) increases by 1 unit?
   a)  1.8  
   b)  10.8  
   c)  11.4  
   d)  4.2  
   e)  7.2  

35. (multiple regression with interaction) Consider the printout for an interaction regression between a response variable \( y \) and two explanatory variables \( x_1 \) and \( x_2 \).

\begin{center}
\begin{tabular}{lcccc}
\hline
  & df & SS & MS & F & Significance F \\
\hline
Regression & 3 & 3393.677324 & 1131.225775 & 9391.974782 & 2.11084E-11 \\
Residual & 6 & 0.722675987 & 0.120445998 & & \\
Total & 9 & 3394.4 & & & \\
\hline
\end{tabular}
\end{center}

a. Write the prediction equation for the interaction model.
b. Test the overall utility of the interaction model using the global F-test at \( \alpha = .05 \).
c. Test the hypothesis (at \( \alpha = .05 \)) that \( x_1 \) and \( x_2 \) interact positively.
d. Estimate the change in \( y \) for each additional 1-unit increase in \( x_1 \) when \( x_2 = 9 \).

36. (multiple regression, indicator variables) An elections officer wants to model voter turnout \( y \) in a precinct as a function of type of election: national or state.

36a Write a model for voter turnout, \( y \), as a function of type of election.
   i.  \( y = \beta_0 + \beta_1 x + \epsilon \), where \( x = 1 \) if the election is national, \( x = 0 \) if the election is local.
   ii.  \( y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \), where \( x_1 = 1 \) if the election is national, \( x_1 = 0 \) if the election is not national; \( x_2 = 1 \) if the election is local, \( x_2 = 0 \) if the election is not local.
   iii.  \( y = \beta_0 + \beta_1 x + \epsilon \), where \( x = \) voter turnout.
   iv.  \( y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon \), where \( x = \) voter turnout.

36b Interpret the value of \( \beta_1 \).
   i)  the rate of increase in the voter turnout for national elections
   ii)  the mean voter turnout for national elections
   iii) the difference between the mean voter turnout in national and local elections
   iv)  the difference between the mean voter turnouts in national elections 1 year apart
37. (multiple regression, indicator variables) An elections officer wants to model voter turnout \( y \) in a precinct as a function of the type of precinct: urban, suburban, or rural. The elections officer uses the following regression model:

\[
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon
\]

where

- \( x_1 = 1 \) if urban, 0 if not,
- \( x_2 = 1 \) if suburban, 0 if not.

Interpret the value of \( \beta_2 \).

i) the difference between the mean voter turnout for suburban and rural precincts

ii) the rate of increase in voter turnout \( y \) for suburban precincts

iii) the mean voter turnout for suburban precincts

iv) the difference between the mean voter turnout for suburban and urban precincts

************************************************************

**SOLUTIONS**

1. b. since \( b = r \left( \frac{s_y}{s_x} \right) = .86 \left( \frac{3.4}{5.2} \right) = 1.56 \).  
2. c. since \( r^2 = (.86)^2 = .74 \)

3. 
\[
r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}} = .990.
\]
There is a strong positive linear relationship between heating degree days and gas consumption.

**SCATTERPLOT: Heat. Deg. Days (x), Gas Consump.(y)**

4. a. The correlation we are studying measures the linear relationship between 2 quantitative variables; sex is a categorical variable.

b. \(-1 \leq r \leq 1\) is violated.

c. \( r \) has no units.

5. c. The husband is 4 inches, or 4/2.7 = 1.5 standard deviations above the mean husband height.

The wife's height is predicted to be above average by \( .25 \times 1.5 = .4 \) standard deviations, or .4 \times 2.5 inches = 1 inch. (Recall \( b = r(s_y/s_x) \))

6. 
\[
b_1 = r \frac{s_y}{s_x}; \quad b_0 = \overline{y} - b_1 \overline{x}; \quad b_1 = .202; \quad b_0 = \overline{y} - b_1 \overline{x} = 1.23
\]
7. **b.** $21.6 \pm 2.648 \left( \frac{2.1}{\sqrt{12}} \right)

9. a and c. **10. a. 66.** b. $(8.726)/7 = 1.2465$. c. $t_{48}^* = 2.4066$; confidence coeff. 98%


22. a.) $\bar{y} = 0.20 + 0.50x + 0.002x = 3.8$; ii) $\bar{y} = 0.2 + 0.50x + 0.002x = 1.6$.

b. predicted college GPA will increase by $0.002 \cdot 100 = 0.2$.

c. $\bar{y} = 0.20 + 0.50x_1 + 0.002 \cdot 500 = 0.20 + 0.50x_1 + 1 = 1.2 + 0.50x_1$.

23. a. **violent crime rate** = -786.7533445 + 13.40434162poverty rate + 33.02182927*single parent + 4.041587623*urbanization

b. SSE = 831,122.7794  c. $R^2 = 0.71474936$  d. the violent crime rate will increase by approx. 13.4 per 100,000 citizens.

e. analyze the model by examining the F test statistic for the F test; the hypotheses are $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$, $H_A : \text{at least one of the } \beta_i \text{'s is not zero.}$

From the regression output: $F = \frac{MSE}{SSE} = \frac{694178.4}{17085.46} = 39.25579$; the "Significance F" (that is, the P-value) is less than 0.05. Therefore, reject $H_0$ and conclude that at least one of the $\beta_i$'s is not zero.

24. a. $b_1 = -0.024$ (approximately); this means that each year since 1975 the average difference (husband age – wife age) has decreased by .024.

b. $t = \frac{b_1}{SE(b_1)} = \frac{-0.023956}{0.00059} = -4.35311$

c. The P-value given in the Excel output is always for a 2-tail test; since we are conducting a 1-tail test $H_0 : \beta_1 < 0$, the P-value is $0.0001275$.

d. $(-0.0353697, -0.01254335)$ from the output; notice that the interval is entirely negative.

e. use $\hat{y}_{998} \pm t_{n-2}SE(\hat{\mu}_{998})$;

for the calculations below, note that from the output we have $s_e = 0.18662649$;

$\hat{y}_{998} = 49.90213043 - 0.023956522(1998) = 2.037$

$SE(\hat{\mu}_{998}) = \sqrt{SE^2(b_1) \times (x_9 - \bar{x})^2 + \frac{s_e^2}{n}}$

$t_{n-2}SE(\hat{\mu}_{998}) = 2.074(.03786889) = 0.1532404$

so $\hat{y}_{998} \pm t_{n-2}SE(\hat{\mu}_{998}) = 2.037 \pm 0.1532404 \Rightarrow (1.883796, 2.190204])$

g. use $\hat{y}_{998} \pm t_{n-2}SE(\hat{\mu}_{998})$;

for the calculations below, note that from the output we have $s_e = 0.18662649$;

$\hat{y}_{998} = 49.90213043 - 0.023956522(1998) = 2.037$

$SE(\hat{\mu}_{998}) = \sqrt{SE^2(b_1) \times (x_9 - \bar{x})^2 + \frac{s_e^2}{n}}$

$t_{n-2}SE(\hat{\mu}_{998}) = 2.074(.03786889) = 0.1532404$

so $\hat{y}_{998} \pm t_{n-2}SE(\hat{\mu}_{998}) = 2.037 \pm 0.1532404 \Rightarrow (1.883796, 2.190204])$

25. a. $b_1 = \frac{r_{xx}}{s^2} = .9711 \frac{17.10}{14} = .99949; b_0 = \bar{y} - b_1 \bar{x} = 61.81 - .99949(67.44) = -5.59358$

b. $\hat{y}_{97} = -5.59358 + .99949(97) = 91.36$

c. $s_e = \sqrt{ \frac{SSE}{n-2} } = \sqrt{ \frac{255.748}{14} } = 4.274;$
d. \[ SE(b_1) = \frac{s_{b_1}}{\sqrt{n-1}} = \frac{1.274}{\sqrt{15 \times 16.7}} = .0661; \]

Confidence interval is \( b_1 \pm t_{n-2}SE(b_1) = .99949 \pm 2.145(.0661) = .99949 \pm .14178 \Rightarrow (.85771, 1.14127) \)

e. \( H_0 : \beta_1 = 0 \) vs. \( H_A : \beta_1 \neq 0; \)

test statistic \( t = \frac{b_1}{SE(b_1)} = \frac{.99949}{.0661} = 15.12; \)

For \( \alpha = .05 \), the rejection region is \( t > 2.145 \) and \( t < -2.145 \) \((n-2 = 16 - 2 = 14 \; d.f.)\); the \( P \)-value is 0 to nine decimal places.

Conclusion: since the test statistic is in the rejection region, reject \( H_0 : \beta_1 = 0 \) and conclude that year is useful for predicting the percentage of respondents that will answer “yes” to the question.

26. \( \hat{p} = \frac{40}{1000} = .04; \) test statistic \( z = \frac{.04-.06}{\sqrt{.04 \times .06}} = -2.67; \)

\[ P-value = P(z < -2.67) = .0038. \] Since the \( P \)-value is less than .05, reject \( H_0 \) and conclude that \( p < .06. \)

29. \( H_0 : p = .7, \; H_A : p \neq .7; \)

\[ \hat{p} = \frac{195}{300} = .65; \; \text{SD}(\hat{p}) = \sqrt{\frac{(1)(.3)}{300}} = .02646; \]

test statistic \( z = -1.89; \) \( P-value = 2P(z > | -1.89|) = 2(.0294) = .0588 \)

do not reject \( H_0; \) there is no evidence to support the claim that the proportion of graduates who earn more than \$45,000 differs significantly from .70.

30. \( H_0 : p = .5, \; H_A : p < .5, \) where \( p \) is the proportion of candies that are green.

\[ \hat{p} = \frac{15}{30} = .30; \; \text{SD}(\hat{p}) = \sqrt{\frac{(.5)(1-.5)}{30}} = \frac{.50}{\sqrt{30}} = .0707; \]

test statistic: \( z = \frac{.30-.50}{.0707} = -2.83 \)

\[ P-value = P(z < -.283) = .0023 \]

Conclusion: since the \( P \)-value is less than .05, we reject the null hypothesis \( H_0 : p = .5 \) and conclude that the proportion green candies is less than .5.

31. i) \[ B \]

ii) \[ D \]

iii) \[ p = \frac{48}{108} = .1176; \; \text{SD}(\hat{p}) = \sqrt{\frac{(.05)(1-.05)}{108}} = \sqrt{\frac{.0025}{108}} = .0108; \]

\[ z = \frac{.1176-.05}{.0108} = 6.26; \]

iv) \[ J \]

32. a) \[ C \]

b) \[ B \]

c) \[ A \]

Test statistic is \( F = \frac{MSR}{MSE} \); since \( n = 20 \) and \( k = 2, \)

\[ \text{MSE} = \frac{SSE}{n-k-1} = \frac{3.87}{17} = .2277 \]

To determine \( MSR, \) note that since

\[ SST = \sum_{i=1}^{20} (y_i - \bar{y})^2 = 37.14, \] and \( SST = SSR + SSE \) (that is, Total Sum of Squares = Sum of Squares for Regression + Sum of Squares of the Residuals; see Annotated Excel Regression Output in the Lecture Handouts section of our class webpage), then

\[ SSR = SST - SSE = 37.14 - 3.87 = 33.27. \] So \[ \text{MSR} = \frac{SSR}{k} = \frac{33.27}{2} = 16.635 \] and

\[ F = \frac{MSR}{MSE} = \frac{16.635}{.2277} = 73.06 \]

33a iv) \[ C \]

b) \[ B \]

c) \[ A \]

35a \[ \hat{y} = 16.7220 - 3.0373x_1 - 1.0465x_2 + 4.0717x_1x_2 \]

35b Test the null hypothesis \( H_0 : \beta_1 = \beta_2 = \beta_3 = 0. \) The \( F \)-statistic is 9.394 with \( P \)-value 2.11 x 10^{-11}; since the \( P \)-value < .05, reject \( H_0 \) and conclude that the model is useful for predicting \( y. \)

35c Test \( H_0 : \beta_3 = 0 \) vs. \( H_A : \beta_3 > 0. \) The test statistic is \( t = 9.17 \) and the 2-tailed \( P \)-value = 9.48 x 10^{-5}, so the \( P \)-value for this 1-tailed test is \( P \)-value = 4.74 x 10^{-5}. Reject \( H_0 : \beta_3 = 0 \) and conclude that \( x_1 \) and \( x_2 \) interact positively.

35d 33.608 36a iii) 36b iii) 37 i)