You MUST work alone – no tutors; no help from classmates. Email me or see me with questions. You will receive a score of 0 if this rule is violated.

EPE/EDP 660 EXAM II: Summer 2013

[5 points] Minitab (or other approved software) output must be included. It must be clearly labeled, with all answers clearly identified. In addition, you must include a copy of your session window. Do NOT include a copy of the worksheet.

Directions: Read each question before responding. In order to receive partial credit, work must be shown.

PART A (17 POINTS): MULTIPLE CHOICE/ FILL IN THE BLANK (with best choice) {1 point per blank}

1. The _____ measures the direction and strength of the linear association between two variables.
   A. Standard Deviation
   B. Alpha
   C. Type II Error
   D. Correlation

2. Predicting y when the x values are outside the range of experimentation is _____.
   A. Extension
   B. Type I Error
   C. Extrapolation
   D. Type II Error

3. Which of the following assumptions about the error term are true? _____
   A. It is equal to zero
   B. It is random
   C. Standard deviation equals 1
   D. All of the above
4. When two independent variables are highly correlated with one another, it is reasonable to suspect an issue with _____.
   A. Random Error
   B. Multicollinearity
   C. Sample Size
   D. Extrapolation

5. Predicting an outcome (dependent variable) based upon several independent variables simultaneously is known as _____ regression
   A. Simple Linear
   B. Cross sectional
   C. Multiple
   D. Quadratic

6. Two models are _____ if both contain the same terms and one has at least one additional term.
   A. Nested
   B. Cross Listed
   C. Interacting
   D. Mutually Exclusive

True or False: Write A for True and B for False.

7. The ε is a random variable with mean = 1 and variance σ^2. _____
   A. True
   B. False

8. To test if all of the slope parameters are zero, we use an F–test. _____
   A. True
   B. False
9. The value of SST does not change with the model, as it depends \textit{only} on the values of the dependent variable $y$. _____
A. True
B. False

10. SSE \emph{increases} as variables are added to a model, and SSM \emph{decreases} by the same amount. _____
A. True
B. False

11. Once an interaction has been deemed important in a model, we can remove any associated first-order terms in the model if their $p$-values are not significant. _____
A. True
B. False

| Table A. |
|------------------|---|---|---|
| Source of Variation | DF | SS | MS |
| Model              | (a) | SSM | MSM |
| Error              | (b) | SSE | MSE |
| Total              | n-1 | (c) |     |

Complete Table A, presented above in the box, by filling in the missing components.

12. Fill in the (a) component. _____.

13. Fill in the (b) component. _____.

14. Fill in the (c) component _____.
Equation A: \( E(y) = \beta_0 + \beta_1 x + \beta_2 x^2 \)

Based on Equation A, presented in the box above, label each of the following components:

15. What is the \( y \)-intercept when \( x \) is 0? ______
   A. \( E(y) \)
   B. \( \beta_0 \)
   C. \( \beta_1 \)
   D. \( \beta_2 \)

16. What is the shift parameter? ______
   A. \( E(y) \)
   B. \( \beta_0 \)
   C. \( \beta_1 \)
   D. \( \beta_2 \)

17. What is the rate of curvature? ______
   A. \( E(y) \)
   B. \( \beta_0 \)
   C. \( \beta_1 \)
   D. \( \beta_2 \)
PART B: Data Analysis (31 POINTS)

Detailed interviews were conducted with over 1,000 street vendors in the city of Puebla, Mexico, in order to study the factors influencing vendors’ incomes (World Development, Feb. 1998). Vendors were defined as individuals working in the street, and included vendors with carts and stands on wheels and excluded beggars, drug dealers, and prostitutes. The researchers collected data on gender, age, hours worked per day, annual earnings, and education level. A subset of these data appears in the table; the data set (a sample reduced) that you will be working with is posted as Take Home 2.

<table>
<thead>
<tr>
<th>Vendor Number</th>
<th>Annual Earnings</th>
<th>Age</th>
<th>Hours worked per day</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>2841</td>
<td>29</td>
<td>12</td>
<td>M</td>
</tr>
<tr>
<td>53</td>
<td>1876</td>
<td>21</td>
<td>8</td>
<td>F</td>
</tr>
<tr>
<td>263</td>
<td>3065</td>
<td>40</td>
<td>11</td>
<td>M</td>
</tr>
<tr>
<td>281</td>
<td>3670</td>
<td>50</td>
<td>11</td>
<td>F</td>
</tr>
</tbody>
</table>

1. For each variable above, use Minitab to describe it. You may use descriptive statistics, a graphical summary or even a frequency table. Choose descriptive statistics with level of measurement in mind. {5 points}

2. Produce a matrix plot and correlation estimates for the interval/ratio variables. {3 points}

3. Compute a simple linear regression with independent variable, hours worked per day, to estimate mean annual earnings. Be sure to include the ANOVA table and 4-in-1 residuals plot. {2 points}

4. Produce a fitted line plot for the equation produced in #3 with a 95% prediction interval. {2 points}

5. Run the analysis as a multiple regression, least-squares regression equation, R-square, and coefficient estimates for estimating mean annual earnings as a function of age (x_1) and hours worked (x_2). Be sure to include the ANOVA table and 4-in-1 residuals plot. {2 points}

6. Re-run the model in #5, so that it includes the interaction term (you may choose to create a new variable in the worksheet or just create the necessary interaction in your Minitab equation). Be sure to include the ANOVA table and 4-in-1 residuals plot. {3 points}
   a. Plot the potential interaction. {2 points}

7. Run a regression analysis to fit the quadratic model for estimating mean annual earnings as a function of age (x_1) and hours worked squared (x_2)^2 (you may choose to create the squared term in the worksheet or just create the necessary squared term in your Minitab equation). Be sure to include the ANOVA table and 4-in-1 residuals plot. {4 points}

8. Compute the regression equation, R-square and coefficient estimates for the complete second order model, \( E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_1 x_2 + \beta_5 x_1^2 x_2 \), for estimating mean annual earnings as a function of age (x_1) and hours worked (x_2). Be sure to include the ANOVA table and 4-in-1 residuals plot. {4 points}

9. Create a dummy (indicator) variable for Gender. Compute the first order least-squares regression equation, R-square and coefficient estimates for estimating mean annual earnings as a function of hours worked (x_2) and gender (x_3). Be sure to include the ANOVA table and 4-in-1 residuals plot. {4 points}
Part C: Short Answer. Use your analysis from part B to help complete this section. (45 points)

1. How many variables are there in the Take-Home data set? Identify the level of measurement for each variable and explain your reasoning. If an interval or ratio variable, discuss the distribution of the variable. Which variable appears to have the greatest variation? Explain. {8 points}

2. Consider B#3 – Based on the matrix plot and common knowledge (and any additional information you feel would support your response), do you feel that simple linear regression is a sound choice of analysis in this setting? Explain. {3 points}

3. Consider B#5 – Minitab reports both the R-square and the R-square (adjusted) values for models. Explain why R-square (adjusted) is a better estimate as compared to R-square for this model. {3 points}

4. Consider B#3 and B#5 – Using the ANOVA component of the output, report Sums of Squares Model (SSM), Sums of Squares Error (SSE), and Sums of Squares Total (SST) for the simple linear regression model B#3 and the multiple regression model B#5. Provide a brief explanation for SSE and SSM. Explain why SST is the same for both models. {5 points}

5. Consider B#5 – Do you conclude that both age and hours worked are statistically significant predictors of annual earnings? Explain. {4 points}

6. Consider B#6 – Test the null hypothesis test that the interaction term is not a statistically significant predictor of annual earnings. Test using α= .05. Is this a reasonable test to conduct? Explain. Is there any reason to believe that an interaction term is meaningful in this model? Explain {5 points}

7. Consider B#7 – Is this model fitting a concave up or concave down model? Explain. {3 points}

8. Consider B#8 –
   a. Conduct a hypothesis test of the global utility of the model. Give the degrees of freedom and the p-value for this test. What do you conclude at α= .05? {3 points}

   b. How would you test the usefulness of the interaction and quadratic terms in the model for predicting annual earning? {3 points} (You do not have to actually conduct the test.)

9. Consider B#9 – Explain the meaning for the coefficient for gender. {3 points}

10. Which model would you select as the best fitting, considering parsimony? Defend your choice. {5 points}