Instructions: This exam contains 34 Multiple Choice questions. Each question is worth 3 points, for a total of 102 points, so there is a bonus question on the exam. One point will be deducted if you do NOT correctly bubble in your name, UFID number and Test Form Code on the scantron sheet. Select the best answer among the alternatives given.

You may write whatever you want on this test, but only the answers bubbled in the scantron sheet will be graded. This test MUST BE SUBMITTED to the instructors together with the scantron sheet for you to receive a grade on the exam. Honor pledge: "On my honor, I have neither given nor received unauthorized aid on this examination."

Signature: __________________________

1. In Multiple Regression we want:
   a) the predictors to be correlated with the response
   b) the response variables to be correlated with each other
   c) no correlations between any of the variables
   d) the predictor variables to be correlated with each other
   e) both a) and b)

2. In Multiple Regression, $R^2=100\%$:
   a) would definitely be a good thing, since the purpose is to maximize $R^2$
   b) could indicate that we have too many predictor variables
   c) is impossible since we can't predict a variable without error
   d) may happen in cases when we have several response variables
   e) all of the above

3. Multiple Regression, in general:
   a) is not as straightforward as Simple Linear Regression even when done with a statistical computing package
   b) may require the investigator to narrow down the number of predictor variables in the model
   c) can have different interpretations of the coefficients, depending on the type of predictor variables
   d) all of the above
   e) none of the above

4. The ANOVA test in Regression is always a test to determine:
   a) if the slope is significantly different from zero
   b) if there are any good predictors in the model
   c) both a) and b)
   d) neither a) nor b)

5. A Normal Probability Plot of the residuals:
   a) can be used to check the assumptions of Normality and constant variance
   b) should show an S-shaped curve if the assumptions are satisfied
   c) both a) and b)
   d) neither a) nor b)

6. The Least Squares Regression method:
   a) is used to find the fitted equation in simple linear regression but not in multiple regression.
   b) is the best method to fit the model because it's not much affected by outliers.
   c) finds a fitted equation that minimizes the sum of squared residuals.
   d) all of the above
   e) none of the above

7. Data was collected to predict a student's height based on their gender and each of their parent's height. We suspect that both parents' height will influence the student's height, but that female students will be closer in height to their mother's and male students will be closer to their father's. In order to allow the parent of the same gender as the student to have a higher influence, we should consider adding:
   a) squared terms
   b) dummy variables
   c) interaction terms
   d) multicollinearity
   e) influential points
Questions 8 – 12 The three graphs below represent three data sets that differ only by one observation.

8. Which of the graphs represents the data set that will have the strongest correlation?
   a) graph (i)   b) graph (ii)   c) graph (iii)   d) all will be equal   e) impossible to tell

9. Which of the graphs represents a data set that has an influential outlier?
   a) graph (i)   b) graph (ii)   c) graph (iii)   d) graphs (ii) and (iii)   e) graphs (i) and (iii)

10. Which of the graphs represents a data set with an outlier that weakens the correlation but does not affect the slope much?
    a) graph (i)   b) graph (ii)   c) graph (iii)   d) graphs (i) and (ii)   e) graphs (i) and (iii)

11. Which of the graphs represents a data set that will have a negative slope on its least squares regression line?
    a) graph (i)   b) graph (ii)   c) graph (iii)   d) graphs (i) and (iii)   e) none of them

12. Which of the graphs represents a data set that is good for conducting a regression analysis?
    a) graph (i)   b) graph (ii)   c) graph (iii)   d) all of them   e) none of them

13. Which of the following problems cannot be detected by a residual plot?
    a) outliers   b) non-constant variance   (c) multicollinearity   d) influential points   e) dummy variables

14. If the plot on the right represents the fitted line for the original data set, which of the following graphs is the residual plot for the same data?
    a)   b)   c)   d)   e) Not a problem

15. In the regression model $Y = \alpha + \beta_1 x + \beta_2 x^2 + \epsilon$ the change in $Y$ for a one unit increase in $x$:  
    a) will always be the same amount, $\alpha$  
    b) will always be the same amount, $\beta$  
    c) will depend on the error term  
    d) will depend on the level of $x$
16. In an ANOVA table for Regression, a larger test statistic:
   a) is always associated with a larger p-value ×
   b) implies that all the predictors are good ×
   c) gives stronger evidence that the model good ✓
   d) all of the above
e) none of the above

17. The null hypothesis for ANOVA in Regression states that:
   a) all the parameters are equal to zero
   b) the coefficients of all the predictors are equal to zero ×
   c) all the parameters are equal to each other
d) the coefficients of all the predictors are equal to each other ×
e) all of the above

18. Dummy variables in Regression:
   a) are coded with zeroes and ones
   b) can be coded with any numbers ✗
   c) may be words or categories
   d) all of the above
   e) none of the above

Questions 19 – 34 House prices can vary considerably depending on several variables. We have regression output for various models that use data on houses recently sold in Gainesville to predict the selling price of a house based on some combination of the following: the size of the house (measured in square feet), the size of the lot (also in square feet), the number of bedrooms, the number of bathrooms, and the location of the house (1 if the house is in the NW quadrant of Gainesville, 0 otherwise). Questions always refer to the output on that page, unless specified otherwise.

Best Subsets Regression: price versus size, lot, ...

<table>
<thead>
<tr>
<th>Vars</th>
<th>R-Sq</th>
<th>R-Sq(adj)</th>
<th>Mallows</th>
<th>Cp</th>
<th>Set</th>
<th>s</th>
<th>s</th>
<th>WW</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>58.0</td>
<td>57.5</td>
<td>58.0</td>
<td>36730</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>50.9</td>
<td>50.4</td>
<td>83.6</td>
<td>39672</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>71.1</td>
<td>70.5</td>
<td>11.7</td>
<td>30588</td>
<td>X X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>72.9</td>
<td>72.1</td>
<td>7.1</td>
<td>29775</td>
<td>X X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>72.7</td>
<td>71.8</td>
<td>8.0</td>
<td>29912</td>
<td>X X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>73.8</td>
<td>72.7</td>
<td>5.8</td>
<td>29429</td>
<td>X X X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>73.5</td>
<td>72.4</td>
<td>6.9</td>
<td>29999</td>
<td>X X X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>74.3</td>
<td>72.9</td>
<td>6.2</td>
<td>29327</td>
<td>X X X X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>74.0</td>
<td>72.7</td>
<td>7.1</td>
<td>29468</td>
<td>X X X X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>74.6</td>
<td>73.0</td>
<td>7.0</td>
<td>29303</td>
<td>X X X X X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

19. Based on the output above, which variable is the single best predictor of price?
   a) size
   b) lot
   c) bedrooms
   d) baths
   e) NW

20. Of all the models that appear on this output, the best one has:
   a) 6 predictors, because it has the highest R²
   b) 6 predictors, because it has the highest R² adjusted
   c) 1 predictor, because it has the highest S
   d) 1 predictor, because it has the highest Mallow’s Cp
   e) 1 predictor, because it has the highest p-value

21. Best Subsets Regression is a method that:
   a) starts by finding the best predictor and adds new predictors to the model one at a time
   b) starts by including all the predictors and eliminates them from the model one at a time
   c) fits every model with one, two, and up to p predictors and prints the best ones of each size
   d) fits every model with a different response variable, and prints the best ones for each predictor
OUTPUT 1 - Regression Analysis: price versus Bedrooms, NW

<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>2259</td>
<td>26542</td>
<td></td>
<td>0.932</td>
</tr>
<tr>
<td>Bedrooms</td>
<td>33518</td>
<td>7850</td>
<td>0.000</td>
<td>✓</td>
</tr>
<tr>
<td>NW</td>
<td>32292</td>
<td>11887</td>
<td>0.008</td>
<td>✓</td>
</tr>
</tbody>
</table>

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>60815122229</td>
<td>30407561115</td>
<td>0.000</td>
<td>✓</td>
</tr>
<tr>
<td>Residual Error</td>
<td>97</td>
<td>2.53617E+11</td>
<td>26141612344</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>99</td>
<td>3.14433E+11</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

OUTPUT 2 - Regression Analysis: price versus Bedrooms, NW, Bd*NW

INTERACTION

<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>137100</td>
<td>73698</td>
<td>0.066</td>
<td></td>
</tr>
<tr>
<td>Bedrooms</td>
<td>-9700</td>
<td>23399</td>
<td>0.679</td>
<td></td>
</tr>
<tr>
<td>NW</td>
<td>-14186</td>
<td>77775</td>
<td>0.132</td>
<td></td>
</tr>
<tr>
<td>Bd*NW</td>
<td>48525</td>
<td>24794</td>
<td>0.053</td>
<td></td>
</tr>
</tbody>
</table>

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>3</td>
<td>70546004385</td>
<td>23.52E+9</td>
<td>0.000</td>
<td>✓</td>
</tr>
<tr>
<td>Residual Error</td>
<td>96</td>
<td>2.43887E+11</td>
<td>2.54E+9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>99</td>
<td>3.14433E+11</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

22. What type of graph would the regression models that have been fitted in Output 1 and Output 2 give?
   a) both models give parallel lines
   b) output 1 gives parallel lines and output 2 gives non-parallel lines
   c) both models give non-parallel lines
   d) output 1 gives non-parallel lines and output 2 gives parallel lines
   e) none of the above

23. How many predictor variables are in each of the models fitted in Output 1 and Output 2?
   a) both models have 3 predictors
   b) output 1 has 3 predictors and output 2 has 2 predictors
   c) both models have 2 predictors
   d) output 1 has 2 predictors and output 2 has 3 predictors
   e) none of the above

24. Are the models fitted in Output 1 and Output 2 any good?
   a) both models are good
   b) both models are bad
   c) output 1 is good, but in output 2 the model should be revised by eliminating two of the variables
   d) output 2 is good, but in output 1 the model should be revised by eliminating one of the variables
   e) both c) and d) are correct

25. What is the baseline for the model fitted in Output 2?
   a) prices
   b) bedrooms
   c) NW
   d) not NW
   e) $137,100

26. What is the standard deviation of the points around the fitted values for the model fitted in Output 1?
   a) 9.26
   b) 51133
   c) 48493
   d) 1.96
   e) -1.52

27. What is the test statistic to determine if any of the predictors are good in the model fitted in Output 2?
   a) 9.26
   b) 51133
   c) 48493
   d) 1.96
   e) -1.52

28. What is the test statistic to determine if there is a significant effect of NW in the model fitted in Output 2?
   a) 9.26
   b) 51133
   c) 48493
   d) 1.96
   e) -1.52

29. Use the model in Output 2 to predict the selling price of a 4 bedroom house located in NW Gainesville:
   a) $98,300
   b) $75,914
   c) $174,214
   d) $292,400
   e) $331,200

\[
\text{Price} = 157100 + 9700(4) - 118186(1) + 48525(4)(1)
\]
Regression Analysis: price versus size, lot, Bedrooms, Baths, NW

<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>-258.05</td>
<td>163.04</td>
<td>-1.58</td>
<td>0.117</td>
</tr>
<tr>
<td>size</td>
<td>51.466</td>
<td>6.680</td>
<td>7.65</td>
<td>0.000</td>
</tr>
<tr>
<td>lot</td>
<td>2.5215</td>
<td>0.4335</td>
<td>5.82</td>
<td>0.000</td>
</tr>
<tr>
<td>Bedrooms</td>
<td>-4868</td>
<td>5621</td>
<td>-0.87</td>
<td>0.385</td>
</tr>
<tr>
<td>Baths</td>
<td>14.714</td>
<td>7665</td>
<td>1.92</td>
<td>0.056</td>
</tr>
<tr>
<td>NW</td>
<td>15649</td>
<td>7247</td>
<td>2.16</td>
<td>0.033</td>
</tr>
</tbody>
</table>

Predicted Values for New Observations
New Obs | Fit  | SE Fit | 95% CI          | 95% PI          |
1       | 193091 | 6542   | (176130, 210051) | (132173, 254008) |

Values of Predictors for New Observations
New Obs | size | lot | Bedrooms | Baths | NW |
1       | 20000| 30000| 4.00     | 3.00  | 1.00 |

30. How should the coefficient of NW be interpreted in this model?
   a) On average, NW Gainesville homes will cost $15649 more than in other areas, everything else being equal.
   b) We expect prices to increase by $7247 as NW increases by one unit, everything else being equal.
   c) For each extra dollar price increase we predict NW will increase by 2.16 units, everything else being equal.
   d) We should not interpret the coefficient of NW because there are no houses with zero NW.
   e) We should not interpret the coefficient of NW because it is not a significant predictor in this model.

31. How should the coefficient of bedrooms be interpreted in this model?
   a) Selling prices of houses are expected to increase by $5621 for every bedroom, keeping all other variables constant.
   b) Selling prices of houses are expected to decrease by $4868 for every bedroom, keeping all other variables constant.
   c) On average, the number of bedroom increases by .87 when the selling price increases by $1.
   d) We should not interpret the coefficient of bedrooms because there are no houses with zero bedrooms.
   e) We should not interpret the coefficient of bedrooms because it is not a significant predictor in this model.

32. We are interested in a 2,000 sqft house, with 4 bedrooms, 3 baths, located in a 30,000 sqft lot in NW Gainesville.
Using this output we can say:
   a) these types of homes will cost, on average, around $193 thousand.
   b) these types of homes will cost, on average, between $176 and $210 thousand approximately.
   c) 95% of homes with these characteristics will cost between $132 and $254 thousand approximately.
   d) all of the above
   e) none of the above

33. Looking at the residual plots for this regression analysis we can say that:
   a) the assumption of Normality has been clearly violated. 
   b) the assumption of constant variance has been clearly violated. 
   c) the assumption of randomness has been clearly violated. 
   d) all of the above
   e) none of the above

34. Using ALL outputs provided – can we say that bedroom is a good predictor of price?
   a) Yes, it is a significant predictor of the selling price of a house so it should always be included in the model.
   b) Yes, but there is multicollinearity with other predictors so it appears not significant in some models.
   c) No, number of bedrooms is the worst of all the predictor, so we should never use it to predict selling price.
   d) No, we would be extrapolating if we tried to use a variable that is significant only sometimes.
   e) No, the variable bedroom is an influential point that can change the results when removed from the model.