STATISTICS 110/201 PRACTICE FINAL EXAM KEY (REGRESSION ONLY)

Questions 1 to 5: There is a downloadable Stata package that produces *sequential sums of squares* for regression. In other words, the SS is built up as each variable is added, in the order they are given in the command. The last page of this exam gives output for the following situation. The data consist of the 68 houses from Appendix C7 that have Quality = 1. Y = Sales Price of the house ("salesprice" on the output), and the three predictor variables are:

X_1 = Square Feet divided by 100 = "sqft100" on the output
X_2 = Number of bedrooms = "bedrooms" on the output
X_3 = Lot Size in square feet= "lotsize" on the output

The model used did not involve any transformations; it is $E\{Y_i\} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}$. Notice in the output that the model was fit twice, with the variables in two different orders, but we will keep the designation of X_1, etc as defined above. In other words, we define X_1 = Square Feet divided by 100, and so on, no matter what order they appear in the Stata command.

NOTE: In case you aren't familiar with this notation, 1.18e+11 = 1.18×10^{11}, as an example.

1. Write the estimated regression equation for the full model with all 3 variables, filling in numbers for the coefficients.

$$\hat{Y} = 290558.1 + 9893.524X_1 - 36018.68X_2 + 2.513X_3$$

2. Carry out a test of the null hypothesis H_0: $\beta_1 = \beta_2 = \beta_3 = 0$. State the test statistic, the *p*-value, and your conclusion about whether or not to reject the null hypothesis using $\alpha = .05$.

Test statistic: $F = 6.69$
p-value = 0.0005
Reject the null hypothesis (because $0.0005 < 0.05$)

3. Give numerical values for each of the following, where X_1, X_2 and X_3 are defined above:

   a. $SSR(X_2)$ = *4.02×10^7 (from the "bedrooms" row and Seq.SS column for the 2nd version)*

   b. $SSR(X_3| X_1, X_2)$ = *6.19×10^{10} (from the "lotsize" row for either version)*

   c. $SSR(X_1|X_2)$ = *1.87×10^{11} (find this in the "sqft100" row for the 2nd version)*

   d. The *p*-value for testing H_0: $\beta_3 = 0$, *given* that X_1 and X_2 are in the model = *0.0291*

   e. The *p*-value for testing H_0: $\beta_2 = 0$, *given* that the other X's are *not* in the model = *0.9548*

   (note that the F value is only .003)*

4. Give the numerical values (from the output) for two different test statistics for testing H_0: $\beta_3 = 0$ (given that X_1 and X_2 are in the model). Then explain in words, in the context of the real estate situation, what this hypothesis is testing.

   $t = 2.23$
   $F = 4.989$
   (The above two test statistics are equivalent and give a corresponding p-value of 0.029)

   The hypothesis is testing whether the predictor variable Lot Size needs to be in the model, given that square feet and number of bedrooms are already in the model. Or you could say that it tests whether Lot Size is a statistically significant predictor variable, given that square feet and number of bedrooms are in the model.

5. If we were to fit the simple linear regression model using bedrooms as the only predictor, would the result be that number of bedrooms is a significant predictor of Sales Price? Explain how you know, using information provided in the output.

   No, number of bedrooms by itself is not a significant predictor. The appropriate test is given in the "bedrooms" row of the last table. The test statistic is $F = 0.0032$ and the p-value is 0.9548, which is clearly much larger than 0.05.

-----END OF QUESTIONS BASED ON THE STATA OUTPUT-----

6. When examining case diagnostics in multiple regression, under what circumstance is it acceptable to remove a case that is clearly a Y outlier?

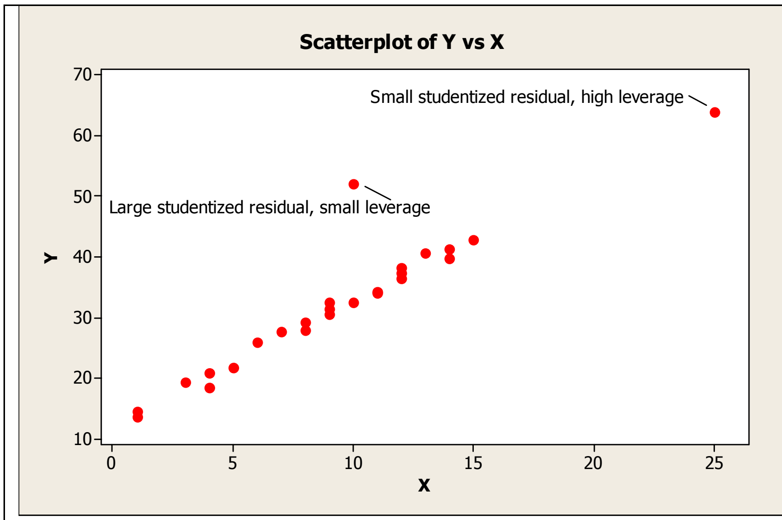
   It is only acceptable to remove a Y outlier if it is clearly a mistake.

7. Give two circumstances in which it is acceptable to remove one or more cases that are outliers in the X variables.

   1. *One of more of the X values for the case is clearly a mistake.*

   2. *If several cases have X values different from the rest of the data, and the prediction does not work well for cases with those combinations of X value, then it's likely that the cases belong to a different population and the relationship between Y and the predictor variables is not the same as it is for the main population. The cases can be removed, but the model would not apply in the future for predicting Y when X has values in that area.*

8. Draw a scatter plot of Y versus X showing points for a simple linear regression analysis, illustrating a case that has a small studentized residual but high leverage, and a case that has a large studentized residual but small leverage. Make sure you label which point is which.



In general, leverage has to do with how far X_i is from $\bar{X}$. So the point with high leverage should be far from the center of your X values, and the point with small leverage should be close to the center of the X's. The point with small studentized residual should fall close to the regression line, while the point with large studentized residual should fall far above or below it. The graph on the left shows one such set of points.

9. Suppose you have four possible predictor variables (X_1, X_2, X_3, and X_4) that could be used in a regression analysis. You run a forward selection procedure, and the variables are entered as follows: <u>Step 1</u>: X_2 <u>Step 2</u>: X_4 <u>Step 3</u>: X_1 <u>Step 4</u>: X_3
   In other words, after Step 1, the model is $E\{Y\}=\beta_0 + \beta_1 X_2$
   After Step 2, the model is $E\{Y\}=\beta_0 + \beta_1 X_2 + \beta_2 X_4$
   And so on…
   You also run an all subsets regression analysis using R^2 as the criterion for the "best" model for each possible number of predictors (1, 2, 3, 4). Would the same models result from this analysis as from the forward stepwise procedure? In other words, would "all subsets regression" definitely identify the following as the best models for 1, 2, 3, and 4 variables? Circle Yes or No in each case.

   a. β_0 + 1 variable, best model would be $E\{Y\}=\beta_0 + \beta_1 X_2$ **YES**

   b. β_0 + 2 variables, best model would be $E\{Y\}=\beta_0 + \beta_1 X_2 + \beta_2 X_4$ **NO**

   c. β_0 + 3 variables, best model would be $E\{Y\}=\beta_0 + \beta_1 X_2 + \beta_2 X_4 + \beta_3 X_1$ **NO**

   d. β_0 + 4 variables, best model would be $E\{Y\}=\beta_0 + \beta_1 X_2 + \beta_2 X_4 + \beta_3 X_1 + \beta_4 X_3$ **YES**

10. An international company is worried that employees in a certain job at its headquarters in Country A are not being given raises at the same rate as employees in the same job at its headquarters in Country B. Using a random sample of employees from each country, a regression model is fit with:
Y = employee salary
X_1 = length of time employee has worked for the company
X_2 = 1 if employee is in Country A, and 0 if employee is in Country B.

New employees, who have $X_1 = 0$, all start at the same salary, so the company is not interested in fitting a model with different intercepts, only with different slopes.

a. Write the full and reduced models for determining whether or not the slopes are different for employees in the two countries, using the variable definitions above and standard notation.

Full model $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_1 X_2$

Reduced model: $E\{Y\} = \beta_0 + \beta_1 X_1$

b. For the full model, write the row of the X matrix for an employee with 10 years of experience in Country A, and the row of the X matrix for an employee with 12 years of experience in Country B. You should write these using numbers, not symbols.

For 10 years experience in Country A: [1 10 10]

For 12 years experience in Country B: [1 12 0]

MULTIPLE CHOICE QUESTIONS
Circle the best answer.

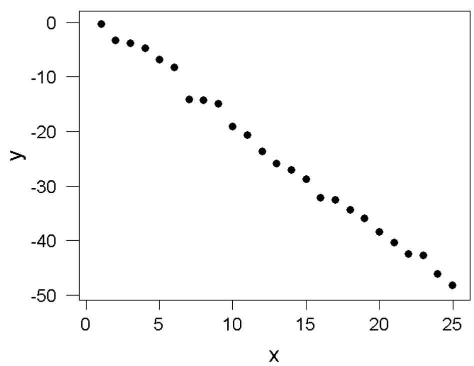
1. In a linear regression analysis with the usual assumptions (stated on page 218 and other places in the text), which one of the following quantities is the same for all individual units in the analysis?
A. Leverage h_{ii}
B. $s\{Y_i\}$
C. $s\{e_i\}$
D. $s\{\hat{Y}_i\}$

2. A regression line is used for all of the following *except* one. Which one is *not* a valid use of a regression line?
A. to estimate the average value of Y at a specified value of X.
B. to predict the value of Y for an individual, given that individual's X-value.
C. to estimate the change in Y for a one-unit change in X.
D. to determine if a change in X causes a change in Y.

3. Which choice is *not* an appropriate description of $\hat{Y}$ in a regression equation?
   A. Estimated response
   B. Predicted response
   C. Estimated average response
   D. Observed response

4. Which of the following is the *best* way to determine whether or not there is a statistically significant linear relationship between two quantitative variables?
   A. Compute a regression line from a sample and see if the sample slope is 0.
   B. Compute the correlation coefficient and see if it is greater than 0.5 or less than −0.5.
   C. Conduct a test of the null hypothesis that the population slope is 0.
   D. Conduct a test of the null hypothesis that the population intercept is 0.

5. Shown below is a scatterplot of Y versus X.



   Which choice is most likely to be the approximate value of R^2?
   A. −99.5%
   B. 2.0%
   C. 50.0%
   D. 99.5%

6. In a regression model with p – 1 predictor variables chosen from a set of P – 1 possible predictor variables, which of the following indicates that bias is *not* a problem with the model?
   A. Mallow's $C_p \leq p$ (small p)
   B. Mallow's $C_p \leq P$ (cap P)
   C. Mallow's $C_p > p$ (small p)
   D. Mallow's $C_p > P$ (cap P)

7. Which of the following case diagnostic measures is based on *Y* values only (and not *X* values)?
   A. Cook's Distance
   B. Studentized deleted residual
   C. DFFITS
   D. None of the above – they all use the X values and the Y values

8. Which of the following methods is the most appropriate for testing H_0: $\beta_k = 0$ versus Ha: $\beta_k > 0$?
   A. A t-test
   B. An F-test
   C. A test of a full versus reduced model
   D. All of the above are equally good.

9. Which of the following is *not* a valid null hypothesis?
   A. $H_0: \beta_1 = 0$
   B. $H_0: \beta_1 = \beta_2$
   C. $H_0: b_1 = b_2 = 0$
   D. All of the above *are* valid null hypotheses

10. Which of the following can never be 0 (unless the population standard deviation $\sigma = 0$)?
   A. The estimated intercept, b_0
   B. A studentized deleted residual, t_i
   C. The variance of the prediction error, $\sigma^2\{pred\}$
   D. The estimate of $E\{Y_h\}$, $\hat{Y}_h$

Here is the Stata Output for Questions 1 to 5.

```
reg_ss salesprice sqft bedrooms lotsize

      Source |       SS       df       MS              Number of obs =      67
-------------+------------------------------           F(  3,    63) =    6.69
       Model |  2.4910e+11     3  8.3035e+10           Prob > F      =  0.0005
    Residual |  7.8158e+11    63  1.2406e+10           R-squared     =  0.2417
-------------+------------------------------           Adj R-squared =  0.2056
       Total |  1.0307e+12    66  1.5616e+10           Root MSE      =  1.1e+05

------------------------------------------------------------------------------
  salesprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     sqft100 |   9893.524   2508.561     3.94   0.000     4880.564    14906.48
    bedrooms |  -36018.68   14665.17    -2.46   0.017    -65324.68   -6712.675
     lotsize |   2.512511   1.124531     2.23   0.029     .2653148    4.759707
       _cons |   290558.1   75398.79     3.85   0.000     139885.6    441230.5
------------------------------------------------------------------------------

Sequential Sum of Squares for Regression

------------------------------------------------------------------------------
  salesprice |      Coef.     Seq SS     df1     df2          F   Prob > F
-------------+----------------------------------------------------------------
     sqft100 |   6187.993   1.18e+11       1      63   9.474134     0.0031
    bedrooms |  -34701.03   6.96e+10       1      63   5.605086     0.0210
     lotsize |   2.512511   6.19e+10       1      63   4.989326     0.0291
------------------------------------------------------------------------------

reg_ss salesprice bedrooms sqft lotsize

      Source |       SS       df       MS              Number of obs =      67
-------------+------------------------------           F(  3,    63) =    6.69
       Model |  2.4910e+11     3  8.3035e+10           Prob > F      =  0.0005
    Residual |  7.8158e+11    63  1.2406e+10           R-squared     =  0.2417
-------------+------------------------------           Adj R-squared =  0.2056
       Total |  1.0307e+12    66  1.5616e+10           Root MSE      =  1.1e+05

------------------------------------------------------------------------------
  salesprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    bedrooms |  -36018.68   14665.17    -2.46   0.017    -65324.68   -6712.675
     sqft100 |   9893.524   2508.561     3.94   0.000     4880.564    14906.48
     lotsize |   2.512511   1.124531     2.23   0.029     .2653148    4.759707
       _cons |   290558.1   75398.79     3.85   0.000     139885.6    441230.5
------------------------------------------------------------------------------

Sequential Sum of Squares for Regression

------------------------------------------------------------------------------
  salesprice |      Coef.     Seq SS     df1     df2          F   Prob > F
-------------+----------------------------------------------------------------
    bedrooms |  -668.8945   4.02e+07       1      63   .0032419     0.9548
     sqft100 |   9739.083   1.87e+11       1      63   15.07598     0.0003
     lotsize |   2.512511   6.19e+10       1      63   4.989326     0.0291
------------------------------------------------------------------------------
```