UNIVERSITY OF CAMBRIDGE
Judge Business School

Updated: November 14, 2011
Lab Instructor: Michael Freeman
Contact: mef35@cam.ac.uk

Contest Quiz 5

Question Sheet

In this quiz we will focus on constructing regression models and performing basic diagnostic tests.

NOTE: Please ensure that any numeric result you produce is rounded to TWO decimal places using the R function: `round(x,2)`, where `x` is the number you wish to round. We cannot guarantee that a number that is not rounded according to these specific instructions will be correct, and you may be penalised.

Question 1

Install (if necessary) and load the `car` package. Select and store the Ericksen data set in a new variable (e.g. `crime.dat <- Ericksen`). This data set contains information on the demographic and characteristics of different cities and states in the US.

We would like to determine the extent to which the percentage of minorities, `minority`, the percentage of `poverty`, the percentage having English `language` difficulties, the percentage aged 25 or older who haven't finished `highschool`, the percentage of `housing` in small, multiunit buildings, and whether the observation is a `city` or state has an influence on the serious `crime` rate per 1000 population. Produce this multiple linear regression as described.

I) Produce a residual plot. Based on this plot do you believe there is evidence to suspect heteroskedasticity?

  (a) Yes, since the spread of the residuals tends to increase as we move right in the plot.
  (b) Yes, since the spread of the residuals tends to decrease as we move right in the plot.
  (c) No, since the spread of the residuals increases, then decreases, then increases again as we move right in the plot.
  (d) No, since the spread of the residuals decreases, then increases, then decreases again as we move right in the plot.
  (e) No, since the spread of the residuals remains about the same as we move right in the plot.

II) Now produce a histogram and QQplot. Based on these plots, do you feel that normality (or a lack of normality) is a big issue and should cause us to reject this model?
  (a) Yes (b) No

III) Based on your answer to Part I) you may or may not feel a transformation is necessary. What is the value of the transformation parameter for the Box-Cox method?

IV) Based on your answer to Part III ONLY, perform any transformation you feel is necessary (if at all). Return the Adjusted R-squared value of either your original model or, if you performed a transformation, of your transformed model.

V) Re-plot the residual plot for your preferred model in Part IV). Which of the following do you feel is most appropriate:

(a) The residual plot is identical to the one I produced in Part I.

(b) The residual plot has improved the issue that I identified in Part I, and I can see no other potential problems in the plot.

(c) The residual plot has improved the issue that I identified in Part I, but I can still see other potential problems in the plot.

(d) None of the above.

Question 2

Using the multiple regression model described in the 2nd paragraph of Question 1, we are now interested in exploring potential multicollinearity between the variables.

I) Produce a matrix of correlations for the 7 variables in your model. What is the correlation between `crime` and `highschool`?

II) What is the correlation between `crime` and `city`?

III) Now use the `plotcorr` function with the color specifications as given to you in the lab session. How many of these show weak correlation (i.e. are white in appearance)? Note: only consider those in the lower left diagonal of the matrix.

IV) Now calculate the VIFs. What is the largest VIF?

V) What is the average VIF?

VI) Do you suspect that there may exist multicollinearity?
(a) Yes (b) No

Question 3

Install (if necessary) and load the `car` package. Select and store the Anscombe data set in a new variable (e.g. `educ.dat <- Anscombe`). This data set contains information on the per-capita `education` expenditure, per-capita `income`, number of `young` people, per 1000, and number living in an `urban` area, per 1000.

We would like to determine the extent to which the other variables have an influence on the per-capita education expenditure. Produce a multiple linear regression. We now wish to attempt to improve this regression by considering transformations to the dependent and independent variables. By considering only a `log` transformation of the dependent variable, and squared transformations of the independent variables, come up with the best model you can. For the purpose of this exercise, consider the 'best' model to be the one with the highest Adjusted R-squared, and with each of the independent variables individually significant at the 10% level using the `summary` function (except for the Intercept term, which you must include but that does not need to be significant).

NOTE: For this problem please give all results to two **significant** figures using the R function: `signif(x,2)`, where `x` is the number you wish to round. We cannot guarantee that a number that is not rounded according to these specific instructions will be correct, and you may be penalised.

I-VI) Report the value of each coefficient (excluding the intercept), in the following order: `income`, `income^2`, `young`, `young^2`, `urban`, `urban^2`. If the variable is not included in your final model then assign it a value of 0 in your answer.

VII) What is the improvement in the value of your Adjusted R-squared from the initial model you fitted?

Question 4

Return to the multiple regression model described in the 2nd paragraph of Question 1 and used in Question 2.

I) From your model you might suspect that one or more of the independent variables should be removed. What would be an appropriate test here?
(a) z-test (b) t-test (c) neither of these

II) Return the p-value produced by an appropriate hypothesis test under the null that this/these variable(s) have coefficient(s) equal to 0.

III) Based on your answer to Part II) fit a new model (if appropriate), without transforming any of the variables. What is the difference in the Adjusted R-squared between this new model and the original model?

Question 5

Install (if necessary) and load the `car` package. Select and store the Leinhardt data set in a new variable (e.g. `mort.dat <- Leinhardt`). This data set contains information on 105 countries, detailing the per-capita `income`, `infant` mortality rate per 1000 live births, the `region` in which the country exists, and whether or not the country is `oil` exporting.

We would like to determine the extent to which the other variables have an influence on the infant mortality rate. Produce a multiple linear regression model. By considering the following potential transformations:
Dependent Variable : `log`
Independent Variables : `^2`, `log`
and also by considering interaction terms, come up with the best model you can fit. For the purpose of this exercise, consider the 'best' model to be the one with the highest Adjusted R-squared, and with each of the independent variables individually significant at the 5% level using the `summary` function (except for the Intercept term, which you must include but that does not need to be significant).

I) Compare the Adjusted R-squared of your best model against that of the original model. In which region does this difference lie?
(a) 0 to 0.1 (b) 0.1 to 0.2 (c) 0.2 to 0.3 (d) 0.3 to 0.4 (e) greater than 0.4

II) Provide the Adjusted R-squared value of your best model.

III) Type the R code of your model (e.g. `lm(infant~.)`) into the text box. I will check all of these as well as those in Part II and award points to those who arrive at my best model, and bonus points to those who beat my best model.