Homework Assignment 3

Max H. Farrell – Chicago Booth
BUS41100 Applied Regression Analysis

Due at the beginning of week 4

1 Infant Nutrition

This question involves data from a study on the "nutrition of infants and preschool children in the north central region of the United States of America".[1] It is available on the course web page as `nutrition.csv`, and contains 72 observations of boys' weight/height ratio (`woh`) for equally spaced values of `age` in months.

(a) Plot the data ($Y =$ `woh`), and overlay the least squares line and a 95% prediction interval in the range of the data. Comment on the goodness of fit.

(b) Plot the residuals from the above fit and comment on any patterns you see. Based on this plot, how would you change the model to better fit the data? Further justify your answer with a statistical test. Plot your updated regression and 95% prediction interval over a scatterplot of the data.

(c) Plot the residuals from new fit and compare to the plot in part **(b)**.

(d) The authors of the study have reason to believe that the observations fall into to groups: (1) the first seven boys and (2) the remaining 65. By introducing an appropriate dummy variable and interaction term, find the least squares fit of these lines. Plot them and their corresponding predictive intervals in such a way as they cover *only* their respective `age` ranges (i.e., so that they do not overlap). Include the simple linear regression and prediction interval from **(a)** in your plot. Comment on the differences you see.

(e) Plot the residuals from new fit and compare to the plot in parts **(b)** and **(c)**.

(f) Of the three, which model do you prefer? Why?

2 Beef – It's What's for Dinner

In 1988, US cattle producers voted on whether or not to each pay a dollar per head towards the marketing campaigns of the American Beef Council. At the time of this vote, the council's TV campaign featured a voice-over by actor Robert Mitchum, using the theme "Beef – it's what's for dinner." To understand the vote results (it passed), the Montana state cattlemen's association looked at the effect of the physical size of the farm and the value of the farms' gross revenue on voter preference. The data (in the file `beef.csv` on the course website) consist of the vote results (% `YES`), average `SIZE` of farm (hundreds of acres), and average `VAL` of products sold annually by each farm (in $ thousands) for each of Montana's 56 counties.

(a) Plot the data and comment on what you see. How will this effect our analysis?

[1] by E.S. Eppright, H.M. Fox, B.A. Fryer, G.H. Lamkin, V.M. Vivian and E.S. Fuller in *World Review of Nutrition and Dietetics*, **14**, 1972, pp. 269–332.

(b) Fit a regression model for `YES` with both `SIZE` and log(`VAL`) as covariates. Interpret the results. What regression assumptions might we have violated here?

(c) Find a better model: does the effect of `SIZE` change depending on log(`VAL`)? What is your estimate of the effect on `YES` of a unit change in `SIZE`? Interpret your conclusion.

3 Crime Statistics

In this question we consider crime-related and demographic statistics for 47 US states in 1960, available as `crime.csv` on the course web page, and via:

```
> library(MASS)
> data(UScrime)
```

The data were collected from the FBI's Uniform Crime Report and other government agencies to determine how the Crime Rate (`CR`, offenses per million population) depends on thirteen socio-economic variables. We shall focus on a subset including residents' average years of education (`Ed`), labor force participation (`LF`), and median income (`W`). (For a full description run `?UScrime` at the `R` prompt. A few variables have different names: their `GDP` is our `W`, their `y` is our our `CR` $\times 10$.)

(a) Present a visual summary of the data. How does the crime rate relate to these three potential explanatory variables?

(b) Consider the regression of crime rate onto each of the three explanatory variables (`Ed`, `LF`, and `W`), individually in turn. Do you find any significant relationships? Any which are surprising?

(c) A continental US state not in our sample had a median income of \$2750 in 1960 (i.e., `W` = 275), but the crime rate recordings were not considered accurate enough for inclusion. What is a 90% prediction interval for the unknown crime rate in this state? Is there anything disturbing about this interval?

(d) Consider now the MLR of crime rate onto all of the three explanatory variables (`Ed`, `LF`, and `W`. Compare your results to what you found in **(b)**. Explain any differences/similarities you find.

(e) Now consider the variable `S`, an indicator if the state is in the South (0 = No, 1 = Yes). Add interactions of `S` with each of `Ed` and `W` to your model. Compute the partial effects of `Ed` and `W` on crime in the southern and northern states. Give a confidence interval for each partial effect. (That's four partial effects total: two for northern states, two for southern states.) Interpret and discuss both the values of the four partial effects and their intervals/significance. To form the confidence intervals, you can follow the steps given here:

`http://stats.idre.ucla.edu/r/faq/how-can-i-test-contrasts-in-r/`

`http://rpubs.com/djcava/lincom`.

4 Newspaper Circulation

Data were collected on the average Sunday and daily (i.e., weekday) circulations (in thousands) for 48 of the top 50 newspapers in the United States for the period March–September, 1993. See the `newspaper.csv` file on the course web site.

(a) Construct a scatter plot of Sunday circulation versus daily circulation. Does the plot suggest a linear relationship between the variables? Do you think this is a plausible relationship?

(b) Fit a regression line predicting Sunday circulation from daily circulation.

(c) What do β_0 and β_1 represent in this model? Be precise.

(i) Is there any (statistically) significant relationship between Sunday circulation and daily circulation? Justify your answer by a statistical test. Fully describe the test you are using, include null and alternative hypothesis, test statistic, and critical value.

(ii) What is special in this context about the value $\beta_1 = 1$? Given this data, is $\beta_1 = 1$ plausible? Justify your answer statistically.

(iii) Test the null hypothesis of $\beta_1 = 1$ against a two-sided alternative, with robust standard errors. What is the conclusion you reach? What is the p-value associated with this test?

(d) Suppose that you are proposing to add a Sunday edition of a newspaper with a weekday circulation of 225,000 copies. What would you tell advertisers is the expected Sunday circulation? What is the standard deviation of this expectation? What would you say when they ask you to predict a likely range of possible Sunday circulation numbers?

(e) Argue that working with the logarithm of the circulation(s) might be better than using the raw numbers. Fit the corresponding log-log regression model. Compare and contrast the fit and the predictive interval obtained for the Sunday edition of a newspaper with a weekly circulation of 225,000 copies.