STAT425 Homework #3
Fall 2015 Oct 18, 2015

Homework 3

Due: Oct 27, 2015

Problem 1. The State of Florida played an important role in the 2000 Presidential Election[1]. One of the crucial issues is whether the butterfly design of the ballot paper misled many Al Gore's supporters into voting for Patrick Buchanan, which consequently caused Gore lose the election. It is reasonable to predict the Buchanan vote (Y) using the Bush vote (X), since both are conservative.

1. (2pt) Fit a simple linear regression (Y vs X), excluding the sample from the Palm Beach County since it is suspected to be an outlier. Plot the data and add the fitted regression line.

2. (6pt) Plot residuals versus X to check for non-constant variance. Also perform the Breusch-Pagan test against heteroskedasticity. What's your conclusion?

3. (2pt) Since X and Y are both counts, we can consider to fit a linear regression model on the transformed data, $\log Y$ vs $\log X$. Use the command "`powerTransform`" to determine whether the log transformation is appropriate. *Hint:* check whether 0 is in the CI for the suggested λ values.

```
library(alr3)
tranxy = powerTransform(cbind(Y, X) ~ 1)
```

4. (6pt) Fit a simple linear regression of $\log Y$ vs $\log X$, excluding the sample from the Palm Beach County. Plot the data and add the fitted regression line, and plot residuals versus $\log X$. Which model is more reasonable, the new model (with log transformation) or the original one? Besides graphical displays, you can also 1) compare the R^2 from the two models, and 2) check whether the Breusch-Pagan test is still significant.

5. (4pt) Form a 95% prediction interval for Buchanan's vote in Palm Beach. If your interval is for log(votes), you should transform it back to the original scale. Do you think Buchanan's vote in Palm Beach is a statistically plausible outcome, assuming the linear model is appropriate?

6. (2pt) If we didn't know that the ballot design in Palm Beach is different from the one used in other counties, we would fit the regression model using all samples, and then apply the outlier-test on each county with the Bonferroni correction. Now, is Palm Beach still an outlier?

[1] You can do a google search to get more background information about the Palm Beach County "butterfly ballot" problem in election 2000.

Problem 2. Use the `infmort` data from the Faraway book.

1. (8pt) Fit a linear regression model for the infant mortality in terms of the other variables, `log(income)`, `region`, and `oil`.

   You can create a new data frame with `income` in log-scale.

   ```
   mydata = infmort
   mydata$income = log(mydata$income)
   ```

   Interpret the regression coefficients from the fitted model. Display the fitted regression line (or lines) on top of the data on two figures: one figure for "oil export" countries and one for "no oil export" countries.

   Report any outliers and high-influential points. Are all observations used by `R`? (Check the sample size and the degree of freedom from the linear model.)

2. (2pt) Check for large leverage points.

3. (2pt) Check for outliers.

4. (2pt) Check for high influential points.

5. (4pt) Check whether there are any interactions between `income` (in log-scale) and the other two categorical predictors. You may find the p-values from the following F tests useful.

   ```
   anova(lm(mortality ~ income+region+oil+ income:oil+income:region, mydata))
   anova(lm(mortality ~ income+region+oil+ income:region+ income:oil, mydata))
   ```

Problem 3. Recall the `savings` data from the textbook, where 5 variables, `sr`, `pop15`, `pop75`, `dpi`, and `ddpi`, are recorded for each country. Suppose you have fitted a linear regression model to predict `sr` based on the other 4 predictors, and then you lost the original `sr` vector, but you still have the 4 predictors and the fitted value $\hat{y}_i$, which are saved in the new data frame `newsavings`, and also $\hat{\sigma}$ from the original linear regression model.

```
g=lm(sr~pop15+pop75+dpi+ddpi, data = savings)
new savings = savings
newsavings$sr = g$fitted
sigma.hat = summary(g)$sigma
remove(g)
```

Answer the following questions based on `newsavings` and `sigma.hat`. For each question below, it is not enough to provide just a numerical answer; you should provide details, such as derivations and `R` code, so we could understand how you obtain your answer.

1. (4pt) What's the adjusted R^2 for the original regression model?

2. (8pt) What are the estimated LS coefficients for the 4 predictors from the original regression model?

3. (4pt) What's the p-value for testing whether the coefficient with `pop75` is equal to zero (in the original regression model)?

4. (4p) Suppose we want to test whether we could remove the two population-related predictors, `pop75` and `pop15` from the original linear model. What's the F-stat for testing the following two models, and what's the corresponding p-value, and what's your decision?

Problem 4. The term *regression* was first used in connection with the work of Sir Francis Galton on inheritance of characteristics. In a paper on "Typical Laws of Heredity," delivered to the Royal Institution on February 9, 1877, Galton discussed some experiments he had investigated using sweet peas. By comparing the sweet peas produced by parent plants to those produced by offspring plants, he could observe the inheritance from generation to generation. One obvious characteristic of sweet peas is their diameter. To obtain data, Galton categorized plants according to the typical diameter of the peas they produced. For each of the 7 size classes (15 to 21 hundredths of an inch), he arranged for each of 9 of his friends to grow 10 plants from seed in each size class; however, 2 of the crops were total failures. Galton's data were later published by Karl Pearson (1914), as given in the table below. In the table, only the mean diameters of the offspring seed are given along with respective standard deviations; sample sizes are unknown.

Diameter of Parent Peas	Mean Diameter of Offspring Peas	Standard Deviation
21	17.26	1.988
20	17.07	1.938
19	16.37	1.896
18	16.40	2.037
17	16.13	1.654
16	16.17	1.594
15	15.98	1.763

Table 1: Galton's data.

1. (4pt) Draw the scatter plot of Y = mean offspring diameter versus X = parent diameter. Assuming that the standard deviations are exactly correct, compute the weighted regression of Y on X. Add the fitted line on your scatter plot. Also add the fitted OLS line on your scatter plot.

   When showing two lines in the same plot, you should plot them using different line types and also use legend to explain which one is the OLS line and which one is the WLS line.

2. (2pt) Using WLS, estimate the third quartile for the diameter of the offspring peas produced by peas whose diameter is 20.5.

3. (6pt) Galton wanted to know if characteristics of the parent plant such as size were passed on to the offspring plants. Perfect inheritance would correspond to the slope parameter $\beta_1 = 1$, while $\beta_1 < 1$ would suggest that the offspring are "reverting" toward "what may

be roughly and perhaps fairly described as the average ancestral type" (Galton, 1885). Test the hypothesis $\beta_1 = 1$ versus $\beta_1 < 1$. State your test statistic, its distribution under H_0, and the p-value.

4. (6pt) Obtain a test for lack of fit of the straight-line model. State your test statistic, its distribution under H_0, and the p-value.