STAT 224 / PBHS 324 Problem Set 4 Fall 2017

Due Tuesday, Nov 7th by 10am. Possible points 73.

1. [22 pts total] Modified from Exercises 5.4

Perform a thorough analysis of the Education Expenditures data in Tables 5.12, 5.13, and 5.14. You are expected to examine the relationship between Y and X_1, X_2, X_3 and Region over time.
The data can be downloaded from
`http://statistics.uchicago.edu/~collins/data/RABE5/P151-153.txt`

or read directly into R

```
rawData <- read.delim("http://statistics.uchicago.edu/~collins/data/RABE5/P151-153.txt")
```

(a) [1 pt] Create appropriate indicator variables for the region variable.
(b) [1 pt] Treat the year variable as ordinal, not nominal. More specifically, recode the variable as 0 (1960), 10 (1970), and 15 (1975). Note: This is different from the textbook suggestion of 0/1 indicator variables in Section 5.7.
(c) [2 pts] Fit a model to the data. In your model, include X_1, X_2, X_3, year, region indicator variables, and interaction effects to allow the coefficients ("slopes") for X_1, X_2, X_3 to vary by year.
(d) [3 pts] Plot the studentized (standardized) residuals vs. the fitted values ($\widehat{y}$). Draw a normal probability plot (quantile plot) of the studentized (standardized) residuals. Which model assumptions appear to be violated? Explain.
(e) [2 pts] The problems discovered in part (d) can sometimes be alleviated by changing the response from y to $\log(y)$. Create a new response variable: $\log(y)$. Fit the same model as in part (c), except using the new response variable: $\log(y)$. Note: Log base 10 of the expenditures is perhaps easier to interpret than natural log.
(f) [4 pts] Check basic model assumptions (linear relationship to Y, error variance, normality, influence, etc.)
(g) [2 pts] Test the overall effects of X_1, X_2, X_3 on Y. Specify the hypothesis to be tested, the test used and your conclusions at the 5% significance level.
(h) [2 pts] Test whether the effects of X_1, X_2, X_3 remain unchanged over time. Specify the hypothesis to be tested, the test used and your conclusions at the 5% significance level.
(i) [3 pts] Based on findings in (h) decide whether separate regressions by year interval need to be reported. Report coefficients for X variables separately by year (i.e., fit a separate model for each year).
(j) [2 pts] Compare the estimated coefficients X_1, X_2, X_3 for the models in parts (e) and (i). Show how the coefficient estimates from part (e) can be used to find the coefficients you observed in part (i).

2. [15 pts total] Modified from Exercises 5.7

Three types of fertilizer are to be tested to see which one yields more corn crop. Forty similar plots of land were available for testing purposes. The 40 plots are divided at random into four groups, 10 plots in each group. Fertilizer 1 was applied to each of the 10 corn plots in Group 1. Similarly, Fertilizers 2 and 3 were applied to the plots in Groups 2 and

3, respectively. The corn plants in Group 4 were not given any fertilizer; it will serve as the control group. Table 5.17 gives the corn yield y_{ij} for each of the 40 plots. Data can be loaded as follows:

```
Stata
. import delim http://statistics.uchicago.edu/~collins/data/RABE5/P158, clear

R
fertilizerData <- read.delim("http://statistics.uchicago.edu/~collins/data/RABE5/P158.txt")
```

(a) [1 pt] Create three indicator variables, F_1, F_2, F_3, one for each of the three fertilizer groups (to compare against control)
(b) [1 pt] Fit the model $y_{ij} = \mu_0 + \mu_1 F_{i1} + \mu_2 F_{i2} + \mu_3 F_{i3} + \epsilon_{ij}$.
(c) [4 pts] Test the hypothesis that none of the three types of fertilizer has an effect on corn crops. Specify the hypothesis to be tested, the test used, and your conclusions at the 5% significance level.
(d) [4 pts] Test the hypothesis that the three types of fertilizer have equal effects on corn crop. Specify the hypothesis to be tested, the test used and your conclusions at the 5% significance level.
(e) [5 pts] Irrespective of the results in (d), test whether there is a common effect of fertilizer, call it μ_F, relative to control.

4. [16 pts total] Exercise 5.6 PARTS a-f

The price of a car is thought to depend on the horsepower of the engine and the country where the car is made. The variable Country has four categories: USA, Japan, Germany, and Others. To include the variable Country in a regression equation, three indicator variables are created, one for USA, another for Japan, and the third for Germany. In addition, there are three interaction variables between horsepower and each of the three Country categories (HP*USA, HP*Japan, and HP*Germany). Some regression outputs when fitting three models to the data is shown in Table 5.16 (for those of you who do not have the new version of the textbook, Table 5.16 is on the last page of this homework). The usual regression assumptions hold.
(a) [3 pts] Compute the correlation coefficient between the price and the horsepower.
(b) [1 pts] What is the least squares estimated price of an American car with a 100 horsepower engine?
(c) [3 pts] Holding the horsepower fixed, which country has the least expensive car? Why?
(d) [3 pts] Test whether there is an interaction between Country and horsepower. Specify the null and alternative hypotheses, test statistics, and conclusions.
(e) [3 pts] Given the horsepower of the car, test whether the Country is an important predictor of the price of a car. Specify the null and alternative hypotheses, test statistics, and conclusions.
(f) [3 pts] Would you recommend that the number of categories of Country be reduced? If so, which categories can be joined together to form one category?

5. [10 pts total] Box-Cox Analysis

In class, we saw that a natural log transform of price was sufficient for using a linear model predicting price with age of Port wine. Data can be loaded as follows:

```
Stata
. import delim http://statistics.uchicago.edu/~collins/data/STAT224other/wine.txt, clear

R
wineData <- read.delim("http://statistics.uchicago.edu/~collins/data/STAT224other/wine.txt")
```

(a) [4 pts] Run the Box-Cox analysis for transformation of the response variable. Obtain the suggested λ parameter. Depending on which section of the course you are taking, the boxcox method was demonstrated in Stata or R during lecture. (In R, use the `boxcox` function in the `MASS` package. The default plot will show you the range of lambda values against the log likelihood.)
(b) [6 pts] Transform Y (price) according the Box-Cox suggested transformation and perform the regression. Compare to the results for using log(price) as the response variable.

6. [10 pts total] Diamond Data

Continuing with the diamond data from previous homework. Although we saw some issues with model assumptions (which may be a function of the small sample at lower carat weight), for now ignore and do the following.
(a) [5pts] Test for a carat by color (coded as ordinal numeric) interaction effect, and decide if this term is needed in the model.
(b) [5pts] Variable X1 is diamond cut, with 4 rating categories. Create appropriate indicator variables and evaluate whether this categorical factor adds to price prediction in this dataset.

Table 5.16 Some Regression Outputs When Fitting Three Models to the Car Data

Model 1

Source	Sum of Squares	df	Mean Square	F-Test
Regression	4604.7	1	4604.7	253
Residual	1604.44	88	18.2323	

Variable	Coefficient	s.e.	t-Test	p-value
Constant	–6.107	1.487	–4.11	0.0001
Horsepower	0.169	0.011	15.9	0.0001

Model 2

Source	Sum of Squares	df	Mean Square	F-Test
Regression	4818.84	4	1204.71	73.7
Residual	1390.31	85	16.3566	

Variable	Coefficient	s.e.	t-Test	p-value
Constant	–4.117	1.582	–2.6	0.0109
Horsepower	0.174	0.011	16.6	0.0001
USA	–3.162	1.351	–2.34	0.0216
Japan	–3.818	1.357	–2.81	0.0061
Germany	0.311	1.871	0.166	0.8682

Model 3

Source	Sum of Squares	df	Mean Square	F-Test
Regression	4889.3	7	698.471	43.4
Residual	1319.85	82	16.0957	

Variable	Coefficient	s.e.	t-Test	p-value
Constant	–10.882	4.216	–2.58	0.0116
Horsepower	0.237	0.038	6.21	0.0001
USA	2.076	4.916	0.42	0.6740
Japan	4.755	4.685	1.01	0.3131
Germany	11.774	9.235	1.28	0.2059
HP*USA	–0.052	0.042	–1.23	0.2204
HP*Japan	–0.077	0.041	–1.88	0.0631
HP*Germany	–0.095	0.066	–1.43	0.1560