STA 6207 – Homework 4 - Fall 2019

Q.1. NHL Height/Weight Analysis

The dataset **nhl_ht_wt.csv** contains weights (Y, in pounds) and heights (X, in inches) for 717 National Hockey League players for the 2013/2014 season. Complete the following parts (treating this as a sample from a conceptual population of potential athletes).

- Plot Weight (Y) versus Height (X)
- Fit a Simple Linear Regression model in matrix form relating Weight (Y) to (X). Obtain the Regression coefficients, standard errors, t-tests, Analysis of Variance, and r^2
- Fit a Simple Linear Regression using a standard statistical software package.
- Plot the residuals versus fitted values
- Obtain a normal probability plot of residuals
- Test the hypothesis that the errors are normally distributed with the Shapiro-Wilk test.
- Test the hypothesis that the error variance is constant with the Breush-Pagan test. (Using direct computations and statistical software package).
- Test the hypothesis that the relationship between Weight and Height is linear with the F-test for lack-of-fit.
- Estimate the multiplicative bias factor for β_1 due to measurement error (since the heights in the dataset are rounded to the nearest inch).
- If there is evidence of non-normality or non-constant variance of errors, obtain a Box-Cox transformation, and repeat the previous parts.

Q.2. Ship Emissions at Ports in Australia

The dataset **ship_emissions2.csv** contains data from 34 ports in Australia. Consider relating SO2 (Y) to Fuel (X).

- Plot: Y versus X, log(Y) vs X, Y vs log(X), and log(Y) vs log(X). which appears to be the "best" linear relation?
- Fit a simple linear regression, based on your choice above.
- Test the hypothesis that the errors are normally distributed with the Shapiro-Wilk test.
- Test the hypothesis that the error variance is constant with the Breush-Pagan test.
- Based on whichever of the two models that fits best:
    - Obtain the studentized deleted residuals and determine whether any ports are outliers
    - Obtain the leverage values and identify any potentially influential ports
    - Obtain the DFFITS and identify any influential ports.
    - Obtain the DFBETAS and identify any influential ports.
    - Obtain Cook's D and identify any influential ports.
    - Obtain the COVRATIO and identify any influential ports.

Q.3. A simple regression model is fit, relating a dependent variable Y, to an independent variable X. You are given the following data and summary statistics.

X	Y	Y-hat_j	Y-bar_j			sum(X)	sum(Y)
0	15					90	225
0	13						
0	17					SS_XX	SS_YY
10	25					600	642
10	27						
10	23					SS_XY	
20	34					600	
20	32						
20	39						

p.3.a. Give the fitted equation, based on the simple linear regression model:

$\hat{\beta}_1 =$ _____ $\hat{\beta}_0 =$ _____ $\hat{Y} =$ _____

p.3.b. Fill in the 3rd and 4th columns of the table above.

p.3.c. Compute the Pure Error Sum of Squares and degrees of freedom:

SSPE = _____ $df_{PE} =$ _____

p.3.d. Compute the Lack-of-Fit Sum of Squares and degrees of freedom:

SSLF = _____ $df_{LF} =$ _____

p.3.e. Test H_0: Model is Linear vs H_A: Model is not Linear at $\alpha = 0.05$ significance level.

Test Statistic _____ Rejection Region _____

Q.4. A simple linear regression model is fit, based on n=5 individuals. The data and the projection matrix are given below:

X		Y	P				
1	2	8	0.344	0.303	-0.129	0.262	0.221
1	4	12	0.303	0.274	-0.035	0.244	0.215
1	25	24	-0.129	-0.035	0.953	0.059	0.153
1	6	14	0.262	0.244	0.059	0.226	0.209
1	8	18	0.221	0.215	0.153	0.209	0.203

p.4.a. Give the leverage values for each observation. Do any exceed twice the average of the leverage values?

Observation 1 _____ Obs 2 _____ Obs 3 _____ Obs 4 _____ Obs 5 _____

p.4.b. Give $\boldsymbol{\beta}$, based on the following results

X'X		X'Y	INV(X'X)	
5	45	76	0.438	-0.026
45	745	892	-0.026	0.003

p.4.c. Compute SSE and S_e (Note: $\mathbf{Y'Y} = 1304 \quad \mathbf{Y'PY} = 1282.45$)

p.4.d. The following table contains the fitted values with and without each observation, residual standard deviation when that observation was not included in the regression, and the regression coefficients when that observation was not included in the regression.

Y-hat	Y-hat(-i)	S_i	beta0_i	beta1_i
10.92	12.45	2.07	11.41	0.52
12.14	12.19	3.28	9.76	0.61
24.99	45.00	0.63	5.00	1.60
13.36	9.46	3.24	9.46	0.62
14.59	13.72	1.86	8.72	0.62

p.4.d.i. Compute DFFITS for the fifth observation

p.4.d.ii. Compute DFBETAS0 for the first observation

p.4.d.iii. Compute DFBETAS1 for the third observation

Q.5. For the F-test for Lack-of-Fit, where:

$$H_0 : E\{Y_{ij}\} = \beta_0 + \beta_1 X_j \qquad H_A : E\{Y_{ij}\} = \mu_j \neq \beta_0 + \beta_1 X_j \quad j = 1,...,c; \quad i = 1,...,n_j$$

with: Residual $\equiv$ $Y_{ij} - \hat{Y}_j$ Pure Error $\equiv$ $Y_{ij} - \bar{Y}_j$ Lack of Fit $\equiv$ $\bar{Y}_j - \hat{Y}_j$ $\qquad \hat{Y}_j = \hat{\beta}_0 + \hat{\beta}_1 X_j \quad \bar{Y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{ij}$

Show: $\sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(Y_{ij} - \hat{Y}_j\right)^2 = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(Y_{ij} - \bar{Y}_j\right)^2 + \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(\bar{Y}_j - \hat{Y}_j\right)^2$

Q.6. EXPERIMENTS ON OYSTERS

A study was conducted to measure the effects of several factors on the growth of oyster shells the variables under study are:

- Response: growth (mm in shell width)
- Predictor 1:Food Concentration
- Predictor 2: Flow speed

The authors fit 2 models:

COMPLETE MODEL:

$$\text{E(Width)} = \beta_0^* + \beta_1^*(\text{Food conc}) + \beta_2^*(\text{Flow}) + \beta_3^*\left(\text{Flow}^2\right) + \beta_4^*(\text{Conc x Flow})$$

SOURCE	*DF*	SS
REGRESSION	4	101.68
RESIDUAL	15	37.35
TOTAL	19	139.03

PARAMETER	*ESTIMATE*	*STANDARD ERROR*
INTERCEPT	0.96	N/A
FOOD CONC	2.52	0.785
FLOW	1.72	0.595
FLOW²	-0.10	0.064
CONC X FLOW	-0.19	0.204

REDUCED MODEL

$$\mathrm{E(Width)} = \beta_0^{**} + \beta_1^{**}(\mathrm{Food\,conc}) + \beta_2^{**}(\mathrm{Flow})$$

SOURCE	*DF*	SS
REGRESSION	2	93.33
RESIDUAL	17	45.70
TOTAL	19	139.03

PARAMETER	*ESTIMATE*	*STANDARD ERROR*
INTERCEPT	2.41	N/A
FOOD CONC	1.98	0.549
FLOW	0.67	0.144

- FOR EACH MODEL, GIVE THE FITTED VALUE (PREDICTION) WHEN FOOD CONC=8 AND FLOW=2.5
- GIVE THE COEFFICIENT OF DETERMINATION FOR EACH MODEL
- FOR THE COMPLETE MODEL, TEST H_0:$\beta_1=\beta_2=\beta_3=\beta_4=0$
- FOR THE COMPLETE MODEL, TEST H_0: $\beta_i=0$ (i=1,2,3,4)
- FOR THE COMPLETE MODEL, TEST H_0:$\beta_3=\beta_4=0$

SOURCE: LENIHAN, H.S., C.H. PETERSON, J.M. ALLEN (1996). "DOES FLOW SPEED ALSO HAVE A DIRECT EFFECT ON GROWTH OF ACTIVE SUSPENSION-FEEDERS: AN EXPERIMENTAL TEST ON OYSTERS," *LIMNOLOGY AND OCEANOGRAPHY*, VOL.41,#6, PP. 1359-1366

Q.7. A study obtained mortgage yields in n=18 U.S. metropolitan areas in the 1960s. The researcher obtained the following variables and fit a linear regression model to see which factors (variables) were associated with yield (each variable was obtained for each metro area):

- Y = Mortgage Yield (Interest Rate as a %)
- X_1 = Average Loan/Mortgage Ratio (High Values $\Rightarrow$ Low Down Payments/Higher Risk)
- X_2 = Distance from Boston (in miles) – (Most of population was in Northeast in the 1960s)
- X_3 = Savings per unit built (Measure of Available capital versus building rate)
- X_4 = Savings per capita
- X_5 = Population increase from 1950 to 1960 (%)
- X_6 = Percent of first mortgage from inter-regional banks (Measures flow of money from outside SMSA)

For parts a) and b), obtain the regression through a Regression package (e.g. R's lm function) and in matrix form. Conduct all tests at α = 0.05 significance level. For all parts, formally give your results, as well as computer output.

a) Fit the full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \varepsilon$.
   i. Test whether any of the independent variables are associated with mortgage yield. That is, test $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$. What proportion of variation in Y is "explained" by the independent variables?
   ii. Obtain the parameter estimates and t-tests for the individual partial regression coefficient and test $H_0 : \beta_j = 0$ individually for each variable (controlling for all others).
   iii. Obtain the partial sum of squares for each independent variable, and conduct the F-tests for $H_0 : \beta_j = 0$ individually for each variable (controlling for all others). Show that this is equivalent to the t-tests in the previous part.

b) Test whether X_2 (Distance from Boston), X_5 (Population increase from 1950 to 1960), and X_6 (Percent of first mortgage from inter-regional banks) are associated with mortgage yield, after controlling for X_1,X_3, and X_4. That is, test $H_0 : \beta_2 = \beta_5 = \beta_6 = 0$. Use the matrix form, as well as using a linear regression procedure.

c) Obtain $R(\beta_1), R(\beta_3 \mid \beta_1), R(\beta_4 \mid \beta_1, \beta_3), R(\beta_2, \beta_5, \beta_6 \mid \beta_1, \beta_3, \beta_4)$

d) Obtain $R^2(X_1), R^2(X_3 \mid X_1), R^2(X_4 \mid X_1, X_3), R^2(X_2, X_5, X_6 \mid X_1, X_3, X_4)$