**P8100 – Applied Regression I**

**Homework #9 – due Thursday 4/20/2017 1pm the latest!**

**NO late homework will be accepted.**

If you upload the HW, please make sure that you upload PDF that is not a very large file.

* HW has to be TYPED, **indicating your name and UNI on top**, and STAPLED.
* Problems must be ordered, starting with Problem #1.
* Any question (a sentence ending with a question mark) must be answered with a full and meaningful sentence using the words of the problem!
* Don’t forget to **SHOW KEY STEPS OF YOUR WORK**! You get partial credit for showing work. A final answer is not enough for full credit.
* Any hypothesis test must include a null and alternative hypothesis, test statistics, degrees of freedom (if applicable), decision and interpretation.
* Write in proper and understandable English.
* You can work in groups or discuss the problems with your classmates. However, your **written answers and solutions must be strictly your own and cannot be an exact copy of your classmates answers!**

**Required readings for week 12 (4/20) discussion and quiz:**

1. Lin, Y.C., et al., *Dairy calcium is related to changes in body composition during a two-year exercise intervention in young women.* J Am Coll Nutr, 2000. **19**(6): p. 754-60.

**Problem 1 (continuing problem 2 from homework 8):**

Import the following simulated data into SAS:

**data** HW8;

input SUBJECT AGE SEX HGT BONEM WGT ASTHMA FEV;

cards;

1 24 1 175 253 62 1 4.7

2 36 1 172 240 63.6 1 4.3

3 28 0 171 270 60.2 1 3.5

4 25 1 166 231 57.3 1 4

5 26 0 166 231 58.1 1 3.2

6 22 1 176 241 60.6 1 4.7

7 27 1 185 270 64.9 1 4.3

8 27 1 171 247 58.4 1 4.7

9 36 1 185 264 66.9 1 5.2

10 24 1 182 270 61 1 4.2

11 26 1 180 250 61.6 1 3.5

12 29 1 163 238 59.2 1 3.2

13 33 0 180 248 66.4 1 2.6

14 31 1 180 245 66.8 0 2

15 30 1 180 250 61.9 1 4

16 22 1 168 231 56.3 1 3.9

17 27 1 168 259 60 0 3

18 46 1 178 245 69.7 1 4.5

19 36 1 173 235 63.9 1 1.7

20 32 0 156 254 59.2 0 2.5

21 52 1 175 237 70.3 1 5.1

22 61 0 162 272 71.1 1 4.2

23 64 0 149 222 64.4 1 4.6

24 25 1 185 264 64.7 1 4.7

25 34 1 179 257 65 1 4.7

26 38 1 169 249 63.5 1 4.9

27 27 0 172 241 61.6 0 2.5

28 26 0 176 238 60.5 0 2.9

29 48 1 179 267 69.6 1 4.8

30 42 0 156 275 59.6 1 2.8

31 57 1 167 249 69.7 1 4.1

32 26 0 163 273 55.7 0 2.1

33 34 1 169 258 62.6 1 4.2

34 51 0 152 253 64.4 0 2.9

35 44 0 158 233 61.3 1 3.1

36 24 0 167 256 56.2 0 2

37 34 0 172 275 63 1 3

38 51 0 176 256 67.8 1 3.8

39 68 1 195 264 79.4 1 4.9

40 32 0 174 237 63.4 1 3.5

41 41 1 196 272 73.7 1 4.2

42 57 0 165 245 69.1 1 3.4

43 46 1 204 279 74.1 1 8.5

44 47 1 168 267 67.7 1 4.2

45 51 0 154 261 62 0 2.1

46 34 1 198 267 72.4 1 4.5

47 32 0 172 263 61.8 1 3.9

48 64 1 172 244 74.5 1 4.7

49 51 0 176 233 70.3 1 3.8

50 39 0 159 225 60.9 1 3.4

;

**run**;

The data were collected over random sample of subjects and contain variables with the following explanation:

FEV = ‘Forced Expiratory Volume'

SUBJECT = ‘Subject ID'

AGE = ‘Age of a subject in years'

SEX = 'SEX of a subject: 0=Female'

HGT = ‘Height of a subject in cm’

BONEM = ‘Bone Mass’

Asthma = ‘Subject having any signs of asthma: 0=Yes’

WGT = ‘Weight of a subject in kg’

Using SAS, fit a linear model that describes the Forced Expiratory Volume (FEV) as a function of main effects of Height, Weight, Age, Bone Mass, Asthma status and Gender.

1. Run the following code:

ods graphics on;

**proc** **reg** data=hw8;

model FEV= hgt wgt age bonem asthma sex / r influence spec tol vif;

output out=hw8\_output p=yhat\_i r=e\_i H=h\_ii student=r\_i rstudent=r\_i\_ext DFFITS=dffits\_i cookd=D\_i;

**run**;

**quit**;

ods graphics off;

Write down the estimated model that models the Forced Expiratory Volume (FEV) as a function of main effects of Height, Weight, Age, Bone Mass, Asthma status and Gender. Attach appropriate SAS table.

1. Asses whether there is a potential of collinearity in the model using tolerance and VIF. Attach appropriate SAS table.
2. To further assess whether the collinearity is due to just 2 variables associated with each other, compute a correlation matrix using SAS PROC CORR for all of the predictors in the model. Report the correlation matrix. Attach appropriate SAS table.
3. Between which two predictors is the correlation the strongest? Test, whether the correlation is significant using the appropriate test on level of significance 5%. Write down the null and alternative hypothesis, test statistics and corresponding degrees of freedom (if applicable), critical value or p-value and interpretation of the conclusion.
4. If you think that there is collinearity present in the initial model analyzed in part a), suggest different models (more than one) for the FEV that would avoid the collinearity. Write down all the potential population models that you suggest.
5. Analyze all the models that you suggested in part d) that avoid the collinearity with the VIF and TOL option and for each, report the SAS Parameter Estimates table. Are the models free of collinearity problems?
6. Among the models that you suggested in d) and analyzed in e), pick the one that has the highest Adjusted Coefficient of Determination as your final new model. Report the final new model fitted equation and its Coefficient of Determination and Adjusted Coefficient of Dermination.
7. Using PROC UNIVARIATE, plot the qqplot and histogram of the residuals in your final new model from part f). Attach the plots.
8. Assess your final model from part f) for influential outliers. Are any subjects in your final model that are influential outliers?
9. Test whether the residuals in your final model from part f) are heteroscedastic on level of significance 5%. Write down the null and alternative hypothesis, test statistics and corresponding degrees of freedom, critical value or p-value and interpretation of the conclusion. Attach appropriate SAS table.

**Problem 2:**  
Import the following simulated data into SAS:

**data** airlinecost;

input Airline $ length speed day\_time pop\_served tot\_cost revenue load capacity assets funds assets\_adj;

cards;

AllAmerican 57 133 6.10 20200 116.3 0.96 0.400 2.400 21.13 3.21 17.92

American 270 216 6.93 56928 43.0 3.98 0.689 5.776 1436.53 165.22 1271.31

Bonanza 100 140 4.45 183 141.5 0.79 0.358 2.207 6.65 0.01 6.64

Braniff 176 182 6.60 11869 50.6 2.57 0.557 4.614 160.30 5.81 154.49

Capital 142 167 7.47 41097 51.0 2.68 0.510 5.255 195.02 6.06 188.96

Central 51 134 4.67 1757 318.5 0.35 0.167 2.096 14.02 0.01 14.01

CS 175 175 8.60 18000 59.2 2.17 0.558 3.889 114.16 3.11 111.05

Colonial 112 150 6.87 13500 77.0 1.68 0.505 3.327 35.34 0.79 34.55

Continental 131 179 6.50 3831 62.3 1.70 0.537 3.166 49.74 3.49 46.25

Delta 174 191 7.60 13119 45.3 2.74 0.598 4.582 174.32 3.87 170.45

Eastern 182 187 9.50 44000 42.6 3.07 0.528 5.814 1042.58 187.84 854.74

Empire 59 143 4.83 451 112.4 0.69 0.313 2.204 4.71 0.02 4.69

Frontier 81 141 7.47 2500 125.2 0.72 0.398 1.809 17.66 0.12 17.54

LakeCentral 73 142 3.85 5405 169.3 0.48 0.212 2.264 6.32 0.01 6.31

MidContinent 144 167 6.20 6725 64.8 1.53 0.565 2.708 76.12 2.44 73.68

Mohawk 79 137 5.88 9250 100.5 1.02 0.476 2.143 14.54 0.09 14.45

National 199 207 8.28 23431 42.9 3.12 0.567 5.503 181.27 16.14 165.13

Northeast 94 150 6.77 16000 81.1 1.19 0.597 1.988 49.38 2.11 47.27

Northwest 271 202 6.88 27000 56.7 4.30 0.570 7.544 471.50 7.98 463.52

Piedmont 90 153 8.13 3362 75.4 1.08 0.449 2.405 21.60 3.10 18.50

Pioneer 89 150 7.05 2050 71.3 1.16 0.483 2.402 19.11 0.01 19.10

Southern 67 147 6.15 2337 150.1 0.60 0.317 1.893 12.44 0.01 12.43

Southwest 54 127 5.85 5143 78.9 1.18 0.491 2.403 16.06 0.08 15.98

TransTexas 78 147 6.08 2600 130.3 0.65 0.372 1.747 11.10 0.42 10.68

TWA 293 193 6.40 51500 46.2 3.62 0.670 5.403 1217.98 120.07 1097.91

United 270 207 6.63 48913 42.3 3.75 0.630 5.952 1127.25 79.98 1047.27

WestCoast 58 135 4.50 1850 103.3 0.82 0.421 1.948 10.98 0.01 10.97

Western 172 181 6.67 6250 44.2 2.68 0.656 4.085 137.39 17.02 120.37

Inland 147 158 9.38 1000 64.7 1.54 0.575 2.678 16.61 0.03 16.58

Wiggins 45 116 2.35 2500 820.9 0.07 0.166 0.422 2.03 0.62 1.41

Wisconsin 69 142 5.72 6500 130.9 0.81 0.430 1.884 10.80 0.01 10.79

;

**run**;

The data were collected in 1950 of airline cost data relating ‘Cost per revenue ton-mile’ to other variables:

AIRLINE = ‘Airline name’

LENGTH = ‘Average length of flights (miles)’

SPEED = ‘Average speed of plane (miles per hour)’

DAY\_TIME = ‘Average daily flight time per plane (hours)’

POP\_SERVED = ‘Population served (in 1000)’

TOT\_COST = ‘Total operating cost (cents per revenue ton-mile)’

REVENUE = ‘Revenue Tons per Aircraft mile’

LOAD = ‘Ton-Mile load factor (proportions)’

CAPACITY = ‘Available capacity (tons per mile)’

ASSETS = ‘Total Assets (in $100,000)’

FUNDS = ‘Investments and special funds (in $100,000)’

ASSETS\_ADJ = ‘Adjusted assets (in $100,000) = assets-funds’

Using SAS, our task is to create a model for estimating ‘Revenue’ outcome. Revenue tons per aircraft mile is a metric used in the freight industry. It accounts for the revenue brought in for the movement of one ton of goods over the distance of one mile. Revenue tons per mile is an important determinant of profitability in the freight business.

**BIVARIATE ANALYSES:**

1. Run the code for importing the data. Click on the bottom ‘Explorer’ tab in the left window, double-click on ‘Libraries’ and then on ‘Work’. Double click on AIRLINECOST data set and inspect it. Why there is a dollar sign ‘$’ after variable AIRLINE in the data input statement? Don’t forget to close the AIRLINECOST window at the end.
2. Write down the population model that models the REVENUE as a function of main effects of variable LENGTH.
3. Populate the following table below. For each predictor, run a simple regression with the outcome REVENUE.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Predictor | Slope estimate, T-test statistic, p-value | Coefficient of Determination | Sample correlation b/w predictor & outcome | Adj. R-square |
| LENGTH |  |  |  |  |
| SPEED |  |  |  |  |
| DAY\_TIME |  |  |  |  |
| POP\_SERVED |  |  |  |  |
| TOT\_COST |  |  |  |  |
| LOAD |  |  |  |  |
| CAPACITY |  |  |  |  |
| ASSETS |  |  |  |  |
| FUNDS |  |  |  |  |

1. Which predictor alone explains the largest portion of variance in REVENUE outcome?
2. Which predictor alone explains the smallest portion of variance in REVENUE outcome?

**ADJUSTED ANALYSES:**

1. Populate the table below. For each set of predictors, run a multiple regression model with the outcome REVENUE.

|  |  |  |
| --- | --- | --- |
| Predictor | Coefficient of Determination | Adj. R-square |
| LENGTH, CAPACITY (2 highest R2 from part c) |  |  |
| LENGTH, CAPACITY, SPEED |  |  |
| LENGTH, CAPACITY, DAY\_TIME |  |  |
| LENGTH, CAPACITY, POP\_SERVED |  |  |
| LENGTH, CAPACITY, TOT\_COST |  |  |
| LENGTH, CAPACITY, LOAD |  |  |
| LENGTH, CAPACITY, ASSETS |  |  |
| LENGTH, CAPACITY, FUNDS |  |  |

1. Based on your answers in part f) which predictor (from SPEED, DAY\_TIME, POP\_SERVED, TOT\_COST, LOAD, ASSETS and FUNDS) is a ‘reasonable’ third predictor to be added to model that contains already LENGTH and CAPACITY based on Adjusted R2? In other words, from the models that contain 3 predictors, which one is the best based on Adjusted R2?
2. Write down the estimated model for the three predictors that you selected in part g).
3. Interpret the coefficient of determination for the model that you selected in part g).
4. Using the formula from the lecture, compute by hand Adjusted R2 for model that you selected in part g). Compare it to the SAS output.



1. Interpret the adjusted coefficient of determination (computed above) for the model that you selected in part g).

**OUTLIERS INVESTIGATION:**

1. Using the 9 figures plot “Fit Diagnostic” that is produced SAS PROC REC, identify whether there is an influential outlier airline. If you decide that there is, write down its name.

1. In order to compare results with and without the outlier, manually delete the airline that you have identified as an outlier in part l) from the dataset and run the same model that you selected in part g) on the new dataset (without the outlier).
2. Using the 9 figures plot “Fit Diagnostic” that is produced SAS PROC REC, again identify whether there is an influential outlier airline in model from part m). If you decide that there is, write down its name.
3. Formally perform Kolmogorov-Smirnov test on residuals from model in part m), attach appropriate SAS table.
4. Formally perform White test on residuals from model in part m), attach appropriate SAS table.
5. Look at the residual plots for model in part m) (residuals vs. predicted and residuals by regressors) that are automatically created by PROC REG. Do you see any ‘nonlinear’ behavior? If you do, identify which predictor might be causing it.
6. Discuss in 1-3 sentences whether there might be any potential violations of assumption of ‘independence of observations’ or whether you think it is safe to assume the observations (airlines) are independent.