**Ximan Liu**

**Data Analysis And Regression**

**Assignment-3** | **Total Points: 30 pts for DSC 423 / 20 pts for DSC 323**

**Due Date: 4/19/2021 by 11:59 pm**

Note:

* All assignments should be submitted in a **single MS WORD format**, no PDFs or any other file types will be accepted. If you submit any other file type, it will not be graded.
* No extensions will be given unless for a documented reason specified in the syllabus, no late assignments past the due date even a couple of minutes late will be accepted as you have an extra day (8-days) to submit your assignments.
* Submitting work that is not yours is grounds for an automatic ‘F’ for the entire course – this includes taking content and ideas from others or consulting others to complete your deliverables other than your instructor.
* SAS software and virtual server stalls, gets slow and crashes; so start early and keep multiple backups in multiple places/mediums. Late submission or inability to do the assignment due to server and/or software issues will not be accepted. Any issues relating with SAS, contact IS using the phone number provided in the syllabus, I won’t be able to help you with DePaul software related issues.

***Note: For all questions, immaterial if whether the relevant output is asked to be attached or not, make sure to include it. Also, it is important to include the sign (negative/positive or increase/decrease, and units of measurements e.g. $ or $ 99 million,%, etc.) otherwise points will be deducted.***

**Problem 1 [5 pts] – ONLY for GRADUATES**

A university career center collects information on the job status and starting salary of graduating seniors. Data recently collected over a two-year period included over 900 seniors who had found employment at the time of graduation. The information was used to model starting salary Y as a function of two qualitative independent variables: COLLEGE at four levels {Business, Engineering, Liberal Arts, Nursing} and SEX (male and female).

1. Define the dummy variables to include college (use Business as your baseline) in a regression model for starting salary Y

|  |  |  |  |
| --- | --- | --- | --- |
| **COLLEGE** | **Z1** | **Z2** | **Z3** |
| **Business** | **0** | **0** | **0** |
| **Engineering** | **1** | **0** | **0** |
| **Liberal Arts** | **0** | **1** | **0** |
| **Nursing** | **0** | **0** | **1** |

**Three dummy variables for college.**

**Define:**

**Z1 = 1 if college = Engineering; Z1 = 0 otherwise;**

**Z2 = 1 if college = Liberal Arts; Z2 = 0 otherwise;**

**Z3 = 1 if college = Nursing; Z3 = 0 otherwise;**

**Z1 = 0 and Z2 = 0 and Z3 = 0 if college = Business.**

1. Write down the general regression model relating starting salary Y to both college and sex.

**Y = β0 + β1 Z1 + β2 Z2 + β3 Z3 + β4 DSEX + e**

**Dummy variable DSEX = 1 if sex = ‘male’**

1. How would your model change if students in Engineering have the same starting salary as students in Business? Show the final regression model.

**Engineering: YE = β0 + β1(1) + β2(0) + β3(0) = β0 + β1**

**Business: YB = β0 + β1(0) + β2(0) + β3(0) = β0**

**Because YE = YB, hence β0 + β1 = β0, so β1 = 0**

**With that result, we can tell that Engineering(Z1) does not have effect on Y, hence it can be removed in the model.**

**All in all, Y = β0 + β2 Z2 + β3 Z3 + β4 DSEX + e**

**Problem 2 [5 pts] – ONLY for GRADUATES**

Answer the following Questions:

1. Why the number of dummy variables used in regression model is always equal to the number of levels -1? (1 point)

**The number of dummy variables required to represent a particular categorical variable depends on the number of values that the categorical variable can assume. To represent a categorical variable that can assume k different values, a researcher would need to define k - 1 dummy variables. The number of levels is captured by the intercept and is specified when the dummy variables are all set to zero. If the number of dummy variables used in regression model is always equal to the number of levels, it will cause the regression to fail. A kth dummy variable is redundant; it carries no new information. And it creates a severe multicollinearity problem for the analysis. Using k dummy variables when only k - 1 dummy variables are required is known as the dummy variable trap. Also, there is always a baseline level as reference.**

1. Why doesn't it make sense to analyze the residual plots for a dummy variable? Use an appropriate image to explain your answer. (1 point)

A picture containing shape

Description automatically generated

**The residual plots represent residuals on vertical axis and independent variable on the horizontal axis. Dummy variables only have 0 and 1, hence the graph looks like above, which is meaningless to analyze residuals.**

1. Why are linear regression used? List 3 reasons. (3 points)

**(1) determining the strength of predictors.**

**(2) forecasting an effect.**

**(3) trend forecasting.**

**Problem 3 [5 pts] – to be answered by everyone**

You will continue the analysis of the banking.txt dataset that was analyzed in Assignment 2 – data file is attached. Answer this question based on your final model from assignment-2.

1. Analyze the residuals of the regression model you found in your previous assignment. Include the residual plots. Discuss your findings.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, line chart

Description automatically generated

**The scatter plots of Predicted Value, Income and Age are not randomly scattered around the zero line, in that case we cannot say they have constant variance. The shape of three scatter plots also violates independence assumption.**

**The residual plots for Balance, Age, and Income represents it is normal distribution and linearly.**

**There are possible outliers in the first three graphs because studentized residuals are greater than 2 or less than -2.**

1. Conduct a global F-test for overall model adequacy. Write down the test hypotheses and test statistic and discuss conclusions. Include the relevant output.

Table

Description automatically generated

**Null hypothesis:**

**Ho: β1 = β2 = β3 = … = βk = 0**

**Alternative hypothesis:**

**Ha: At least one coefficient βj ≠ 0**

**Test statistic:**

**F = Model / Error = 3521526788 / 6038289 = 583.199444081**

**Therefore, F = 583.199444081 and with p-value less than 0.05 (at alpha=0.05). The null hypothesis of no association between balance(y) and age & income(x-var) is rejected. At least one x-variable has a significant effect on changes in balance. F-test gives strong support to the fitted model.**

1. Copy and paste your FULL SAS code into the word document along with your answers.

**PROC IMPORT datafile="C:\Users\XLIU115\Desktop\Assignment3\banking.txt" out=salary replace;**

**getnames=yes;**

**RUN;**

**PROC CORR;**

**var Balance Age Income;**

**RUN;**

**PROC REG CORR;**

**title "Regression Model - Remove Education";**

**model Balance = Age Income;**

**RUN;**

**PROC REG;**

**title "Residual Plots - Balance, Age and Income";**

**model Balance = Age Income;**

**plot student.\*predicted.;**

**plot student.\*(Age Income);**

**plot npp.\*student.;**

**RUN;**

**Problem 4 [15pts] – to be answered by everyone**

A national homebuilder builds single-family homes and condominium style townhouses.

The file housesales.txt provides information on the selling price (PRICE), lot cost (COST), type of home (HOME) (SF=single family home or T=condominium style) and region of the country (REGION) (M=Midwest, S=south) for closings during one month.

1. Define the dummy variables for region and home (write them down here), and create them in SAS.

|  |  |
| --- | --- |
| **REGION** | **R1** |
| **M** | **0** |
| **S** | **1** |

|  |  |
| --- | --- |
| **HOME** | **H1** |
| **SF** | **0** |
| **T** | **1** |

**Two dummy variables.**

**Define:**

**H1 = 1 if HOME = condominium style; H1 = 0 if HOME = single family home otherwise;**

**R1 = 1 if REGION = south; R1 = 0 if REGION = Midwest otherwise.**

1. Analyze the association between selling price and each individual attribute (cost, home and region) using appropriate statistics and graphs. Discuss your findings. Include the relevant output.

Calendar

Description automatically generated

Table

Description automatically generated

**The scatter plot represents the positive linear regression relationship between selling price and cost. There is a strong association and almost no outliers. The correlation value is 0.7263 and quite close to 1 which means the assumption above is correct.**

**The residual plots for dummy variables are meaningless because they only have values of 0 and 1.**

1. Fit an adequate regression model for sales price as a function of lot cost, region of country, and type of home. Remove the terms that are not significant. The final model should only contain variables that are significantly associated with sale price. Write down the model equation. Include the relevant output.

Table

Description automatically generated

Table

Description automatically generated

**Sales Price = 126,123 + 3.50527 \* Cost – 72,566 \* H1 – 9,081.58833 \* R1**

**(where H1 = 1 if HOME = ‘T’ and R1 = 1 if REGION = ‘S’)**

**The result of T-test shows that Cost and H1 both have important effect on sales price with p-value of them are less than 0.0001.**

**The p-value for R1 is 0.3599 which is greater than 0.05, in that case we cannot say R1 does not have effect on sales price.**

**The second graph shows the situation after deleting dummy variable R1.**

**This time we have below as final model:**

**Sales Price = 116,426 + 3.65986 \* Cost - 73,847 \* H1**

**We try T-test again and find both cost and H1 have significant associated with sale price with both p-value of them are less than 0.0001.**

1. Conduct a global F-test for overall model adequacy. Write down the test hypotheses and test statistic and discuss conclusions. Include the relevant output.

Table

Description automatically generated

**Null hypothesis:**

**Ho: β1 = β2 = β3 = … = βk = 0**

**Alternative hypothesis:**

**Ha: At least one coefficient βj ≠ 0**

**Test statistic:**

**F = Mean Square Model / Mean Square Error = 3.67352E11 / 2384392143 = 154.065261907**

**Therefore, F = 154.065261907 and with p-value less than 0.05 (at alpha=0.05). The null hypothesis of no association between sales price(y) and cost & H1(x-var) is rejected. At least one x-variable has a significant effect on changes in sales price. F-test gives strong support to the fitted model.**

1. Analyze model residuals to check if assumptions on data are satisfied. Discuss your findings. Include the relevant output.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart

Description automatically generated

Chart, line chart

Description automatically generated

**Dummy variable H1 is qualitative variable with scattering along 0 and 1, which is meaningless for residual analysis.**

**The graph of Cost and Predicted Value are not randomly scattered around zero line, which no constant variance. Also, they have funnel shapes, which violate independence assumption.**

**CDF of Studentized Residual graph represents the model is normal distributed with a linear line.**

**Probably there are outliers because studentized residuals in cost exist values greater than 3 or less than -3.**

1. Discuss what the regression model indicates for the relationship between price and home type (i.e. interpret the coefficient values).

**Sales Price(Y) = 116,426 + 3.65986 \* Cost - 73,847 \* H1**

**Assume cost is fixed. Then, Sales Price(Y) = β0 + β1 \* H1**

**YSF = β0 + β1(0) = β0**

**YT = β0 + β1(1) = β0 + β1**

**YT - YSF = β1**

**Therefore, β0 is $116,426 is intercept value. In that case, sales price starts at $116,426 irrelevant with cost and home. Assume all the other variables constant, then condominium style will decrease the sale price by $73,847 compared to single family home.**

1. Use the regression analysis to determine whether mean sale prices are different for the two regions? Explain.

**The p-value for R1 is 0.3599 which greater than 0.05. That shows that dummy variable region is not a significant variable to sales price. Hence it has been removed from the model. Neither the south nor the midwest has a significant effect on the results.**

1. Copy and paste your FULL SAS code into the word document along with your answers.

TITLE "Analysis - HouseSales";

**PROC** **IMPORT** datafile="C:\Users\XLIU115\Desktop\Assignment3\HouseSales.txt" out=Price replace;

getnames=yes;

delimiter='09'x;

**RUN**;

**DATA** Price;

set Price;

H1=(Type="T");

R1=(Region="M");

**RUN**;

**PROC** **PRINT** data=Price;

**RUN**;

**PROC** **SGSCATTER**;

title "Scatterplot Matrix - Selling Price and Other Individual Attributes";

matrix Price Cost H1 R1;

**RUN**;

**PROC** **REG** CORR;

title "Residual Plots - Selling Price, Cost, Home and Region";

model Price= Cost H1 R1;

**RUN**;

**PROC** **REG**;

model Price= Cost H1 R1;

**RUN**;

**PROC** **REG**;

model Price= Cost H1;

**RUN**;

**PROC** **SGSCATTER**;

title "Scatterplot Matrix - Selling Price, Cost and Home";

matrix Price Cost H1;

**RUN**;

**PROC** **REG** corr;

title "Residual Plots - Selling Price, Cost and Home";

model Price= Cost H1;

plot student.\*predicted.;

plot student.\*(Cost H1);

plot npp.\*student.;

**RUN**;