Homework 2

Group 4

BUAN 6337 Predictive Analytics using SAS

Professor B.P.S. Murthi

Sept 24, 2020

Lin, Shih-Min

Devineni, Rajasekhar

Talreja, Vinita Govind

Zhou, Wei

Read the data

|  |
| --- |
| **DATA** cc\_data;  INFILE "H:\creditcard\_data.dat" FIRSTOBS =**2**;  INPUT MDR Acc Age Income Avgexp Ownrent Selfempl;  **RUN**; |

**SAS Homework 2**

**Note: This homework is to be done in groups**.

Income and Expenditure Data. 100 Cross Section Observations

Source: Greene (1992)

MDR = Number of Derogatory Reports

Acc = Credit card application accepted (1=yes),

Age = Age in years+ 12ths of a year,

Income = Income, divided by 10,000 ,

Avgexp = Avg. monthly credit card expenditure,

Ownrent = OwnRent, individual owns (1) or rents (0) home.

Selfempl = Self employed (1=yes, 0=no)

We are interested in modeling “Avgexp” as a function of the other variables.

1. Find the basic statistics (means and frequencies) for each of the variables. Comment on what you learned.

|  |
| --- |
| **Proc** **means**;  var Age Income Avgexp;  **Run**;  **Proc** **freq;**  table MDR Acc Ownrent Selfempl;  **Run**; |
|  |
| The average age in the data is 32 years.  The average monthly credit card expenditure accounts for 0.6% of average income.  Most people have credit cards. The proportion of people who have cards is 73% in the dataset. 64% of the people rent home and 95% of people were hired by others, which means they are not self-employed.  According to the output, can find that most people, who have credit card, also rent home, and were hired by others.  82% of the people have no Derogatory Reports. |

1. Find the correlations between the independent variables. What did you learn? Which variables had the highest correlations?

|  |
| --- |
| **PROC** **CORR** DATA=cc\_data;  VAR Avgexp MDR Acc Age Income Ownrent Selfempl;  **RUN**; |
|  |
| All the independent variables have statistically significant linear relationship with at least one other variable.  Correlation between variables:  MDR and Acc – 0.40  Income & Ownrent – 0.34  Age & Ownrent – 0.28  Age & Income – 0.26  Income & Selfempl – 0.20  The variable with highest correlation is MDR and Acc |

1. Find the best regression model (use PROC REG) to explain average monthly credit card expenditure. Comment on all model results including (F-test, R-sq, Adjusted R-sq, coefficients, t-values and explain in plain English the meaning of the coefficients).

|  |
| --- |
| **PROC** **REG** data = cc\_data PLOTS = none;  MODEL Avgexp = MDR Acc Age Income Ownrent Selfempl;  **RUN**;  /\*Adj. R-sq value of 0.26 \*/  **PROC** **REG** data = cc\_data PLOTS = none;  MODEL Avgexp = MDR Acc Income Ownrent Selfempl;  **RUN**;  /\*Best model with highest Adj. R-sq value of 0.28 \*/  **PROC** **REG** data = cc\_data ;  MODEL Avgexp = Acc Income;  **RUN**; |
|  |
| * The Avg. monthly credit card expenditure increases by 248.5 units if that person's credit card application is accepted. * For every 10,000 increase in Income of a person, their Avg. monthly credit card expenditure increases by 66.7 units * t-value for Acc & Income says that they have non-zero effect on Avgexp   for 95% of the times   * p- value for Acc & Income are <0.05, hence we can say that they are significant at confidence level >95% * R-sq value says that this model explains 29% of variance in Avgexp variable * Adj. R-sq says that it increased from previous model, 0.26 to .28 as the result of removing non-significant variables. Hence this is a better model. * F- Test values says that there is at least one coefficient with non-zero effect on Avgexp in this model.   F-Test:  H0: All the independent variables do not affect dependent variable.  H1: There exists at least one independent variable that affects dependent variable. |

1. Which variable is the most important in explaining “AvgExp”? How did you find this? Rank all explanatory variables in terms of importance.

|  |
| --- |
| /\*Q4\*/  **PROC** **REG** data = cc\_int PLOTS = none;  MODEL Avgexp = Acc Income/stb vif collin influence;  **RUN**;  **PROC** **REG** data = cc\_int PLOTS = none;  MODEL Avgexp = MDR Acc Age Income Ownrent Selfempl/stb vif collin influence;  **RUN**; |
|  |
| The most important variable to explain "Avgexp" is "Acc", this is found using Standerdized Estimates from the regression. Ranking all explanatory variables in terms of importance.  1. Acc  2. Income  3. Ownrent  4. MDR  5. Selfempl  6. Age |

1. Test for the presence of nonlinearity effect for one X variable. Comment on what you find.

|  |
| --- |
| **data** cc\_int; set cc\_data;  income\_2 = Income\*Income;  age\_2 = Age\*Age;  MDR\_2 = MDR\*MDR;  **PROC** **REG** data = cc\_int PLOTS = none;  MODEL Avgexp = Acc Income income\_2;  **RUN**; |
|  |
| Checked for nonlinearity effect for Income and MDR but found that all of them have linear effect on Avgexp |

1. Test for one interaction effect between any two explanatory variables. Comment on what you find and the meaning of the interaction coefficient.

|  |
| --- |
| **data** cc\_int; set cc\_data;  income\_acc = Income\*Acc;  income\_mdr = Income\*MDR;  **PROC** **REG** data = cc\_int PLOTS = none;  MODEL Avgexp = Acc Income MDR income\_mdr income\_acc;  **RUN**; |
|  |
| Checked for interaction variables on Income, Acc and Income, MDR  found that all of them are not statistically significant. |

1. Perform diagnostics for multicollinearity (VIF, COLLIN) as discussed in class. Comment on your findings whether there is multicollinearity in the model.

|  |
| --- |
| **PROC** **REG** data = cc\_int PLOTS = none;  MODEL Avgexp = Acc Income/stb vif collin influence;  **RUN**; |
|  |
| Since the variation Inflation is <10, we can say that there is no multicollinearity  The highest Condition Index value<100, hence we can say that there is no multicollinearity |

1. Test for heteroscedasticity using White test (use PROC MODEL code). Estimate a weighted least squares model on this data if you find evidence of heteroscedasticity. Comment on how the WLS results are different compared to the OLS regression results.

|  |
| --- |
| /\*  White Test  H0: Errors are Homogenous  H1: Errors are Heteroskedastic  \*/  **PROC** **MODEL** DATA = cc\_int;  params b0 b1 b2;  Avgexp = b0 + b1\*Acc + b2\*Income;  fit Avgexp/white;  **RUN**; |
| A screenshot of a cell phone  Description automatically generated |
| White test from the Heteroskedasticity Test cannot the reject the null hypothesis that  the errors are homogenous despite few outliers in the residuals, hence we can conclude that the standard errors are reliable and there is no heteroskedasticity. |

1. Send your SAS program and output files by email. Remember to put your group number on each document.