**DSC423: Data Analysis And Regression / DSC 324: Data Analysis & Statistical Software II**

**Assignment-2** | **Total points: 15 for DSC 324 / 20 for DSC 423**

**Due Date: 04/16/2019 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 [15 pts] – to be answered by everyone**

The file banking.txt attached to this assignment provides data acquired from banking and census records for different zip codes in the bank’s current market. Such information can be useful in targeting advertising for new customers or for choosing locations for branch offices. The data show

* median age of the population (AGE)
* median income (INCOME) in $
* average bank balance (BALANCE) in $
* median years of education (EDUCATION)

In this exercise you are asked to apply regression analysis techniques to describe the effect of age education and income on average account balance.

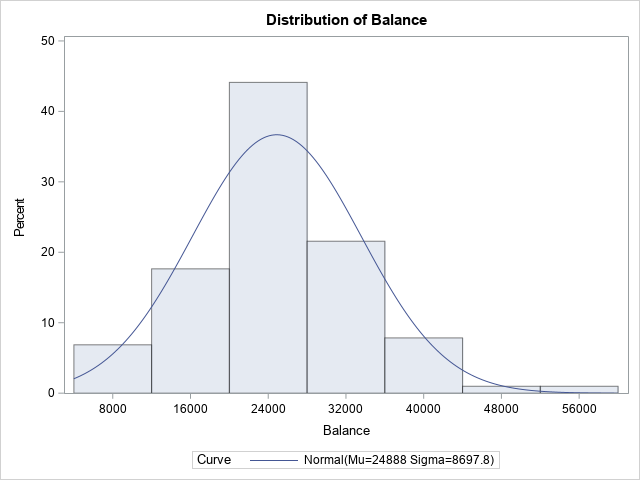
1. Analyze the distribution of average account balance using histogram, and compute appropriate descriptive statistics. Write a paragraph describing distribution of Balance and use appropriate descriptive statistics to describe center and spread of the distribution. Discuss your findings. Also, do you see any outliers? Include the histogram.

|  |
| --- |
| descriptives |

The MEANS Procedure

| **Analysis Variable : Balance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Mean** | **Minimum** | **25th Pctl** | **50th Pctl** | **75th Pctl** | **Maximum** |
| 24887.88 | 5956.00 | 20020.00 | 24660.50 | 29198.00 | 56569.00 |
| histogram | | | | | |

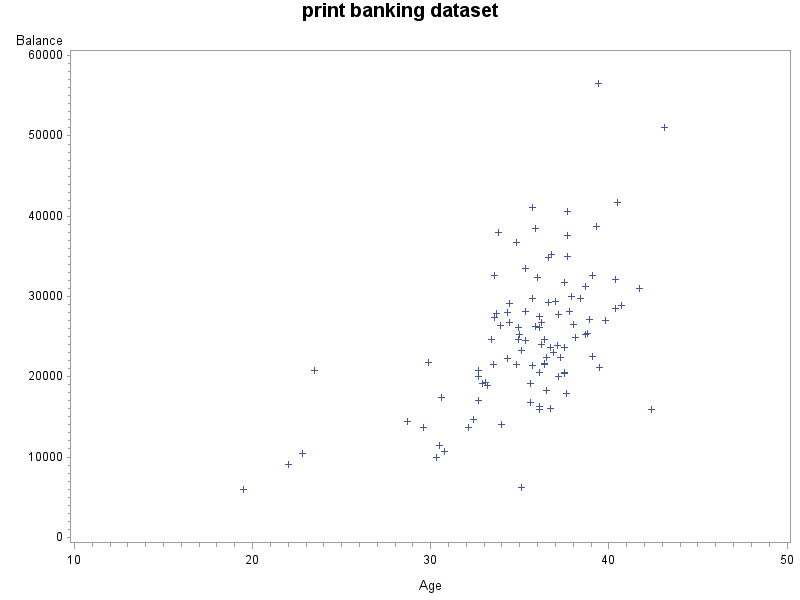
The UNIVARIATE Procedure

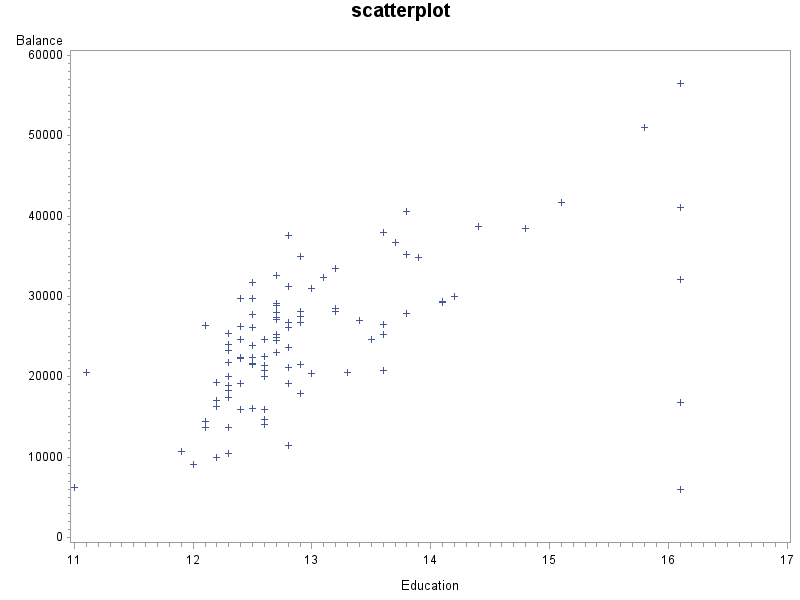


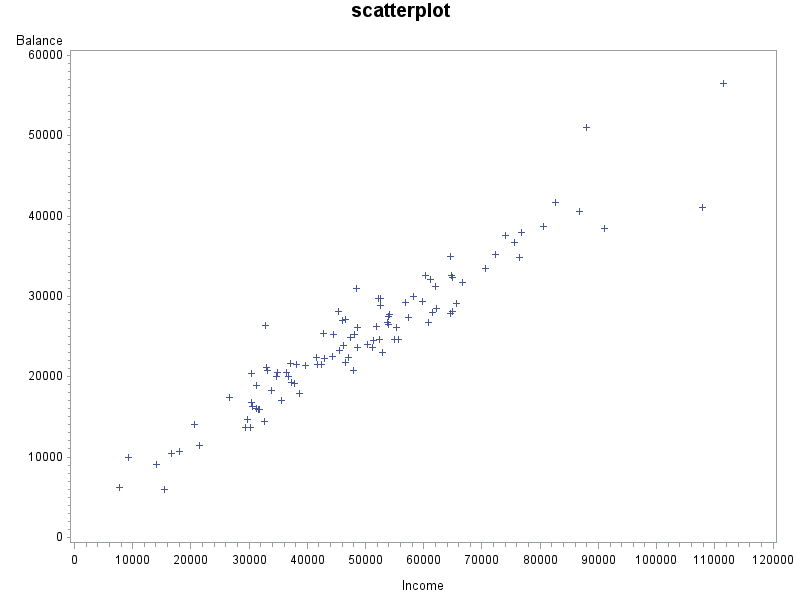
The Mean of account balance lays at $24,887. The inner quartiles show a range of $20,020 to $29,198, with a low of $5956 and a high of $56,569. Although the max was almost near $60k, a majority of the sample was far below that in the mid-$20k. There do not seem to be any apparent outliers, but one can argue that $56,000 may be a possible outlier. The distribution spread seems to be a little right-skewed, with $24,660.50 being the median.

1. Create scatterplots to visualize the associations between bank balance and the other variables. Discuss the patterns displayed by the scatterplot. Also, do the associations appear to be linear? (You can create scatterplots or a matrix plot). Include the scatterplots.

Upon assessing the scatterplots below, we can safely assume that all include a positive correlation. The association between bank balance and income seem to be linear and have a high positive correlation. Education and balance have some potential outliers and doesn’t clearly define a linear trend. Balance and age cannot be said to have a clear linear association, but this can be arguable.







1. Compute correlation values of bank balance vs the other variables. Interpret the correlation values, and discuss which pairs of variables appear to be strongly associated. Include the relevant output that shows the correlation values.

| **Pearson Correlation Coefficients, N = 102  Prob > |r| under H0: Rho=0** | | | | |
| --- | --- | --- | --- | --- |
|  | **Balance** | **Age** | **Education** | **Income** |
| **Balance** | |  | | --- | | 1.00000 | |  | | |  | | --- | | 0.56547 | | <.0001 | | |  | | --- | | 0.55488 | | <.0001 | | |  | | --- | | 0.95168 | | <.0001 | |
| **Age** | |  | | --- | | 0.56547 | | <.0001 | | |  | | --- | | 1.00000 | |  | | |  | | --- | | 0.17341 | | 0.0813 | | |  | | --- | | 0.47715 | | <.0001 | |
| **Education** | |  | | --- | | 0.55488 | | <.0001 | | |  | | --- | | 0.17341 | | 0.0813 | | |  | | --- | | 1.00000 | |  | | |  | | --- | | 0.57539 | | <.0001 | |
| **Income** | |  | | --- | | 0.95168 | | <.0001 | | |  | | --- | | 0.47715 | | <.0001 | | |  | | --- | | 0.57539 | | <.0001 | | |  | | --- | | 1.00000 | |

Income and Balance have a correlation value of .95168 and this is a very high correlation. Education and balance sit at 1.000 which is at the highest of all the variable associations.

1. What is the dependent variable and what are the independent variables in this regression analysis?

Dependent Variable is Balance and Independent are income education and age

|  |
| --- |
| print banking dataset |

The REG Procedure

Model: MODEL1

Dependent Variable: Balance

|  |  |
| --- | --- |
| **Number of Observations Read** | 102 |
| **Number of Observations Used** | 102 |

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 3 | 7048628348 | 2349542783 | 388.80 | <.0001 |
| **Error** | 98 | 592215797 | 6043018 |  |  |
| **Corrected Total** | 101 | 7640844145 |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 2458.25514 | **R-Square** | 0.9225 |
| **Dependent Mean** | 24888 | **Adj R-Sq** | 0.9201 |
| **Coeff Var** | 9.87732 |  |  |

| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | **1** | -9539.94542 | 4423.05947 | -2.16 | 0.0335 |
| **Age** | **1** | 332.50007 | 72.33549 | 4.60 | <.0001 |
| **Education** | **1** | 288.65551 | 300.53376 | 0.96 | 0.3392 |
| **Income** | **1** | 0.38705 | 0.01748 | 22.14 | <.0001 |

1. Use SAS to fit a regression model to predict balance from age, education and income. Analyze the model parameters. Which predictors have a significant effect on balance? Use the t-tests on the parameters for alpha=0.05. Include the relevant regression output.

Age and income would have a significant effect on balance because the p-value is less than .05.

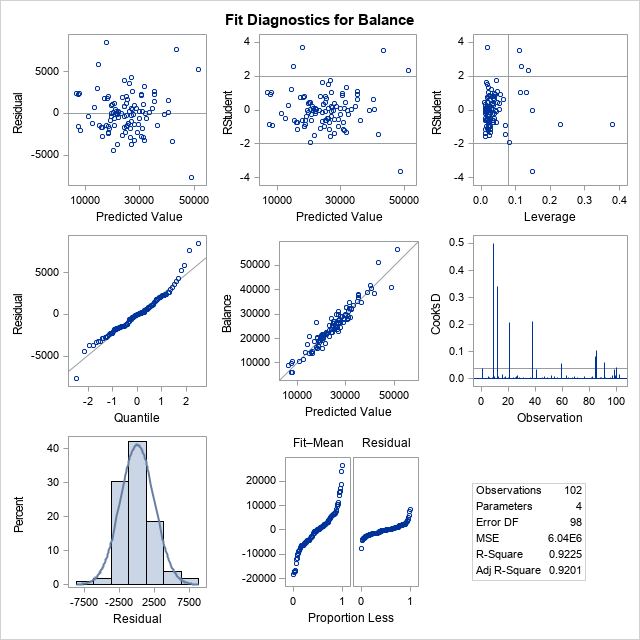
Balance = -9539.94542 + 332.50007 age+288.65551 education+.38705 income

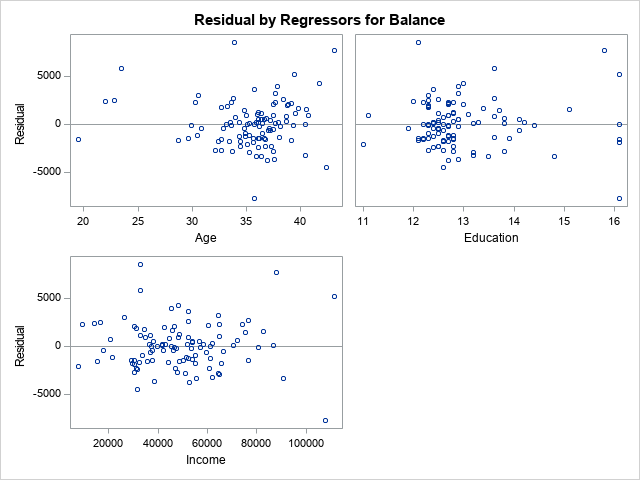
|  |
| --- |
| regression model |

The REG Procedure

Model: MODEL1

Dependent Variable: Balance





|  |
| --- |
| print banking dataset |

The REG Procedure

Model: MODEL1

Dependent Variable: Balance

|  |  |
| --- | --- |
| **Number of Observations Read** | 102 |
| **Number of Observations Used** | 102 |

| **Analysis of Variance** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| **Model** | 2 | 7043053576 | 3521526788 | 583.20 | <.0001 |
| **Error** | 99 | 597790568 | 6038289 |  |  |
| **Corrected Total** | 101 | 7640844145 |  |  |  |

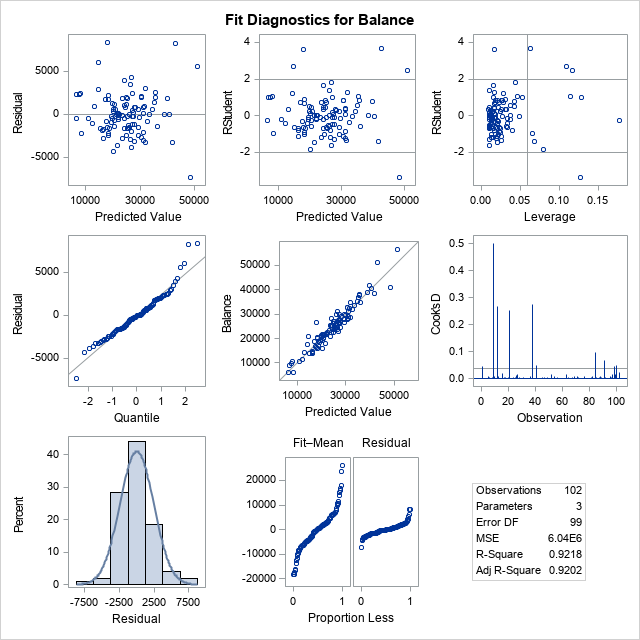
|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 2457.29294 | **R-Square** | 0.9218 |
| **Dependent Mean** | 24888 | **Adj R-Sq** | 0.9202 |
| **Coeff Var** | 9.87345 |  |  |

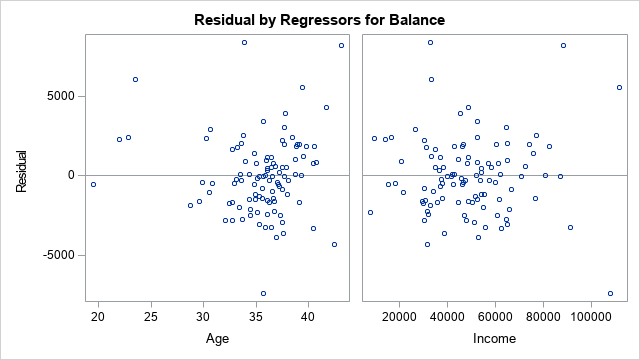
| **Parameter Estimates** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** |
| **Intercept** | **1** | -5912.21531 | 2300.75819 | -2.57 | 0.0117 |
| **Age** | **1** | 322.72377 | 71.58774 | 4.51 | <.0001 |
| **Income** | **1** | 0.39661 | 0.01437 | 27.60 | <.0001 |

1. If one of the predictors is not significant, remove it from the model and refit the new regression model. Write the expression of the newly fitted regression model.

We would remove education because the p-value is 0.3392.

Balance = -5912.21531+322.72377 age + .39661 income





1. Interpret the value of the parameters for the variables in the model.

The Adj-R2 value is increased when removing the education variable. We can conclude that taking out the education variable(deemed insignificant), helped improve the model. Comparing the Adj-R2 value between two models of the same dataset is the most effective way of indicating improvement.

1. Report the value for the R2and Adj-R2 coefficient and describe what it indicates. Include the portion of the output that includes the R2and Adj-R2 coefficient values.

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 2458.25514 | **R-Square** | 0.9225 |
| **Dependent Mean** | 24888 | **Adj R-Sq** | 0.9201 |
| **Coeff Var**  Remove education: | 9.87732 |  |  |
| **Root MSE** | 2457.29294 | **R-Square** | 0.9218 |
| **Dependent Mean** | 24888 | **Adj R-Sq** | 0.9202 |
| **Coeff Var** | 9.87345 |  |  |

The R2 and Adj-R2 values are high, which means that it is a good model. When removing the insignificant value (education), we improved the model and that was shown in Adj-R2 .0001.

1. According to census data, the population for a certain zip code area has median age equal to 34.8 years, median education equal to 12.5 years and median income equal to $42,401.

* Use the final model computed in step (f) above to compute the predicted average balance for the zip code area.

Balance = -5912.21531+322.72377 age + .39661 income

Balance = -5912.21531+322.72377 (34.8) + .39661(42401) =

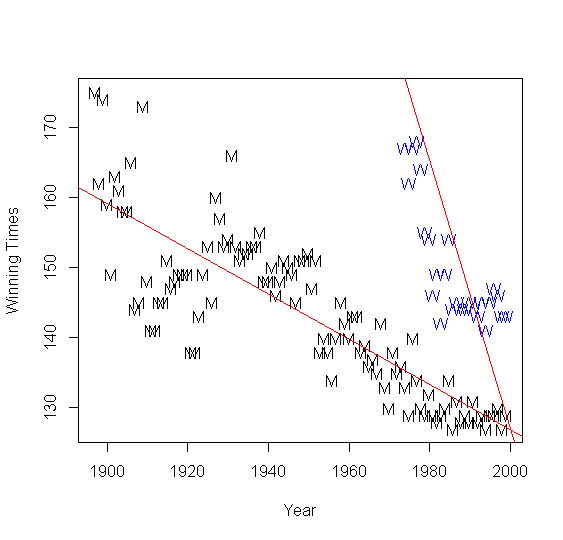
Balance = -5912.21531+11230.7872+ 16816.6606 = **$22,135.2325**

* If the observed average balance for the zip code area is $21,572, what’s the model prediction error?

22135.2325-21572 = $563.2325

1. Copy and paste your FULL SAS code into the word document along with your answers.
2. **proc** **import** datafile = "Banking.txt" out = Banking replace;
3. delimiter = '09'x;
4. getnames = yes;
5. datarow = **2**;
6. **run**;
7. title "print dataset using import";
8. **proc** **print**;
9. **run**;
10. title "descriptives";
11. **proc** **means** mean min p25 p50 p75 max;
12. var Balance;
13. **run**;
14. title "histogram";
15. **proc** **univariate** normal;
16. var Balance;
17. histogram / normal (mu=est sigma=est);
18. **run**;
19. title "scatterplot";
20. **PROC** **GPLOT**;
21. PLOT Balance\*(Age Education Income);
22. **RUN**;
23. title "Correlation";
24. **Proc** **CORR**;
25. var Balance Age Education Income;
26. **run**;
27. title "regression model";
28. **PROC** **reg**;
29. model balance= age income;
30. **run**;

**Problem 2 [5 points] - ONLY for Graduate Students**

Historical data about the Boston Marathon can be found on its website. The graph shows winning times (in minutes) for men and women against the year in which the race was run. Men’s times are represented by “M” and women’s time by “W”. The graph also displays two regression lines of winning times vs year for men and women. There is no dataset for this question, but answer the following questions based on the graph.

1. Consider the men’s winning times, is there evidence of a linear trend? Would you expect the slope of the regression line to be positive or negative? Yes, there is evidence of a linear trend. Negative slope
2. Now let’s consider the winning times for women, is there evidence of a linear trend? Discuss.

Based on the data shown, the winning time for women (marked by “W’S”) shows a linear trend. There seems to be a general linear trend when looking at the data points, but we can base this off of the line of best fit as well. There seems to be a large difference in the amount of data points for women compared to men, but we can say there is indeed evidence of a linear trend.

1. If we fit two separate linear regression models for men’s and women’s winning times, which slope will be greater in absolute value?

The slope for women’s winning time would be greater than of men’s. The “slope” of the line can be mathematically figured out based on the slope formula. The women’s slope seems to take a much sharper angle than of men’s and so the absolute value of the slope will result in women’s winning time being greater than of mens.