Mai Ngo

5/29/2023

DATA ANALYSIS AND REGRESSION – DSC 423

Final Project \_ SAS Code

\*Load the dataset and print first few observation;

**PROC** **IMPORT** datafile="C:\Users\MNGO\OneDrive - DePaul University\salaries.csv" out=salary replace;

DELIMITER = ',';

GETNAMES = YES;

DATAROW = **2**;

**RUN**;

TITLE "Salary Data 10 observation";

**PROC** **PRINT** DATA=salary (OBS=**10**);

**RUN**;

\*Get data types of each attribute;

TITLE "Data Types";

**PROC** **CONTENTS** data=salary;

**RUN**;

\*Checking distribution of each attribute;

\*Starting with predicted value - salary in USD;

TITLE "Distribution of Salary (USD)";

**PROC** **UNIVARIATE** data=salary normal;

VAR salary\_in\_usd;

HISTOGRAM / normal(mu=est sigma=est);

**RUN**;

\*Working Year;

TITLE "Distribution of Working Year - Bar Chart";

**PROC** **FREQ**;

TABLES work\_year;

**RUN**;

**PROC** **SGPLOT** DATA = salary;

VBAR work\_year;

**RUN**;

TITLE "Boxplot: Working Year vs. Salary (USD)";

**PROC** **SORT**;

BY work\_year;

**RUN**;

**PROC** **BOXPLOT**;

PLOT salary\_in\_usd\*work\_year;

**RUN**;

\*Experience Level;

TITLE "Distribution of Experience Level - Bar Chart";

**PROC** **FREQ**;

TABLES experience\_level;

**RUN**;

**PROC** **SGPLOT** DATA = salary;

VBAR experience\_level;

**RUN**;

TITLE "Boxplot: Experience vs. Salary (USD)";

**PROC** **SORT**;

BY experience\_level;

**RUN**;

**PROC** **BOXPLOT**;

PLOT salary\_in\_usd\*experience\_level;

**RUN**;

\*Employment Type;

TITLE "Distribution of Employment Type - Bar Chart";

**PROC** **FREQ**;

TABLES employment\_type;

**RUN**;

**PROC** **SGPLOT** DATA = salary;

VBAR employment\_type;

**RUN**;

\*Salary - Original Currency;

TITLE "Distribution of Salary (Original currency)";

**PROC** **UNIVARIATE** data=salary normal;

VAR salary;

HISTOGRAM / normal(mu=est sigma=est);

**RUN**;

\*Salary Currency;

TITLE "Distribution of Salary Currency - Bar Chart";

**PROC** **FREQ**;

TABLES salary\_currency;

**RUN**;

**PROC** **SGPLOT** DATA = salary;

VBAR salary\_currency;

**RUN**;

\*Employee Residence;

TITLE "Distribution of Employee Residence - Bar Chart";

**PROC** **FREQ**;

TABLES employee\_residence;

**RUN**;

**PROC** **SGPLOT** DATA = salary;

VBAR employee\_residence;

**RUN**;

\*Company Location;

TITLE "Distribution of Company Location - Bar Chart";

**PROC** **FREQ** data=salary;

TABLES company\_location;

**RUN**;

**PROC** **SGPLOT** DATA = salary;

VBAR company\_location;

**RUN**;

\*Company Size;

TITLE "Distribution of Company Size - Bar Chart";

**PROC** **FREQ** data=salary;

TABLES company\_size;

**RUN**;

**PROC** **SGPLOT** DATA = salary;

VBAR company\_size;

**RUN**;

\*Remote Ratio;

TITLE "Convert Remote Ratio to categorical variable";

**DATA** salary;

SET salary;

remote\_ratio = put(remote\_ratio, $CHAR.);

**RUN**;

TITLE "Distribution of Remote Ratio - Bar Chart";

**PROC** **FREQ** data=salary;

TABLES remote\_ratio;

**RUN**;

**PROC** **SGPLOT** DATA = salary;

VBAR remote\_ratio;

**RUN**;

TITLE "Boxplot: Remote Ratio vs. Salary (USD)";

**PROC** **SORT**;

BY remote\_ratio;

**RUN**;

**PROC** **BOXPLOT**;

PLOT salary\_in\_usd\*remote\_ratio;

**RUN**;

\*Transform predicted attribute for final dataset;

TITLE "Transform DV";

**DATA** salary;

SET salary;

sqrtSalary\_in\_usd = sqrt(salary\_in\_usd);

**RUN**;

**PROC** **PRINT** DATA=salary (OBS=**10**);

**RUN**;

TITLE "Distribution of Sqrt Salary (USD)";

**PROC** **UNIVARIATE** data=salary normal;

VAR sqrtSalary\_in\_usd;

HISTOGRAM / normal(mu=est sigma=est);

**RUN**;

\*Correlation between Salary(USD) and Salary(org currency);

TITLE "Salary (USD) vs. Salary (org currency): Before transformation";

**PROC** **SGPLOT** data=salary;

SCATTER x=sqrtSalary\_in\_usd y=salary;

**RUN**;

\*Convert attribute for final dataset;

**DATA** finalSalary;

SET salary;

\*Dummy variable set US value = 1 | Employee Residence and Company Location;

IF employee\_residence = "US" THEN numEmployee\_residence = **1**;

ELSE numEmployee\_residence = **0**;

IF company\_location = "US" THEN numCompany\_location = **1**;

ELSE numCompany\_location = **0**;

\*Dummy variable | Experience Level;

IF experience\_level in ("EN", "MI") THEN experience\_New = **1**;

ELSE experience\_New = **0**;

\*Dummy variable | Employment Type;

IF employment\_type = "FT" THEN employment\_FT = **1**;

ELSE employment\_FT = **0**;

\*Dummy variable | Company Size;

IF company\_size = "M" THEN size\_M = **1**;

ELSE size\_M = **0**;

\*Dummy variable | Remote Ratio;

IF remote\_ratio = "0" THEN remote\_ratioNone = **1**;

ELSE remote\_ratioNone = **0**;

\*Dummmy variable | Salary Currency;

IF salary\_currency = "USD" THEN salary\_currencyUSD = **1**;

ELSE salary\_currencyUSD = **0**;

DROP experience\_level employment\_type employee\_residence company\_location job\_title salary\_currency company\_size remote\_ratio salary\_in\_usd;

**RUN**;

TITLE "Final Salary Data 10 observation";

**PROC** **PRINT** DATA=finalSalary (OBS=**10**);

**RUN**;

TITLE "Descriptive Statistics - Final Dataset";

**PROC** **MEANS** min max median p25 p75;

VAR sqrtSalary\_in\_usd work\_year salary numEmployee\_residence numCompany\_location experience\_New employment\_FT size\_M remote\_ratioNone salary\_currencyUSD;

**RUN**;

TITLE "Correlation Values between Variables";

**PROC** **CORR**;

VAR sqrtSalary\_in\_usd work\_year salary numEmployee\_residence numCompany\_location experience\_New employment\_FT size\_M remote\_ratioNone salary\_currencyUSD;

**RUN**;

TITLE "Regression Result: Full model";

**PROC** **REG** data=finalSalary;

MODEL sqrtSalary\_in\_usd = work\_year salary numEmployee\_residence numCompany\_location experience\_New employment\_FT size\_M remote\_ratioNone salary\_currencyUSD / STB VIF TOL;

**RUN**;

\*Residual plot: Studentized residual vs.Predictors;

TITLE "Residual Plot: Studentized Residual vs. numEmployee\_residence (US)";

PLOT student.\* (work\_year salary numEmployee\_residence numCompany\_location experience\_New employment\_FT size\_M remote\_ratioNone salary\_currencyUSD);

**RUN**;

\*Residual plot: Studentized residual vs. predicted value;

TITLE "Residual Plot: Studentized Residual vs. Predicted Value";

PLOT student.\* predicted.;

**RUN**;

\*Normal probability plot or QQ plot;

TITLE "Normal Probability plot";

PLOT npp.\* student.;

**RUN**;

\*Remove numCompany\_location due to multicollinearity;

TITLE "Regression Result: 2nd model";

**PROC** **REG** data=finalSalary;

MODEL sqrtSalary\_in\_usd = work\_year salary numEmployee\_residence experience\_New employment\_FT size\_M remote\_ratioNone salary\_currencyUSD / STB VIF TOL;

**RUN**;

\*Residual plot: Studentized residual vs.Predictors;

TITLE "Residual Plot: Studentized Residual vs. numEmployee\_residence (US)";

PLOT student.\* (work\_year salary numEmployee\_residence experience\_New employment\_FT size\_M remote\_ratioNone salary\_currencyUSD);

**RUN**;

\*Residual plot: Studentized residual vs. predicted value;

TITLE "Residual Plot: Studentized Residual vs. Predicted Value";

PLOT student.\* predicted.;

**RUN**;

\*Normal probability plot or QQ plot;

TITLE "Normal Probability plot";

PLOT npp.\* student.;

**RUN**;

\*Run model with stepwise selection method;

TITLE "Regression Result: Model Selection Stepwise";

**PROC** **REG** data=finalSalary;

MODEL sqrtSalary\_in\_usd = work\_year salary numEmployee\_residence experience\_New employment\_FT size\_M remote\_ratioNone salary\_currencyUSD / selection = stepwise;

**RUN**;

\*Run third model with slected attributes;

TITLE "Regression Result: 3rd model with outliers/influential points";

**PROC** **REG** data=finalSalary;

MODEL sqrtSalary\_in\_usd = work\_year salary numEmployee\_residence experience\_New employment\_FT salary\_currencyUSD / STB VIF TOL INFLUENCE R;

**RUN**;

\*Residual plot: Studentized residual vs.Predictors;

TITLE "Residual Plot: Studentized Residual vs. numEmployee\_residence (US)";

PLOT student.\* (work\_year salary numEmployee\_residence experience\_New employment\_FT salary\_currencyUSD);

**RUN**;

\*Residual plot: Studentized residual vs. predicted value;

TITLE "Residual Plot: Studentized Residual vs. Predicted Value";

PLOT student.\* predicted.;

**RUN**;

\*Normal probability plot or QQ plot;

TITLE "Normal Probability plot";

PLOT npp.\* student.;

**RUN**;

TITLE "Remove Influential Points and Outliers";

**DATA** finalSalary2;

SET finalSalary;

IF \_n\_ in (**5**, **40**, **44**, **72**, **93**, **112**, **169**, **170**, **185**, **187**, **211**, **288**, **370**, **565**, **577**, **1088**, **1096**, **1186**, **1310**, **1344**, **1373**, **1556**, **1670**, **1672**, **1682**, **1698**, **1924**, **1933**, **1937**, **1952**, **1955**, **1957**, **1963**, **1965**, **1975**, **1991**, **1999**, **2000**, **2013**, **2016**, **2029**, **2033**, **2045**, **2055**, **2096**, **2097**, **2107**, **2108**, **2118**, **2125**, **2126**, **2129**, **2133**, **2136**, **2148**, **2152**, **2155**, **2164**, **2165**, **2168**, **2169**, **2170**, **2192**, **2208**, **2240**, **2257**, **2262**, **2324**, **2339**, **2345**, **2353**, **2359**, **2498**, **2541**, **2550**, **2551**, **2552**, **2600**, **2652**, **2657**, **2700**, **2711**, **2760**, **2798**, **2904**, **3278**, **3310**, **3311**, **3319**, **3475**, **3505**, **3545**, **3560**) then delete;

**RUN**;

TITLE "Regression Result: 4th model";

**PROC** **REG** data=finalSalary2;

MODEL sqrtSalary\_in\_usd = work\_year salary numEmployee\_residence experience\_New employment\_FT salary\_currencyUSD / STB TOL VIF;

**RUN**;

\*Residual plot: Studentized residual vs.Predictors;

TITLE "Residual Plot: Studentized Residual vs. numEmployee\_residence (US)";

PLOT student.\* (work\_year salary numEmployee\_residence experience\_New employment\_FT salary\_currencyUSD);

**RUN**;

\*Residual plot: Studentized residual vs. predicted value;

TITLE "Residual Plot: Studentized Residual vs. Predicted Value";

PLOT student.\* predicted.;

**RUN**;

\*Normal probability plot or QQ plot;

TITLE "Normal Probability plot";

PLOT npp.\* student.;

**RUN**;

\*Run final model;

TITLE "Regression Result: Final model";

**PROC** **REG** data=finalSalary2;

MODEL sqrtSalary\_in\_usd = work\_year salary numEmployee\_residence experience\_New employment\_FT salary\_currencyUSD / STB TOL VIF;

**RUN**;

\*Create prediction dataset;

TITLE "Prediction data";

**DATA** pred;

INPUT work\_year salary numEmployee\_residence experience\_New employment\_FT salary\_currencyUSD;

DATALINES;

2024 100000 1 1 1 1

2024 100000 0 1 1 0

;

**PROC** **PRINT**;

**RUN**;

\*Join prediction dataset with current dataset;

**DATA** predSalary;

SET pred finalSalary2;

**RUN**;

**PROC** **PRINT** DATA=predSalary (OBS=**5**);

**RUN**;

TITLE "Regression Analysis and Confidence Interval for Average Estimate.";

**PROC** **REG** data = predSalary;

MODEL sqrtSalary\_in\_usd = work\_year salary numEmployee\_residence experience\_New employment\_FT salary\_currencyUSD / p clm cli;

**RUN**;

\*Split the data into training and test sets - 80/20;

TITLE "Test and Train Set for Salary Data";

**PROC** **SURVEYSELECT** data = finalSalary2 out = xvalSalary seed = **1997** samprate = **0.8** outall;

**PROC** **PRINT** data = xvalSalary (OBS=**10**);

**RUN**;

\*Generate new predicted value used for cross validation;

TITLE "Create predicted value for cross validation";

**DATA** xvalSalary;

SET xvalSalary;

IF (selected = **1**) then new\_y = sqrtSalary\_in\_usd;

**RUN**;

**PROC** **PRINT** data = xvalSalary (OBS=**10**);

**RUN**;

\*Run full model Stepwise selection using train set;

TITLE "Full model regression with stepwise selection method: Train set";

**PROC** **REG** data=xvalSalary;

MODEL new\_y = work\_year salary numEmployee\_residence numCompany\_location experience\_New employment\_FT size\_M remote\_ratioNone salary\_currencyUSD / selection = stepwise;

**RUN**;

\*Run model with selected attribute using test data;

TITLE "Validation - Test Set";

**PROC** **REG** DATA=xvalSalary;

MODEL new\_y = work\_year salary numEmployee\_residence experience\_New employment\_FT salary\_currencyUSD / STB VIF;

OUTPUT out=out(where=(new\_y=**.**)) p=yhat;

**RUN**;

**PROC** **PRINT** data = out (OBS=**10**);

**RUN**;

\*Compute the difference between observed and predicted values.;

TITLE "Difference between Observed and Predicted - Test Set";

**DATA** out\_sum;

SET out;

diff = sqrtSalary\_in\_usd - yhat;

abs\_diff = abs(diff);

**RUN**;

\*Compute descriptive statistics: RMSE, MAE;

TITLE "Descriptive Statistics: RMSE and MAE - Test Set";

**PROC** **SUMMARY** data=out\_sum;

VAR diff abs\_diff;

OUTPUT out=out\_stats std(diff)=rmse mean(abs\_diff)=mae;

**RUN**;

**PROC** **PRINT** data=out\_stats;

TITLE "Validation statistics for Model";

**RUN**;

\*Computes correlation of observed and predicted values in test set;

TITLE "Correlation between Observed and Predicted values";

**PROC** **CORR** data=out;

var sqrtSalary\_in\_usd yhat;

**RUN**;