STATISTICAL MODELING

HW 1

UMUT BAYRAK
GHENT UNIVERSITY
27.11.2017

1) Introduction

This report analyses if the test score, average of reading and math scores on the Stanford achievement test, utilized by school districts in the USA is affected by the class size (stratio) defined as total number of students divided by the total number of teachers in the same district and the percentage of non-native English speakers among the students in each district (nonep). Government is interested in improving students' success by hiring more teachers but some sceptics think this will increase the costs without producing benefits.

2) Methods

The dataset contains a random sample of California elementary school districts. Test scores, class sizes and the percentage of non-native English speakers among the students in each district (nonep) are continuous variables. To have a better understanding of the data simple descriptive analysis is conducted. Then simple linear regression model and multiple regression models are performed on the data. For the regressions, the response variable (Y) is the test score, the predictors are (X_1, X_2) are class size and the percentage of non-native English speakers among the students. For simple linear regression model only Y and X_1 is used; multiple regression models consist of predictors with and without interaction between them.

3) Results

3.1. Descriptive Statistics:

Variables, test score, class size, and the percentage of non-native English speakers among the students, are summarized via histogram plots and basic statistical measures. All of the variables have 420 observations so there is no missing data in the data-set.

Test score has a mean of 654.1565, median of 654.4500, mode of 616.3000, and standard deviation of 19.0533. Class size has a mean of 19.6404, median of 19.7232, mode of 20.0000, and standard deviation of 1.8918. The percentage of non-native English speakers has a mean of 15.7681, median of 8.7776, mode of 0.0000, and standard deviation of 18.2859. Test score have almost equal mean, median, and mode as well as the class size. Both of the variables have bell shaped curve and their Q-Q plots fit to a line, thus it can be concluded that test score and class size are normally distributed variables. However, distribution of the percentage of non-native English speakers is skewed to the right with skewness 1.4319.

There seems to be a negative correlation between the test score and the percentage of non-native English speakers in Figure 2. However, the association between the test score and class size cannot be easily interpreted from the plot, correlation seems to be weak in the negative direction. Non-parametric regression model, LOESS, is used to discover the relation between those variables. As can be seen in the graph, correlation is negative; red and blue lines represent the best fitting line and best fitting curve respectively.

3.2. Simple Linear Regression

The simple linear regression of test scores on class sizes gives parameter estimates, standard deviations and 95% confidence interval of the coefficients as follows in Fig. 1.

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	698.93295	9.46749	73.82	<.0001	680.32313	717.54278
stratio	stratio	1	-2.27981	0.47983	-4.75	<.0001	-3.22298	-1.33664

Figure 1. Parameter Estimates

If class size becomes zero at any time, test score value becomes 698.93295 (Y-intercept); however, minimum value of class zero is 14.5. Thus intercept does not give any meaningful interpretation for this data-set. On the other side, when class size increases one unit, the test score decreases by -2.27981 on average and vice versa. This relation also states that there is a negative correlation between them. 95% confidence interval for the slope – Class Size – is (-3.22298, -1.33664) and 95 % chance that the estimate of the parameter would be in this interval. Two-sided t-test gives that t-value is -4.75, and the p-value is close to zero which is lower than desired significance value 0.05, null hypothesis that claims to slope is zero can be rejected under these conditions.

Simple Linear Regression Model assumes that the relationship is linear, normally distributed residuals and homoscedasticity in the variance. Distribution of Residuals vs Predictor and Residuals vs Fitted Values plots reflect that points are randomly distributed around 0 and the linear relationship between test score and class size assumption holds. Second assumption is variance of residuals should be constant in each level of predictor. Most of the variance seemed to be constant however I can say that this assumption holds but not as strong as linearity assumption. Normality assumption holds since Q-Q suggests that residuals are fitted to a line. Since the distribution and homoscedasticity assumption is not that strong, simple linear regression model may not be the most explanatory model for this data-set.

3.3. Multiple Linear Regression

When multiple linear regression model without interaction between predictors is applied to the data it can be seen below that coefficient of regression are negative. A unit increase in class size and the percentage of non-native English speakers (nonep) makes the test score decreases by -1.10130 and -0.64978 respectively on average. 95% confidence intervals of predictors are given as follows (-1.84880, -0.35379) for class size; (-0.72711, -0.5244) for nonep. Both p-values are below significance level 0.05, so the null hypothesis claiming that coefficients are zero is rejected.

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	686.03225	7.41131	92.57	<.0001	671.46406	700.60044
stratio	stratio	1	-1.10130	0.38028	-2.90	0.0040	-1.84880	-0.35379
nonep	nonep	1	-0.64978	0.03934	-16.52	<.0001	-0.72711	-0.57244

Figure 2. Parameter Estimates of Regression Coefficients

Distribution of residuals vs predictor and residuals vs fitted Values show that variables are randomly distributed. Q-Q plot of residuals makes the normality assumption hold and variance of the residuals are nearly constant which makes in total all assumptions hold for multiple linear regression model.

In the simple linear regression model, the coefficient of class size is -2.27981, however in the multiple case, the same coefficient decreases to -1.10130. The effect of having another predictor for the data (confounding), in this case this is the percentage of non-native English speakers, lowers the weight of the main predictor to almost in its half which shows that, simple linear regression model for this dataset is not sufficiently good to predict the change in test score.

Adding the interaction term into the model gives the regression coefficients as follows; Intercept is 686.33852 with 95% confidence interval (667,85599, 704.82106), Class Size is -1.11702 with 95% confidence interval (-2.06553, -0.16850), the percentage of non-native English speakers is -0.67291 with 95% confidence interval (-1.53385, 0.18803) and lastly the interaction term is 0.00116 with 95% confidence interval (-0.041190, -0.04422). The parameter estimate results show that the interaction between class size and the percentage of non-native English speakers does not really affect the test score because the regression coefficient is quite low comparing to class size's and nonep's individually. Also, the effect of the class size increased by 0.01. The residual distribution of the class size and nonep seem to be not changed. Which makes the random distribution of the residuals assumption hold.

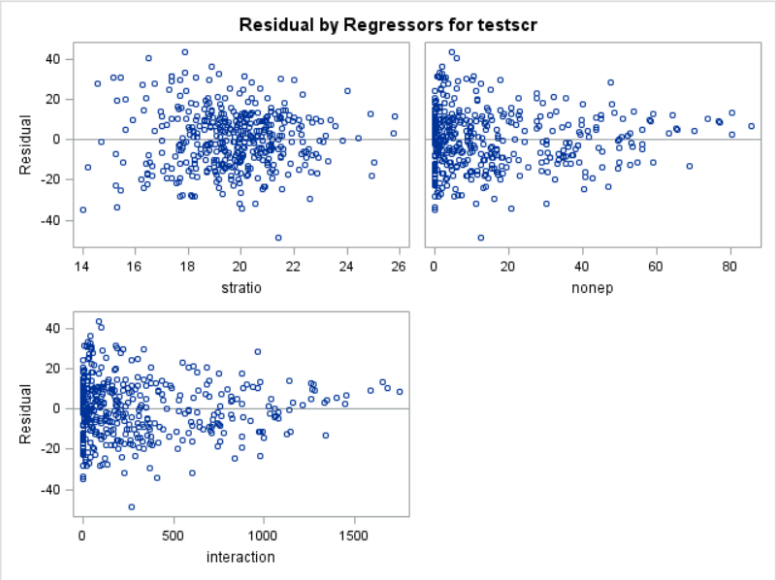

Figure 3. Distribution of residuals

Prediction interval of a school district which has a student-teacher ratio of 20 and 50% of non-native English speakers is given by the formula, Sample estimate ± (t-multiplier × standard error):

$$\hat{y}_h \pm t_{(\alpha/2,n-p)} \times \sqrt{MSE + [\mathrm{se}(\hat{y}_h)]^2}$$

If the test is repeated many times the test score is predicted to be in the interval for test score is (602.894, 660.134) with a fit of 631.514 and 95% chance.

4) Discussion and conclusions

From the given random sample data of California elementary school districts, test score is negatively correlated with class size and the percentage of non-native English speakers among students which can be interpreted as any increase in the class size will cause a negative impact on the test score on average. For example, assume that there are 100 students and 4 teachers in a district and class size is 25; when a teacher added to that district class size decreases to 20 and with its coefficient being --1.11702 the test score is increased 5.60 points on average and if the percentage of non-native English speakers among students is increased the test score tends to be decreasing by -0.67291 but the effect on the average test score is less than the class size.

Thus, in the light of the outcomes of the analysis, there seems to be a negative association between test score and class size however it does not mean that the increase in the test score is caused by decrease in the class size, there might be other factors affecting the increase the test score. In order to measure the effect of the class size, the data should be divided into two groups as test and control groups. After that class size, the percentage of the non-native English speakers among students and the interaction terms should be measured individually in a time interval to assess if there is a causality occurring between them or not.

A. Appendix

A.1. Descriptive Statistics

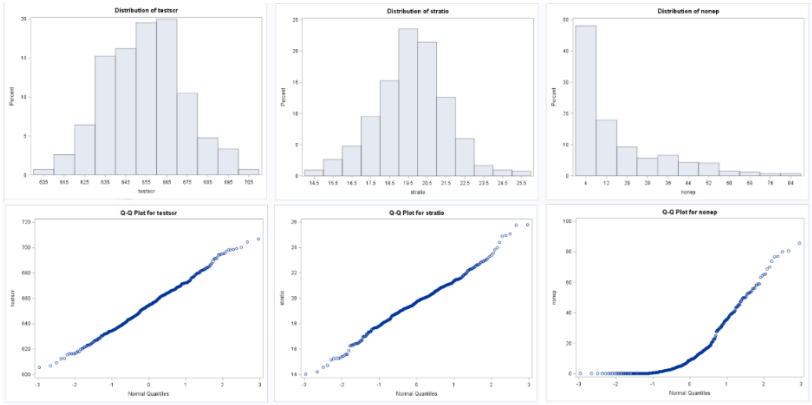

Figure 1. Distributions of Variables (Test score, Class size, and the percentage of non-native English speakers)

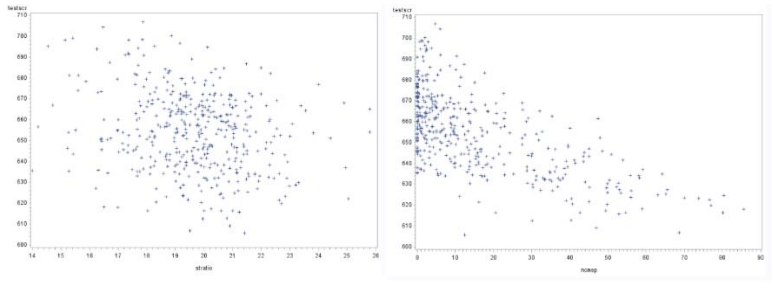

Figure 2. Scatter Plots of Test Score vs. Class Size and Test Score vs. the percentage of non-native English speakers

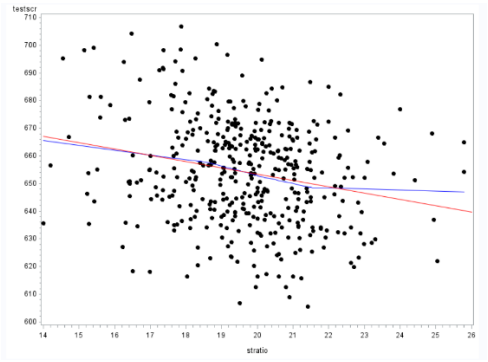

Figure 3. Scatter Plot of Test Score vs Class Size

A.2. Simple Linear Regression Model

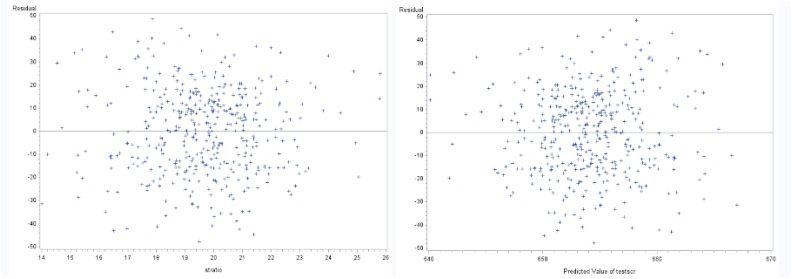

Figure 5. Distribution of Residuals vs Predictor and Residuals vs Fitted Values

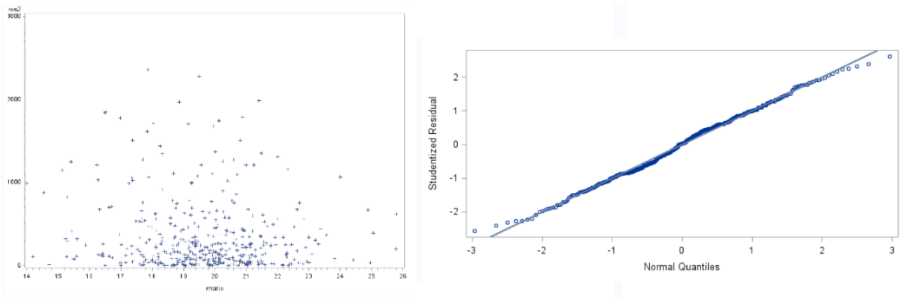

Figure 6. Variance of residuals and Q-Q plot of residuals

A.3. Multiple Linear Regression

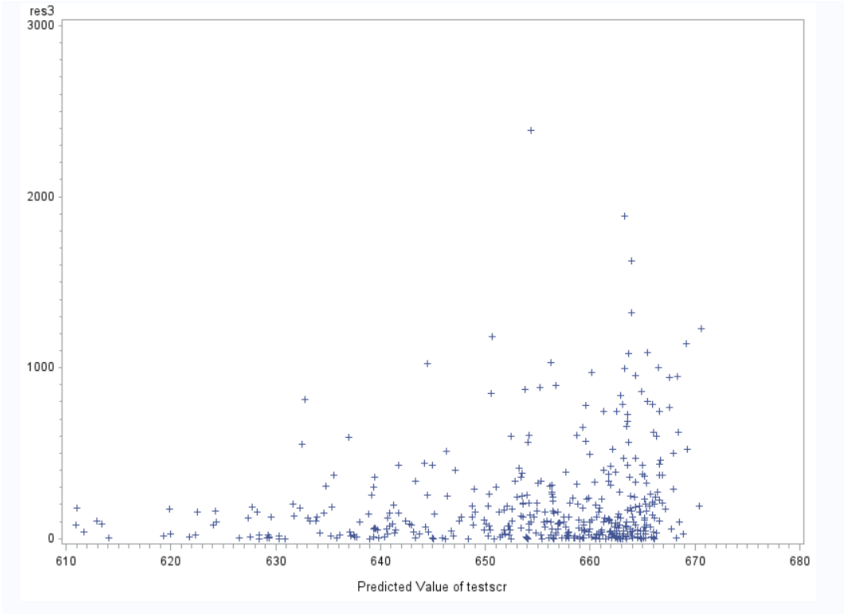

Figure 7: Distribution of Variance of Residuals

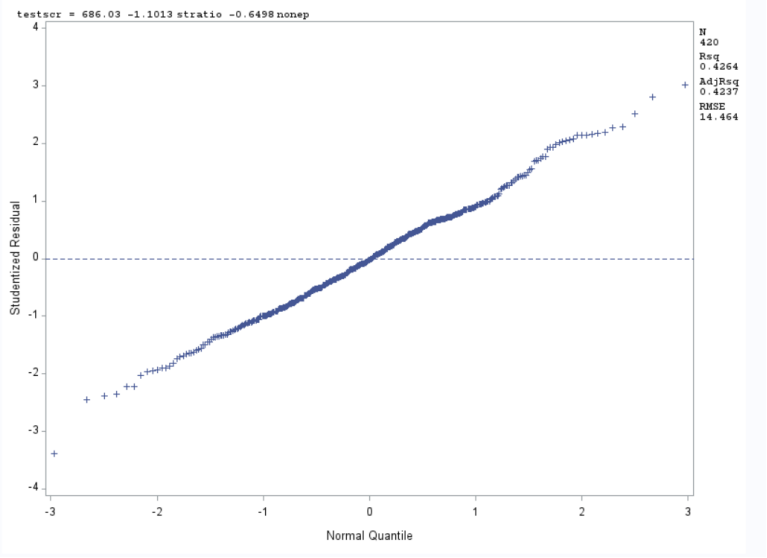

Figure 8: Q-Q Plot of Studentized residuals

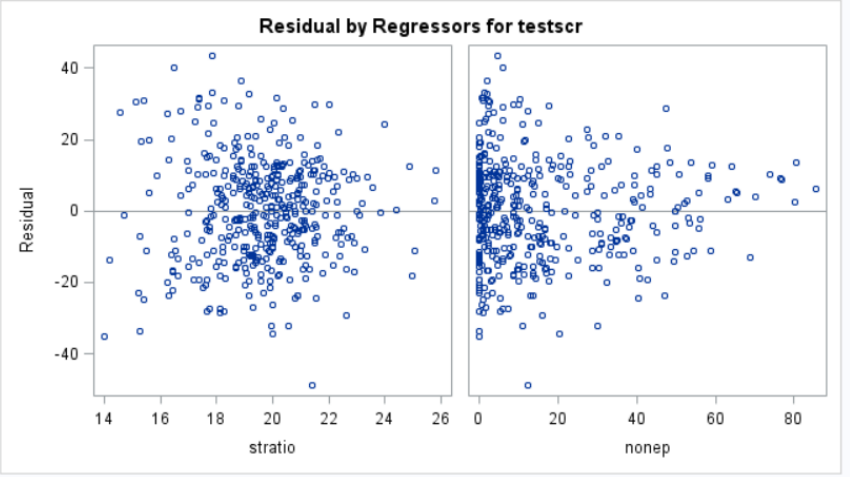

Figure 9: Distribution of Residuals

A. 4. SAS Code

```
PROC IMPORT DATAFILE = 'H:\StatisticalModelling\caschoolvar3.xls' OUT =
SchoolRaw DBMS = XLS;
RUN;

/* Descriptive Statistics for the Q.1's first part, for each variable*/

PROC UNIVARIATE PLOT DATA=Schoolraw ;
/*The plot option produces a stem-and-leaf plot and a boxplot*/

VAR testscr stratio nonep;
/*In the var statement, we can provide the names of the variables to be
investigated*/

HISTOGRAM testscr stratio nonep;
/*The histogram statement produces a histogram*/

QQPLOT testscr stratio nonep;
/*The qqplot statment produces a qqplot*/

RUN;


/* Bivariate Analysis for the Q.1's second part */

PROC GPLOT DATA = schoolraw;
PLOT testscr*stratio;
RUN;

PROC GPLOT DATA = schoolraw;
PLOT testscr*nonep;
RUN;

/* Exploring the linear relationship (1) */

PROC LOESS DATA = schoolraw;
MODEL testscr = stratio / smooth=0.75;
ODS OUTPUT OutputStatistics=Results;
RUN;
```

```
PROC SORT DATA=Results;
BY stratio;
RUN;


proc gplot data=Results;
plot  DepVar*stratio Pred*stratio / overlay;
symbol1 c=red v=dot h=0.2 mode=include;
symbol2 c=black i=join w=2 mode=include;
run;

goptions reset=all;
symbol1 v=dot i=none c=black;
symbol2 v=none i=join c=blue;
symbol3 v=none i=r c=red;
proc gplot data=results;
 plot DepVar*stratio=1 pred*stratio=2 DepVar*stratio=3 / overlay;
run;
quit;
/* This will give parameter estimates, standard Deviations and 95% CI
clb will produce confidence intervals for the beta's
clm gives confidence intervals for the mean outcome (at each
observed x) */
PROC REG DATA=schoolraw;
   MODEL testscr = stratio / CLB CLM;
RUN;


PROC REG DATA=schoolraw;
   MODEL testscr = stratio;
   output out=resid p=pman r=rman student=student;
RUN;

/* Residuals versus predictor: */
PROC GPLOT DATA=resid;
PLOT rman*stratio /vref=0;
RUN;

/* Residuals versus fitted values: */
PROC GPLOT DATA=resid;
PLOT rman*pman /vref=0;
RUN;

/* Homosdestacity */

DATA resid2;
SET resid;
res2 = rman*rman;
RUN;

goptions reset=all;
PROC GPLOT DATA=resid2;
PLOT res2*stratio;
RUN;
QUIT;


PROC UNIVARIATE PLOT DATA=resid;
VAR student;
RUN;
```

```
proc reg data=schoolraw;
model testscr = stratio / r;
run;

/* Multiple regression starts here */

PROC REG DATA = schoolraw;
MODEL testscr = stratio nonep / CLB CLM;
RUN;

/*checking assumptions*/
PROC REG DATA=schoolraw;
   MODEL testscr = stratio nonep;
   output out=resid2 p=pman r=rman student=student;
RUN;

DATA resid3;
SET resid2;
res3 = rman*rman;
RUN;

goptions reset=all;
PROC GPLOT DATAA = resid3;
  PLOT res3*pman;
RUN;
QUIT;

PROC REG DATA = schoolraw;
   MODEL testscr = stratio nonep;
   PLOT student.*nqq. ;
  RUN;

PROC REG DATA = schoolraw;
MODEL testscr = stratio nonep / lackfit;
RUN;

/* Creating the interaction term in another Dataset */

DATA schoolrawInt;
SET schoolraw;
interaction = stratio*nonep;
RUN;

PROC REG DATA = schoolrawInt;
MODEL testscr = stratio nonep interaction / CLB CLM;
OUTPUT OUT = ResidInt p=pman r=rman student=student;
RUN;

PROC LOESS DATA = schoolrawInt;
MODEL testscr = stratio nonep interaction/ smooth=0.75;
ODS OUTPUT OutputStatistics=ResultsInt;
RUN;

PROC SORT DATA=ResultsInt;
BY stratio;
RUN;

proc reg data=schoolrawInt;
model testscr = stratio nonep interaction ;
plot testscr*stratio / pred conf;
```

```
run;
```