

Analysis of Continuous Data HW-4

Omkar Kulkarni

The dataset caschool.xlsx contains a random sample of California elementary school districts. The data consists of test scores (Y: testscr), class sizes (X1: stratio) and the percentage of non-native English speakers among the students in each district (X2:nonep).

Y : (testscr) districtwide average score of reading and math scores on the Stanford achievement test

X1: (stratio) the total number of students in the district divided by the number of teachers.

X2: (nonep) the percentage of non native English speakers among the students in each district

The test score is a districtwide average of reading and math scores on the Stanford achievement test, a test utilized by school districts in the USA. The student-teacher ratio, i.e. the total number of students in the district divided by the number of teachers, is used as a measure of the (overall) class size in the district. Policy makers are interested in the question of whether reducing class size, for instance by hiring more teachers, improves the student's education. Skeptics worry that reducing class size will increase costs without producing substantial benefits.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i \text{ --- 1}$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i \text{ --- 2}$$

$$\hat{Y} = b_0 + b_1 X_{i1} + b_2 X_{i2} + b_3 X_{i1} X_{i2} \text{ --- 3}$$

$$\varepsilon_i \sim N(0, \sigma^2)$$



Figure 1,2,3 show the univariate properties of testscr, strratio and nonep respectively. Nonep, X_{i2} is skewed to the right, which is evident from the histogram and QQ plots. Moreover, for bivariate analysis from figure 4 we see the scatter plot. Correlation between testscr, strratio is negative so is between testscr and nonep, whereas correlation between strratio and nonep is positive, which prompts us to check for confounding. The Pearson Correlation Coefficients are evident in table 1.

Now fitting model as per below equation (without interaction)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

we have the estimated values as $\hat{Y} = b_0 + b_1 X_{i1} + b_2 X_{i2}$, and from table 2, we get the values as $b_0 = 686.03$, $b_1 = -1.1$, $b_2 = -0.64$ with confidence intervals at 95% (671.46, 700.60), (-1.85, -0.35), (-0.73, -0.57) respectively, with significant p-values ($p > 0.05$). Both X_{i1} and X_{i2} together are not 0 in our setup hence, we do not have a meaningful interpretation for b_0 . To understand b_1 , it is observed test score (Y) on average decreases by 1.1 units per unit increment of (X_{i1})strratio when nonep is held constant. And for b_2 , it is observed test score (Y) on average decreases by 0.64 units per unit increase of (X_{i2}) nonep when strratio is held constant.

Figure 5 shows the residuals plotted against the two predictors, residual vs stratio looks well scattered but in residuals vs nonep, variances increases as the nonep increases, it is also evident from residual² vs testscr, as testscr increases the variance increases. As mentioned earlier nonep (figure 3) is skewed to the right and we can log transform (we do see few 0's in nonep, so before we log transform we can add 1 to all values, though it induces bias it shall be minor) it to bring it closer to symmetry before regressing.

$$\hat{Y} = a_0 + a_1 X_{i1} + a_2 X'_{i2}$$

Where $X'_{i2} = \log X_{i2}$

Figure 7 shows the residuals plotted against log transformed nonep showing a better random scatter. Figures 8,9,10 show us the linearity, constancy of error variance, no outliers, normality of residuals and hence the model assumptions.

Now, to interpret the parameters $a_0 = 687.56, a_1 = -0.82, a_2 = -8.16$ with confidence interval at 95% (672, 703.1), (-1.63, -0.01), (-9.33, -6.99) respectively. It is observed test score (Y) decreases on average by 0.82 units per unit increment of (X_{i1})stratio when nonep is held constant. And it is observed test score (Y) on average decreases by 8.16 units per unit increase of (X'_{i2}) log (nonep), (10 times increase of nonep) when stratio is held constant.

Now for model with only stratio :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$$

We have $\hat{Y} = b'_0 + b'_1 X_{i1}$

And from table 4, we see $b'_1 = -2.27$, its interpretation being for per unit increase of X_{i1} (stratio) we observe a decrement of 2.27 in average value of \hat{Y} (testscr). But if we look at model with nonep also, we observe $b'_1 = -2.27$ which is more than 10% different from $b_1 = -1.1$ (for model with stratio and nonep), hence there is significant confounding and we have to follow model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$ with the nonep predictor.

Now for model with the interaction term :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X'_{i2} + \beta_3 X_{i1} X'_{i2} + \varepsilon_i$$

After centering, i.e $\widetilde{X}_1 = X_{i1} - \overline{X_{i1}}$ and $\widetilde{X}_2 = X'_{i2} - \overline{X'_{i2}}$ we get

$$\hat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \widetilde{X}_1 + \widehat{\beta}_2 \widetilde{X}_2 + \widehat{\beta}_3 \widetilde{X}_1 \widetilde{X}_2$$

From table 5, it is clear that estimated values for the model $\hat{Y} = b_0 + b_1 X_{i1} + b_2 X'_{i2} + b_3 X_{i1} X'_{i2}$ are $b_1 = -0.90, b_2 = -8.41, b_3 = -1.17$ with 95% confidence intervals (-1.70, -0.10), (-9.57, -7.25), (-1.81, -0.54) with P-values for all being significant, the change in parameters is evident in comparison to table 3. As confidence interval of b_3 does not have 0 in it, it is significant and hence we retain X_{i1} and X_{i2} in model irrespective of the significance of b_1 and b_2 .

To interpret the slopes with the interaction, in $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1 + \hat{\beta}_2 \bar{X}_2 + \hat{\beta}_3 \bar{X}_1 \bar{X}_2$

When X'_{i2} tend to \bar{X}'_{i2} , \bar{X}_2 tends to 0 ; thus $\hat{Y} = b_0 + b_1 \bar{X}_1$ it is observed on an average (\hat{Y}) testscr decreases by 0.9 unit for 1 unit increment of (X_{i1}) strratio when nonep is at its mean value.

When X_{i1} tend to \bar{X}_{i1} , \bar{X}_1 tends to 0 ; thus $\hat{Y} = b_0 + b_2 \bar{X}_2$ it is observed on an average (\hat{Y}) testscr decreases by 8.41 unit for 1 unit increment of (X'_{i2}) log of nonep when stratio is at its mean value.

Graphically representation : To show it graphically we can categorize \bar{X}_2 (nonep) and for different categories plot how \hat{Y} varies as \bar{X}_1 varies.

Table 6, shows the AIC – Akaike Information Criterion, values for different models and model with both the predictors has the lowest value.

And from the below table we see a significant P value for nonep_trans, and hence model with both the predictors has a higher F Value.

Factor	DF	Sum of Squares	Mean Square	F Value	Pr > F	Label
stratio	3	2277.133873	759.044624	3.73	0.0115	stratio
nonep_trans	3	59858	19953	97.99	<.0001	

To predict the score for student-teacher ratio of 20 and 50% of non-native English speakers we have

$$\hat{Y} = b_0 + b_1 X_{i1} + b_2 X_{i2} + b_3 X_{i1} X_{i2}$$


We get Predicted value = 669.68 with prediction interval at 95% as PI (639.64 , 699.72)



We have observed the values of b_1 to be negative consistently in all the models so far and which is significant. Hence there is an 'association' between test scores and stratio , if we reduce student - teacher ratio we expect an increment in test scores.

However, for overall analysis to make the decision the 'cost' of hiring more teachers should also be considered in the model to analyse the 'substantial benefits'.

Bonus question : cross validation for selection of models :

We split the data into different  folds (groups) randomly. Then keeping say i^{th} fold outside the model, and predict the values for the i^{th} fold and calculate the error rate for that particular fold. We repeat this procedure to get error rate for each fold. **Next we combine the error rates by averaging** and get the cross validation error, this represents how the model works if the data we collected is an accurate representation of the population. We repeat this procedure for all the methods under test and then we can select the method with minimum cross validation error.



Theoretical Part: Some Small Questions

Are following statements right or wrong ? Explain.

1. Based on a regression analysis with one regressor, the null hypothesis $H_0 : \beta_1 = 0$ is rejected in favor of the alternative $H_1 : \beta_1 > 0$ with a very small p-value ($p < .0001$). The fitted model will give thus very good predictions.

If null hypothesis $\beta_1 = 0$ is rejected in favor of alternative $H_1 : \beta_1 > 0$, might not necessarily give good predictions. As the prediction interval also depends on the number of observations. For instance, with a significant β_1 in model but with just 3 or 4 observations might not give good predictions.

2. In a study the subject's age is known to be a confounder. The statistician includes age as a covariate in the regression model, but it turns out that the regression coefficient of age is not significant at the 5% level of significance (say, $p = 0.69$). He/she can therefore remove age from the model.

If it is known 'age' is a confounder, it has significant effect on the other predictor. Hence irrespective of the significance of regression coefficient of age we cannot remove it from the model.

In a 2008 paper by Cameriere et al., the authors wrote that 'This model had the lowest Akaike Information Criterion (AIC) value among the considered multiple regression models. Indeed Eq. (1) explained 93% of the total variance ($R^2 = 0.93$).' In which way are the AIC and the R^2 connected? Does a high AIC imply a high R^2 or vice versa?

Explain.

We know that

$$AIC = n \ln \left(\frac{SSE}{n} \right) + 2p$$

Where $2p$ is penalty for large number of parameters. 

$$\text{And } R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \quad (SSR + SSE = SSTO)$$

From the above two equation we can see that, for higher AIC we have a higher SSE but that causes a lower R^2 .

Appendix A – Figures and Tables

Figure 1: Y, testscr properties.

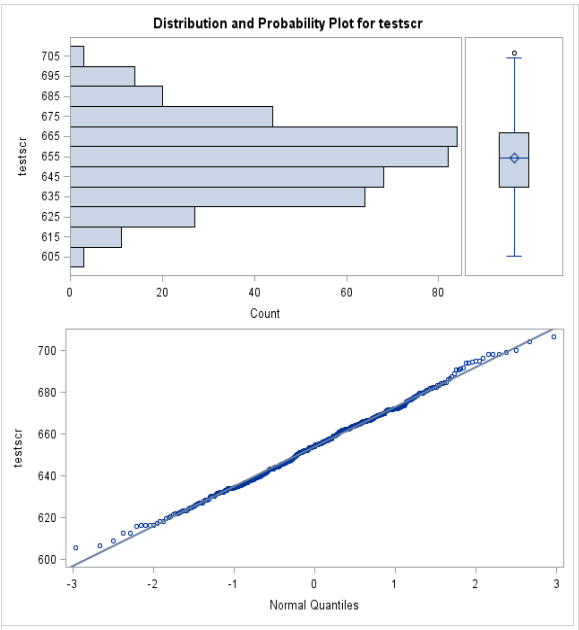


Figure 2: X1, stratio properties.

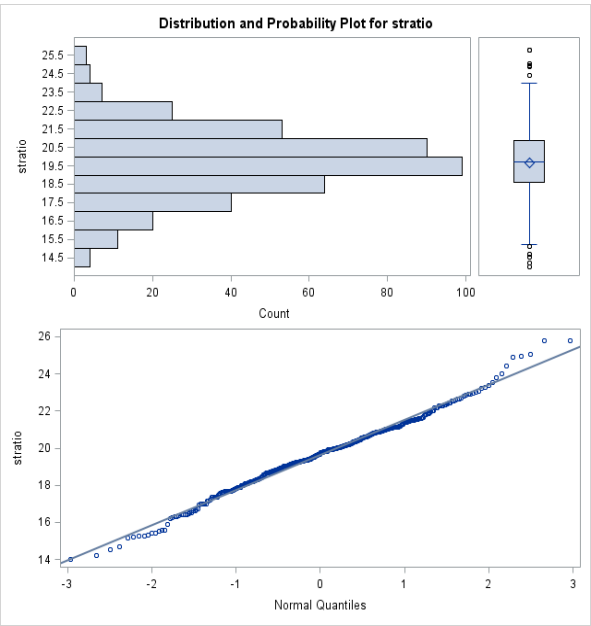
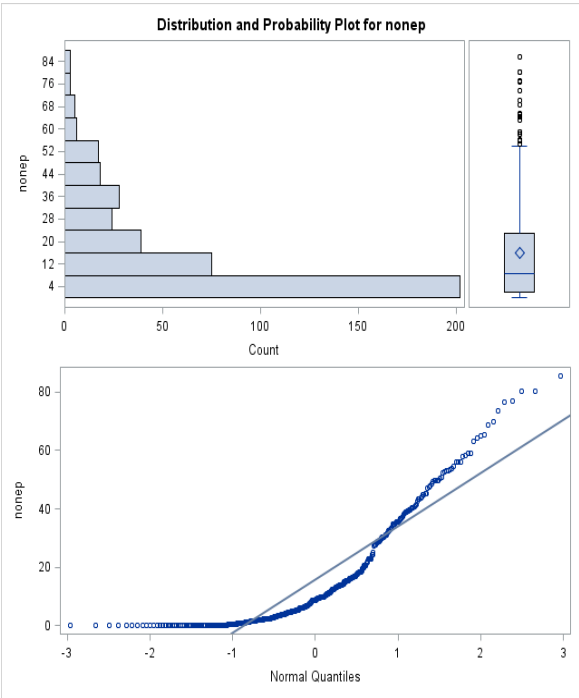


Figure 3 : X2, nonep properties Figure



4 : Scatter Plots - Bivariate

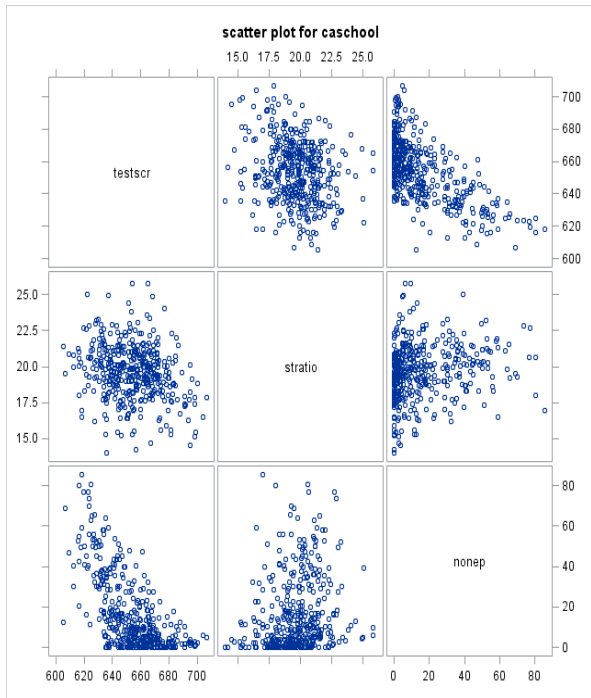


Figure 5 Residual Vs the two predictors

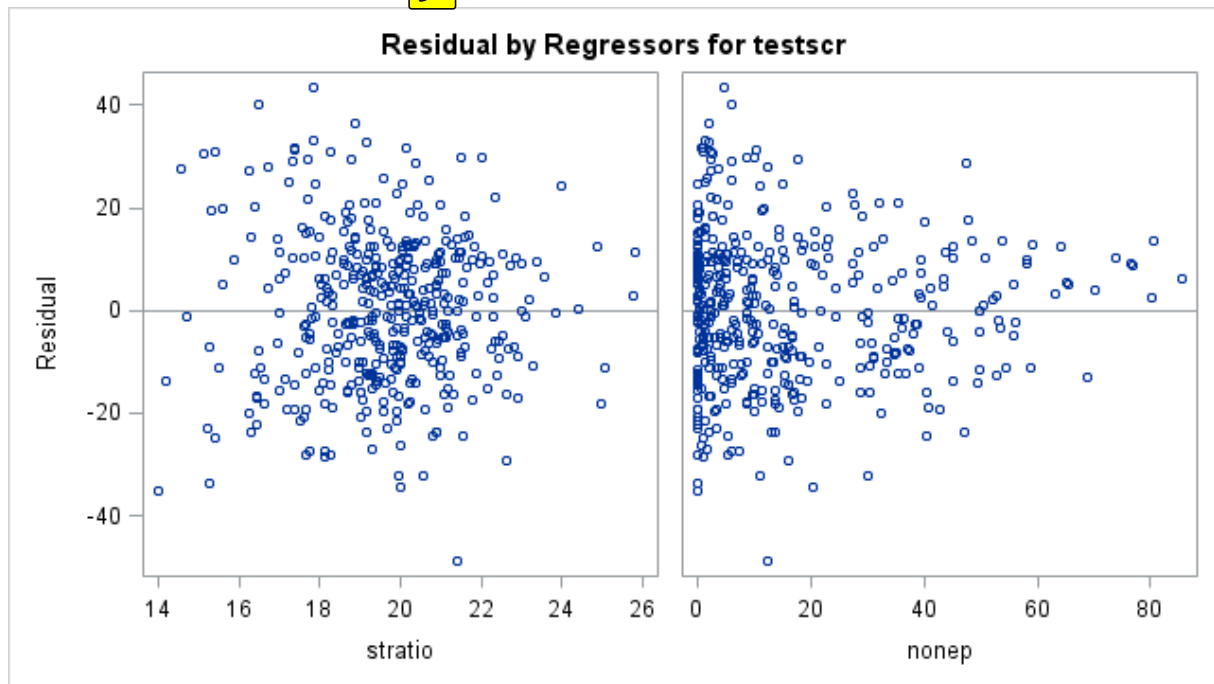
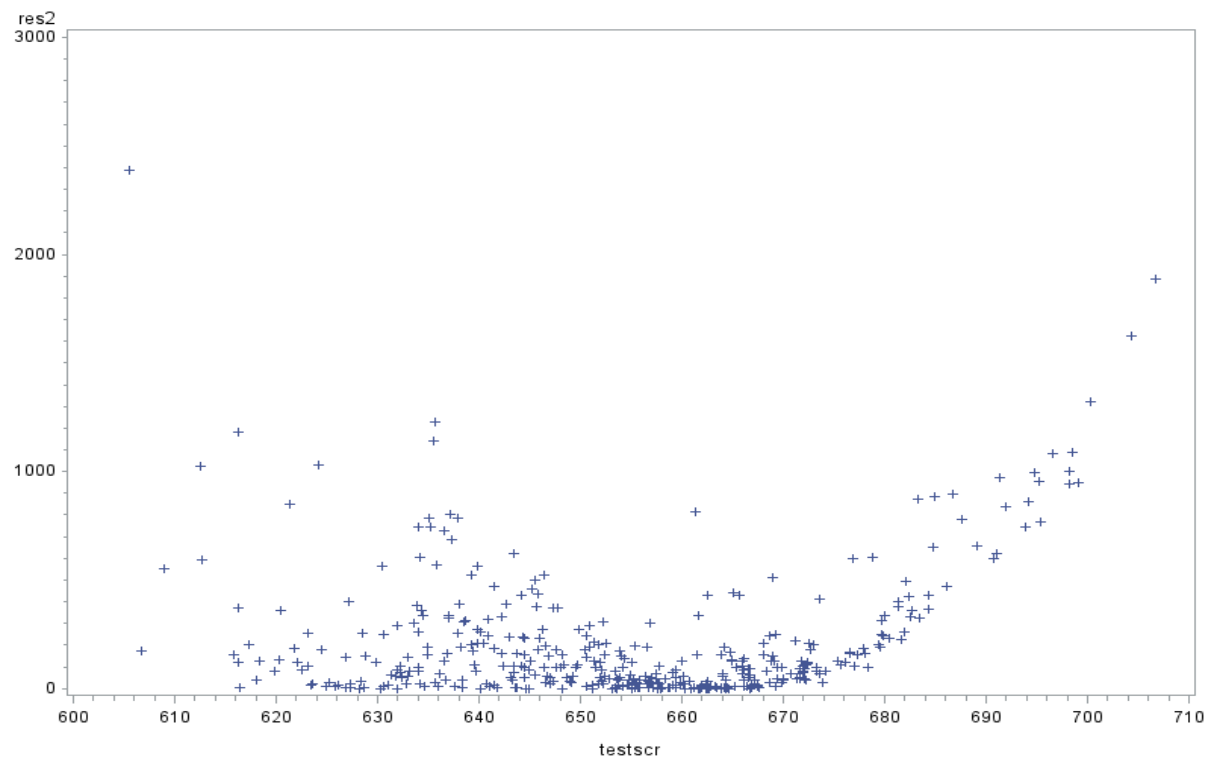


Figure 6 (Residual*Residual) vs (testscr)



Figures for log transformed nonep

Figure 7 Residual plot with log transformed "nonep" showing better random scatter of residuals

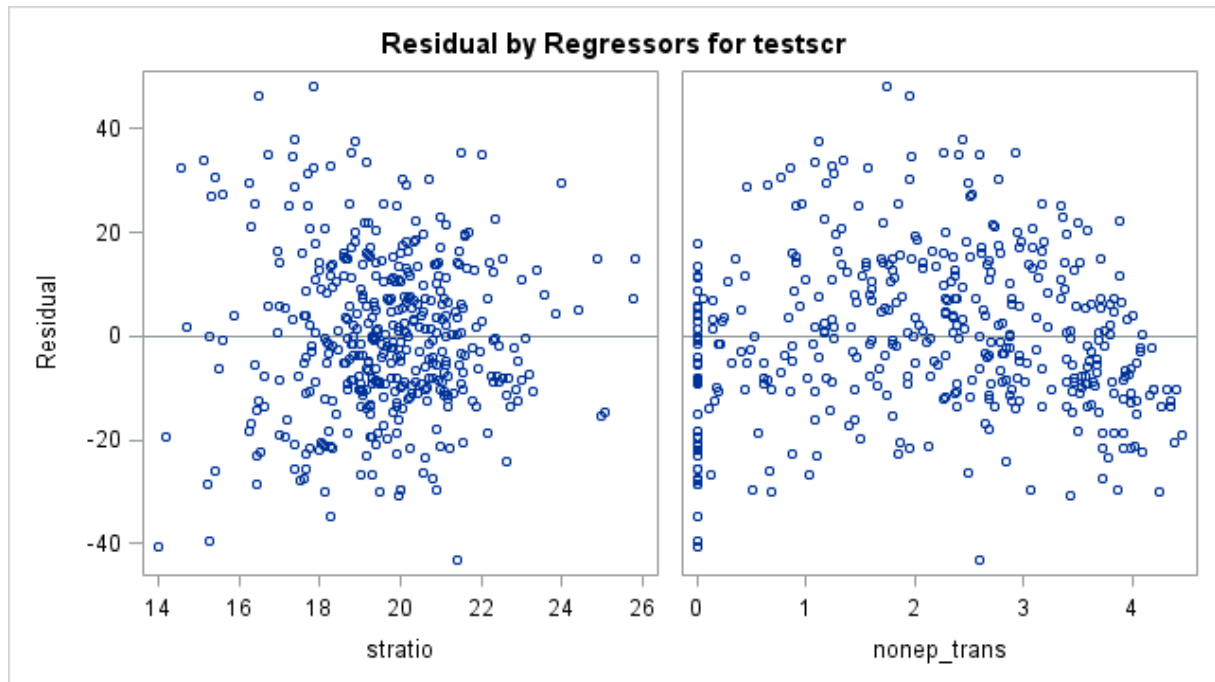


Figure 8 Residual vs Predicted value -- transformed nonep

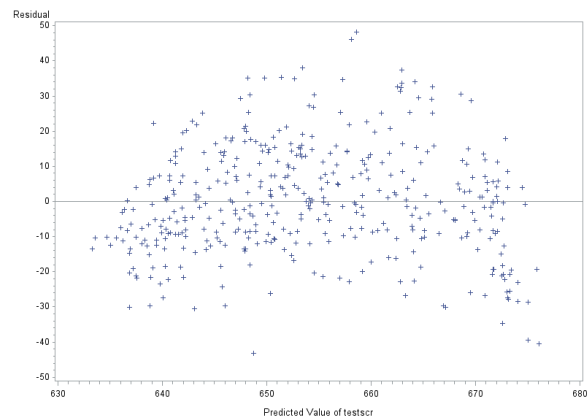


Figure 9 Residual² Vs Predicted Value

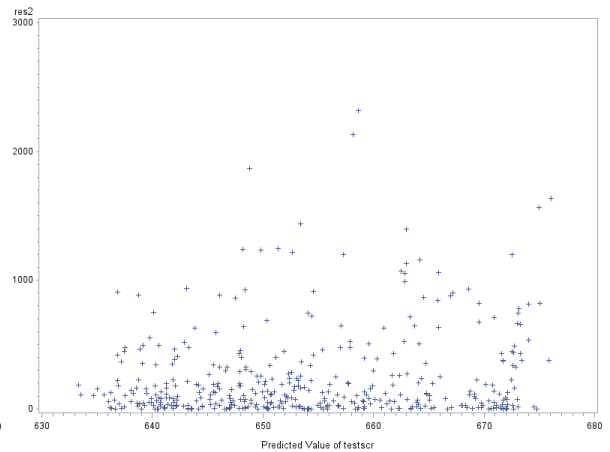
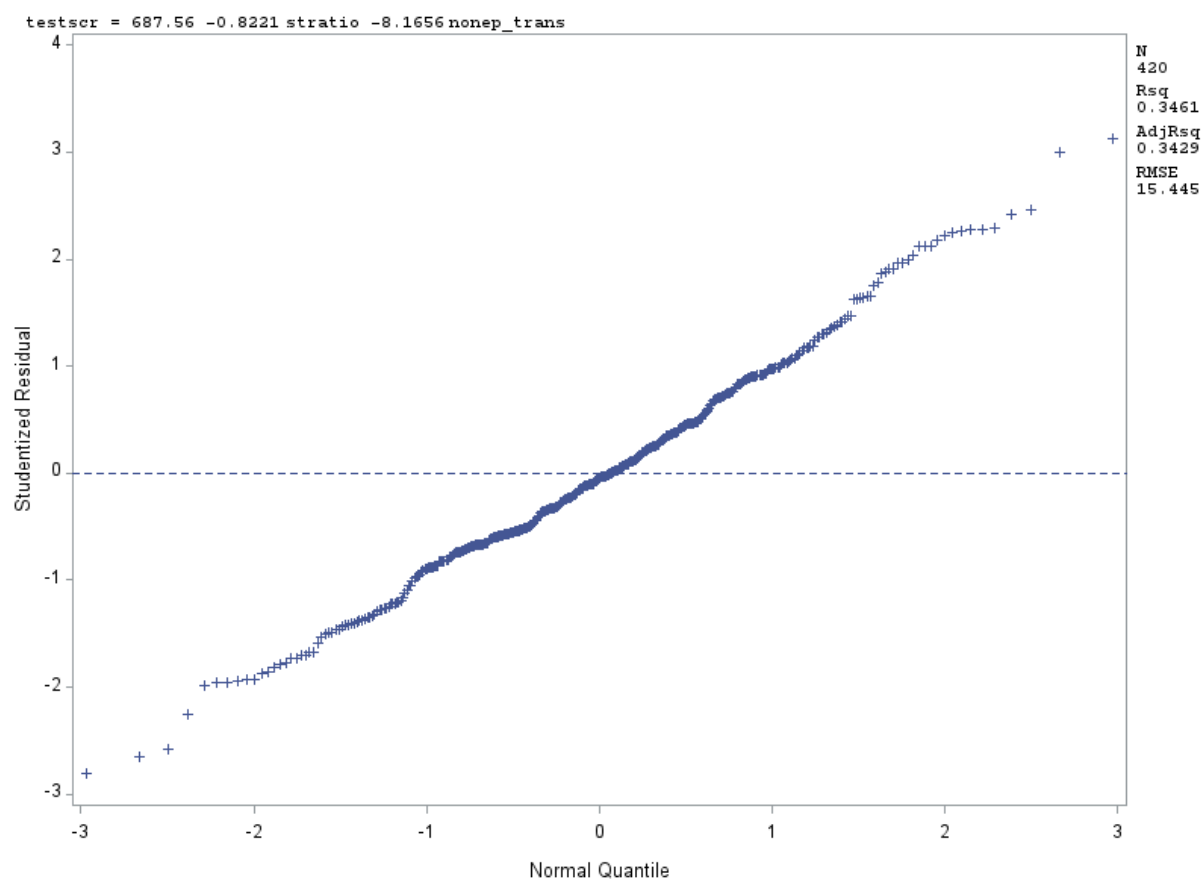


Figure 10 Plot for studentized residual -- transformed nonep



Tables

Table 1 Pearson Correlation Coefficients

Pearson Correlation Coefficients, N = 420 Prob > r under H0: Rho=0			
	testscr	stratio	nonep
testscr	1.00000	-0.22636	-0.64412
testscr		<.0001	<.0001
stratio	-0.22636	1.00000	0.18764
stratio	<.0001		0.0001
nonep	-0.64412	0.18764	1.00000
nonep	<.0001	0.0001	

Table 2 Parameter estimates for model without interaction

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

Table 3 Parameter Estimates for model without interaction and After TRANSFORMING 'nonep'

Parameter Estimates								
Variable	Label	D F	Parameter Estimate	Standard Error	t Value	Pr > t	95% Confidence Limits	
Intercept	Intercept	1	687.56001	7.91305	86.89	<.0001	672.00559	703.11444
stratio	stratio	1	-0.82210	0.41276	-1.99	0.0471	-1.63345	-0.01075
nonep_trans		1	-8.16561	0.59553	-13.71	<.0001	-9.33623	-6.99498

Table 4 Parameter estimates for simple linear regression model with only stratio as explanatory variable

Parameter Estimates								
Variable	Label	D F	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

Table 5 :Parameter Estimates for model WITH interaction and After TRANSFORMING 'nonep'; AFTER CENTERING

Parameter Estimates								
Variable	Label	D F	Parameter Estimate	Standard Error	t Value	Pr > t	95% Confidence Limits	
Intercept	Intercept	1	654.93713	0.77031	850.23	<.0001	653.42295	656.45130
stratio_centered		1	-0.90246	0.40730	-2.22	0.0273	-1.70307	-0.10184
nonep_trans_centered		1	-8.41668	0.59077	-14.25	<.0001	-9.57795	-7.25542
interaction		1	-1.17930	0.32096	-3.67	0.0003	-1.81021	-0.54839

Table 6 AIC values for different models

Number in Model	Adjusted R-Square	R-Square	AIC	SSE	Variables in Model
2	0.3429	0.3461	2302.2897	99470	stratio nonep_trans
1	0.3383	0.3398	2304.2662	100416	nonep_trans
1	0.0490	0.0512	2456.5910	144315	stratio

Table 7 The F-Values and sum of squares for the predictors

Factor	D F	Sum of Squares	Mean Square	F Value	Pr > F	Label
Stratio	3	2277.133873	759.044624	3.73	0.0115	stratio
nonep_trans	3	59858	19953	97.99	<.0001	

Appendix B – SAS Code

```
/*
Author : Omkar Kulkarni
Date : 22/11/2015
Description : Analysis of continious data HW 4

*/

/*import statements not included*/

data caschool;
    set conhw4.caschool;
RUN;

proc univariate plot data = caschool;
    var testscr stratio nonep;
    histogram testscr stratio nonep;
    qqplot testscr stratio nonep;
run;

proc corr data=caschool;
var testscr stratio nonep;
run;

proc sgscatter data=caschool;
title "scatter plot for caschool";
matrix testscr stratio nonep ;
run;

/* regressing without the iinteraction */

proc reg data=caschool;
model testscr = stratio nonep / clb ;
output out=atares r=rman ;
run;

/* transforming 'nonep' */

data caschool_trans;
set caschool;
    nonep_trans = log(nonep + 1 );
run;

proc univariate plot data = caschool_trans;
    var nonep_trans;
    histogram nonep_trans;
    qqplot nonep_trans;
run;

proc reg data=caschool_trans;
model testscr = stratio nonep_trans / clb ;
output out=atares r=rman ;
run;

/*checking assumptions*/
proc reg data=caschool_trans;
    model testscr = stratio nonep_trans /clb;
    output out=resid p=pman r=rman student=student;
```

```

run;

/* Check assumptions - normality */
proc univariate data=resid;
var student;
histogram student;
qqplot student;
run;

proc gplot data=resid;
plot rman*pman /vref=0;
run;

data resid2;
set resid;
res2 = rman*rman;
run;

goptions reset=all;
proc gplot data=resid2;
plot res2*pman;
run;
quit;

proc reg data=caschool_trans;
model testscr = stratio nonep_trans ;
plot student.*nqq. ;
run;

proc reg data=caschool_trans;
model testscr = stratio nonep_trans ;
plot student.*nqq. ;
run;

/*Q3 regression with only stratio*/
proc reg data=caschool_trans;
model testscr = stratio /clb;
output out=resid p=pman r=rman student=student;
run;

proc gplot data=resid;
plot rman*pman /vref=0;
run;

/*Q 4*/
/* Interaction term */
/*centering first*/

data caschool_trans_centered;
set caschool_trans;
stratio_centered = stratio - 19.64;
nonep_trans_centered = nonep_trans - 2.11;
run;

proc reg data=caschool_trans_centered;
model testscr = stratio_centered nonep_trans_centered /clb;
output out=resid p=pman r=rman student=student;
run;

```

```

data resint;
set resid;
interaction = stratio_centered*nonep_trans_centered;
run;

proc reg data=resint;
    model testscr = stratio_centered nonep_trans_centered interaction/clb
partial ;
    run;

/* graphical representation */

/*Q5*/
proc reg data=caschool_trans outest=est;
    model testscr = stratio nonep_trans / selection=adjrsq sse aic ;
    output out=out p=p r=r;
run;

proc rsreg data = caschool_trans;
    model testscr = stratio nonep_trans/ lackfit;
RUN;

PROC GLM data=caschool_trans;
model testscr = stratio nonep_trans / SOLUTION CLPARM;
RUN ;

/*Q 6*/
proc reg data=caschool_trans;
    model testscr = stratio nonep_trans / p cli;
    run;

/* Prediction and 95% prediction interval */
/*create data set for stratio =20 and nonep= 0.5*/
data caschool_pred;
input country $ district $ testscr stratio nonep;
cards;
dummy dummy . 20 0.5
run;

/*merge both datasets and create new one*/
data caschool2;
    set caschool caschool_pred;
Run;

data caschool2_trans;
set caschool2;
    nonep_trans = log(nonep + 1 );
run;

data caschool2_trans;
set caschool2_trans;
    interaction = nonep_trans * stratio;
run;

/*Pointwise prediction and confidence intervals at alpha=0.05 */
proc reg data=caschool2_trans;
model testscr = stratio nonep_trans interaction /p clm cli alpha=0.05;
run;

```