June Suh

Abbie Pearson

Jessica Warren

Oct. 25, 2016

STA4203

Midterm

In this project we will use the datasets m\_train.csv and m\_test.csv available on Blackboard. Fit a linear regression on m\_train, with y as the dependent variable and x1-x12 as predictors.

Answer the following questions.

Codes:

*proc import out=m\_train*

*datafile="/home/jes13j0/Homework/STA 4203/m\_train.csv"*

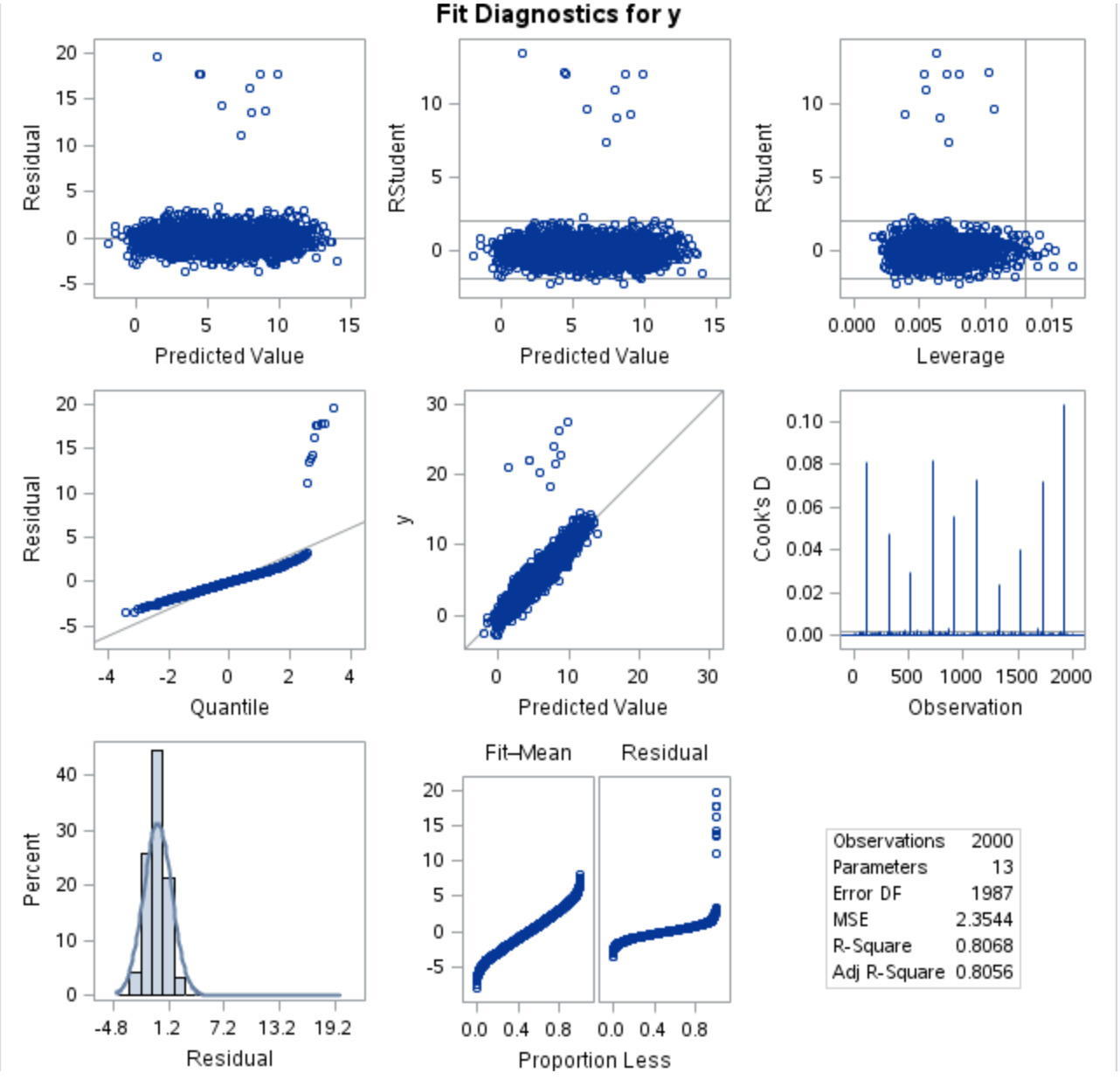
*dbms = csv replace;*

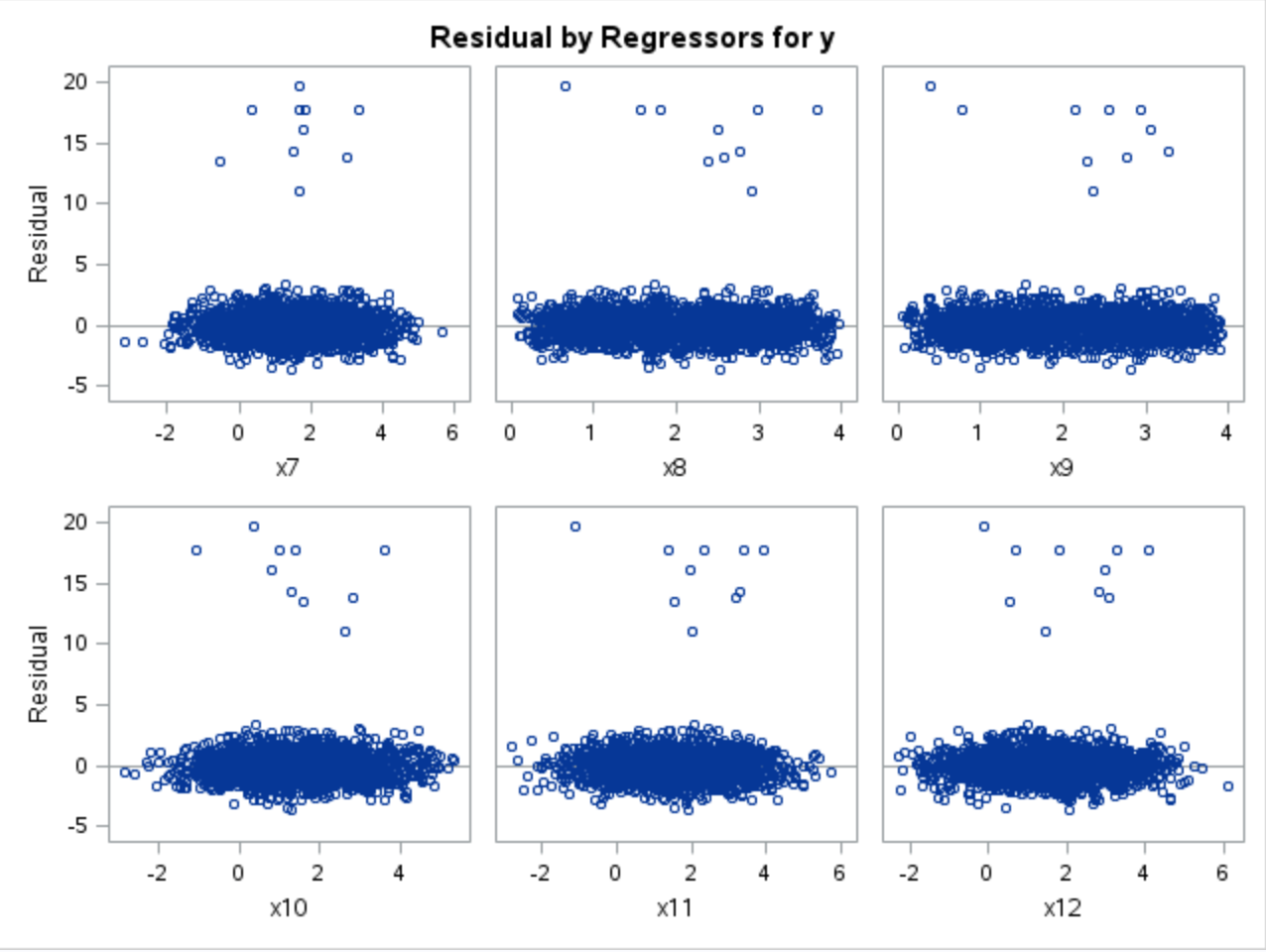
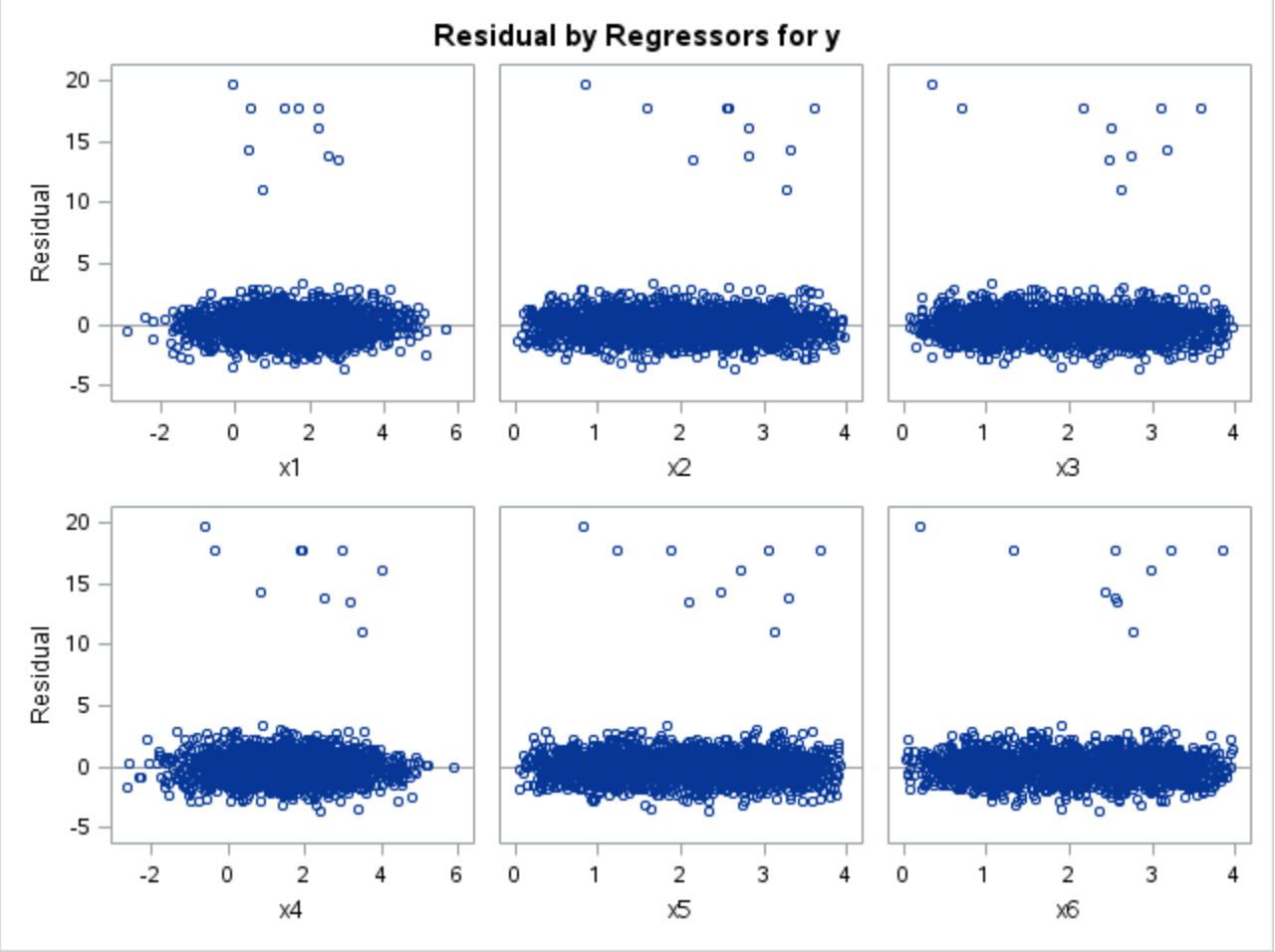
*run;*

*proc reg data=m\_train;*

*model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12;*

*run;*





a) Are the errors normally distributed?

No, the plots are clustered meaning, they do not show enough randomness and normality. In a graph that is normally distributed we want the data to be spread out the same according to x values.

b) Do the errors have constant variance?

No, because the data is not spread apart equally from one another.

c) Are the errors correlated?

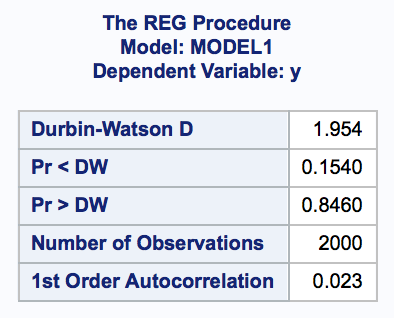
Codes:

*proc reg data=m\_train;*

*model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12/dwprob;*

*run;*

*Quit;*



Since pr<dw is higher than 0.05 the conclusion is that autocorrelation is not present.

d) Draw the partial regression plots.

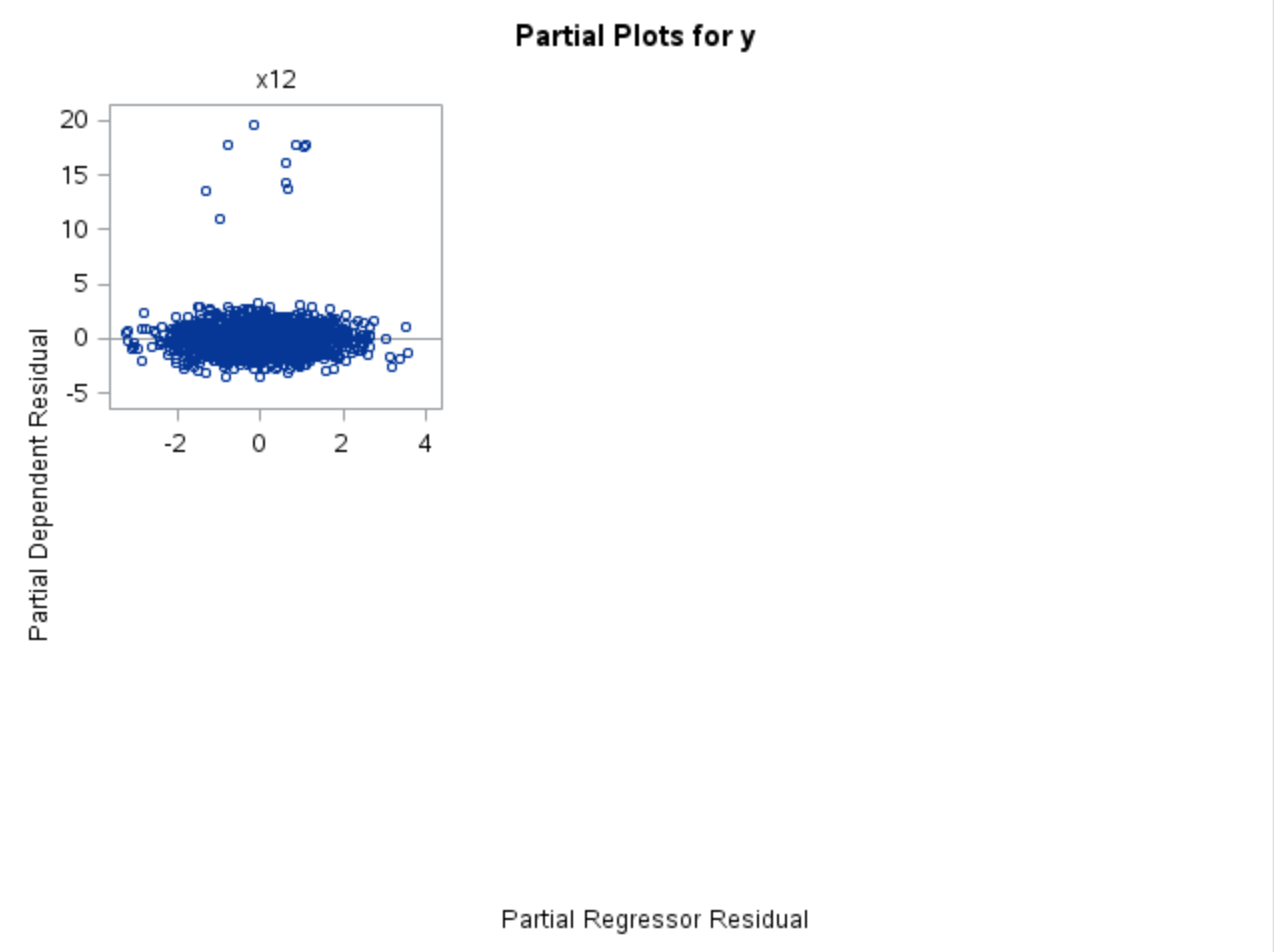
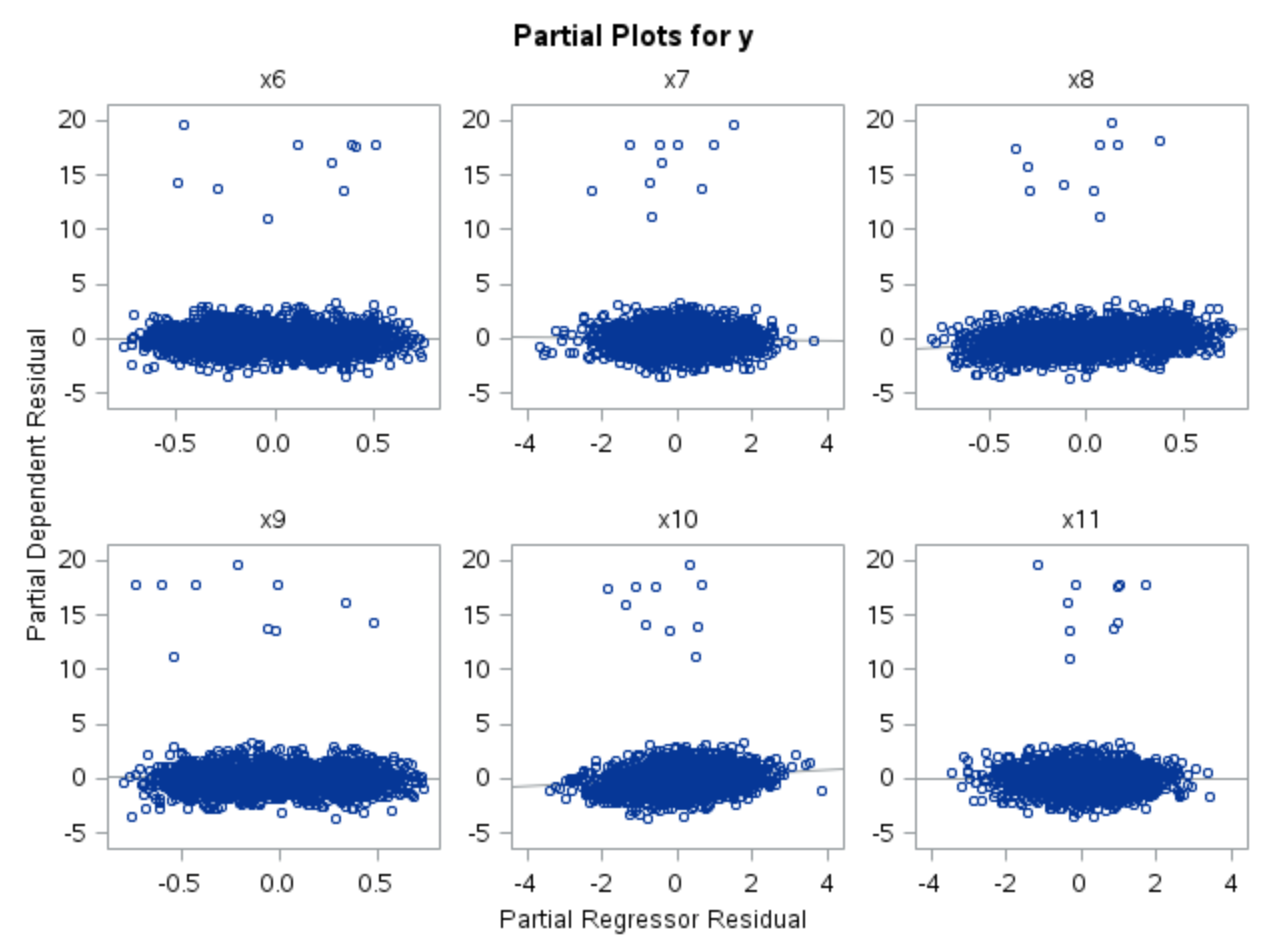
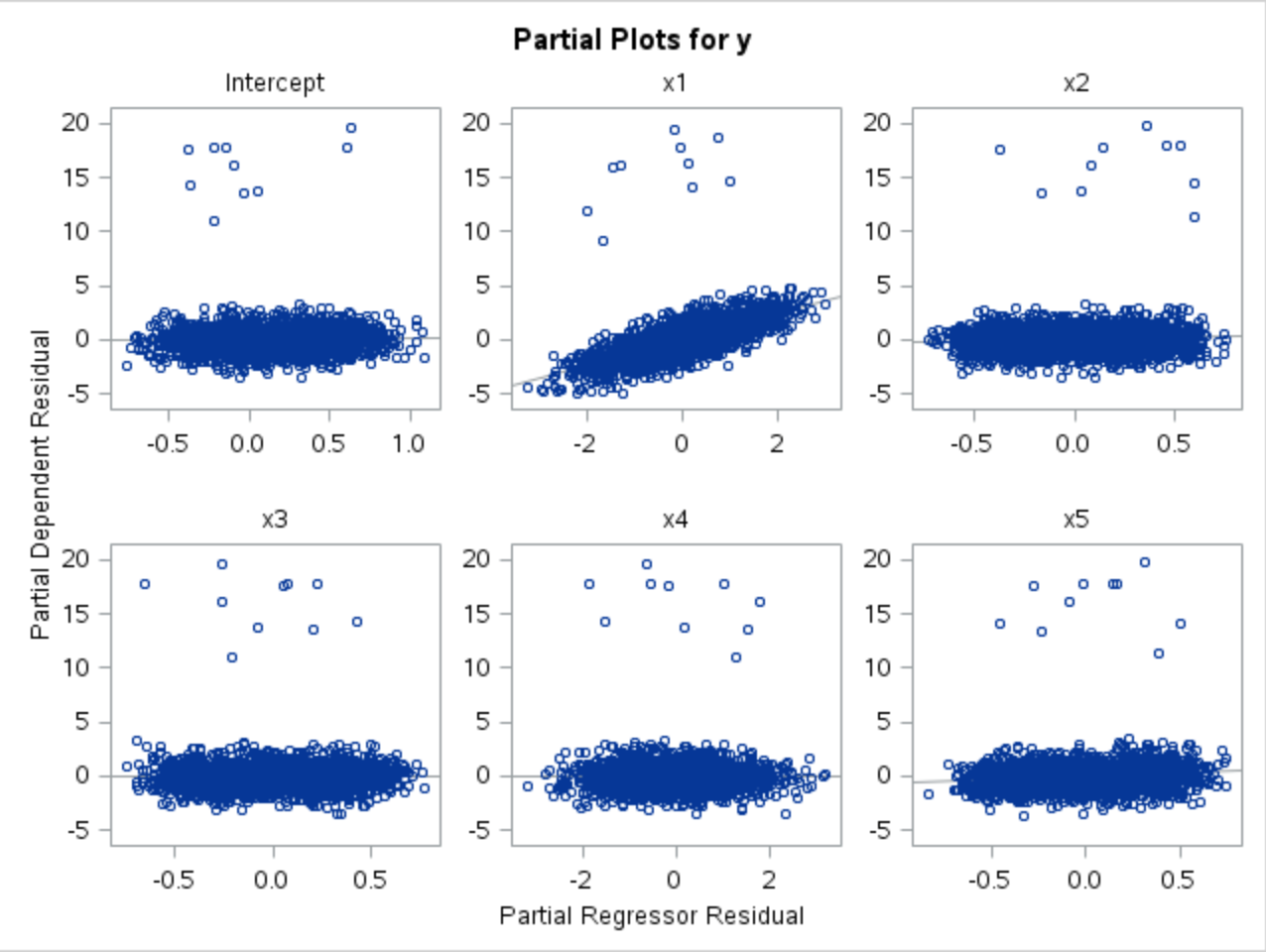
Codes:

*proc reg data=m\_train;*

*model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12/partial;*

*run;*

*quit;*



e) Perform collinearity diagnostics. Is there evidence of collinearity?

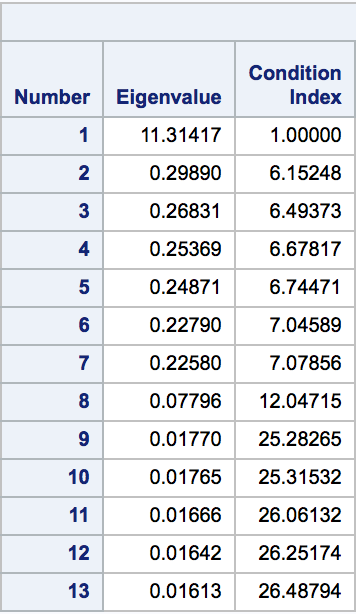
Codes:

*proc reg data=m\_train;*

*model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12/ vif collin;*

*run;*

*quit;*



Vif had some variables with 8+ values which are pretty high

Condition number 26.48794 is less than 30 which does not indicate collinearity. Some evidence

f) Are there any outliers? If so, report how many you found.

Codes:

*proc reg data=m\_train;*

*model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12;*

*output out=new rstudent=rez;*

*run;*

*quit;*

*data quantiles;*

*cutoff=abs(tinv(0.05/(2\*2000),2000-13-1));*

*run;*

*data outliers;*

*do i=1 to 2000 by 1;*

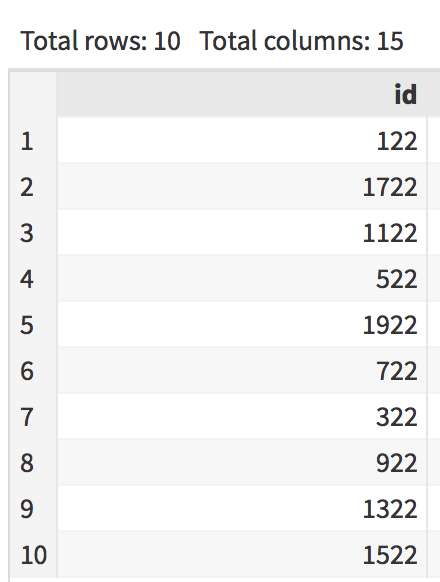
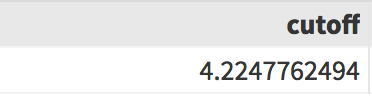
*set new point=i;*

*if (abs(rez)>4.2248) then output;*

*end;*

*stop;*

*run;*



Using the cutoff value of 4.2248, we found 10 outliers from the dataset.

g) Remove all outliers and refit the model on the remaining data and x1-x12 as predictors. Report the R2 .

Codes:

*data outliers1;*

*set new;*

*if (abs(rez)>4.2248) then delete;*

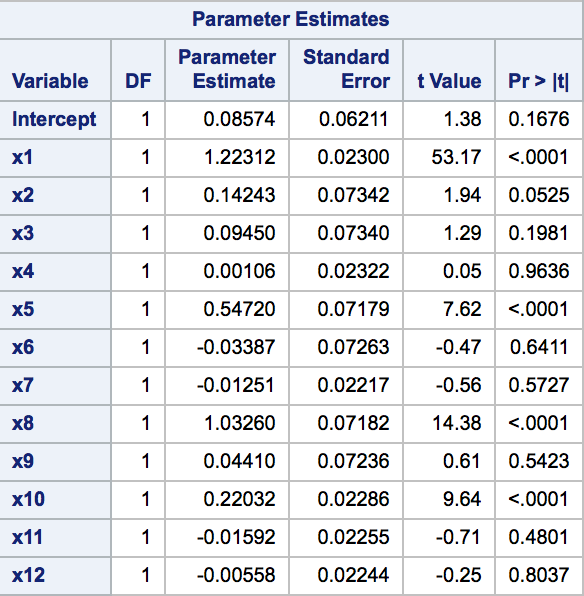
*run;*

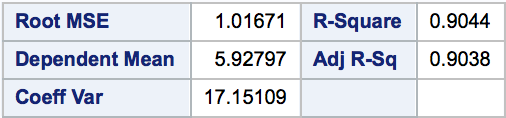
*proc reg data=outliers1;*

*model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12;*

*run;*

*quit;*





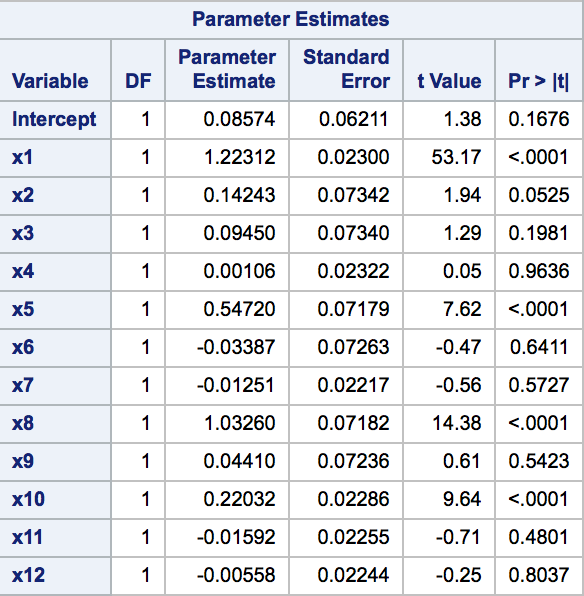
C

Model equation is: 0.08574 + 1.223\*x1 + 0.142\*x2 + 0.095\*x3 + 0.001\*x4 + 0.547\*x5 - 0.034\*x6 - 0.013\*x7 + 1.033\*x8 + 0.044\*x9 + 0.220\*x10 - 0.016\*x11 - 0.006\*x12

R^2 is 0.9044.

h) On the model from g) remove the variables that are not significant at the 0.05 level and refit the linear model (on the data from g). Report the model equation and R2 .

**From g):**

 **We will remove x2, x3, x4, x6, x7, x9, x11, and x12.**

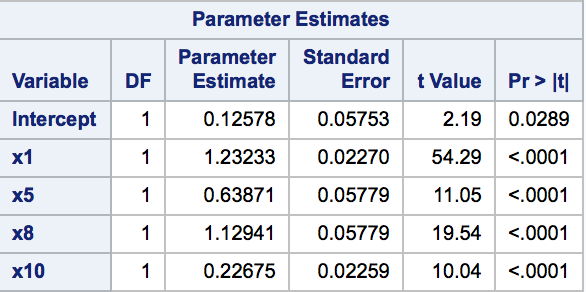
Codes:

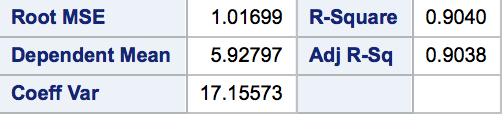
*proc reg data=outliers1;*

*model y=x1 x5 x8 x10;*

*run;*

*quit;*





Model equation is: 0.12578 + 1.232\*x1 + 0.639\*x5 + 1.129\*x8 + 0.227\*x10

R^2 is 0.9040.

i) Use the model from a) to get predicted values yˆi on the m\_test data. Report the MSE = 1/n Pn i=1(yi − yˆi) 2 where n is the number of observations.

Codes:

*proc import out=m\_test*

*datafile="/home/jes13j0/Homework/STA 4203/m\_test.csv"*

*dbms = csv replace;*

*run;*

*proc reg data=m\_train outest=model1;*

*model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12;*

*run;*

*quit;*

*proc score data=m\_test score=model1 out=test1 residual type = parms;*

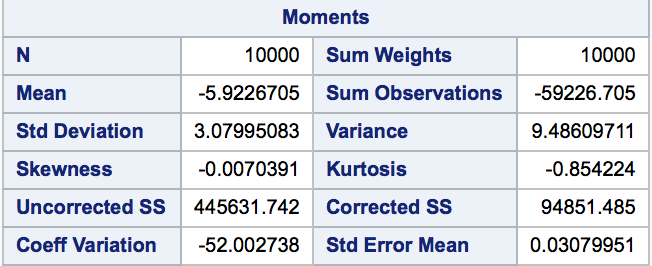
*var x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12;*

*run;*

*proc univariate data=test1;*

*var model1;*

*run;*



MSE = 445631.742/10000 = 44.5632

j) Use the model obtained at g) to get predicted values yˆi on the m\_test data and report the MSE.

Codes:

*proc reg data=outliers1 outest=model2;*

*model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12;*

*run;*

*quit;*

*proc score data=m\_test score=model2 out=test2 residual type = parms;*

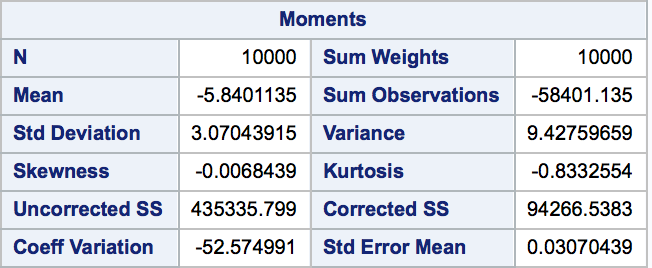
*var x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12;*

*run;*

*proc univariate data=test2;*

*var model1;*

*run;*



MSE = 435335.799/10000 = 43.5336

k) Use the model obtained at h) to get predicted values yˆi on the m\_test data and report the MSE.

Codes:

*proc reg data=outliers1 outest=model3;*

*model y=x1 x5 x8 x10;*

*run;*

*quit;*

*proc score data=m\_test score=model3 out=test3 residual type = parms;*

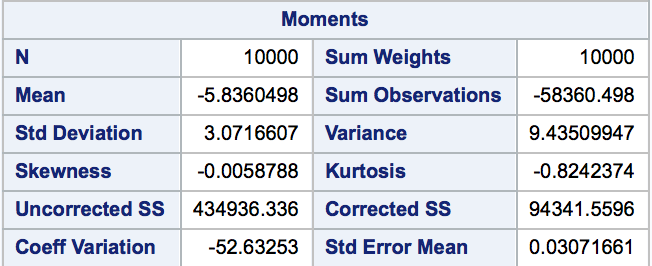
*var x1 x5 x8 x10;*

*run;*

*proc univariate data=test3;*

*var model1;*

*run;*



MSE = 434936.336/10000 = 43.4936

Construct a dataset train1 containing the first 1,000 observations of m\_train. Answer the following questions:

Code:

*Data train1;*

*set m\_train (firstobs = 1 obs = 1000);*

*;*

*run;*

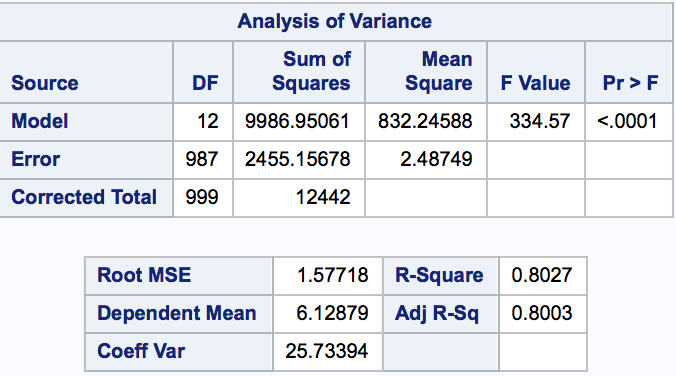
l) Fit a linear regression on train1, with y as the dependent variable and x1-x12 as predictors.Report the R2 .

Codes:

*proc reg data=train1;*

*model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12;*

*run;*



R^2 is 0.8027.

m) Remove all outliers from train1 based on the model from l) and refit the model on the remaining data and x1-x12 as predictors. Then remove the variables that are not significant at the 0.05 level and refit the linear model. Report the model equation and R2 .

Codes:

*proc reg data=train1;*

*model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12;*

*output out=training rstudent=res;*

*run;*

*quit;*

*data quantile;*

*cutoff=abs(tinv(0.05/(2\*1000),1000-13-1));*

*run;*

*data outlier;*

*do i=1 to 1000 by 1;*

*set training point=i;*

*if (abs(res)>4.0736) then output;*

*end;*

*stop;*

*run;*

*data outlier1;*

*set training;*

*if (abs(res)>4.0736) then delete;*

*run;*

*proc reg data=outlier1;*

*model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12;*

*run;*

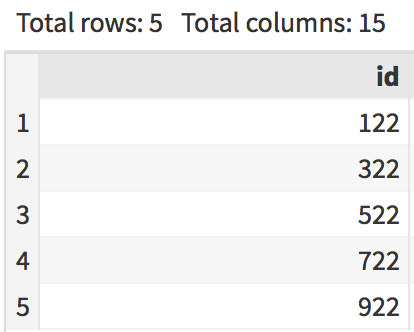
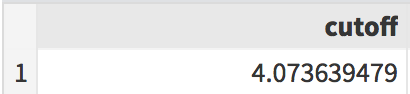
*quit;*

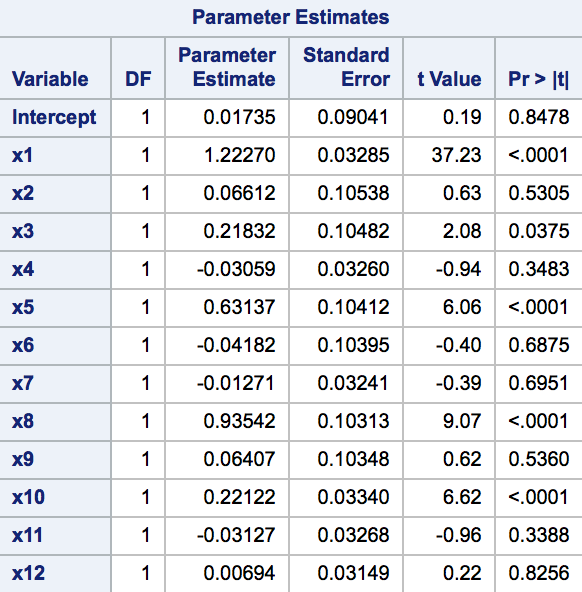
*proc reg data=outlier1;*

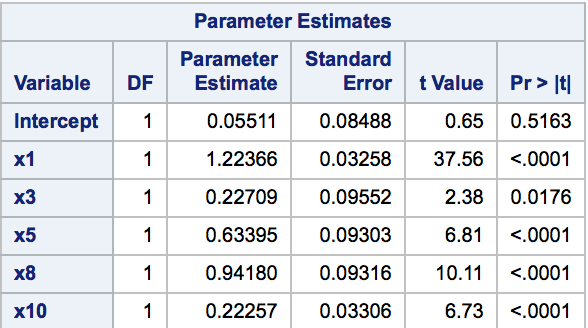
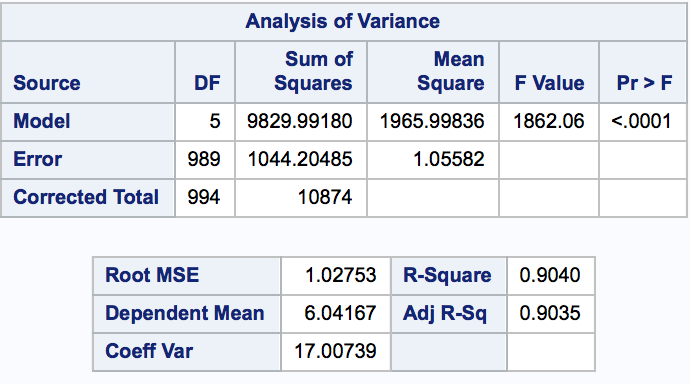
*model y=x1 x3 x5 x8 x10;*

*run;*

*quit;*

(Found 5 outliers)

 (refitted model) **We will remove x2, x4, x6, x7, x9, x11, and x12.**

(final model)

Model equation is: 0.05511 + 1.224\*x1 + 0.227\*x3 + 0.634\*x5 + 0.942\*x8 + 0.223\*x10.

R^2 is 0.9040.

n) Use the model obtained at l) to get predicted values yˆi on the m\_test data and report the MSE.

Codes:

*proc reg data=train1 outest=model4;*

*model y=x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12;*

*run;*

*quit;*

*proc score data=m\_test score=model4 out=test4 residual type = parms;*

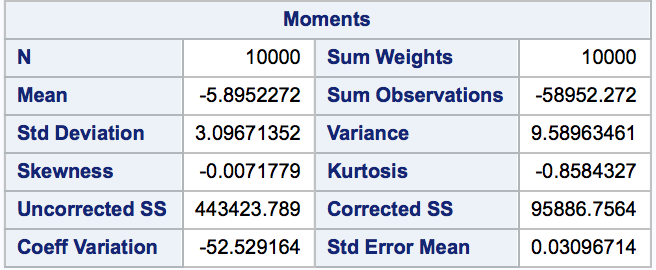
*var x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12;*

*run;*

*proc univariate data=test4;*

*var model1;*

*run;*



MSE = 443423.789 /10000 = 44.3424

o) Use the model obtained at m) to get predicted values yˆi on the m\_test data and report the MSE.

Codes:

*proc reg data=train1 outest=model5;*

*model y=x1 x3 x5 x8 x10;*

*run;*

*quit;*

*proc score data=m\_test score=model5 out=test5 residual type=parms;*

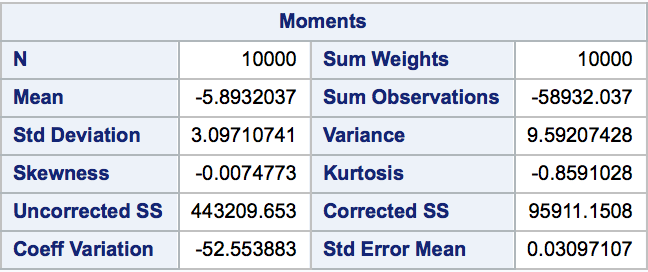
*var x1 x3 x5 x8 x10;*

*run;*

*proc univariate data=test5;*

*var model1;*

*run;*



MSE = 443209.653/10000 = 44.3210

So our steps were correct but we got wrong uncorrected SS from proc univariate no clue why and how it went wrong