June Suh

Abbie Pearson

Jessica Warren

Oct. 2, 2016

STA4203

Homework 5

1. Using the prostate data, fit a model with ***lpsa*** as the response and the other variables as predictors.

a) Code:

proc import out=prostate

datafile="/home/aep120/prostate.csv"

dbms = csv replace;

run;

Proc reg data=prostate;

model lpsa= lcavol lweight svi lbph age lcp pgg45 gleason;

output out=new r=resid p=pred;

run;

quit;

title "Residual vs Predicted value";

axis1 label=(angle=90 height=2 "residual");

axis2 label=(height=2 "predicted");

symbol1 value=dot color=blue;

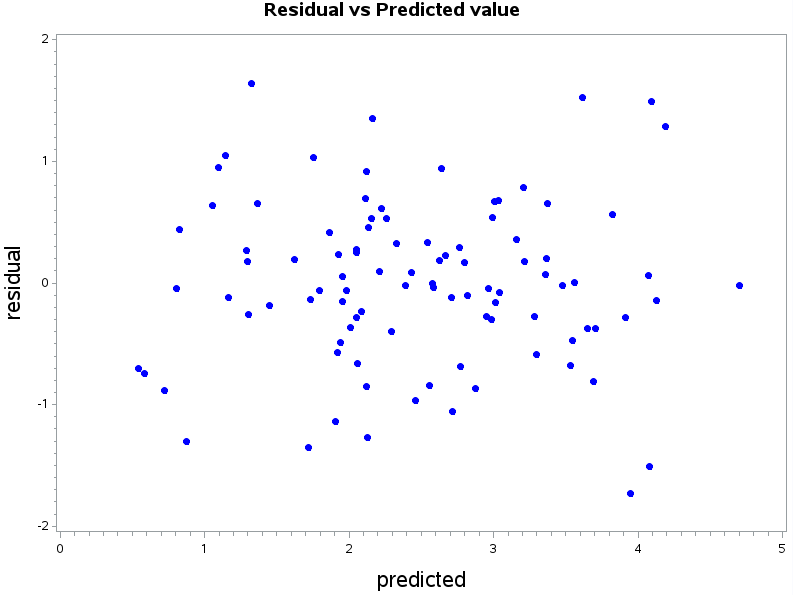
proc gplot data=new;

plot resid\*pred/

vaxis=axis1 haxis=axis2;

Run;

Results:



The Residuals do seem to have constant variance. The residuals are scattered randomly around zero. This indicates that the error terms have a mean of zero which means that this model is valid .

b) Code:

proc univariate data=new;

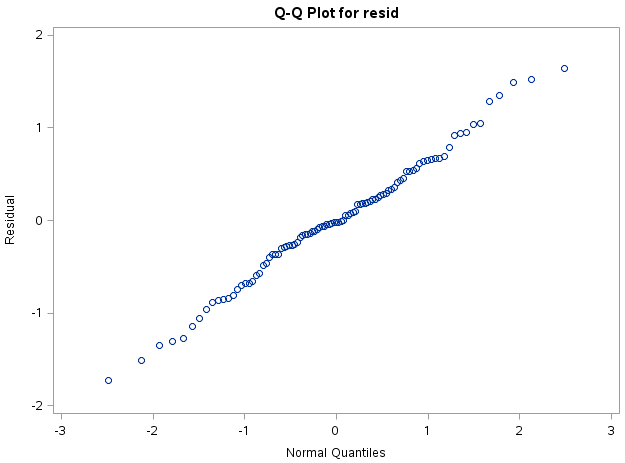
var resid;

qqplot;

run;

Results:

The QQ plot seems linear and relatively normal, which tells us that the observations create normal distributions. So we can assume our sample of the residuals from the *prostate* dataset comes from a set that is normally distributed.



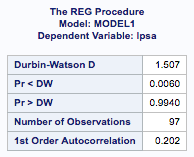
c) Code:

proc reg data=prostate;

model lpsa= lcavol lweight svi lbph age lcp pgg45 gleason /dwprob;

run;

Results:



Since Pr < DW was less than 0.05 the conclusion is that positive autocorrelation is present.

d) Code:

data new;set new;

abs\_resid = abs(resid);

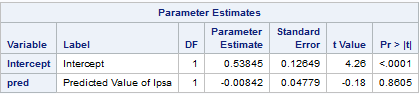
run;

proc reg data=new;

model abs\_resid = pred;

run;

Results:



With the p-value of 0.8605 the predicted values are not statistically significant at alpha value of 0.05.

2. Using the ***divusa*** data, fit a model with divorce as the response and the other variables except year as predictors.

a) Code:

proc import out=divusa

datafile="/home/aep120/divusa.csv"

dbms = csv replace;

run;

Proc reg data=divusa;

model divorce= unemployed femlab marriage birth military;

output out=new2 r=resid p=pred;

run;

quit;

title "Residual vs Predicted value";

axis1 label=(angle=90 height=2 "residual");

axis2 label=(height=2 "predicted");

symbol1 value=dot color=blue;

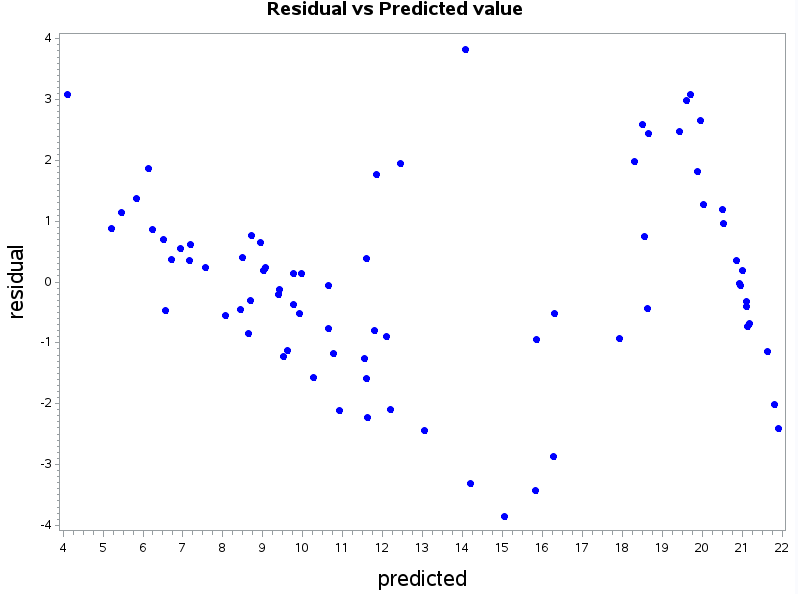
proc gplot data=new2;

plot resid\*pred/

vaxis=axis1 haxis=axis2;

run;

Results:



b) Code:

data sqt\_set;set divusa;

sqt\_div = sqrt(divorce);

run;

Proc reg data=sqt\_set;

model sqt\_div= unemployed femlab marriage birth military;

output out=sqt\_set r=resid p=pred;

run;

quit;

title "Residual vs Predicted value for Square Root";

axis1 label=(angle=90 height=2 "residual");

axis2 label=(height=2 "predicted");

symbol1 value=dot color=blue;

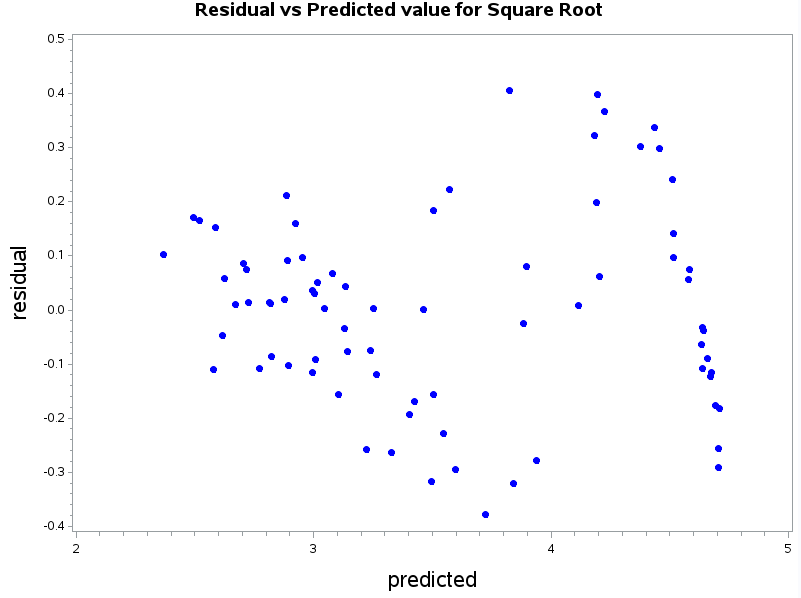
proc gplot data=sqt\_set;

plot resid\*pred/

vaxis=axis1 haxis=axis2;

run;

Results:



c) Code:

data log\_set;set divusa;

log\_div = log(divorce);

run;

Proc reg data=log\_set;

model log\_div= unemployed femlab marriage birth military;

output out=log\_set r=resid p=pred;

run;

quit;

title "Residual vs Predicted value for Natural Log";

axis1 label=(angle=90 height=2 "residual");

axis2 label=(height=2 "predicted");

symbol1 value=dot color=blue;

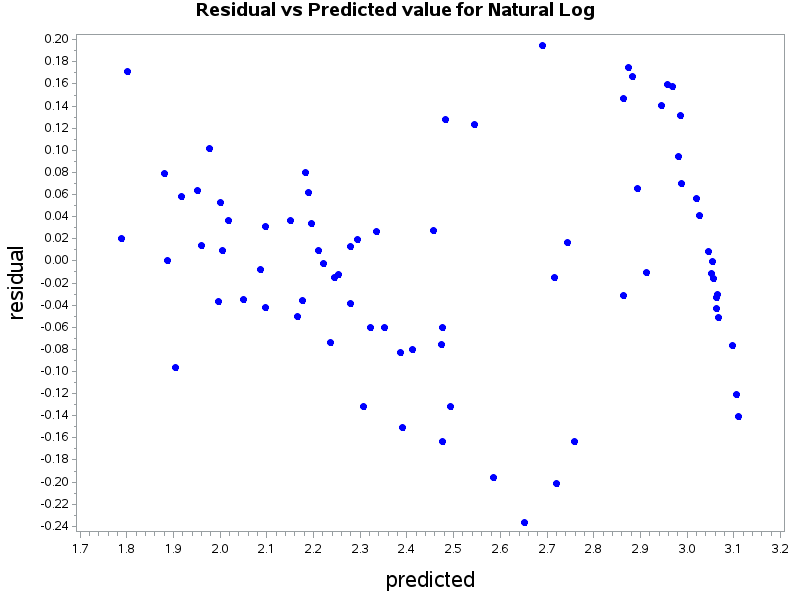
proc gplot data=log\_set;

plot resid\*pred/

vaxis=axis1 haxis=axis2;

run;

Results:



d) Based on the R2 , which of the three models is the best fit?

R2 of original model: 0.9208

R2 of model for sqrt(divorce): 0.9363

R2 of model for log(divorce): 0.9498

The model for log(divorce) has the highest R2 and therefore is the best fit.

e)Code:

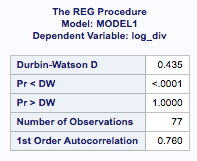
proc reg data=log\_set;

model log\_div= unemployed femlab marriage birth military/dwprob;

run;

quit;

Results:



Since Pr < DW was less than 0.05 the conclusion is that positive autocorrelation is present.