Abbie Pearson

Jessica Warren

June Suh

STA4203

Nov. 1, 2016

Homework 8

1. Researchers at NIST collected pipeline data on ultrasonic measurements of the depths of defects in the Alaska pipeline in the field. The depth of the defects were then remeasured in the laboratory. These measurements were performed in six batches. It turns out that this batch effect is not significant and so can be ignored in what follows. The laboratory measurements are more accurate than the in-field measurements, but more time consuming and expensive. We want to develop a regression equation for correcting the in-field measurements.

a) Fit a regression model with lab as the response and field as the predictor. Plot residuals vs predicted and check for nonconstant variance.

Code:

*proc import out=pipeline*

*datafile="/home/aep120/pipeline.csv"*

*dbms = csv replace;*

*run;*

*proc reg data=pipeline;*

*model lab=field;*

*output out=pipeVals r=resid p=pred;*

*run;*

*quit;*

*title "Residual vs Predicted value";*

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

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

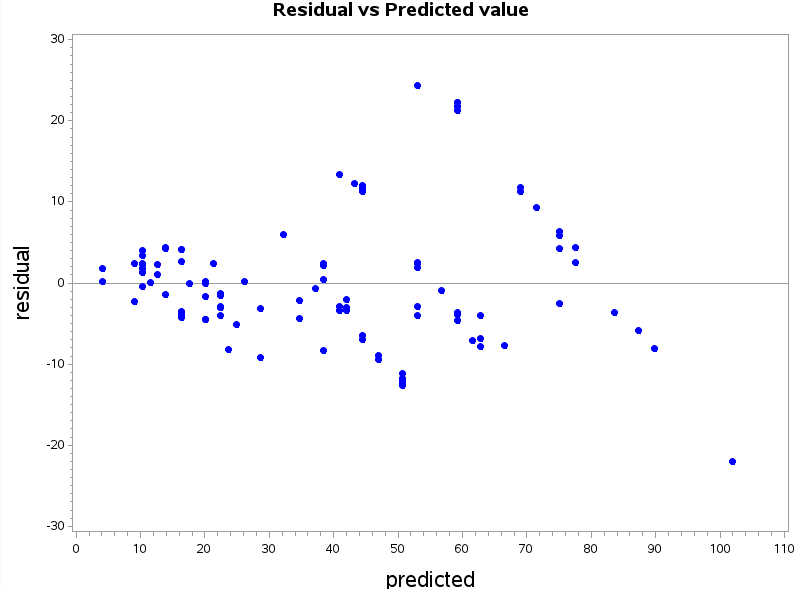
*symbol1 value=dot color=blue;*

*proc gplot data=pipeVals;*

*plot resid\*pred/*

*vaxis=axis1 haxis=axis2;*

*run;*



We observe some points that seem to be outliers or influential on the graph

b) Sort the dataset pipeline by the field variable. Observations with the same field value should be sorted by lab. Write only the SAS code that was used to do the sorting.

Code:

proc sort data=pipeline;

by field lab;

run;

c) Split the range of field into 16 groups of size seven (except for the last group which has only two values). Create a dataset named new with 15 observations. Each observation should contain the variance varlab of the lab values and the mean meanfield of the field values of a group of seven consecutive rows from the sorted pipeline dataset from point b). Draw the scatter plot of the varlab vs. meanfield.

*data d1 d2 d3 d4 d5 d6 d7 d8 d9 d10 d11 d12 d13 d14 d15 d16;*

*set pipeVals;*

*if \_n\_ <= 7 then output d1;*

*else if \_n\_ > 7 and \_n\_ <= 14 then output d2;*

*else if \_n\_ > 14 and \_n\_ <= 21 then output d3;*

*else if \_n\_ > 21 and \_n\_ <= 28 then output d4;*

*else if \_n\_ > 28 and \_n\_ <= 35 then output d5;*

*else if \_n\_ > 35 and \_n\_ <= 42 then output d6;*

*else if \_n\_ > 42 and \_n\_ <= 49 then output d7;*

*else if \_n\_ > 49 and \_n\_ <= 56 then output d8;*

*else if \_n\_ > 56 and \_n\_ <= 63 then output d9;*

*else if \_n\_ > 63 and \_n\_ <= 70 then output d10;*

*else if \_n\_ > 70 and \_n\_ <= 77 then output d11;*

*else if \_n\_ > 77 and \_n\_ <= 84 then output d12;*

*else if \_n\_ > 84 and \_n\_ <= 91 then output d13;*

*else if \_n\_ > 91 and \_n\_ <= 98 then output d14;*

*else if \_n\_ > 98 and \_n\_ <= 105 then output d15;*

*else output d16;*

*run;*

*%macro ave(dat,name,f,l);*

*proc univariate data= &dat;*

*var &f &l;*

*output out=&name var= x varlab mean=meanfield xx;*

*run;*

*%mend ave;*

*%macro what;*

*data test;*

*set*

*%do i=1 %to 15;*

*%ave(d&i ,a&i,field,lab );*

*%end;*

*;*

*run;*

*%mend;*

*%what;*

*%macro combine;*

*data new;*

*set*

*%do i = 1 %to 15;*

*a&i*

*%end;*

*;*

*run;*

*%mend;*

*%combine;*

*data new;set new(keep=meanfield varlab);run;*

*title "Varlab vs. Meanfield";*

*axis1 label=(angle=90 height=2 "varlab");*

*axis2 label=(height=2 "meanfield");*

*symbol1 value=dot color=blue;*

*proc gplot data=new;*

*plot varlab\*meanfield/*

*vaxis=axis1 haxis=axis2;*

*run;*

Data new;

Do i=0 to 16;

n=min (7,107-7\*i);

sumfield=0;

sumlab2=0;

sumlab=0;

Do obsnum=7\*i+1 to 7\*i+n;

Set pipeline point=obsnum;

If \_error\_ then abort;

sumfield =sumfield+field;

sumlab=sumlab+lab;

sumlab2=sumlab2+lab\*lab;

End;

meanfield=sumfield/n;

varlab=(sumlab2-sumlab\*sumlab/n)/(n-1);

logmf=log(meanfield);

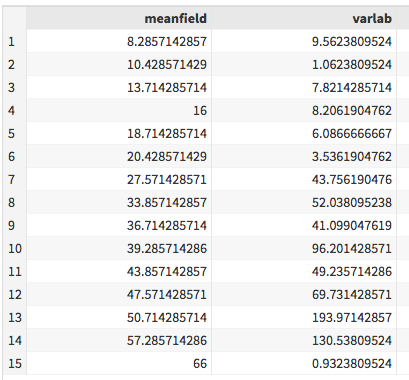
logvl=log(varlab);

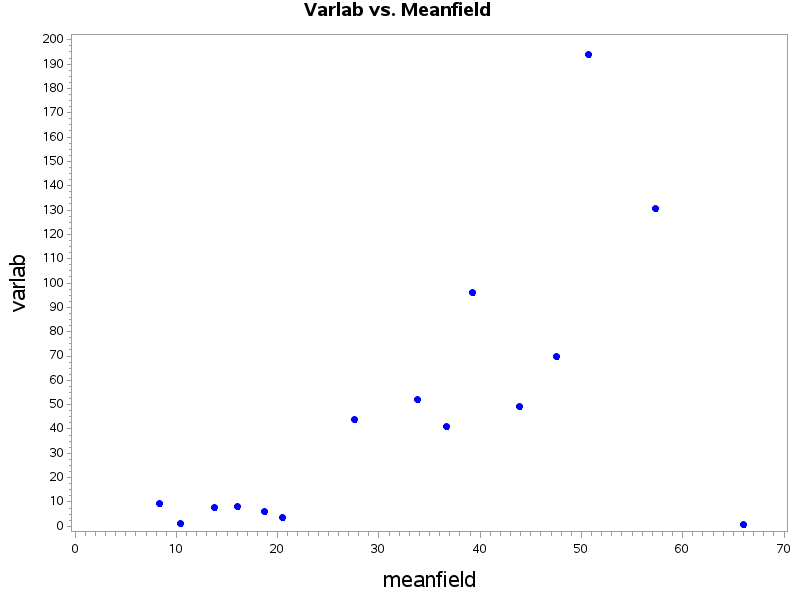
Output;

End;

Stop;

run;





d) Suppose we guess that the variance in the response is linked to the predictor in the following way: varlab = a0meanfielda1 Estimate a0 and a1 by regressing log(varlab) on log (meanfield).

Code:

*data new; set new;*

*logvarLab= log(varlab);*

*logmeanfield = log(meanfield);*

*run;*

*proc reg data=new;*

*model logvarLab = logmeanfield;*

*run;*

*quit;*

) EXPONENTIATE! Get rid of the logs

Equation from regression:

Let and

Therefore and

Proc reg data=new;

Model logvl=logmf;

Run;

Quit; (get parameter of estimates table)

Get the equation and solve for varlab (we have logvarlab so we raise them to e to get rid of log and go on from there)

e) Estimate the variance for each observation in the sorted pipeline data by using field instead of meanfield in the equation from d) above. Draw the scatterplot of the absolute value of the residuals from a) vs field and overlay the curve of the square root of the estimated variance vs field.

Code:

*data varlabset; set new(keep=varlab);*

*do i=1 to 7;*

*output;*

*end;*

*run;*

*data sortedpip;*

*set d1(keep= field)*

*d2(keep= field)*

*d3(keep= field)*

*d4(keep= field)*

*d5(keep= field)*

*d6(keep= field)*

*d7(keep= field)*

*d8(keep= field)*

*d9(keep= field)*

*d10(keep= field)*

*d11(keep= field)*

*d12(keep= field)*

*d13(keep= field)*

*d14(keep= field)*

*d15(keep= field);*

*set varlabset(keep=varlab);*

*run;*

*data sortedpip; set sortedpip;*

*logvar = log(varlab);*

*logfld = log(field);*

*run;*

*proc reg data=sortedpip;*

*model logvar=logfld;*

*run;*

*quit;*

data sortedpip; set sortedpip;

logvar = log(varlab);

logfld = log(field);

run;

proc reg data=sortedpip;

model logvar=logfld;

run;

quit;

data pipeVals; set pipeVals;

absres=abs(resid);

estvar= (0.333208)\*(field\*\*1.31974);

sqrtest=sqrt(estvar);

run;

title "Abs value of Residuals & sqrt of Estimates vs Field";

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

symbol1 value=dot color=blue;

symbol2 interpol=join color=red;

proc gplot data=pipeVals;

plot (absres sqrtest)\*field/ overlay

haxis=axis2;

run;

)

Equation from regression:

Let and

Therefore and

Data pipeline;

Set pipeline;

varlab=0.1985\*field\*\*0.3683(coefficient of logmf from previous question);

Run;

Proc reg data=pipeline;

Model lab=field;

Output out=new1 r=res;

Run;

Quit;

Data pipeline;

Set pipeline;

ares=abs(res);

slab=varlab\*\*0.5;

Run;

axis 1 label=(angle=90 height=0.75);

Symbol1 value=circle color=black height=0.5;

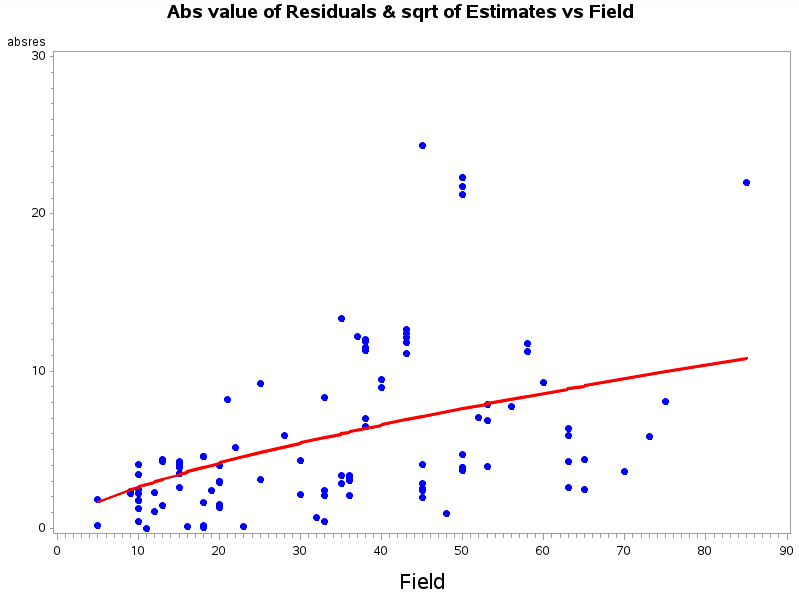
Symbol2 value=none color=red interpol=join;

Proc gplot data=new1;

Plot ares\*field slab\*field/overlay noframe

vaxis=axis1 vminor=1 hminor=0;

run;



f) Use the estimated variance to determine appropriate weights for a WLS fit of lab on field. Report the regression equation and R2 .

Code:

*data pipeVals;*

*set pipeVals;*

*field\_inv = 1/estvar;*

*run;*

*proc reg data=pipeVals;*

*model lab = field;*

*weight field\_inv;*

*run;*

*quit;*

Data pipeline;

Set pipeline;

varlab=0.1995\*field\*\*0.3683;

ivar=1/varlab;

Run;

Proc reg data=pipelines;

Model lab=field;

Weight ivar;

Output out=new1 r=res;

Run;

Quit;

Screen Shot 2016-11-02 at 10.24.36 PM.png

g) An alternative to weighting is transformation. Find transformations on lab and/or field as follows. For the field variable, you may restrict your choice of transformation to identity, square root, log and inverse. For each of the four field transformations, use the Box-Cox method to find an appropriate transformation of the lab, and plot the residuals vs predicted based on the obtained model. Based on the graphs, report the combination of transformations that give the most constant variance

Code:

data trans; set pipeline;

sqrtfield = sqrt(field);

logfield = log(field);

invfield = 1/(field);

run;

proc transreg data = trans;

model boxcox (lab/lambda=-2 to 2

by 0.02 alpha=0.05) = identity(invfield);

run;

data trans; set trans;

transfield = field\*\*.52;

trnsqfield = field\*\*.26;

trnlogfield = field\*\*-.04;

trninvfield = field\*\*-.64;

run;

proc reg data=trans;

model lab = transfield;

output out=trans1 residual=r1 p=p1;

run;

quit;

proc reg data=trans;

model lab = trnsqfield;

output out=trans2 residual=r2 p=p2;

run;

quit;

proc reg data=trans;

model lab = trnlogfield;

output out=trans3 residual=r3 p=p3;

run;

quit;

proc reg data=trans;

model lab = trninvfield;

output out=trans4 residual=r4 p=p4;

run;

quit;

Data pipeline;

Set pipeline;

varlab=0.1995\*field\*\*0.3683;

ivar=1/varlab;

ifield =1/field;

lfield=log(field);

sfield=field\*\*0.5;

Run;

Proc transreg data=pipeline;

Model boxcox(lab/lambda=0 to 1 << zoomed in from -2 to 2 for better graph optional

By 0.1 alpha=0.05)=identity(field);

Run; (lambda is 0.52)

Data pipeline;

Set pipeline;

varlab=0.1995\*field\*\*0.3683;

ivar=1/varlab;

ifield =1/field;

lfield=log(field);

sfield=field\*\*0.5;

slab=lab\*\*0.5;

Run;

Proc reg data=pipeline;

Model slab=field;

Run;

Quit;

Proc transreg data=pipeline;

Model boxcox(lab/lambda=-1 to 0

By 0.1 alpha=0.05)=identity(ifield);

Run; (lambda is -0.5)

Data pipeline;

Set pipeline;

varlab=0.1995\*field\*\*0.3683;

ivar=1/varlab;

ifield =1/field;

lfield=log(field);

sfield=field\*\*0.5;

slab=lab\*\*0.5;

islab=lab\*\*(-0.5);

Run;

Proc reg data=pipeline;

Model islab=ifield;

Run;

quit;

Proc transreg data=pipeline;

Model boxcox(lab/lambda=0 to 1

By 0.1 alpha=0.05)=identity(sfield); << ok i think he is looking at the line in middle of graph

Run;

Data pipeline;

Set pipeline;

varlab=0.1995\*field\*\*0.3683;

ivar=1/varlab;

ifield =1/field;

lfield=log(field);

sfield=field\*\*0.5;

slab=lab\*\*0.5;

islab=lab\*\*(-0.5);

slab=lab\*\*0.25;

Run;

~~~ repeat for all the variables and pick the one with constant variance residual graph ~~

From the graphs below the combination of transformations that give the most constant variance are identity, log and square root.

