Niranjan Poudel
STAT 5100
Spring 2019

Homework 2

Exercise 1.

If clerical errors are frequently made in determining the dollar sales, with number of units sold being measured accurately the relation between number of units sold and dollar sales can still be functional, as long as every unit sold can be associated with a single dollar sales value. If one-unit sold shows two value of dollar sell it cannot be functional. For, example if 50 unit sold show two value of dollar sales of 100 and 150 than it cannot be a functional relationship.

Exercise 2.

I disagree with the student of simple linear regression equation because the error term should have the mean zero which reduces the equation to $E\{Y\} = \beta o + \beta 1X$.

Exercise 3.

Here the given regression function is E{Y} = 20 + .95X with the X value ranging from 40 to 100. The observer is wrong about the conclusion that the training does not raise the production. If we keep the value of X as 40 the value of average Y is 56 and if we keep the value of X as 100 the value is 115. For the given sets of data, the training increases the output. If the value of X was greater than or equal to 400 than there would not be increase on the value. So for the given sets of data on average there is an increase.

Exercise 4.

a. The least squares estimates obtained from SAS are $\beta o = 2.114$ & $\beta 1 = 0.039$

Parameter Estimates

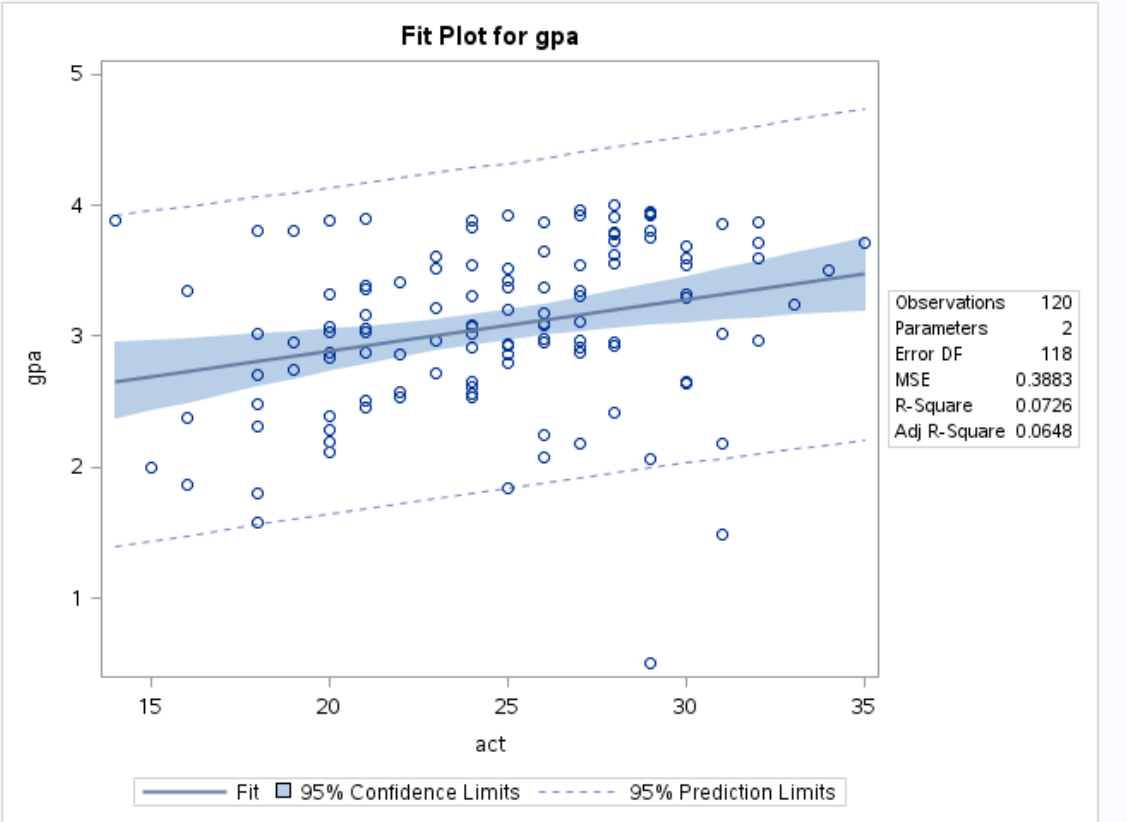
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.11405	0.32089	6.59	<.0001
act	1	0.03883	0.01277	3.04	0.0029

So, the estimated regression function is $E\{Y\} = 2.114 + 0.039X$

b. Here is the plot of estimated regression function and data.

Niranjan Poudel
STAT 5100
Spring 2019

Homework 2



- From the plot it seems like the regression function does not fit the data well because the point does not spread randomly on both sides.

c. Using the estimated equation, the point estimate of mean GPA for student with ACT score X=30 is 3.284.

d. The point estimate of the change in the mean response when the entrance test score increases by one point is β1 which is 0.039.

i.e. $(2.114 + 0.039(X + 1))$- $(2.114 + 0.039X)$
I.e. 0.039

Niranjan Poudel
STAT 5100
Spring 2019

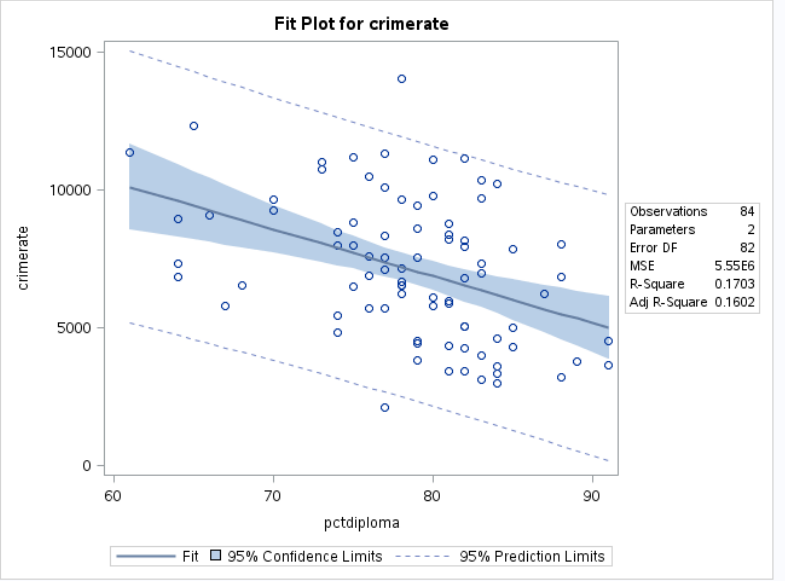
Homework 2

Exercise 5.

a. The estimated linear regression function can be stated as E{Y}=20518-170.575X ……. (1.1) from the following table

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	20518	3277.64269	6.26	<.0001
pctdiploma	1	-170.57519	41.57433	-4.10	<.0001

- The plot is as follow;



- The regression function does not seem to fit the data well.

b.

1) The difference in the point estimate of the difference in the mean crimate of two counties whose high-scholl graduation rates differ by one percentage is decrease by 170.575.

Niranjan Poudel
STAT 5100
Spring 2019

Homework 2

2) The mean crime rate for the in counties with high school graduate percentage X=80 from the above equation (1.1) is 6872.08.

3) For ϵ_{10}, for X=82 , the mean point estimate of E{Y} = 20518-170.575*82
= 20518-13987.15
= 6530.85

Now we have from the output data the actual value of Y is 7932 so the tenth error term ϵ_{10}= 7932-6530.85= 1401.15.

4) σ^2= 5552112 from the following table.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	93462942	93462942	16.83	<.0001
Error	82	455273165	5552112		
Corrected Total	83	548736108			

Exercise 6.

- From the equation (1.9a) we have, $\sum Y_i = nb_0 + b_1 \sum X_i$ now solving this for b_0 we get the following equation;

  i.e, $nb_0 = \sum Y_i - b_1 \sum X_i$

  or, $b_0 = \frac{\sum Y_i - b_1 \sum X_i}{n}$

  i.e, $b_0 = \bar{Y} - b_1 \bar{X}$ (1.2)

- Again from equation (1.9b) we have $\sum XiYi = bo \sum Xi + b_1 \sum X_i^2$ (1.3)

  Putting the value of bo from equation (1.2) into equation (1.3) we get

  $\sum X_i Y_i = (\bar{Y} - b_1 \bar{X}) \sum X_i + b_1 \sum X_i^2$

Niranjan Poudel
STAT 5100
Spring 2019

Homework 2

or, $\sum X_i Y_i = \bar{Y}\sum X_i - b_1\bar{X}\sum X_i + b_1\sum X_i^2$

or, $b_1(\sum X_i^2 - \bar{X}\sum X_i) = \sum X_i Y_i - \bar{Y}\sum X_i$

or, $b_1 = \frac{\sum X_i Y_i - \bar{Y}\sum X_i}{\sum X_i^2 - \bar{X}\sum X_i}$ (1.4)

Now we have given in equation (1.10a) as follow;

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X - \bar{X})^2}$$

$$\text{or, } b_1 = \frac{\sum(X_i Y_i - X_i\bar{Y} - \bar{X}Y_i + \bar{X}\bar{Y})}{\sum\left(X_i^2 - 2X_i\bar{X} - (\bar{X})^2\right)}$$

$$= \frac{\sum X_i Y_i - \bar{Y}\sum X_i - \bar{X}\sum Y_i + n\bar{X}\bar{Y}}{\sum X_i^2 - 2\bar{X}\sum X_i + n(\bar{X})^2}$$

$$= \frac{\sum X_i Y_i - \bar{Y}\sum X_i - \bar{X}\sum Y_i + n\bar{X}\frac{\sum Y_i}{n}}{\sum X_i^2 - 2\bar{X}\sum X_i + n\bar{X}\frac{\sum X_i}{n}}$$

or, $b_1 = \frac{\sum X_i Y_i - \bar{Y}\sum X_i}{\sum X_i^2 - \bar{X}\sum X_i}$ which is same as equation (1.4) which proves that denominator and numerator are equal to those in equation (1.10a) .

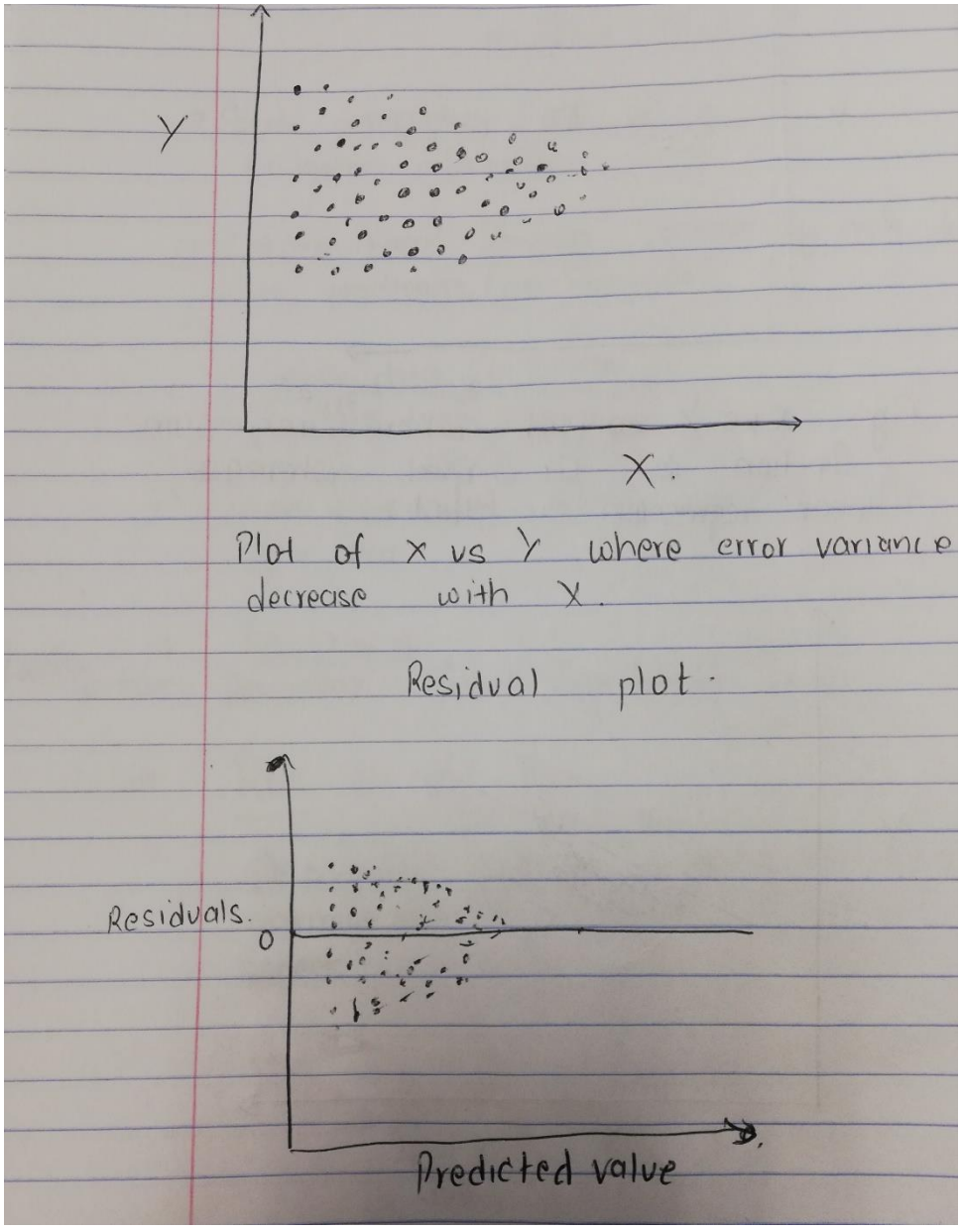
Exercise 7.

3.2. Prepare a prototype residual plot for each of the following cases: (1) error variance decreases with X; (2) true regression function is U shaped, but a linear regression function is fitted.

Niranjan Poudel
STAT 5100
Spring 2019

Homework 2

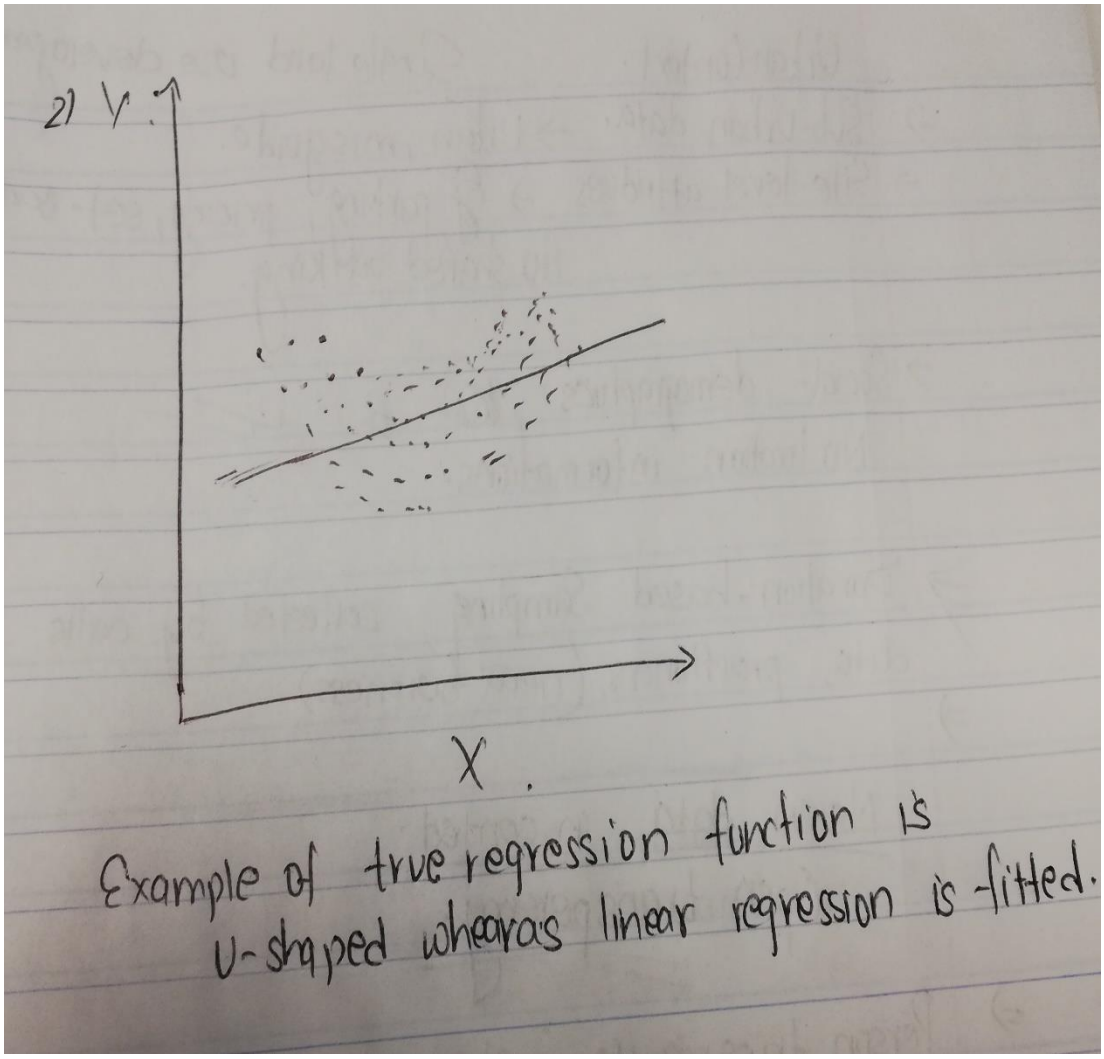
1) Error variance decreases with X



Niranjan Poudel
STAT 5100
Spring 2019

Homework 2

2) True regression function is U shaped, but a linear regression function is fitted.

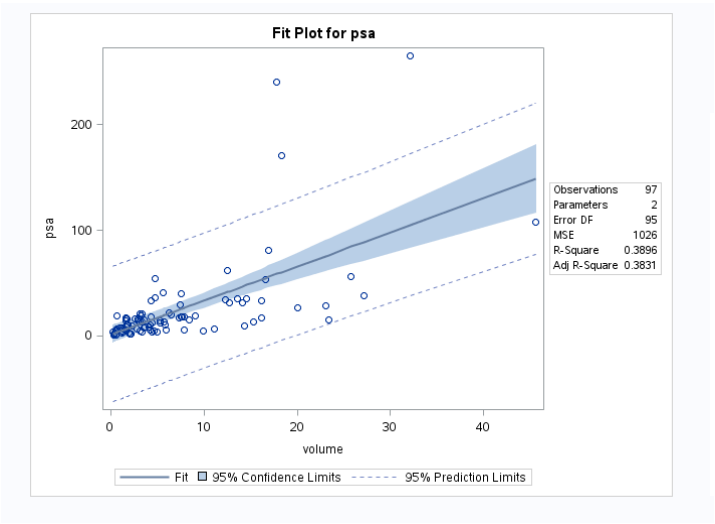


Exercise 8.

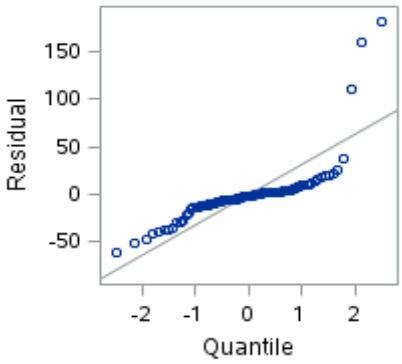
a) A model was fitted to predict PSA from volume. Following are the histogram, normal probability plot, sequence plot, and plot vs. predicted values.

Niranjan Poudel
STAT 5100
Spring 2019

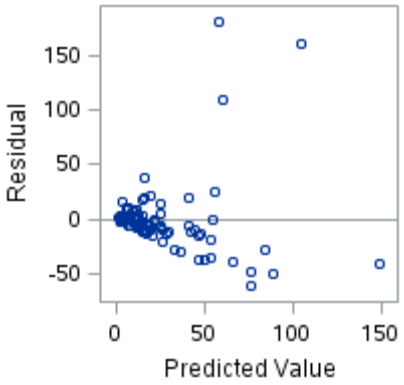
Homework 2



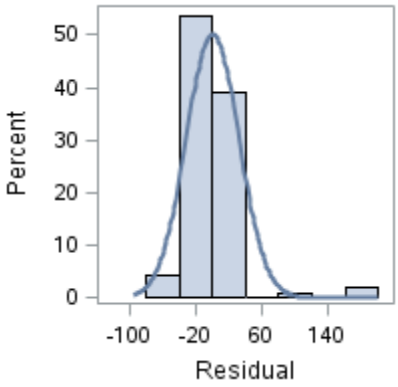
Fit plot for psa



Normal probability plot



Residual plot



Histogram with normal curve

- The residual plot shows there might not be constant variance of errors, the histogram shows the skewness in the error and the normal probability distribution also does not show the normal distribution of the error which suggest the assumptions are violated.
- The plot also shows the data which is unfit for the regression model due to which there needs to be remedial measures.

Following is the Brown-Forsythe and correlation test for normality

Niranjan Poudel
STAT 5100
Spring 2019

Homework 2

P-value for Brown-Forsythe test of constant variance in residual vs. Predicted Value

Obs	t_BF	BF_pvalue
1	3.64116	.000441824

- Here the BF_pvalue is much less than 0.05 which also further strengthen the violation of assumptions that there might not be constant variance.

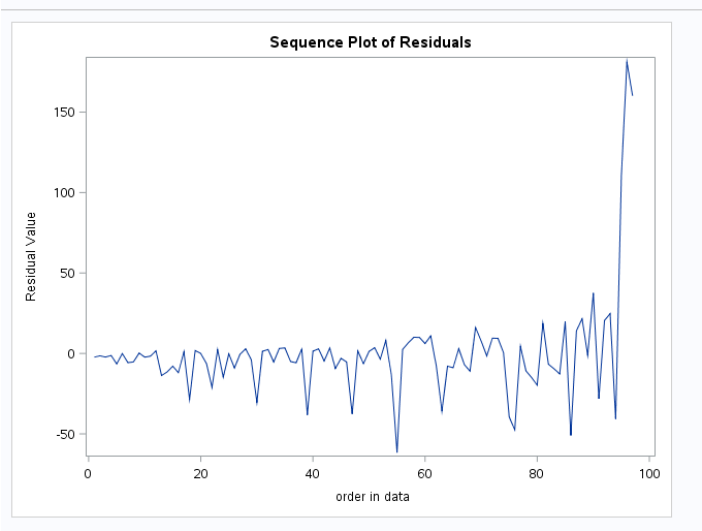
Pearson Correlation Coefficients, N = 97
Prob > |r| under H0: Rho=0

	resid	expectNorm
resid residual	1.00000	0.78208 <.0001
expectNorm	0.78208 <.0001	1.00000

- Looking at the TABLE B.6 provided in **Stat 5100, Chapter 3 – Diagnostics and Remedial Measures** we see the threshold to be around 0.987 whereas from above table the correlation coefficient is 0.782 which is less than 0.986 which suggest that there might be no constant variance of the errors.

Niranjan Poudel
STAT 5100
Spring 2019

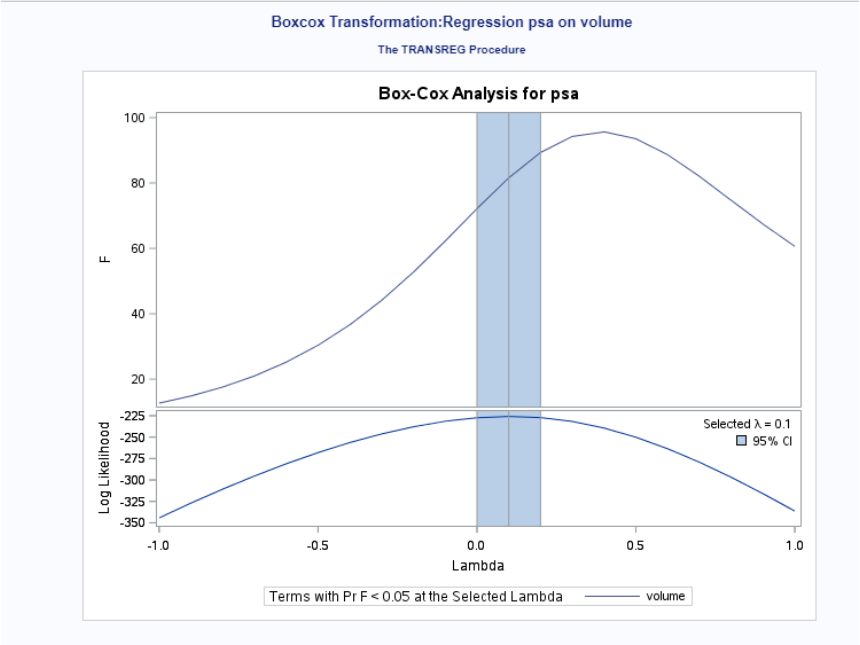
Homework 2



- The sequence plot shows that the data are not in the natural order due to which it is difficult to interpret the plot.

b) As the data seems to violate the assumptions of linear regression we need to apply the remedial measures.

For the process of selection of choosing remedial measure we use Box-cox approach which is as follow.



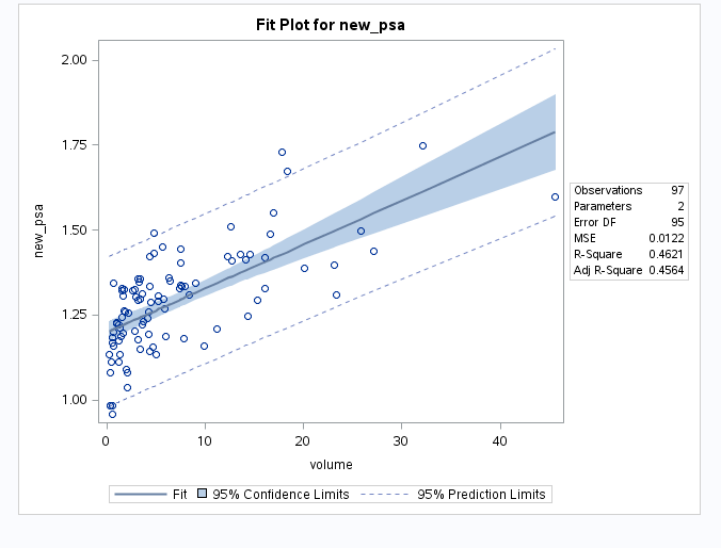
Niranjan Poudel
STAT 5100
Spring 2019

Homework 2

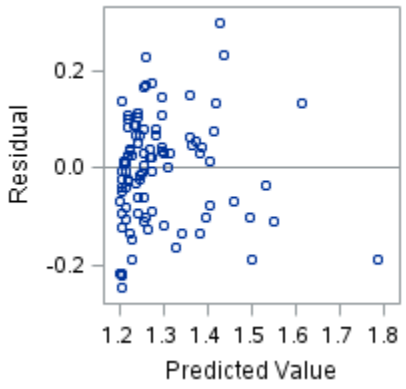
From the Boxcox transformation we can select the value of lambda as 0.1 for the transformation, we run the new regression by making the transformation.

i.e., $Y' = Y^{0.1}$

c. After applying the remedial measure, we again have the following graphical and numerical diagnostics.



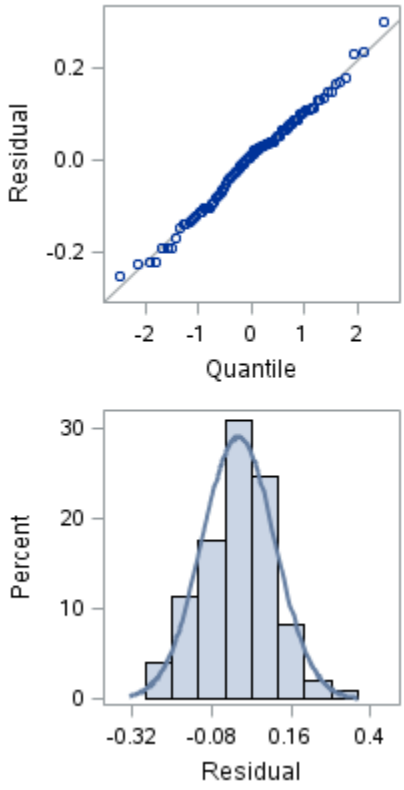
- This shows fitted plot than the previous plot.



Residual pot

Niranjan Poudel
STAT 5100
Spring 2019

Homework 2



Normal probability distribution and Histogram with normal curve.

- From this we can imply that the transformation made the data normally distributed, the residual plot shows the constant variance.

Here are the numerical diagnostics after transformation.

P-value for Brown-Forsythe test of constant variance in residual vs. Predicted Value

Obs	t_BF	BF_pvalue
1	1.01952	0.31054

Niranjan Poudel
STAT 5100
Spring 2019

Homework 2

Pearson Correlation Coefficients, N = 97 Prob > |r| under H0: Rho=0		
	resid	expectNorm
resid residual	1.00000	0.99575 <.0001
expectNorm	0.99575 <.0001	1.00000

- The numerical analysis shows the BF_pvalue as 0.31054 which is greater than 0.05 which tells there is no significance evidence to prove there is no non-constant variance.
- Also, the correlation coefficient is 0.996 which is greater than 0.986 which suggest that there is no enough evidence to confirm there is not normality.

Following is the sequence plot after the transformation.

The sequence plot shows there is no natural order of the data.

Niranjan Poudel
STAT 5100
Spring 2019

Homework 2

d. Now the new equation becomes $Y^{0.1}= 1.199+0.013X$ …… (a)
from the following table;

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.19946	0.01501	79.91	<.0001
volume	1	0.01290	0.00143	9.03	<.0001

Now if the patient has the cancer volume of 20cc putting this value of 'X' in equation (a) we get $Y^{0.1}= 1.459$ and solving for 'Y' we get Y=43.707.
Therefore, the psa level for the patient with cancer volume 20 cc is 43.70

Appendix: SAS Code

```
/* Exercise 4 */
data college;   input gpa act @@; cards;
 3.897 21 3.885 14 3.778 28 2.540 22 3.028 21 3.865 31 2.962 32 3.961
27 0.500 29 3.178 26
 3.310 24 3.538 30 3.083 24 3.013 24 3.245 33 2.963 27 3.522 25 3.013
31 2.947 25 2.118 20
 2.563 24 3.357 21 3.731 28 3.925 27 3.556 28 3.101 26 2.420 28 2.579
22 3.871 26 3.060 21
 3.927 25 2.375 16 2.929 28 3.375 26 2.857 22 3.072 24 3.381 21 3.290
30 3.549 27 3.646 26
 2.978 26 2.654 30 2.540 24 2.250 26 2.069 29 2.617 24 2.183 31 2.000
15 2.952 19 3.806 18
 2.871 27 3.352 16 3.305 27 2.952 26 3.547 24 3.691 30 3.160 21 2.194
20 3.323 30 3.936 29
 2.922 25 2.716 23 3.370 25 3.606 23 2.642 30 2.452 21 2.655 24 3.714
32 1.806 18 3.516 23
 3.039 20 2.966 23 2.482 18 2.700 18 3.920 29 2.834 20 3.222 23 3.084
26 4.000 28 3.511 34
 3.323 20 3.072 20 2.079 26 3.875 32 3.208 25 2.920 27 3.345 27 3.956
29 3.808 19 2.506 21
 3.886 24 2.183 27 3.429 25 3.024 18 3.750 29 3.833 24 3.113 27 2.875
21 2.747 19 2.311 18
 1.841 25 1.583 18 2.879 20 3.591 32 2.914 24 3.716 35 2.800 25 3.621
28 3.792 28 2.867 25
 3.419 22 3.600 30 2.394 20 2.286 20 1.486 31 3.885 20 3.800 29 3.914
28 1.860 16 2.948 28
;
run;

proc reg data=college;
       model gpa=act;
       title1 'Simple linear regression';
run;




/* Exercise 5 */
data crime;   input crimerate pctdiploma @@; cards;
   8487 74   8179 82   8362 81   8220 81   6246 87   9100 66   6561 68
5873 81   7993 74
   7932 82   6491 75   6816 82   9639 78   4595 84   5037 82   4427 79
```

```
6226 78  10768 73
   8335 77  12311 65  10104 77  10503 76   7562 79   8593 79   7133 78
10205 84  14016 78
   5959 81   3764 89   4297 85   7562 77   4844 74   5777 80   3599 84
3219 88  11187 75
   2105 77   6650 78  11371 61   4517 91   7348 83   5696 77   4995 85
9248 70   6860 88
   9776 80   4280 82  11154 82   3442 82   9674 70   7309 64   4530 79
4017 83   7122 77
   5689 76   6109 80   3343 84   5029 82   4330 81   5425 74   8769 81
6880 76   6538 78
   6521 78   9423 79   9697 83   3805 79   3134 83   3433 81   2979 84
6836 64   5804 67
   7986 75  10994 73  11322 77   8937 64   8807 75  11087 80  10355 83
7858 85   3632 91
   8040 88   6981 83   7582 76
;
run;

proc reg data=crime;
       model crimerate=pctdiploma;
       title1 'Linear regression';
run;



/* Exercise 8 */
data prostate;   input psa volume @@; cards;
   0.651  0.5599   0.852  0.3716   0.852  0.6005   0.852  0.3012
1.448  2.1170   2.160  0.3499
   2.160  2.0959   2.340  1.9937   2.858  0.4584   2.858  1.2461
3.561  1.2840   3.561  0.2592
   3.561  5.0028   3.857  4.3929   4.055  3.3535   4.263  4.6646
4.349  0.6570   4.437  9.8749
   4.759  0.5712   4.953  1.1972   5.155  3.1582   5.259  7.8460
5.474  0.5827   5.529  5.9299
   5.641  1.4770   5.871  4.2631   6.050  1.6653   6.172  0.6703
6.360  2.8292   6.619 11.1340
   6.821  1.3364   7.463  1.1972   7.463  3.5966   7.538  1.0101
7.768  0.9900   8.085  3.7062
   8.671  4.1371   8.935  1.5841   9.116 14.2963   9.777  2.2255
9.974  1.8589  10.074  4.2207
  10.278  1.7860  10.697  5.8709  12.429  4.4371  12.807  5.2593
13.066 15.3329  13.066  3.1899
  13.330  5.7546  13.330  3.3872  14.296  2.9743  14.585  5.2593
14.585  1.6653  14.732  8.4149
```

```
  14.880 23.3361  15.180  3.5609  16.281  2.6379  16.281  1.5841
16.610  1.7160  16.610  2.8864
  17.116  1.5841  17.288  7.3891  17.288 16.1190  17.814  7.6141
17.814  7.9248  17.993  4.3060
  18.541  7.5383  19.298  9.0250  19.298  0.6376  19.492  3.2871
20.287  6.4237  20.905  3.1899
  21.328  3.3535  21.758  6.2965  26.576 20.0855  28.219 23.1039
29.666  7.4633  31.187 12.6797
  31.817 14.1540  33.448 16.1190  33.784  4.3492  34.124 12.3049
35.517 13.5991  35.517 14.5851
  36.234  4.7588  37.713 27.1126  39.646  7.5383  40.854  5.6407
53.517 16.6099  54.055  4.7588
  56.261 25.7903  62.178 12.5535  80.640 16.9455 107.770 45.6042
170.716 18.3568 239.847 17.8143
 265.072 32.1367
;
run;

proc reg data=prostate;
       model psa=volume;
       output out=out1 r=resid p=pred;
       title1 'Exercise 8 linear regression';
run;

data temp; set out1;
       order = _n_;
proc sgplot data =temp;
       series x=order y=resid / lineattrs=(pattern=solid);
       xaxis label='order in data';
       yaxis label='Residual Value';
       title1 'Sequence Plot of Residuals';
run;


%macro resid_num_diag(dataset,datavar,label='requested
variable',predvar=' ',predlabel='predicted variable'); title; data
shortfourplotdataset; set &dataset;  label &datavar = &label;  if
&datavar ne .; run; proc means data=shortfourplotdataset noprint;  var
&datavar; output out=shortfourplotoutset N=nval mean=meanval; data
shortfourplotoutset; set shortfourplotoutset;  xn=nval; CALL
SYMPUT('nval',xn);   xmean=meanval; CALL SYMPUT('meanval',xmean);
%global nvalue; %let nvalue=&nval; %global meanvalue; %let
meanvalue=&meanval; run; %if &predvar ne ' '  %then %do;     data
shortfourplotdataset; set shortfourplotdataset;       label &predvar =
&predlabel;     proc sort data=shortfourplotdataset
out=shortfourplottemp;        by descending &predvar;     data
```

```
shortfourplottemp; set shortfourplottemp;       shortfourplotorder =
_n_;       shortfourplotgroup = 1-(shortfourplotorder <
ceil(&nvalue/2));     proc means data=shortfourplottemp median
noprint;       by shortfourplotgroup;       var &datavar;       output
out=shortfourplotouttemp median=medresid;     run;    data
shortfourplottempnew; merge shortfourplottemp shortfourplotouttemp; by
shortfourplotgroup;       d = abs(&datavar-medresid);     run;
       run;     proc ttest data=shortfourplottempnew plots=none;
class shortfourplotgroup;       var d;       ods output
TTests=shortfourplotBFtemp; title1 '(Ignore this nuisance output)';
       run;   run;     data shortfourplotBFtemp2; set
shortfourplotBFtemp;       if method = 'Pooled';       t_BF =
abs(tValue);       BF_pvalue = probt;       keep t_BF BF_pvalue;
proc print data=shortfourplotBFTemp2;       title1 'P-value for Brown-
Forsythe test of constant variance';       title2 'in ' &label ' vs. '
&predlabel;     run; %end; proc sort data=shortfourplotdataset
out=shortfourplottemp;     by &datavar; data shortfourplottemp; set
shortfourplottemp;   n=&nvalue;   expectNorm = probit((_n_-
.375)/(n+.25)); proc corr data=shortfourplottemp;   var &datavar
expectNorm;   title1 'Output for correlation test of normality of '
&label;   title2 '(Check text Table B.6 for threshold)'; run; title;
quit; %mend resid_num_diag;

/* Call shortcut macro */
%resid_num_diag(dataset=out1, datavar=resid, label='residual',
predvar=pred, predlabel='Predicted Value');

proc transreg data= prostate;
       model boxcox(psa / lamda =-1 to 1 by 0.1)
              = identity (volume);
       title1 'Boxcox Transformation:Regression psa on volume';
run;

data prostate; set prostate;
       log_psa = log(psa);
       new_psa = psa**0.1;
run;

/* Psa power to 0.1*/
proc reg data=prostate;
       model new_psa=volume;
       output out=out2 r=resid p=pred;
       title1 'Simple model for psa power to 0.1';
run;
data temp; set out2;
       order = _n_;
```

```
proc sgplot data =temp;
       series x=order y=resid / lineattrs=(pattern=solid);
       xaxis label='order in data';
       yaxis label='Residual Value';
       title1 'Sequence Plot of Residuals after transformation';
run;

%macro resid_num_diag(dataset,datavar,label='requested
variable',predvar=' ',predlabel='predicted variable'); title; data
shortfourplotdataset; set &dataset;  label &datavar = &label;  if
&datavar ne .; run; proc means data=shortfourplotdataset noprint;  var
&datavar; output out=shortfourplotoutset N=nval mean=meanval; data
shortfourplotoutset; set shortfourplotoutset;  xn=nval; CALL
SYMPUT('nval',xn);   xmean=meanval; CALL SYMPUT('meanval',xmean);
%global nvalue; %let nvalue=&nval; %global meanvalue; %let
meanvalue=&meanval; run; %if &predvar ne ' '  %then %do;     data
shortfourplotdataset; set shortfourplotdataset;       label &predvar =
&predlabel;     proc sort data=shortfourplotdataset
out=shortfourplottemp;        by descending &predvar;     data
shortfourplottemp; set shortfourplottemp;       shortfourplotorder =
_n_;       shortfourplotgroup = 1-(shortfourplotorder <
ceil(&nvalue/2));     proc means data=shortfourplottemp median
noprint;       by shortfourplotgroup;       var &datavar;       output
out=shortfourplotouttemp median=medresid;     run;    data
shortfourplottempnew; merge shortfourplottemp shortfourplotouttemp; by
shortfourplotgroup;       d = abs(&datavar-medresid);     run;
       run;     proc ttest data=shortfourplottempnew plots=none;
class shortfourplotgroup;       var d;       ods output
TTests=shortfourplotBFtemp; title1 '(Ignore this nuisance output)';
       run;   run;     data shortfourplotBFtemp2; set
shortfourplotBFtemp;       if method = 'Pooled';       t_BF =
abs(tValue);       BF_pvalue = probt;       keep t_BF BF_pvalue;
proc print data=shortfourplotBFTemp2;       title1 'P-value for Brown-
Forsythe test of constant variance';       title2 'in ' &label ' vs. '
&predlabel;     run; %end; proc sort data=shortfourplotdataset
out=shortfourplottemp;     by &datavar; data shortfourplottemp; set
shortfourplottemp;   n=&nvalue;   expectNorm = probit((_n_-
.375)/(n+.25)); proc corr data=shortfourplottemp;   var &datavar
expectNorm;   title1 'Output for correlation test of normality of '
&label;   title2 '(Check text Table B.6 for threshold)'; run; title;
quit; %mend resid_num_diag;
/* Call shortcut macro */
%resid_num_diag(dataset=out2, datavar=resid, label='residual',
predvar=pred, predlabel='Predicted Value');

/* Log */
```

```
proc reg data=prostate;
 model log_psa = volume;
 output out=out3 r=resid p=pred;
 title1 'Simple model for log psa data';
run;
```