Andrew Curtis

STAT 5100

Spring 2022

Homework #2

Exercise 1

If there were no clerical errors made, then the dollar sales have a functional relationship with number of units sold. Each unit has a price and thus can be related in that manner. However, once clerical errors are made, and not in a systematic fashion, the number of units sold and dollar sales no longer have a functional relationship. This new relationship can be estimated using statistics.

Exercise 2

I do not agree with the student. Their mistake was that they wrote the expected value of Yi is equal to the simple linear regression formula. It should be written as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Exercise 3

The training is certainly having a positive impact on production output over the stated range of x-values. The x's represent production before training. Using x = 40 as an example, after training it is predicted that production output would be y = 58. For x = 100, y = 115. It can be seen that, over this range it will only have a positive impact. In the problem, the observer is not considering the beta-naught intercept term.

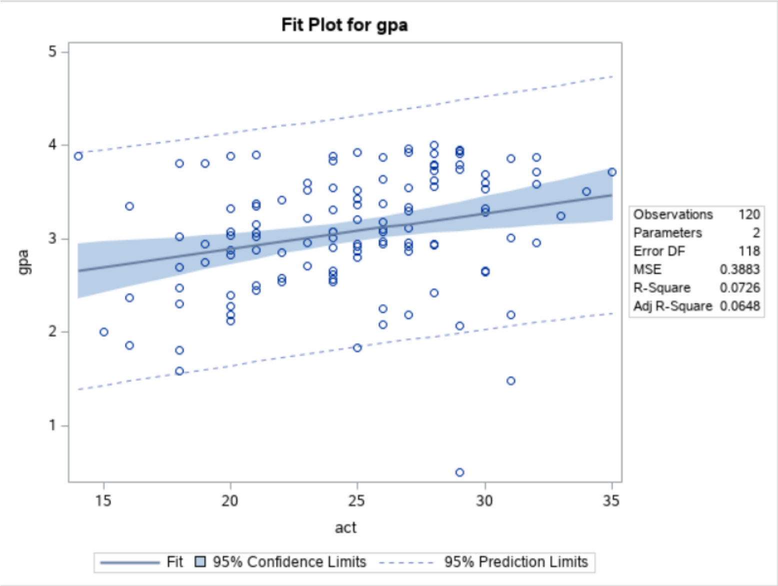
Exercise 4

a)

Via SAS, Estimates for: $\beta_0 = 2.11405$ $\beta_1 = 0.03883$

Estimated regression function: $E\{Y_i\} = 2.11405 + 0.03883X_i$

b)



Observations	120
Parameters	2
Error DF	118
MSE	0.3883
R-Square	0.0726
Adj R-Square	0.0648

The data does have a positive correlation between gpa and act scores. However, the data is quite spread and looking at the confidence intervals, most points do not fall close to our regression line. There are large residual values for the data points.

c)

$E(Y_{30}) = 3.27895$

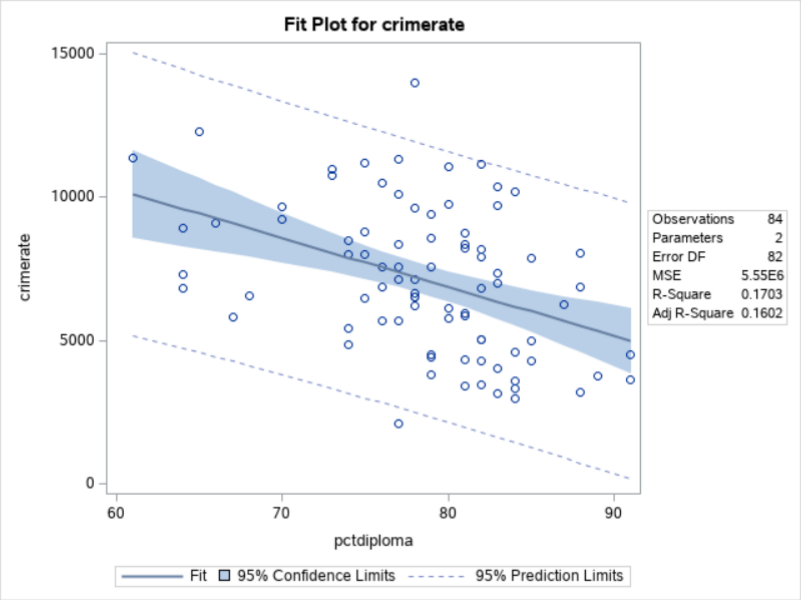
d)

The point estimate of the change in the mean for a 1-point increase in X is just the slope of the regression line, which is equal to 0.03883.

Exercise 5

a)

Estimated regression function: $E\{Y_i\} = 20518 - 170.57519X_i$



Above is the simple linear regression plot of crime rate (per 100,000 people) vs. percentage of population with at least H.S. diploma.

There does appear to be a negative correlation between crime rate and pct diploma. Although the residuals are large, and most points do not fall within the 95% confidence limits.

b)

1. The difference in mean crime rate for two countries that differ by one percentage point in percentage with diploma is simple the slope of the regression line, this is equal to about -170.58.

2. $E(Y_{80}) = 6871.985$

3. $\varepsilon_{10} = -1401.166$ $\quad 4\left(\sigma^{\wedge 2}\right) = 4(5552112) = 22208448$

Notice above I took it to mean 4 times the mean square error since it was talking about the error for that data point. Didn’t think it was referring to the variance.

Exercise 6

Starting with $\sum Y_i = nb_0 + b_1 \sum X_i$

solve for b_0, $b_0 = \frac{1}{n}(\sum Y_i - b_1 \sum X_i) = Y^{bar} - b_1 X^{bar}$

Plug previous result into, $\sum X_i Y_i = b_0 \sum X_i + b_1 \sum X_i^2$

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

Multiply through $\sum X_i$ term and solve for b_1,

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

Next multiply the top and bottom of the equation by 1/n,

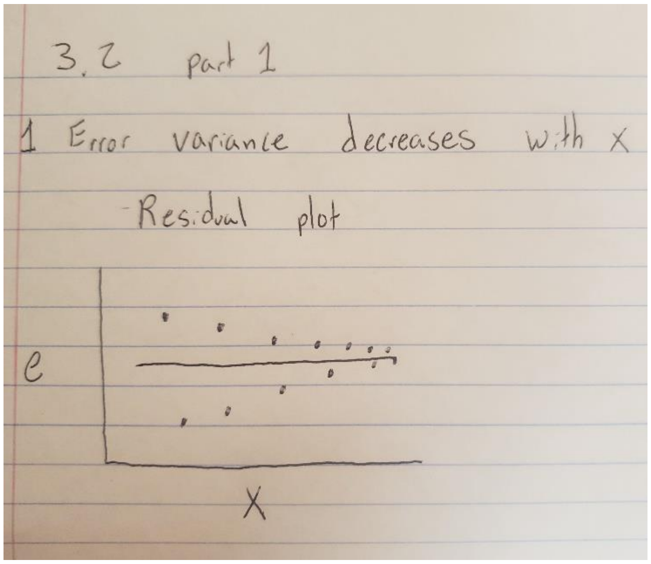
$$b_1 = \frac{(\sum X_i Y_i - Y^{bar} \sum X_i)/n}{(\sum X_i^2 - X^{bar} \sum X_i)/n}$$

Using the fact that $Y^{bar} = \frac{\sum Y_i}{n}$ and similarly for $X^{bar} = \frac{\sum X_i}{n}$, substitute in and simplify to get,

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

Exercise 7

1.



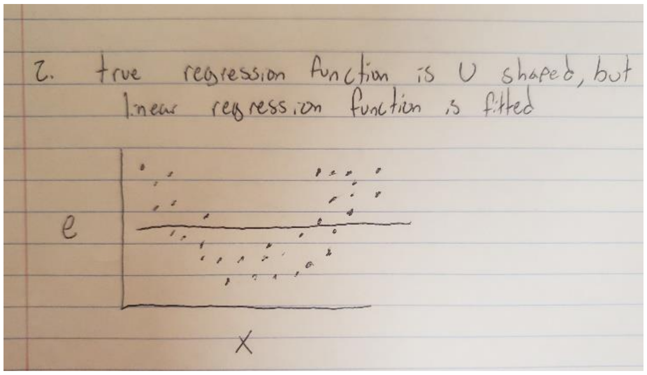
3.2 part 1

1 Error variance decreases with x

-Residual plot

The above graphic shows a residual plot with error variance decreasing with x.

2.



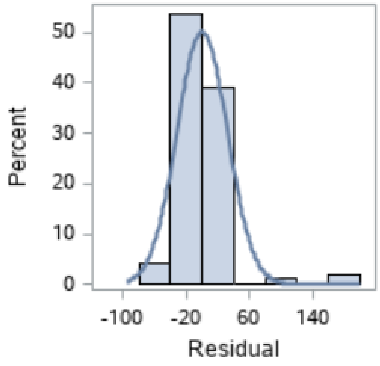
2. true regression function is U shaped, but linear regression function is fitted

The above graphic shows a residual plot where the true regression function should be U – shaped but instead a linear regression function is fitted.

Exercise 8

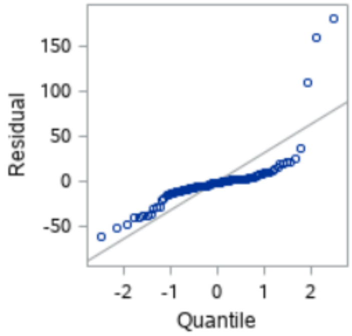
a)

i.

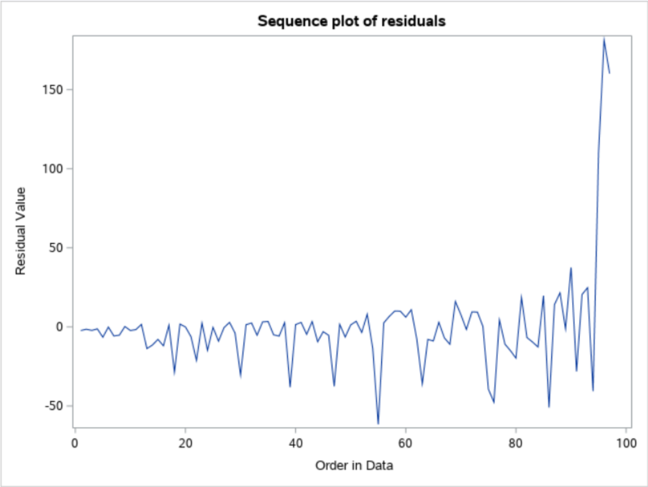


Histogram of residuals

ii.

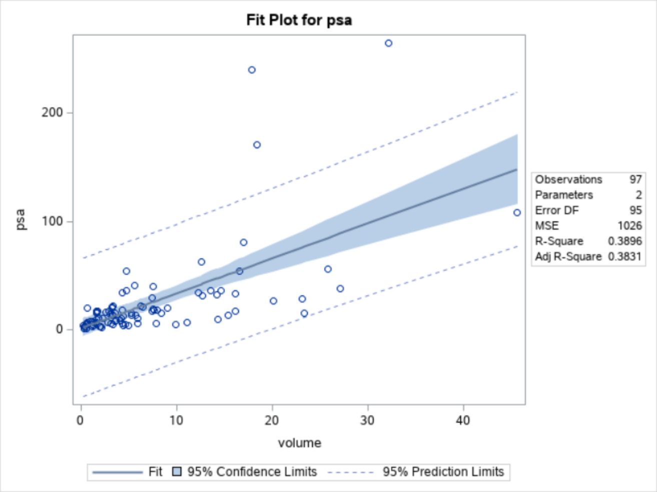


Normal probability plot



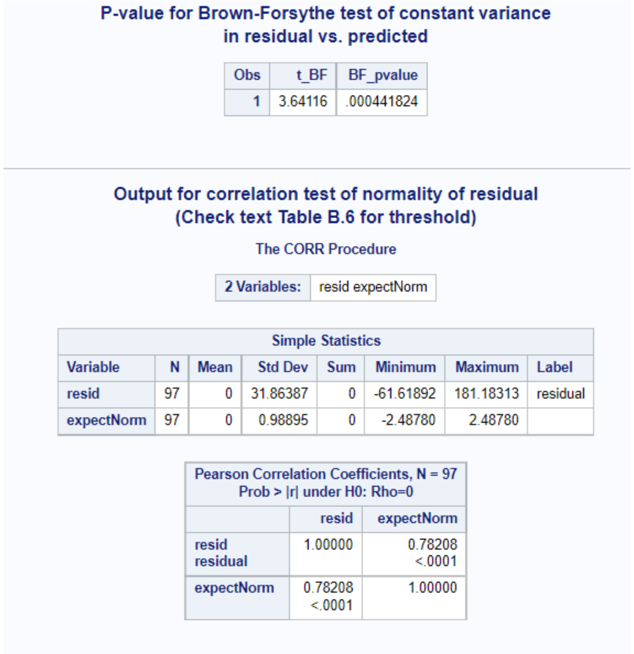
Sequence plot

iii.



Plot vs. predicted values

iv.



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

Obs	t_BF	BF_pvalue
1	3.64116	.000441824

Output for correlation test of normality of residual (Check text Table B.6 for threshold)

The CORR Procedure

2 Variables:	resid expectNorm

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
resid	97	0	31.86387	0	-61.61892	181.18313	residual
expectNorm	97	0	0.98895	0	-2.48780	2.48780	

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

Above is results for Brown-Forsythe test and correlation test for normality v.

For part i, the histogram of residuals does not look very normal.

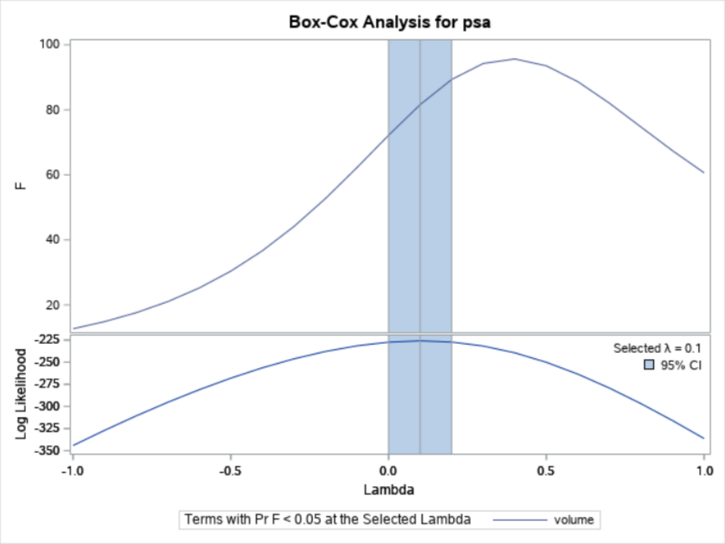
For ii, the normal probability plot of residuals the observed residual values differ significantly from what is expected and look to have a long tail with some skew. The sequence plot may have some consistent behavior.

For iii, the data points appear to have increasing variance as the volume increases with some outliers.

For iv, the Brown-Forscythe test shows a small p-value indicating that we should reject null hypothesis that variance of errors is consistent. For the correlation test of normality, the expect norm value is much less than about .986 as listed in table B.6 in the book. Thus, null hypothesis that errors are normal should be rejected.

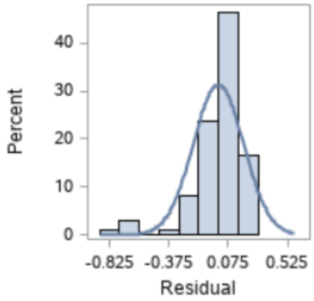
b)

Choosing the Box-Cox transformation and regressing psa on cancer volume, also choosing negative inverse square root.



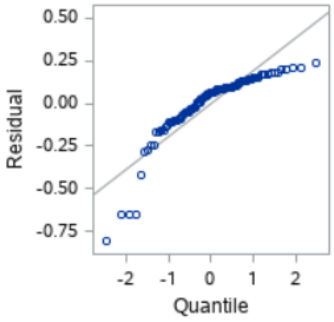
c)

i.

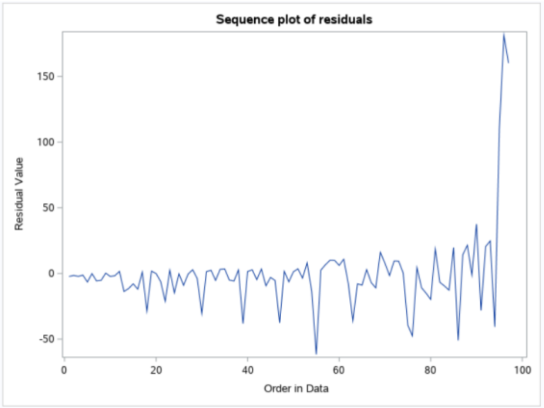


Histogram of residuals

ii.

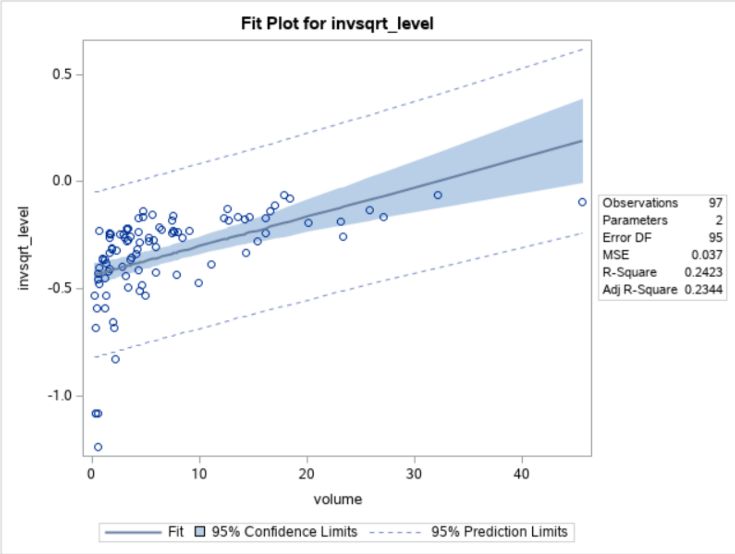


Normal probability plot



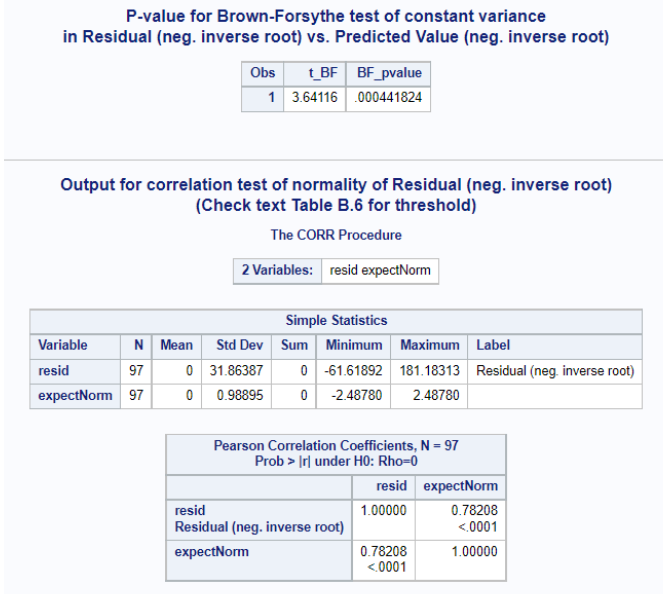
Sequence plot

iii.



Plot vs. predicted values

iv.



P-value for Brown-Forsythe test of constant variance
in Residual (neg. inverse root) vs. Predicted Value (neg. inverse root)

Obs	t_BF	BF_pvalue
1	3.64116	.000441824

Output for correlation test of normality of Residual (neg. inverse root)
(Check text Table B.6 for threshold)

The CORR Procedure

2 Variables:	resid expectNorm

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
resid	97	0	31.86387	0	-61.61892	181.18313	Residual (neg. inverse root)
expectNorm	97	0	0.98895	0	-2.48780	2.48780	

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

	resid	expectNorm
resid Residual (neg. inverse root)	1.00000	0.78208 <.0001
expectNorm	0.78208 <.0001	1.00000

Above is results for Brown-Forsythe test and correlation test for normality v.

For part i, the histogram of the residuals appears more normal than before but does appear to have a left skew.

For ii, the normal probability plot of residuals the observed residual values differ less than before, but results are consistent with a left skew as well. Again, the sequence plot may have some consistent behavior.

For iii, the data points appear to be much tighter to our confidence intervals than before, but there are still many outside and some outliers.

For iv, the Brown-Forscythe test shows a small p-value indicating that we should reject null hypothesis that variance of errors is consistent even with the remedial method being applied. For the correlation test of normality, the expect norm value is much less than about .986 as listed in table B.6 in the book. Thus, null hypothesis that errors are normal should be rejected. Of note, the values did not improve after the remedial method.

d)

Transformed remedial method SLR equation E$\{Y^i\}$ = -0.43817 + 0.01374X_i

For a patient with 20cc volume, transformed $Y_{20} = -0.16337$

Leading to an original predicted value of 2.474081 for psa.

Appendix SAS code

I will attach this additionally as a pdf. Seems easier to look at.