STAT5120—Week 3 Homework, Allen Baumgarten

1. Open the GPA/ACT data from CH03PR03.txt. This data set is the dataset from Chapter 1 on GPA vs. ACT, but includes observations on two additional variables, namely intelligence test scores (third column) and high school class rank percentile (fourth column). We want to know which of the three explanatory variables (ACT, intelligence test score, high school class rank percentile) can best be used to make a linear model for predicting GPA. So you will build and compare three simple linear regression models for

1. GPA vs. ACT
2. GPA vs. intelligence test score
3. GPA vs. class rank percentile

So for each of these three cases, do the following:

(a) obtain the linear model (lm(y ~ x)) and output (summary()) and compare the R^2 values.

For Model 1, GPA vs. ACT Scores:

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.11405    0.32089   6.588  1.3e-09 ***
ch03pr03[, 2]  0.03883    0.01277   3.040  0.00292 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6231 on 118 degrees of freedom
Multiple R-squared:  0.07262,	Adjusted R-squared:  0.06476
F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917
```

For Model 2, GPA vs. Intelligence Scores:

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.873921   0.345709  -5.421  3.2e-07 ***
ch03pr03[, 3]  0.041944   0.002915  14.389  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3899 on 118 degrees of freedom
Multiple R-squared:  0.637,	Adjusted R-squared:  0.6339
F-statistic:   207 on 1 and 118 DF,  p-value: < 2.2e-16
```

For Model 3, GPA vs. Class Rank Percentile:

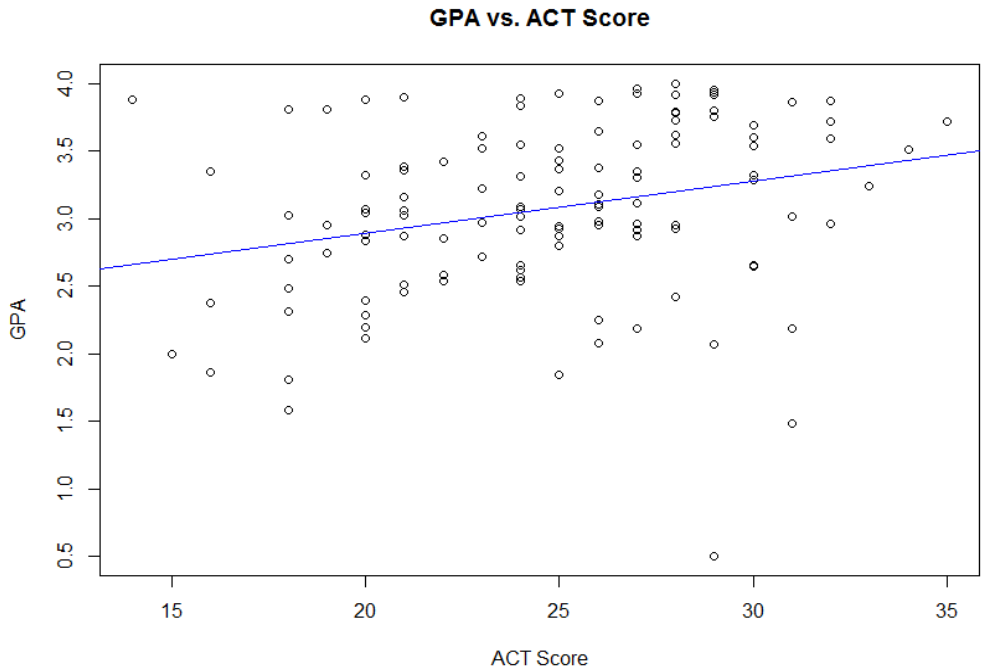
```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.306901   0.185497  12.436  < 2e-16 ***
ch03pr03[, 4] 0.010417   0.002406   4.329 3.15e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6011 on 118 degrees of freedom
Multiple R-squared:  0.1371,	Adjusted R-squared:  0.1298
F-statistic: 18.74 on 1 and 118 DF,  p-value: 3.153e-05
```

Comparing the R^2 statistics, we find that Model 2 best accounts for the variation in y at ~64%:

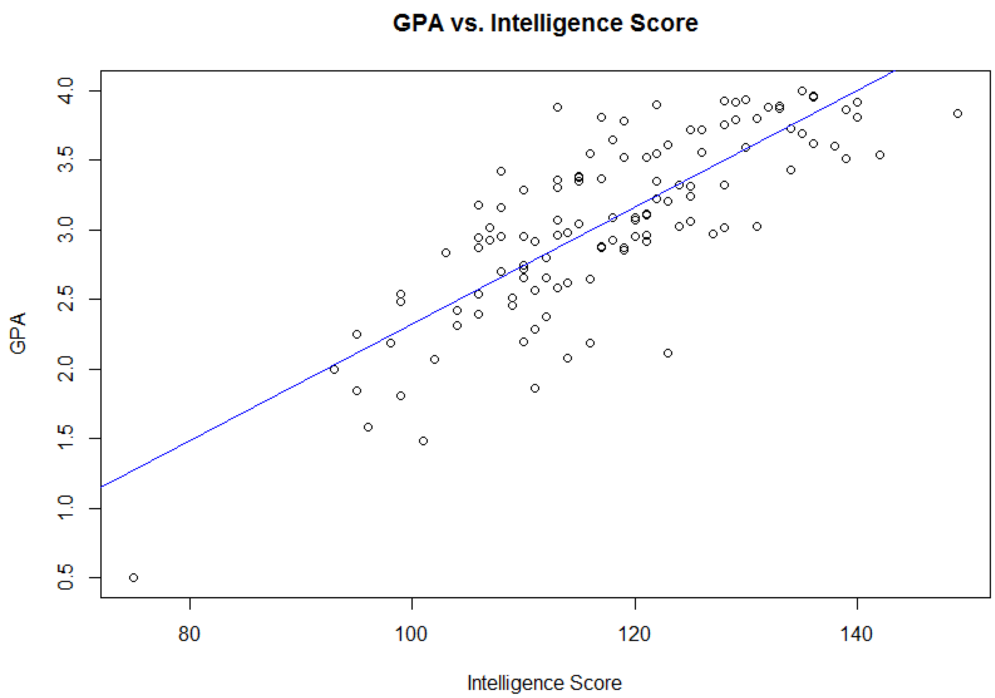
Model 1 R^2: 0.07262
Model 2 R^2: 0.637
Model 3 R^2: 0.1371

(b) make scatter plots that include the regression lines. Identify any potential outliers and influential observations. Decide whether or not to remove them before moving on.

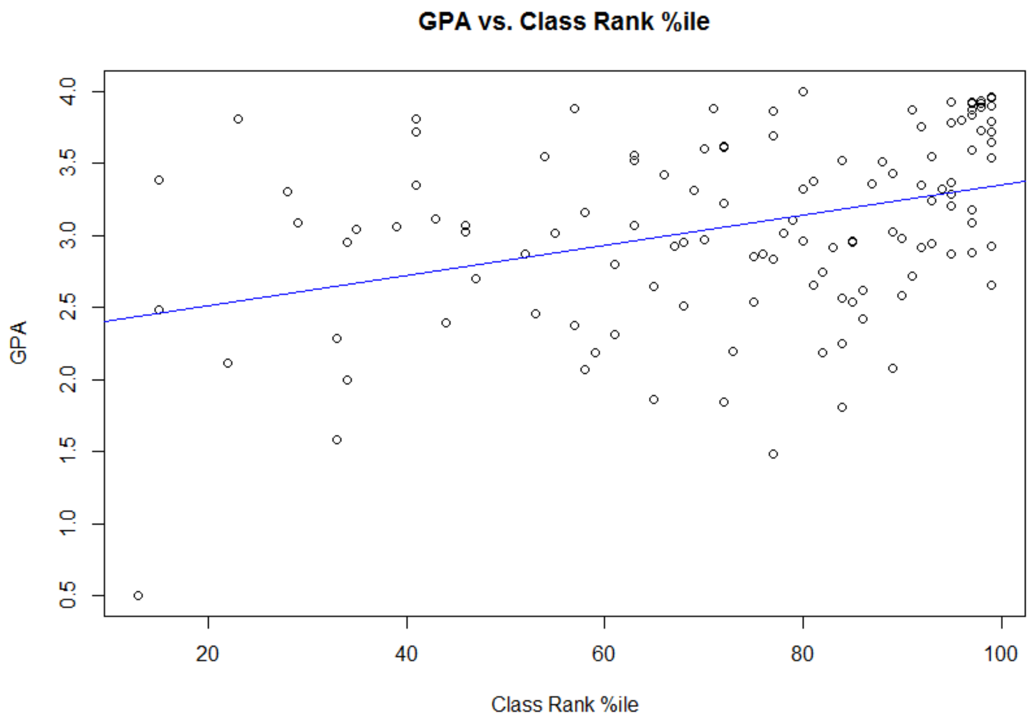
Scatter plot for Model 1:



Scatter plot for Model 2:



Scatter plot for Model 3:



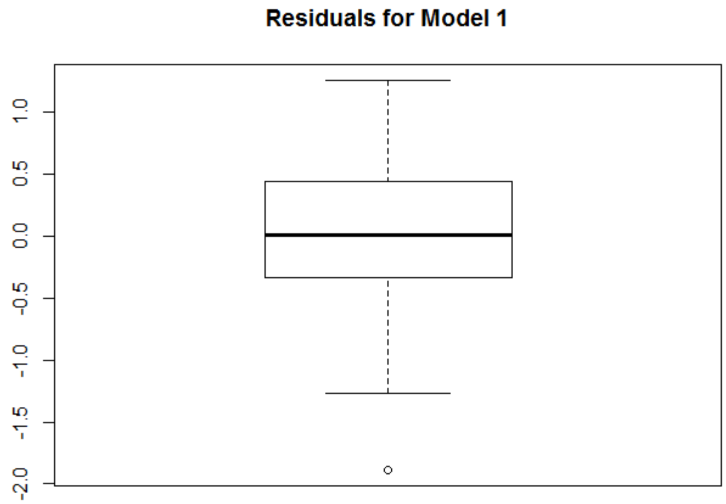
At least one outlier point was observed, viz., in Model 1 with an ACT Score of only 27 or 28 and a correspondingly low GPA score. There was also an outlier in Model 2 with a very low Intelligence score at around 75. Finally, there is a very low Class Rank score in Model 3 at around 10 to 15. Investigation of the dataset reveals that these three observed outliers are actually the same point (row 9) in the dataset. Concluding, I would elect to remove just one point, viz., the one with the low ACT score as per Model 1. This will mean in effect removing only one point from the dataset, leaving us with n = 119.

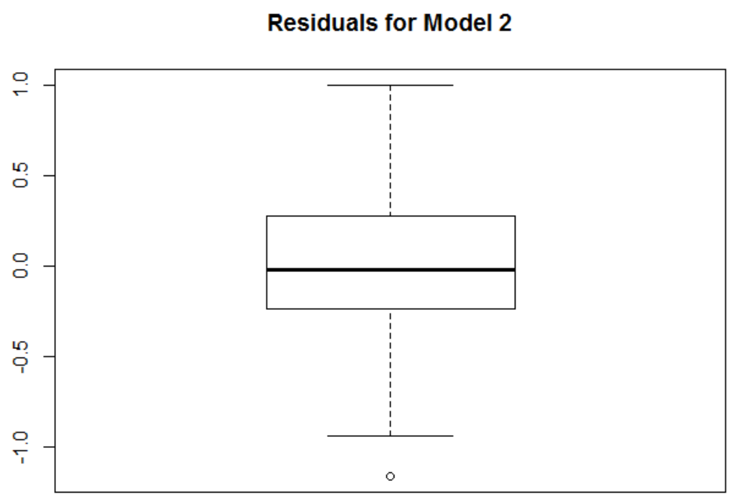
** R code was scripted to remove row 9 from the dataset based on instructions given in subquestion b above. Thus, subquestions c through h will be answered based on this new reduced dataset at n = 119 **

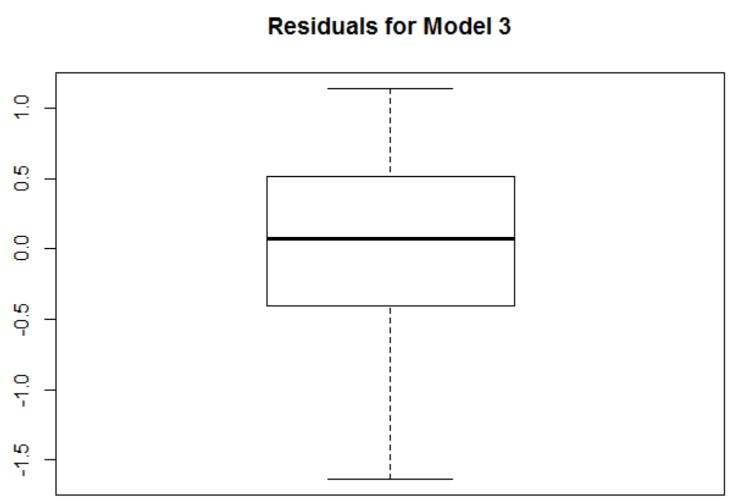
(c) check for normality of the residuals with Shapiro-Wilk tests (shapiro.test()). Based on reduced dataset, Shapiro-Wilks tests were run on the residuals of these three reduced models:

Test for Model 1 (GPA vs. ACT): W = 0.98454, p-value = 0.191
Test for Model 2 (GPA vs. Intell): W = 0.98973, p-value = 0.5172
Test for Model 3 (GPA vs. Class Rank): W = 0.97492, p-value = 0.02517

(d) make boxplots and histograms of the residuals to help check for normality. Boxplots constructed:

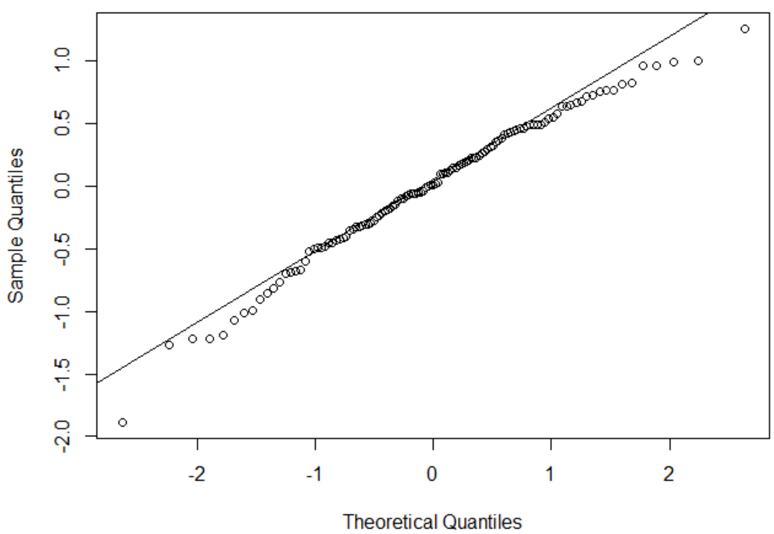




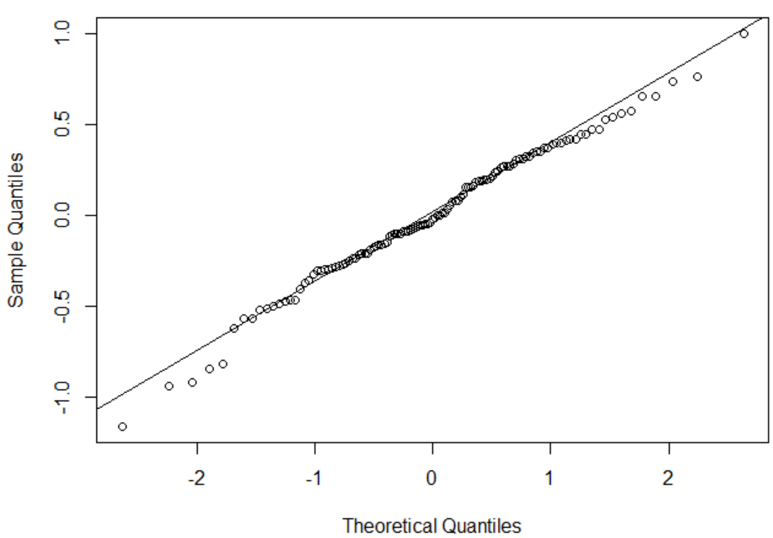


(e) make normal probability plots for the residuals (to check for normality). Normlity plots constructed:

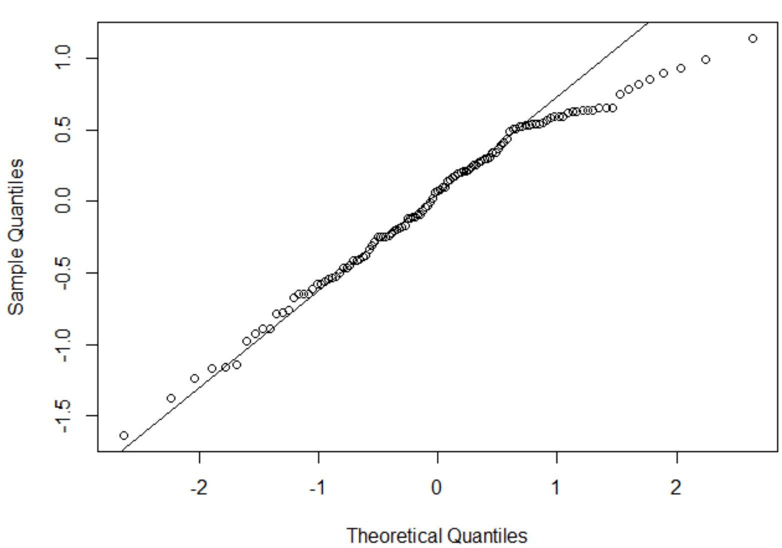
Residuals for Model 1



Residuals for Model 2



Residuals for Model 3



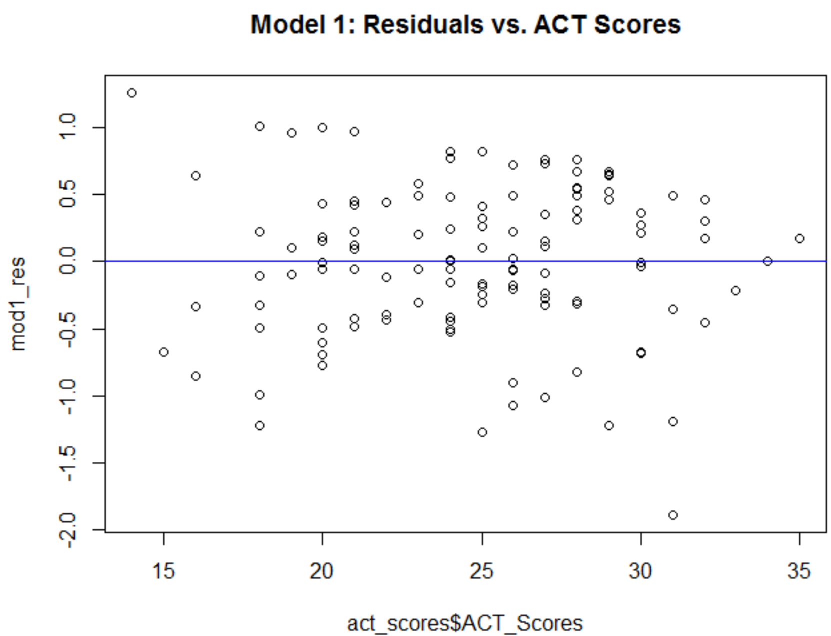
(f) Split the data set into two groups: students with ACT scores less than 26 and students with ACT scores at least 26. Run Levene's test for equality of variance of the residuals (leveneTest() requires the car package) on the response for these two groups. Remark on what you observe. The command subset() can be useful for splitting data sets. Cut data into two groups based on ACT Scores and then ran Levene’s test. A high p-value was observed, indicating we cannot (should not) reject the H_0 that equal variance is present:

```
Levene's Test for Homogeneity of Variance (center = median)
        Df  F value   Pr(>F)
group   1   0.1376    0.7114
```

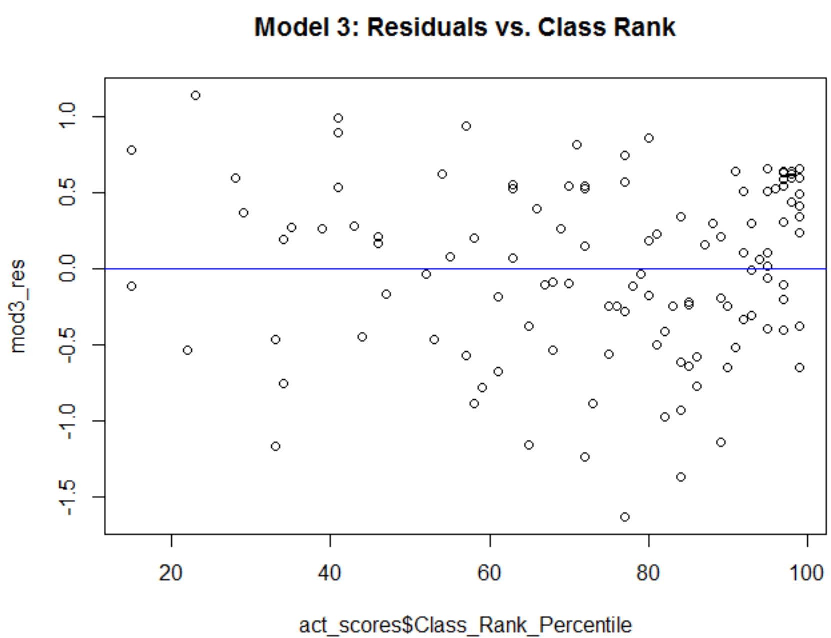
(g) Repeat part (f) for the intelligence test score model, splitting at < 120 and 120. Do the residual variances for the two groups appear to differ? Partitioned data for Model 2 into two groups, <120 and 120>=. Performed a Levene’s test and the p-value was significant as shown below. The H_0 would be that residual variances are equal and the H_A is that we would reject H_0. The significant p-value leads me to reject H_0 in favor of HA:

```
Levene's Test for Homogeneity of Variance (center = median)
        Df  F value   Pr(>F)
group   1   4.1622    0.04359 *
      117
```

(g) plot the residuals vs. the predictor variable values to check for equality of variance. Constructed three scatter plots[1] to check for equality of variances in these three reduced models:



[1] All of which resembled probabilities of the Miami Dolphins’ wins/losses these past few decades (sorry).



(h) Remark on which of the three explanatory variables seems to be the most useful in building a linear model for predicting first-year GPA. Residual plots for each of these models raise some concerns, prompting the need for either data transformations and/or perhaps curvilinear modeling. The residual plot for Model 1 shows the residuals appeargin "trend" down from the upper left quadrant to the lower right. The residual plot for Model 2 seems to show the residuals curving up from the left bottom, into the center, and then back down into the bottom right area of the graph. Residuals in Model 3 appear to have some tight clustering in the top right quadrant. I would recommend transforms, the inclusion of other variables, or non-linear models.

2. Adapted from ALSM 3.15. A chemist wanted to model the evolution of a solution concentration over time. To do this, she randomly assigned three solutions to measure after one hour, three solutions to measure after three hours, three to measure after five hours, three to measure after seven hours, and three to measure after nine hours. The data are in CH03PR15.

(a) Find the equation of the least-squares regression line. $\hat{y} = 2.5753 + (-0.3240x)$

(b) Run the F-test for lack of linear fit. Use the significance level $\alpha = 0.025$. State your p-value and provide your conclusion. Utilized the anova() functionality in R on each of these 3 models. Low p-values would seem to indicate that we cannot reject H_0 hypothesis that there is a linear fit for each of the three models. I would have to investigate, however, as to why findings in my residual plots seem to show some concerns as stated earlier.

Model 1:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
act_scores$ACT_Scores	1	4.536	4.5364	14.2290	0.0002771 ***
Residuals	117	38.188	0.3264		
Lack of fit	19	6.944	0.3655	1.1463	0.3195527
Pure Error	98	31.244	0.3188		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Model 2:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
act_scores$Intelligence_Score	1	25.4590	25.4590	195.4448	<2e-16 ***
Residuals	117	17.2651	0.1476		
Lack of fit	43	7.6257	0.1773	1.3614	0.121
Pure Error	74	9.6394	0.1303		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Model 3:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
act_scores$Class_Rank_Percentile	1	4.135	4.1353	14.0469	0.0003938 ***
Residuals	117	38.589	0.3298		
Lack of fit	55	20.336	0.3697	1.2559	0.1914652
Pure Error	62	18.253	0.2944		

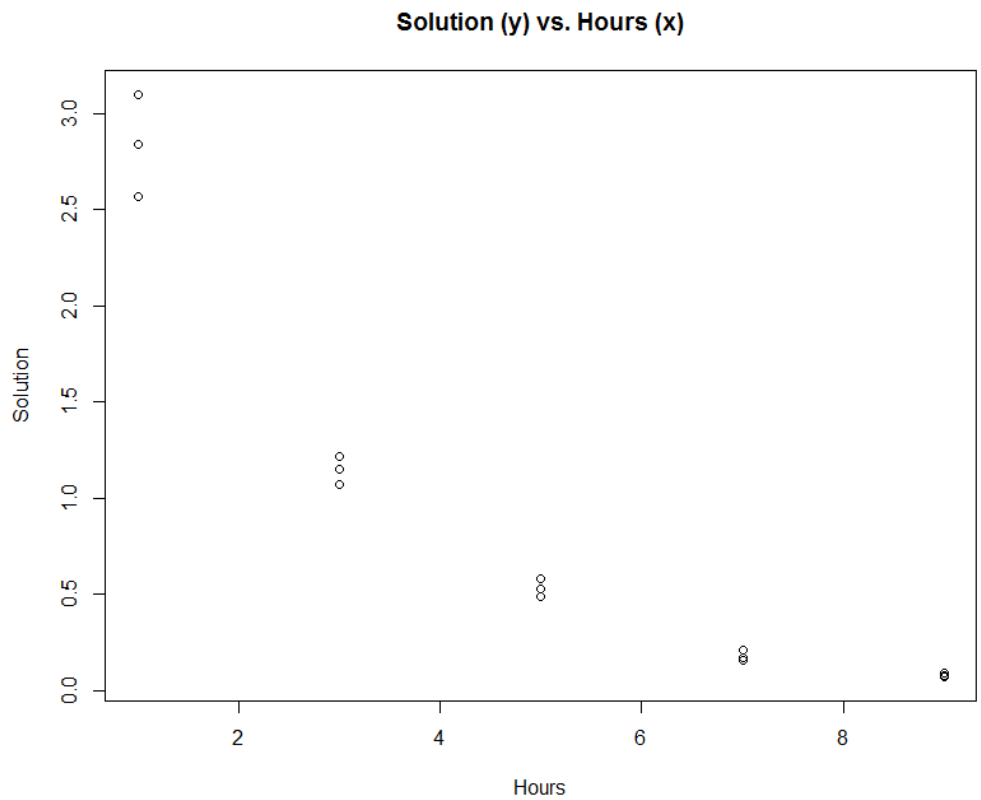
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(c) When a lack-of-linear fit test indicates there is a lack of linear fit, does it suggest exactly what kind of function would be appropriate? Explain. The F-test does not specify which function would be appropriate per se, only that a linear one is not. It can, however, be used to explore other possibilities. Our text states, "The general linear test approach just explained can be used to test the appropriateness of other functions. Only the degrees of freedom for SSLF will need to be modified."[2]

3. Adapted from ALSM 3.16. Use the data from the previous problem (CH03PR15).

(a) Make a scatterplot of the data, with concentration as the response variable. Based on the scatterplot, what kind of data transformation do you suggest to adjust for the non-constant variance and/or non-linearity? Scatter plot constructed:

[2] Kutner, Michael H., Christopher J. Nachtsheim, John Neter, and William Li, *Applied Linear Statistical Models*, 5th ed., (McGraw Hill Education (India) Edition, 2013), 127.



(b) Use the log-likelihood method in R we used for selecting the Box-Cox, and apply the transformation to the data. Then build a new regression model (using lm()) for the transformed data, make a fitted-line plot, and discuss how the new relationship compares with the old one.

Ran Box-Cox transformation (see R script in appendix [at the end of homework, not my intestinal track]) and fitted a new regression equation:

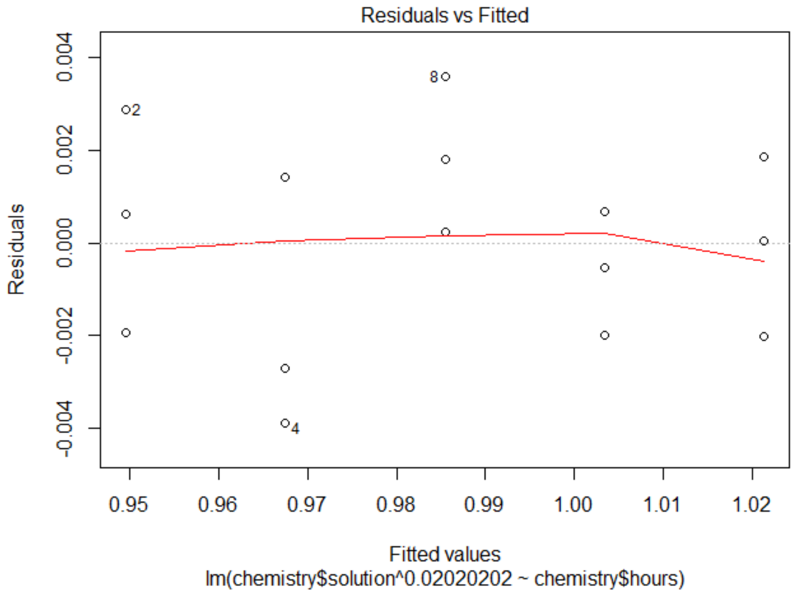
Coefficients:

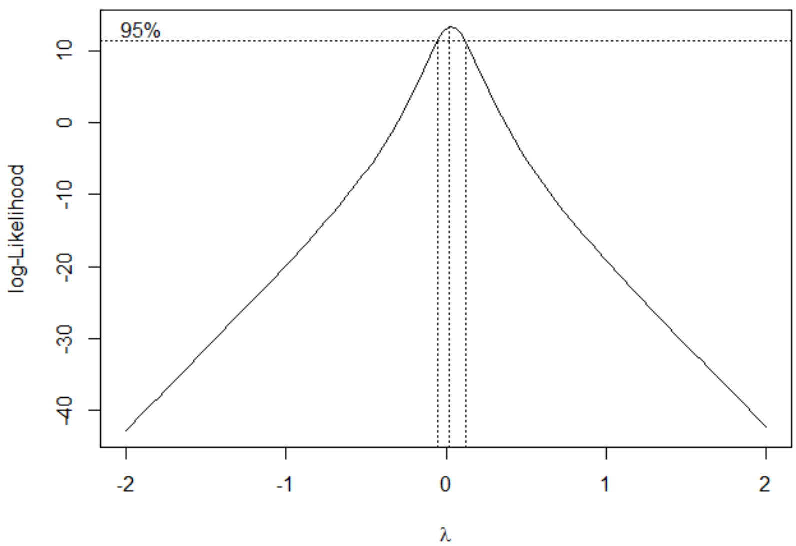
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.0302214	0.0011764	875.73	< 2e-16 ***
chemistry$hours	-0.0089532	0.0002048	-43.72	1.7e-15 ***

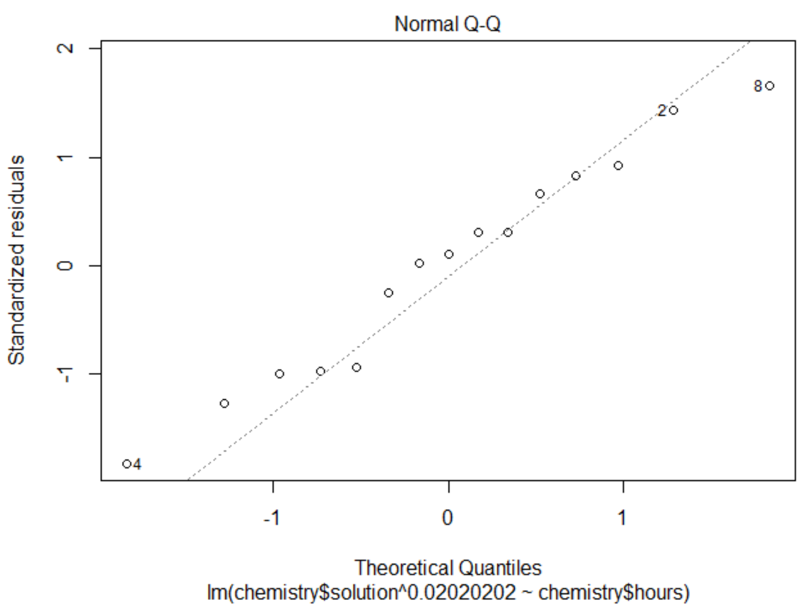
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002243 on 13 degrees of freedom
Multiple R-squared: **0.9932**, Adjusted R-squared: 0.9927
F-statistic: 1911 on 1 and 13 DF, p-value: 1.701e-15

We see that this new transformed model takes on a very high R^2 value of .99, making this a better-fitting model than with our non-transformed values earlier. A QQ plot below also shows a much better fit for the standardized residuals.


Residuals vs Fitted
Residuals
0.004
0.002
0.000
-0.002
-0.004
0.95
0.96
0.97
0.98
0.99
1.00
1.01
1.02
Fitted values
lm(chemistry$solution^0.02020202 ~ chemistry$hours)


95%
log-Likelihood
10
0
-10
-20
-30
-40
-2
-1
0
1
2
λ


Normal Q-Q
Standardized residuals
2
1
0
-1
-1
0
1
Theoretical Quantiles
lm(chemistry$solution^0.02020202 ~ chemistry$hours)

(c) Apply the log10 transformation and get the new regression line equation. Plot this model on a scatterplot of the transformed data. Compare the results of the Box-Cox transformation with those of the log10 transformation. Which do you like better and why?

Regressed on chemistry data transformed to log10:

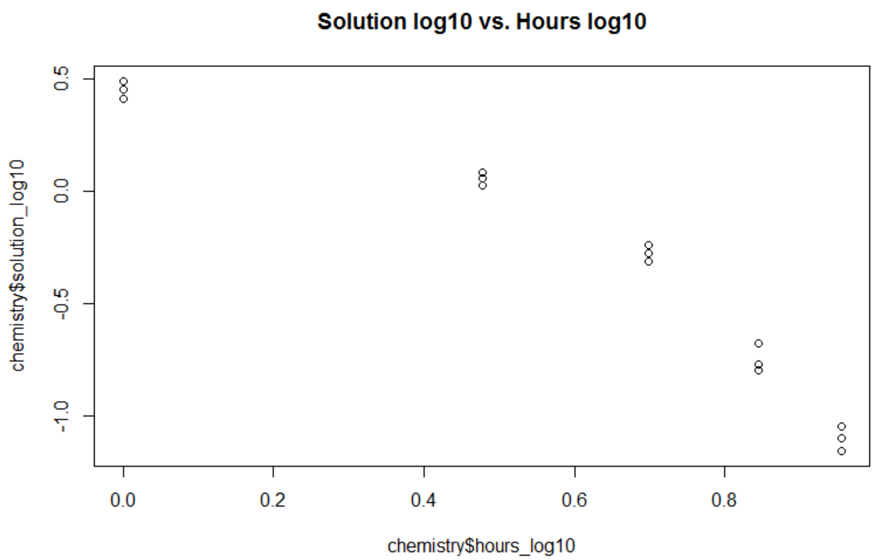
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6022	0.1012	5.95	4.82e-05 ***
chemistry$hours_log10	-1.5532	0.1479	-10.50	1.02e-07 ***

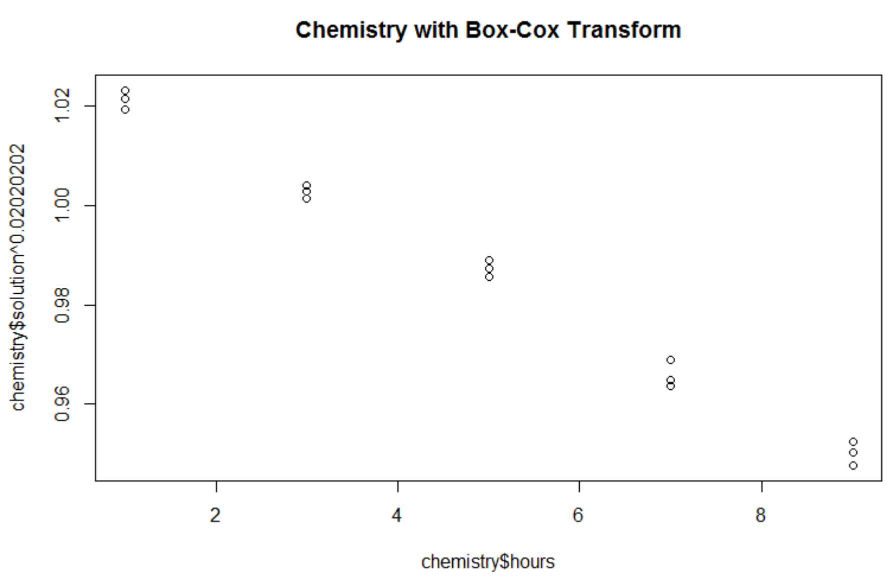
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1935 on 13 degrees of freedom
Multiple R-squared: **0.8945**, Adjusted R-squared: 0.8864
F-statistic: 110.3 on 1 and 13 DF, p-value: 1.017e-07

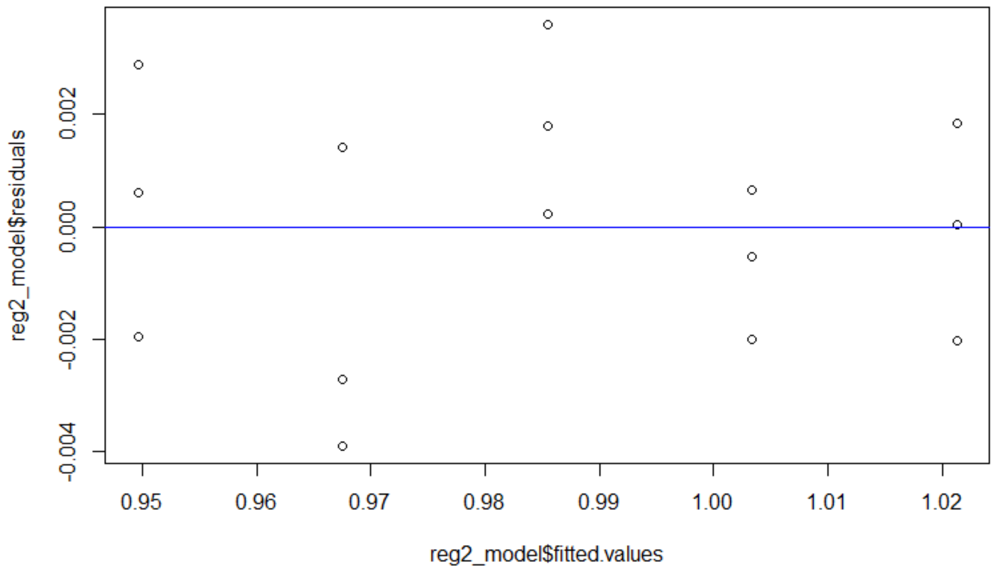
The Box-Cox transform results in a higher R^2 value compared to our log10 transform. However, a scatter plot of the log10 transformed values shown below reveals that there remains a curvature to the data:

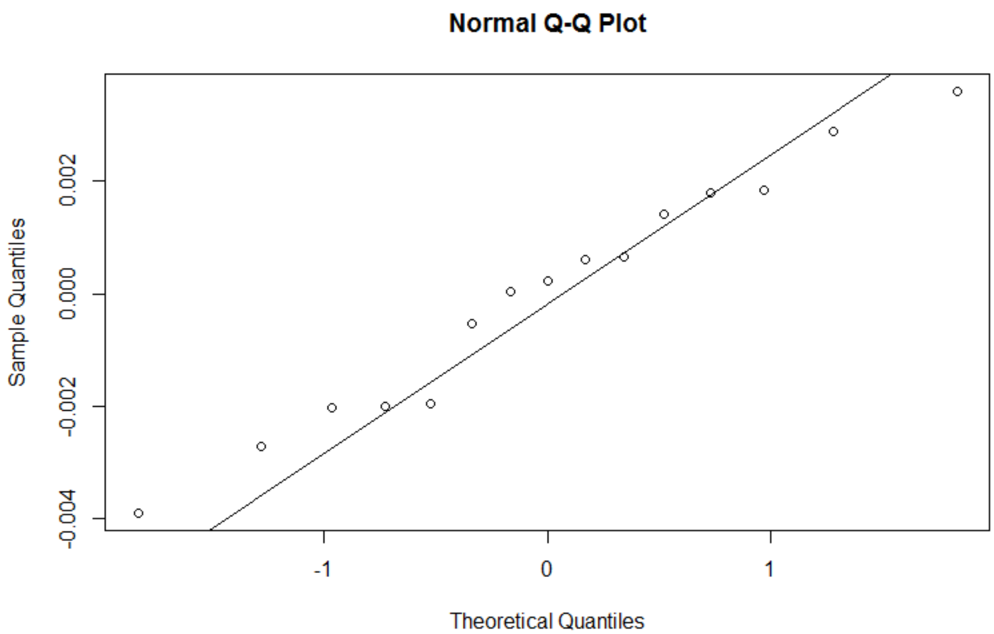


A scatter plot of the Box-Cox transformation shows a good linear fit:



(d) Plot the residuals against the fitted values. Also make a normal probability plot. What do these plots indicate? These plots indicate that the residuals are fitted well, suggesting the Box-Cox transformation model fits the data well:

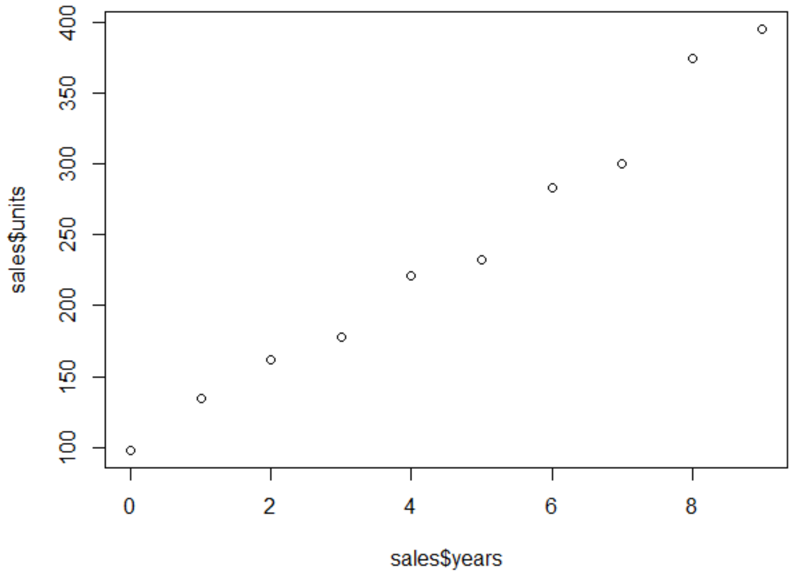




(e) Express your estimated regression functions in the original units. I am assuming that I am supposed to be converting back the transformed betas of the model I thought was the better of the two. In this case, I chose the Box-Cox model. Therefore, our y-intercept of would be expressed as 1.0302214^(1/.0202) = 4.366436 and our slope would be expressed as 0.0089532^(1/.0202) = -4.100024e-102

4. Adapted from ALSM 3.17. A marketing manager studied annual product sales figures over a ten year period. The data (years and sales in thousands of units) are in the file CH03PR17.

(a) Make a scatterplot. Is the linearity assumption reasonable? I am no Sir Ronald Fisher but I would tentatively conclude based on the scatter plot blow that there a linearity assumption between units (in thousands) and years is reasonable:



(b) Apply the maximum likelihood Box-Cox method (like we did in the Trees example) to get an appropriate power transformation of the response (sales). What is the value of SSE in this case? A regression model was generated on the Box-Cox transformed data, transformed at ^.050505:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	10.50006	0.22092	47.53	4.25e-11	***
boxcox_x	1.11723	0.04138	27.00	3.81e-09	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3759 on 8 degrees of freedom
Multiple R-squared: **0.9891**, Adjusted R-squared: 0.9878
F-statistic: 728.9 on 1 and 8 DF, p-value: 3.815e-09

Anova results showing the SSE:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
boxcox_x	1	102.98	102.977	728.9	3.815e-09	***
Residuals	8	1.13	0.141			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(c) Try using the square-root transformation and get a new regression line.

Coefficients:

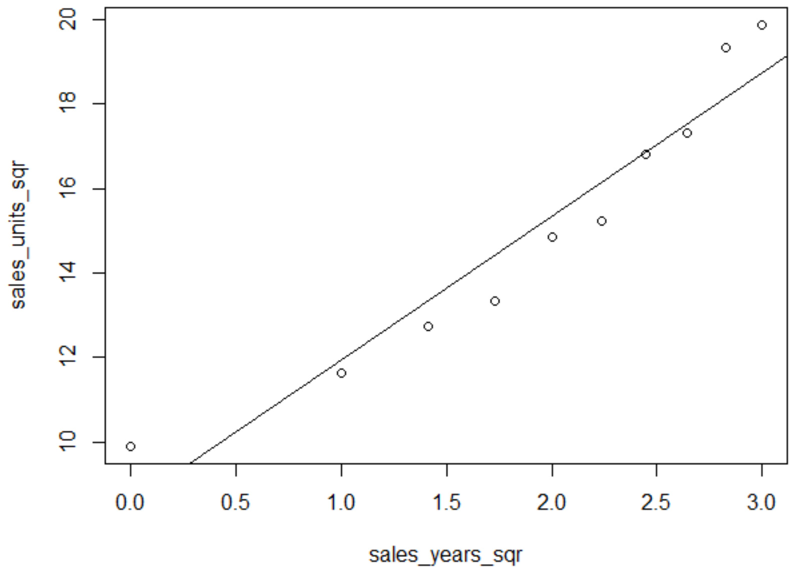
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	8.5410	0.7293	11.711	2.58e-06	***
sales_years_sqr	3.3996	0.3438	9.888	9.23e-06	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

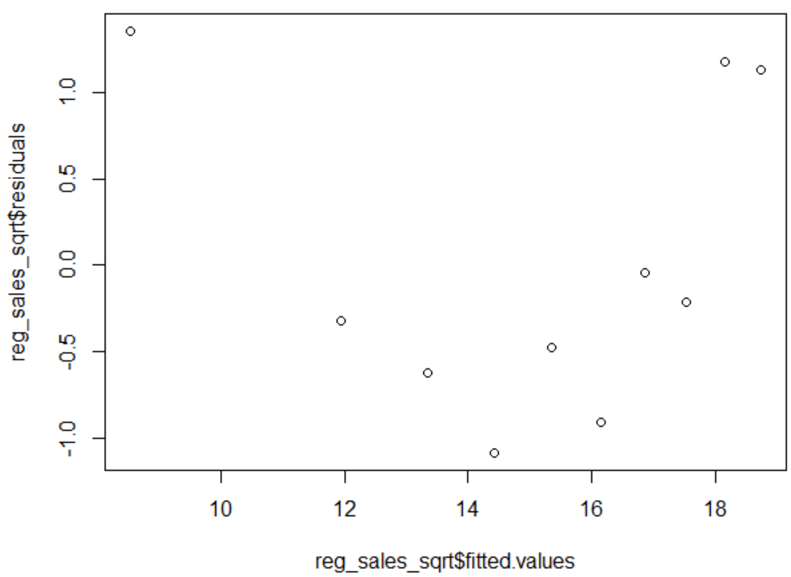
Residual standard error: 0.9557 on 8 degrees of freedom
Multiple R-squared: **0.9244**, Adjusted R-squared: 0.9149
F-statistic: 97.78 on 1 and 8 DF, p-value: 9.229e-06

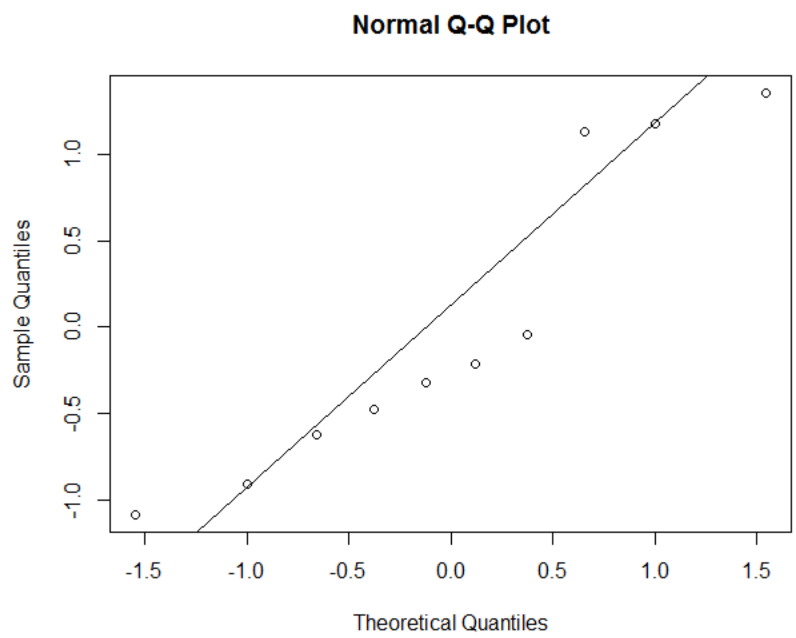
(d) Plot the regression line from the previous part on a scatterplot of the transformed data. Does this line seem to fit the transformed data well?

I would argue that the regression line does NOT fit that well: the line sits above the inside 7 points and below the outside 1 and 2, respectively. There seems to be a slight curvature to this data that a line does not fit:



(e) Make a plot of the residuals vs. fits. Also make a normal probability plot. What do these plots indicate for your transformed data? These plots both confirm my suspicions about a curvature in the data points. (I love it when I'm right, which doesn't happen often):

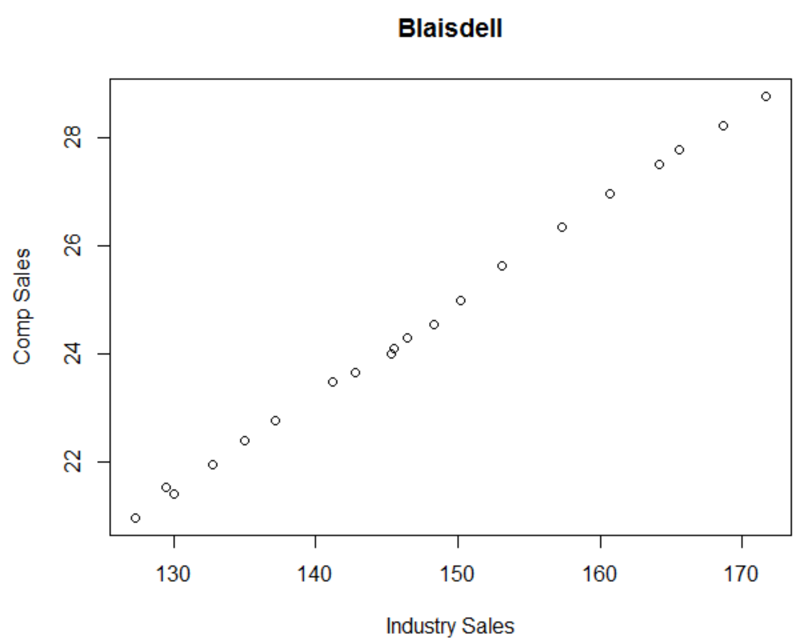




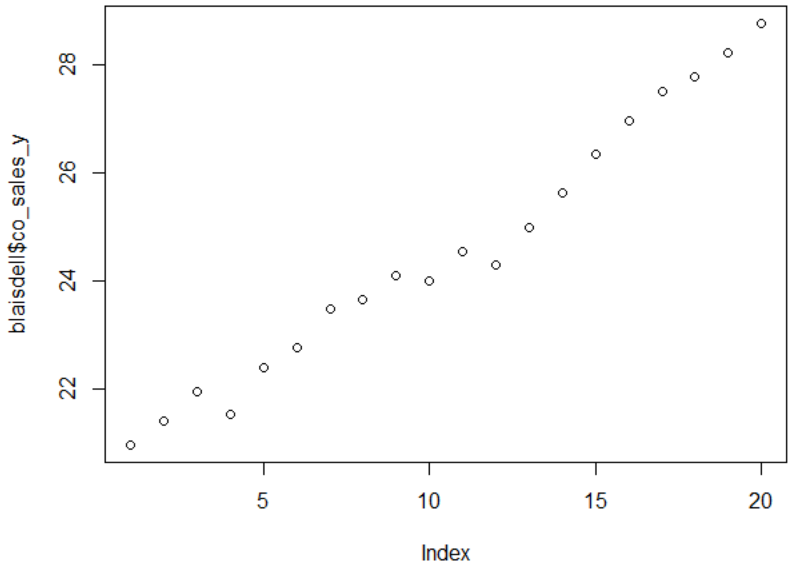
(f) Express the regression models in the original units. The y-intercept is 8.5410^2 = 72.94868 and the slope is 3.3996^2 = 11.55728

5. The Blaisdell Company wanted to use industry sales to predict its sales. Adjusted quarterly sales data for 1998 – 2002 are in the ALSM data set CH12TA02. The first column of data are observations of Blaisdell's sales, and the second column contain industry sales. The very first column, the one with the row numbers, is a time index: 1 means first quarter of 1998, 2 means 2nd quarter of 1998, etc.

(a) Make a scatter plot of company sales using industry sales as the predictor (with R of course). Describe the apparent relationship between the two variables. Loaded data for the Blaisdell Company per the text book, ch. 12, p. 489. Scatter plot constructed for a preliminary examination of the data:



(b) Make a scatter plot of company sales versus the time index. You might have to create a new column for the time values or figure out how to reference the row numbers in R.



(c) Use R to run a Neumann-Durbin-Watson to check for autocorrelation of company sales over time:

H_0 : = 0 The null hypothesis is that there is no autocorrelation
H_A : 6 = 0 The alternative is that we reject the Null hypothesis

The Durbin-Watson test does show a relatively high degree of autocorrelation, as does our scatter above in subquestion b

lag	Autocorrelation	D-W Statistic	p-value
1	0.6260046	0.7347256	0

Alternative hypothesis: rho != 0

(d) How would you suggest to proceed in modelling the relationship between these variables? A few remedial measures are possible,[3] including the addition of predictor variables and transformations. I would look for additional predictors first before transforming: if additional predictors are available, additional insights could be gained as opposed to performing data transforms alone which would not provide any insights.

6. Complete the following lack-of-fit ANOVA table:

Source	DF	SS	MS	F^*	p-value
Regression	??	34.783	??	??	??
Residual	??	??	??		
Lack-of-Fit	5	??	??	??	??
Pure Error	??	2.110	??		
Total	21	41.85			

[3] Ibid., 490ff.

Table filled in:

Source	DF	SS	MS	F*	p-value
Regression	1	34.783	34.783	98.438	0.000
Residual	20	7.07	.35		
Lack-of-Ffit	5	4.96	.99	7.048	0.000
Pure Error	15	2.110	.141		
Total	21	41.85			

APPENDIX: R SCRIPTS USED

QUESTION 1:

```
> library(xlsx)
> ch03pr03 <- read.xlsx("c:/Users/allen.baumgarten/Documents/CH03PR03.xlsx", sheetName = "Sheet1")
> ch03pr03_reg1 <- lm(ch03pr03[,1] ~ ch03pr03[,2])
> ch03pr03_reg2 <- lm(ch03pr03[,1] ~ ch03pr03[,3])
> ch03pr03_reg3 <- lm(ch03pr03[,1] ~ ch03pr03[,4])
> summary(ch03pr03_reg1 <- lm(ch03pr03[,1] ~ ch03pr03[,2]))

Call:
lm(formula = ch03pr03[, 1] ~ ch03pr03[, 2])

Residuals:
     Min       1Q   Median       3Q      Max 
-2.74004 -0.33827  0.04062  0.44064  1.22737 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    2.11405    0.32089   6.588  1.3e-09 ***
ch03pr03[, 2]  0.03883    0.01277   3.040  0.00292 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6231 on 118 degrees of freedom
Multiple R-squared:  0.07262,	Adjusted R-squared:  0.06476 
F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917

> summary(ch03pr03_reg2 <- lm(ch03pr03[,1] ~ ch03pr03[,3]))

Call:
lm(formula = ch03pr03[, 1] ~ ch03pr03[, 3])

Residuals:
    Min      1Q  Median      3Q     Max 
-1.1672 -0.2402 -0.0225  0.2977  1.0193 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -1.873921   0.345709  -5.421  3.2e-07 ***
ch03pr03[, 3]  0.041944   0.002915  14.389  < 2e-16 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3899 on 118 degrees of freedom
Multiple R-squared:  0.637,	Adjusted R-squared:  0.6339 
F-statistic:   207 on 1 and 118 DF,  p-value: < 2.2e-16

> summary(ch03pr03_reg3 <- lm(ch03pr03[,1] ~ ch03pr03[,4]))

Call:
lm(formula = ch03pr03[, 1] ~ ch03pr03[, 4])

Residuals:
     Min       1Q   Median       3Q      Max 
-1.94233 -0.40879  0.05516  0.48679  1.25950 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.306901   0.185497  12.436  < 2e-16 ***
ch03pr03[, 4] 0.010417   0.002406   4.329 3.15e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6011 on 118 degrees of freedom
Multiple R-squared:  0.1371,	Adjusted R-squared:  0.1298 
F-statistic: 18.74 on 1 and 118 DF,  p-value: 3.153e-05

> plot(ch03pr03[,1] ~ ch03pr03[,2], main = "GPA vs. ACT Score", xlab="ACT Score", ylab="GPA")
> abline(ch03pr03_reg1, col="blue")
> plot(ch03pr03[,1] ~ ch03pr03[,3], main = "GPA vs. Intelligence Score", xlab="Intelligence Score", ylab="GPA")
> abline(ch03pr03_reg2, col="blue")
> plot(ch03pr03[,1] ~ ch03pr03[,4], main = "GPA vs. Class Rank %ile", xlab="Class Rank %ile", ylab="GPA")
> abline(ch03pr03_reg3, col="blue")

> nrow(ch03pr03)
[1] 120
> act_scores <- ch03pr03[-9,]
> nrow(act_scores)
[1] 119

#######  3 new reduced models are now built:
> summary(reg_mod1_gpa_act <- lm(act_scores$GPA ~ act_scores$ACT_Scores))
Call:
lm(formula = act_scores$GPA ~ act_scores$ACT_Scores)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.88628 -0.33291  0.00723  0.43701  1.25781 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)            2.01358    0.29494   6.827 4.09e-10 ***
act_scores$ACT_Scores  0.04383    0.01176   3.728 0.000299 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5713 on 117 degrees of freedom
```

```
Multiple R-squared:  0.1062,	Adjusted R-squared:  0.09854
F-statistic:  13.9 on 1 and 117 DF,  p-value: 0.0002988

> summary(reg_mod2_gpa_intel <- lm(act_scores$GPA ~ act_scores$Intelligence_Score))
Call:
lm(formula = act_scores$GPA ~ act_scores$Intelligence_Score)

Residuals:
     Min       1Q   Median       3Q      Max
-1.16390 -0.23990 -0.02005  0.27788  1.00167

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -1.620511   0.360779  -4.492 1.67e-05 ***
act_scores$Intelligence_Score  0.039857   0.003034  13.135  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3841 on 117 degrees of freedom
Multiple R-squared:  0.5959,	Adjusted R-squared:  0.5924
F-statistic: 172.5 on 1 and 117 DF,  p-value: < 2.2e-16

> summary(reg_mod3_gpa_class <- lm(act_scores$GPA ~ act_scores$Class_Rank_Percentile))
Call:
lm(formula = act_scores$GPA ~ act_scores$Class_Rank_Percentile)

Residuals:
     Min       1Q   Median       3Q      Max
-1.63359 -0.40407  0.06992  0.51362  1.13967

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      2.473276   0.183488  13.479  < 2e-16 ***
act_scores$Class_Rank_Percentile 0.008394   0.002370   3.541 0.000573 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5743 on 117 degrees of freedom
Multiple R-squared:  0.09679,	Adjusted R-squared:  0.08907
F-statistic: 12.54 on 1 and 117 DF,  p-value: 0.0005734

> mod1_res <- reg_mod1_gpa_act$residuals
> mod2_res <- reg_mod2_gpa_intel$residuals
> mod3_res <- reg_mod3_gpa_class$residuals

> shapiro.test(mod1_res)
	Shapiro-Wilk normality test
data:  mod1_res
W = 0.98454, p-value = 0.191

> shapiro.test(mod2_res)
	Shapiro-Wilk normality test
data:  mod2_res
W = 0.98973, p-value = 0.5172

> shapiro.test(mod3_res)
```

```
        Shapiro-Wilk normality test
data:  mod3_res
W = 0.97492, p-value = 0.02517

> qqnorm(mod1_res, main="Residuals for Model 1")
> qqline(mod1_res)
> qqnorm(mod2_res, main="Residuals for Model 2")
> qqline(mod2_res)
> qqnorm(mod3_res, main="Residuals for Model 3")
> qqline(mod3_res)

#######  Cut data into 2 groups at less than 26 and 26 or greater:
> library(car)
> act_scores_26_cut <- cut(act_scores$ACT_Scores, c(-Inf,25.999,Inf), labels = c("<26","26>="))
> act_scores_26 <- cbind(act_scores,act_scores_26_cut)
> head(act_scores_26)
    GPA ACT_Scores Intelligence_Score Class_Rank_Percentile act_scores_26_cut
1 3.897         21                122                    99               <26
2 3.885         14                132                    71               <26
3 3.778         28                119                    95              26>=
4 2.540         22                 99                    75               <26
5 3.028         21                131                    46               <26
6 3.865         31                139                    77              26>=

> leveneTest(reg_mod1_gpa_act$residuals ~ act_scores_26$act_scores_26_cut)
Levene's Test for Homogeneity of Variance (center = median)
       Df  F value   Pr(>F)
group   1   0.1376   0.7114
      117
> act_scores_120_cut <- cut(act_scores$Intelligence_Score, c(-Inf,119.999,Inf), labels = c("<120","120>="))
> act_scores_120 <- cbind(act_scores,act_scores_120_cut)
> head(act_scores_120)
    GPA ACT_Scores Intelligence_Score Class_Rank_Percentile act_scores_120_cut
1 3.897         21                122                    99              120>=
2 3.885         14                132                    71              120>=
3 3.778         28                119                    95               <120
4 2.540         22                 99                    75               <120
5 3.028         21                131                    46              120>=
6 3.865         31                139                    77              120>=
> leveneTest(reg_mod2_gpa_intel$residuals ~ act_scores_120$act_scores_120_cut)
Levene's Test for Homogeneity of Variance (center = median)
       Df  F value   Pr(>F)
group   1   4.1622   0.04359 *
      117
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

QUESTION 2:

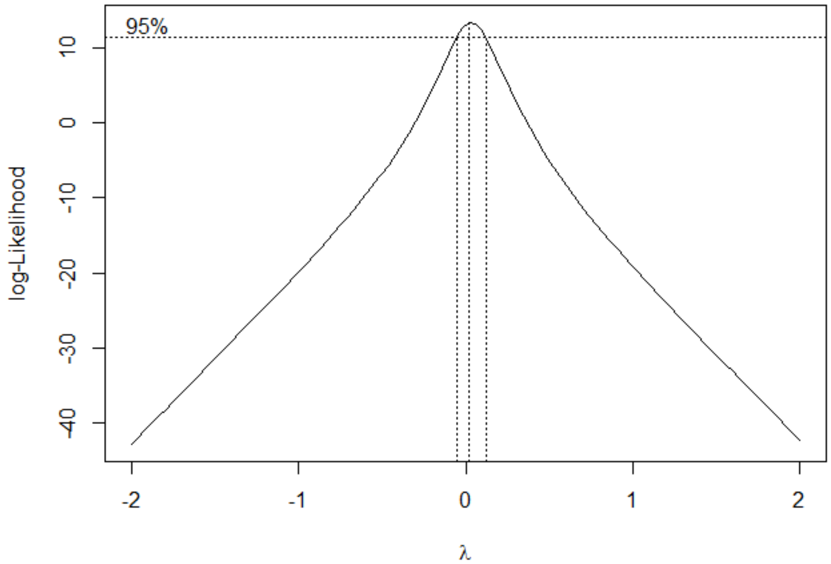
```
> solution <- c(.07, .09, .08, .16, .17, .21, .49, .58, .53, 1.22, 1.15, 1.07, 2.84, 2.57, 3.10)
> hours <- c(9, 9, 9, 7, 7, 7, 5, 5, 5, 3, 3, 3, 1, 1, 1)
> chemistry <- data.frame(solution,hours)
> chemistry
   solution hours
1      0.07     9
2      0.09     9
```

```
3     0.08     9
4     0.16     7
5     0.17     7
6     0.21     7
7     0.49     5
8     0.58     5
9     0.53     5
10    1.22     3
11    1.15     3
12    1.07     3
13    2.84     1
14    2.57     1
15    3.10     1

> reg_chemistry <- lm(chemistry$solution ~ chemistry$hours)
> summary(reg_chemistry)

Call:
lm(formula = chemistry$solution ~ chemistry$hours)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5333 -0.4043 -0.1373  0.4157  0.8487

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.5753     0.2487  10.354 1.20e-07 ***
chemistry$hours  -0.3240     0.0433  -7.483 4.61e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4743 on 13 degrees of freedom
Multiple R-squared:  0.8116,	Adjusted R-squared:  0.7971
F-statistic: 55.99 on 1 and 13 DF,  p-value: 4.611e-06

> install.packages("alr3")
> library(alr3)
> pureErrorAnova(reg_mod1_gpa_act)
Analysis of Variance Table

Response: act_scores$GPA
                   Df  Sum Sq  Mean Sq  F value    Pr(>F)
act_scores$ACT_Scores   1   4.536   4.5364  14.2290  0.0002771 ***
Residuals         117  38.188   0.3264
 Lack of fit       19   6.944   0.3655   1.1463  0.3195527
 Pure Error        98  31.244   0.3188
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> pureErrorAnova(reg_mod2_gpa_intel)
Analysis of Variance Table

Response: act_scores$GPA
                            Df  Sum Sq  Mean Sq  F value    Pr(>F)
act_scores$Intelligence_Score  1  25.4590  25.4590  195.4448  <2e-16 ***
Residuals                    117  17.2651   0.1476
 Lack of fit                  43   7.6257   0.1773   1.3614   0.121
```

```
 Pure Error                     74     9.6394     0.1303
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> pureErrorAnova(reg_mod3_gpa_class)
Analysis of Variance Table

Response: act_scores$GPA
                               Df  Sum Sq   Mean Sq   F value    Pr(>F)
act_scores$Class_Rank_Percentile  1   4.135    4.1353    14.0469   0.0003938 ***
Residuals                     117   38.589    0.3298
 Lack of fit                   55  20.336    0.3697     1.2559    0.1914652
 Pure Error                    62   18.253    0.2944
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

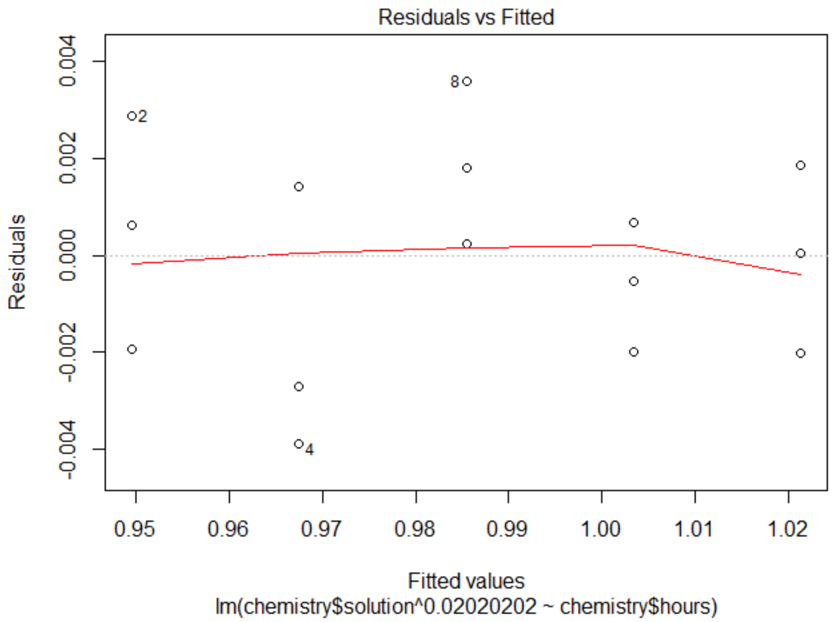
QUESTION 3:

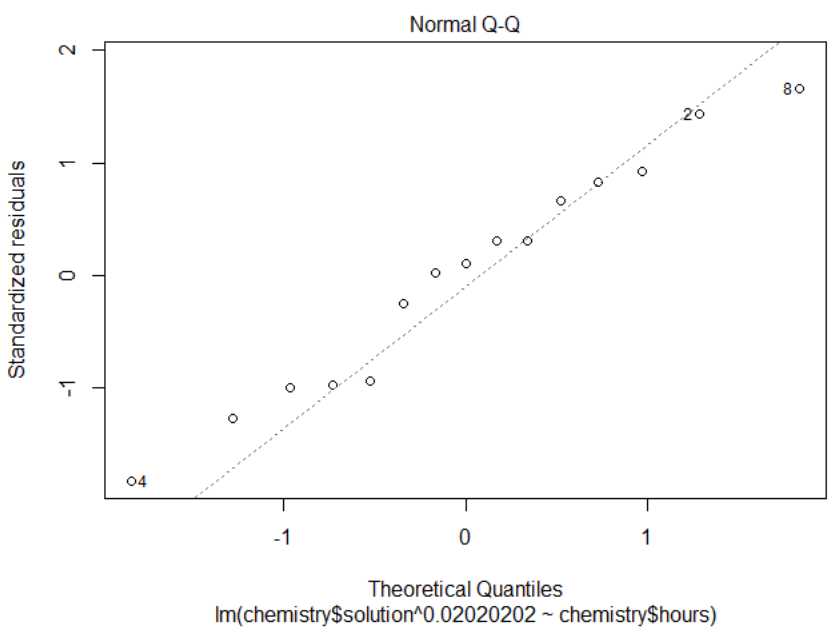
```
> plot(chemistry$solution ~ chemistry$hours, main="Solution (y) vs. Hours (x)", xlab="Hours",ylab="Solution")
> library(MASS)
> trans <- boxcox(chemistry$solution ~ chemistry$hours)
```

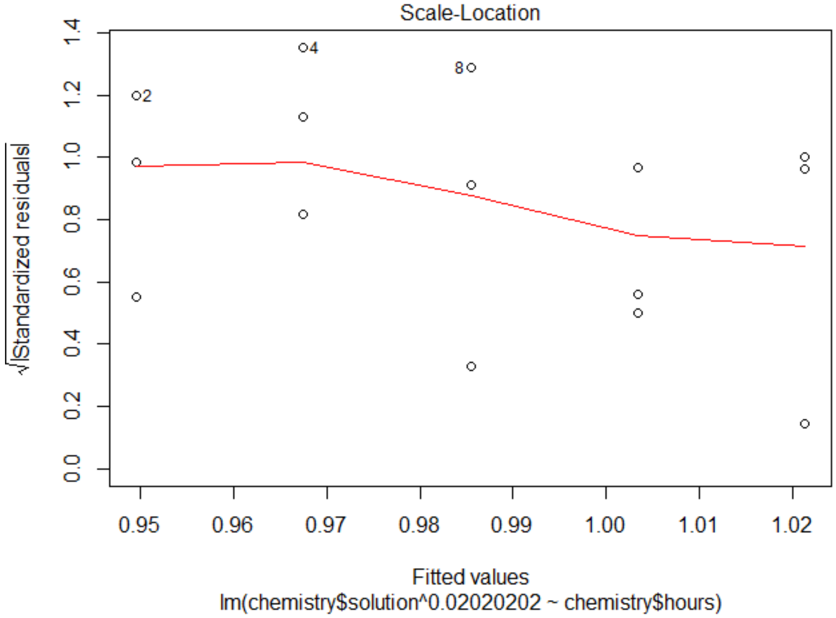


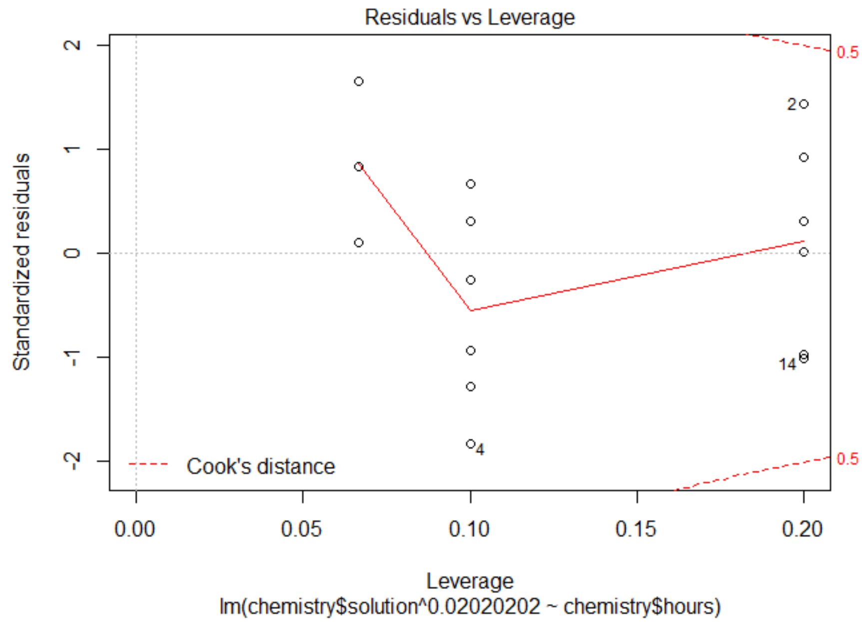

```
> lambda <- trans$x
> loglh <- trans$y
> boxcox <- cbind(lambda, loglh)
> boxcox[order(-loglh),]  # Using log-likelihood to optimize lambda
          lambda      loglh
  [1,]  0.02020202  13.4557358
  [2,]  0.06060606  13.2665123
> reg2_model <- lm(chemistry$solution^0.02020202 ~ chemistry$hours)
> summary(reg2_model)
Call:
lm(formula = chemistry$solution^0.02020202 ~ chemistry$hours)

Residuals:
      Min        1Q     Median        3Q        Max
-0.0038939 -0.0019707  0.0002368  0.0016077  0.0036004

Coefficients:
                  Estimate     Std. Error    t value    Pr(>|t|)
(Intercept)       1.0302214   0.0011764   875.73  < 2e-16 ***
```

```
chemistry$hours   -0.0089532   0.0002048   -43.72    1.7e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002243 on 13 degrees of freedom
Multiple R-squared:  0.9932,	Adjusted R-squared:  0.9927
F-statistic:  1911 on 1 and 13 DF,  p-value: 1.701e-15
> plot(reg2_model)
Hit <Return> to see next plot:
```







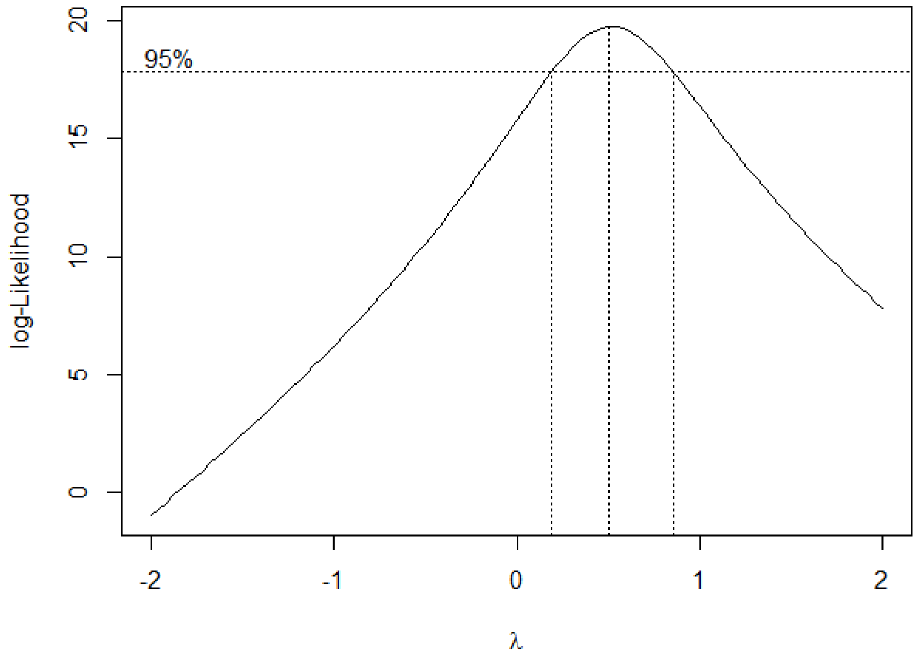



```
> chemistry$solution_log10 <- log10(chemistry$solution)
> chemistry$hours_log10 <- log10(chemistry$hours)
> summary(reg_chemistry_log10 <- lm(chemistry$solution_log10 ~ chemistry$hours_log10))
Call:
lm(formula = chemistry$solution_log10 ~ chemistry$hours_log10)

Residuals:
     Min       1Q   Median       3Q      Max
-0.27494 -0.15732 -0.05912  0.18663  0.24690

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.6022     0.1012    5.95 4.82e-05 ***
```

```
chemistry$hours_log10  -1.5532     0.1479  -10.50 1.02e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1935 on 13 degrees of freedom
Multiple R-squared:  0.8945,	Adjusted R-squared:  0.8864
F-statistic: 110.3 on 1 and 13 DF,  p-value: 1.017e-07

> plot(reg2_model$residuals ~ reg2_model$fitted.values)
> qqnorm(reg2_model$residuals)
> qqline(reg2_model$residuals)
> plot(chemistry$solution_log10 ~ chemistry$hours_log10, main="Solution log10 vs. Hours log10")
> plot(chemistry$solution^0.02020202 ~ chemistry$hours, main="Chemistry with Box-Cox Transform")
> 1.0302214^(1/.0202)
[1] 4.366436
> -0.0089532^(1/.0202)
[1] -4.100024e-102
```

QUESTION 4:

```
> sales <- read.xlsx("C:/Users/allen.baumgarten/Documents/CH03PR17.xlsx",sheetName = "Sheet1")
> sales
   units years
1     98     0
2    135     1
3    162     2
4    178     3
5    221     4
6    232     5
7    283     6
8    300     7
9    374     8
10   395     9
> plot(sales$units ~ sales$years)

> library(MASS)
> boxcox_data <- sales
> boxcox_y <- sales$units
> boxcox_x <- sales$years
> trans <- boxcox(boxcox_y ~ boxcox_x )
> lambda <- trans$x
> loglh <- trans$y
> boxcox <- cbind(lambda, loglh)
> boxcox[order(-loglh),]  # Using log-likelihood to optimize lambda
> lambda_value <- 0.50505051   # input the minimizing x value that maximizes y
> boxcox_reg_model <- lm(dataset$y_variable^lambda_value ~ dataset$x_variable)
```




```
> boxcox_reg_model <- lm(boxcox_y^lambda_value ~ boxcox_x)
> summary(boxcox_reg_model)
Call:
lm(formula = boxcox_y^lambda_value ~ boxcox_x)

Residuals:
    Min       1Q   Median       3Q      Max
-0.49396 -0.31557  0.01724  0.30425  0.48855

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.50006    0.22092   47.53 4.25e-11 ***
boxcox_x     1.11723    0.04138   27.00 3.81e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3759 on 8 degrees of freedom
Multiple R-squared:  0.9891,	Adjusted R-squared:  0.9878
F-statistic: 728.9 on 1 and 8 DF,  p-value: 3.815e-09

> pureErrorAnova(boxcox_reg_model)
Analysis of Variance Table

Response: boxcox_y^lambda_value
          Df Sum Sq Mean Sq F value    Pr(>F)
boxcox_x   1 102.98 102.977   728.9 3.815e-09 ***
Residuals  8   1.13   0.141
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> sales_units_sqr <- sqrt(sales$units)
> sales_years_sqr <- sqrt(sales$years)
> summary(reg_sales_sqrt <- lm(sales_units_sqr ~ sales_years_sqr))
Call:
lm(formula = sales_units_sqr ~ sales_years_sqr)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-1.0876 -0.5841 -0.2683  0.8397  1.3585 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)     8.5410     0.7293  11.711 2.58e-06 ***
sales_years_sqr 3.3996     0.3438   9.888 9.23e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9557 on 8 degrees of freedom
Multiple R-squared:  0.9244,	Adjusted R-squared:  0.9149 
F-statistic: 97.78 on 1 and 8 DF,  p-value: 9.229e-06

> plot(sales_units_sqr ~ sales_years_sqr)
> abline(reg_sales_sqrt)
> plot(reg_sales_sqrt$residuals ~ reg_sales_sqrt$fitted.values)
```

QUESTION 5:

```
> blaisdell <- read.xlsx("c:/Users/allen.baumgarten/Documents/CH12TA02.xlsx", sheetName = "Sheet1")
> blaisdell
   co_sales_y  ind_sales_x
1      20.96       127.3
2      21.40       130.0
3      21.96       132.7
4      21.52       129.4
5      22.39       135.0
6      22.76       137.1
7      23.48       141.2
8      23.66       142.8
9      24.10       145.5
10     24.01       145.3
11     24.54       148.3
12     24.30       146.4
13     25.00       150.2
14     25.64       153.1
15     26.36       157.3
16     26.98       160.7
17     27.52       164.2
18     27.78       165.6
19     28.24       168.7
20     28.78       171.7
> plot(blaisdell$co_sales_y ~ blaisdell$ind_sales_x, main="Blaisdell",ylab="Comp Sales",xlab="Industry Sales")
> plot(blaisdell$co_sales_y)

> summary(reg_blaisdell <- lm(blaisdell$co_sales_y ~ blaisdell$ind_sales_x))
Call:
lm(formula = blaisdell$co_sales_y ~ blaisdell$ind_sales_x)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.149142 -0.054399 -0.000454  0.046425  0.163754 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)          -1.454750   0.214146  -6.793 2.31e-06 ***
blaisdell$ind_sales_x  0.176283   0.001445 122.017  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08606 on 18 degrees of freedom
Multiple R-squared:  0.9988,	Adjusted R-squared:  0.9987
F-statistic: 1.489e+04 on 1 and 18 DF,  p-value: < 2.2e-16

> library(car)
> durbinWatsonTest(reg_blaisdell)
 lag Autocorrelation D-W Statistic p-value
   1       0.6260046     0.7347256       0
 Alternative hypothesis: rho != 0
```