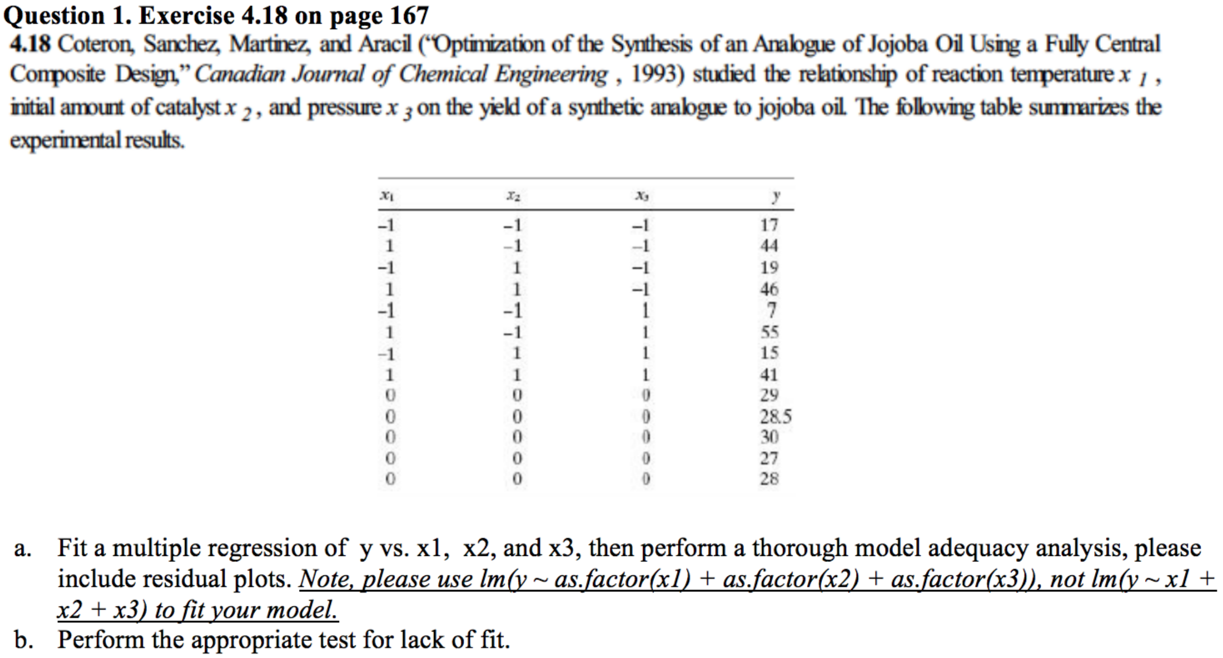
	Q1	Q2	Q3	Q4	Q5
50 Points	10	10	10	10	10

Question 1



Question 1. Exercise 4.18 on page 167

4.18 Coteron, Sanchez, Martinez, and Aracil ("Optimization of the Synthesis of an Analogue of Jojoba Oil Using a Fully Central Composite Design," *Canadian Journal of Chemical Engineering* , 1993) studied the relationship of reaction temperature x_1, initial amount of catalyst x_2, and pressure x_3 on the yield of a synthetic analogue to jojoba oil. The following table summarizes the experimental results.

x_1	x_2	x_3	y
–1	–1	–1	17
1	–1	–1	44
–1	1	–1	19
1	1	–1	46
–1	–1	1	7
1	–1	1	55
–1	1	1	15
1	1	1	41
0	0	0	29
0	0	0	28.5
0	0	0	30
0	0	0	27
0	0	0	28

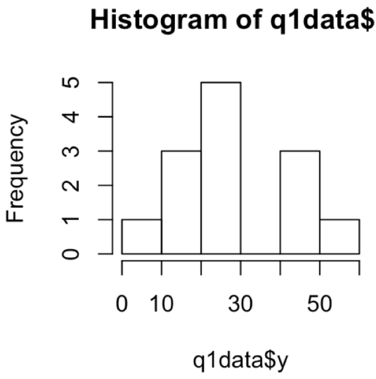
a. Fit a multiple regression of y vs. x1, x2, and x3, then perform a thorough model adequacy analysis, please include residual plots. *Note, please use lm(y ~ as.factor(x1) + as.factor(x2) + as.factor(x3)), not lm(y ~ x1 + x2 + x3) to fit your model.*

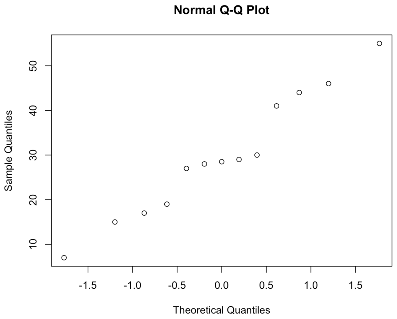
b. Perform the appropriate test for lack of fit.

Q1 Part (a)

The five key assumptions:

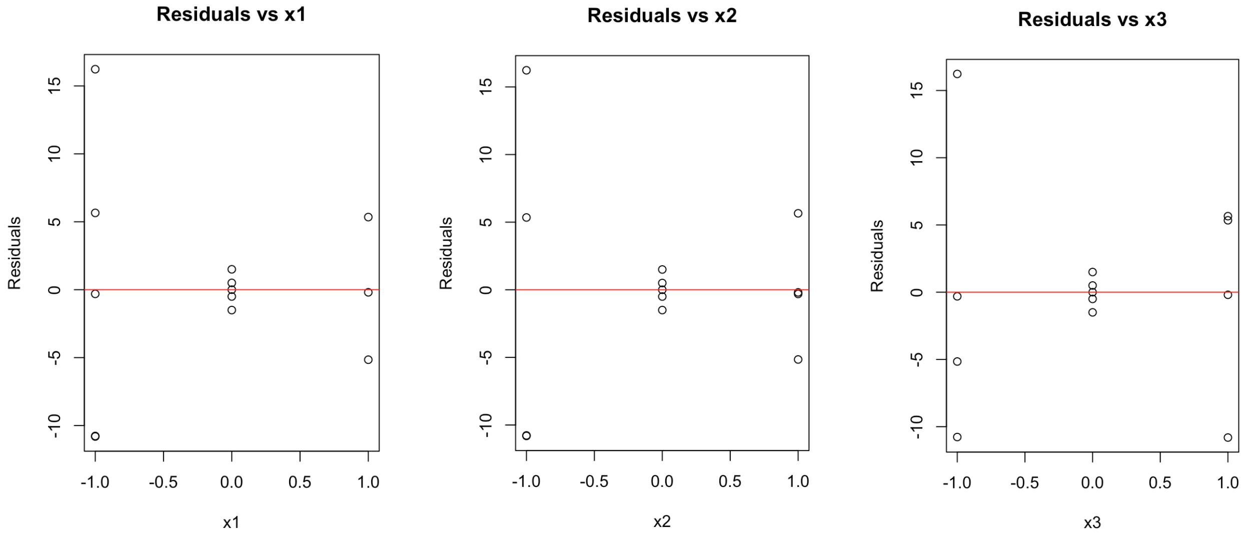
a. Normality - Our response variable(s) (by themselves), residuals (by themselves), and residuals vs regressors, look normal.





The distribution of our response variable seems mostly normal - I don't know what to make of the gap in the middle, though.

b. Independence - Our samples are independent (i.e. the value of one does not affect the value of any other). This is usually the hardest one to test for - usually it's argued from a sampling side. If you plot the residuals vs the predicted values, if they're independent, you should see no pattern.



Our samples appear to be independent - there doesn't seem to be a pattern in the data (but then again, there's only 13 data points).

c. Constant Variance - The residual plots should just be bands (i.e. no funnels, cones, or any weird shape).

It does appear that we have a "bowtie" kind of pattern, so I would assume that we don't have constant variance.

d. $E[\epsilon] = 0$ - This is assumed since that's the way we build our model (i.e. via least squares)

The cluster of residuals for all the residual plots seems to be centered around 0.

e. Linearity - The model actually fits (i.e. the data follows the shape of the model: R^2 is high)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	27.7692	5.3190	5.22	0.0008
as.factor(x1)0	0.7308	6.5577	0.11	0.9140
as.factor(x1)1	31.8462	6.7280	4.73	0.0015
as.factor(x2)1	-8.4615	6.2935	-1.34	0.2157
as.factor(x3)1	-9.9615	6.2935	-1.58	0.1521

Multiple R-squared: 0.7393, Adjusted R-squared: 0.609

F-statistic: 5.672 on 4 and 8 DF, p-value: 0.01828

Only the intercept (***) and as.factor(x1)0 (**) are significant.

I would say the model somewhat fits - it's hard to say since we only have 13 data points. Our R^2 is mediocre, and we have a lot of degrees of freedom relative to our number of samples.

Technically, if we do a hypothesis test and assume H_0: all β 's are 0, then we can disprove the null hypothesis with our β for x_1's 2nd factor (as shown in the table: as.factor(x1)1). But, everything else is insignificant. I'd say the model fits, but barely (needs changes).

Q1 Part (b)

Data:

	x1	x2	x3	y	level
1	-1	-1	-1	17.00	1
2	1	-1	-1	44.00	2
3	-1	1	-1	19.00	3
4	1	1	-1	46.00	4
5	-1	-1	1	7.00	5
6	1	-1	1	55.00	6
7	-1	1	1	15.00	7
8	1	1	1	41.00	8
9	0	0	0	29.00	9
10	0	0	0	28.50	9
11	0	0	0	30.00	9
12	0	0	0	27.00	9
13	0	0	0	28.00	9

Recall:
$\text{SS}_{Res} = \text{SS}_{PE} + \text{SS}_{LOF}$
Our test statistic is:

$$F_0 = \frac{SS_{LOF}/(m-2)}{SS_{PE}/(n-m)} = \frac{MS_{LOF}}{MS_{PE}}$$

If $\text{F}_0 > \text{F}_{m-2,n-m}$, conclude that the regression function is not linear.
From R:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
as.factor(x1)	2	2060.31	1030.15	824.12	0.0000
as.factor(x2)	1	0.50	0.50	0.40	0.5614
as.factor(x3)	1	8.00	8.00	6.40	0.0647
Residuals	8	188.50	23.56		
Lack of fit	4	183.50	45.87	36.70	0.0021
Pure Error	4	5.00	1.25		

Since we have 9 levels of (x_1, x_2, x_3),

$$m = 9$$

$$(29 + 28.5 + 30 + 27 + 28)/5 = 28.5$$

$$(29 - 28.5)^2 + (28.5 - 28.5)^2 + (30 - 28.5)^2 + (27 - 28.5)^2 + (28 - 28.5)^2 = 5(SS_{PE})$$

$$188.50 = 5 + SS_{LOF} \longrightarrow SS_{LOF} = 183.50$$

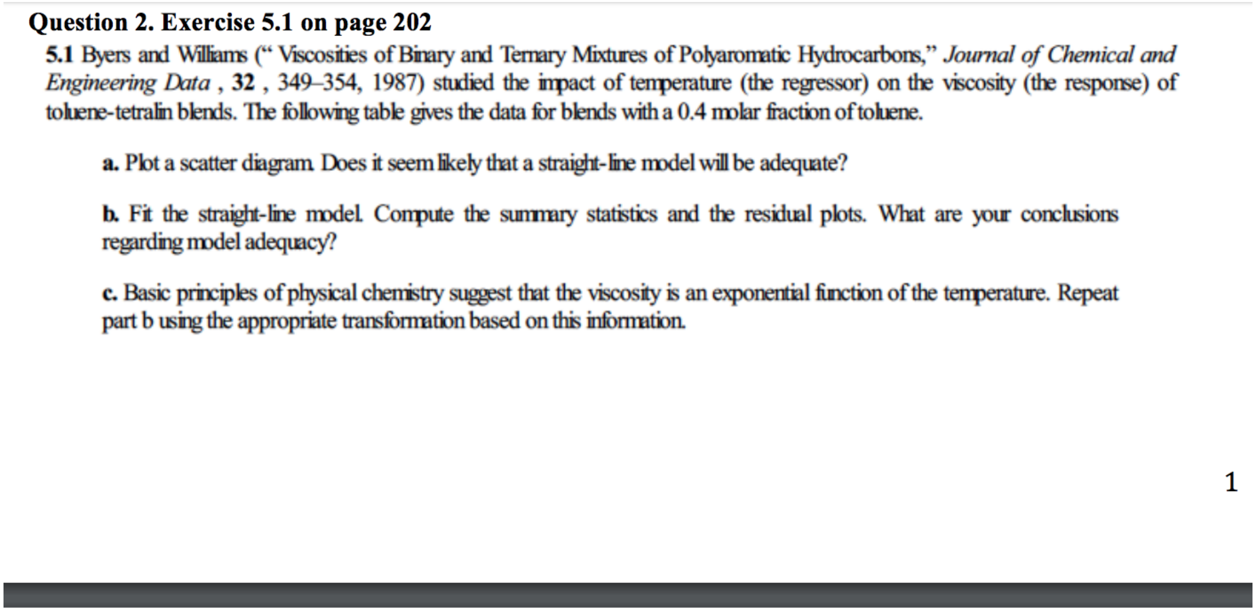
$$F_0 = \frac{183.50/(9-2)}{5/(13-9)} = \frac{26.214}{1.25} = 36.70???$$

Test:

$$36.70 > F_{m-2,n-m} = F_{3,8} = 2.92380 \text{ (even at } \alpha \ = 0.10 \text{ it still fails)}$$

So we conclude that the regression function is not linear.

Question 2



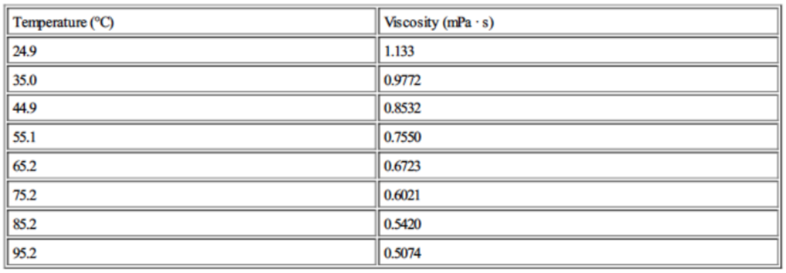
Question 2. Exercise 5.1 on page 202

5.1 Byers and Williams (" Viscosities of Binary and Ternary Mixtures of Polyaromatic Hydrocarbons," *Journal of Chemical and Engineering Data* , **32** , 349–354, 1987) studied the impact of temperature (the regressor) on the viscosity (the response) of toluene-tetralin blends. The following table gives the data for blends with a 0.4 molar fraction of toluene.

a. Plot a scatter diagram. Does it seem likely that a straight-line model will be adequate?

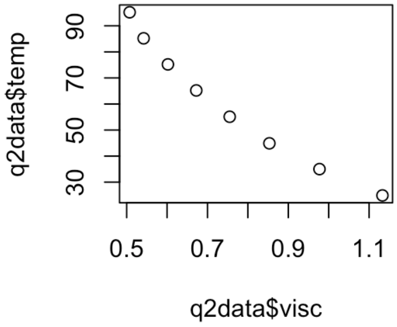
b. Fit the straight-line model. Compute the summary statistics and the residual plots. What are your conclusions regarding model adequacy?

c. Basic principles of physical chemistry suggest that the viscosity is an exponential function of the temperature. Repeat part b using the appropriate transformation based on this information.



Temperature (°C)	Viscosity (mPa · s)
24.9	1.133
35.0	0.9772
44.9	0.8532
55.1	0.7550
65.2	0.6723
75.2	0.6021
85.2	0.5420
95.2	0.5074

Q2 Part (a)



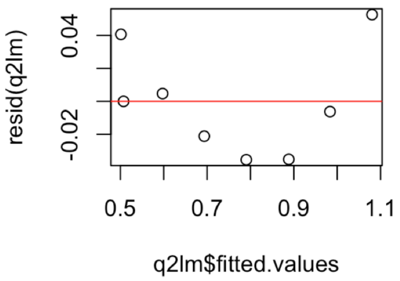
Yes, it seems likely that a straight line model will be adequate.

Q2 Part (b)

```
q2lm <- lm(q2data$visc ~ q2data$temp)
summary(q2lm)
plot(q2lm$fitted.values, resid(q2lm))
abline(h = 0, col="red")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.2815	0.0469	27.34	0.0000
q2data$temp	-0.0088	0.0007	-12.02	0.0000

Residual standard error: 0.04743 on 6 degrees of freedom
Multiple R-squared: 0.9602 Adjusted R-squared: 0.9535

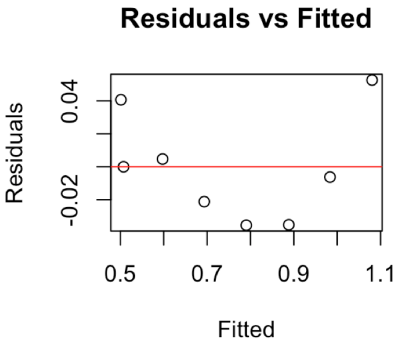


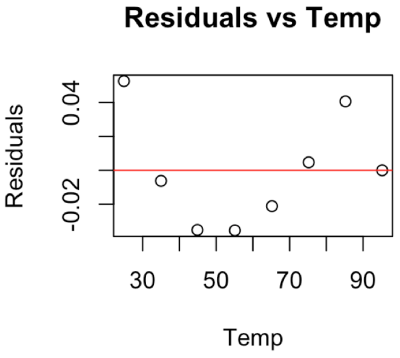
It looks like all 5 of our assumptions hold except for non constant variance for our error term (due to the quadratic pattern).

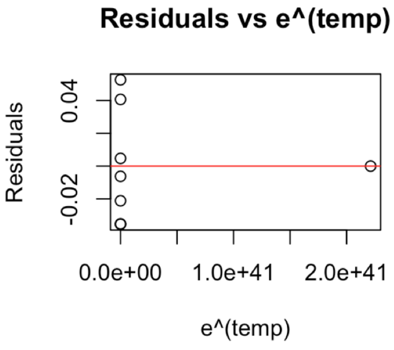
Q2 Part (c)

```
q2lm <- lm(q2data$visc ~ q2data$temp + exp(q2data$temp))
summary(q2lm)
plot(q2lm$fitted.values, resid(q2lm))
plot(q2data$temp, resid(q2lm))
plot(exp(q2data$temp), resid(q2lm))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.3195	0.0424	31.11	0.0000
q2data$temp	-0.0096	0.0007	-13.27	0.0000
exp(q2data$temp)	0.0000	0.0000	2.03	0.0986


Residuals vs Fitted

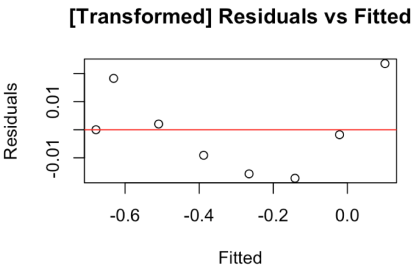

Residuals vs Temp

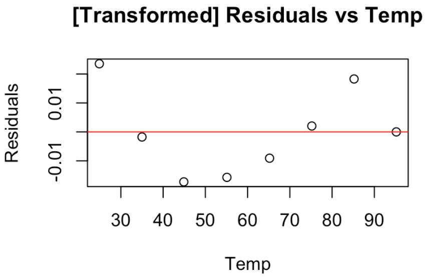

Residuals vs e^(temp)

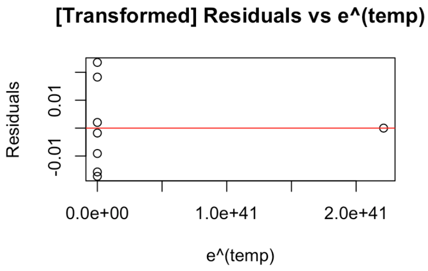
Post-Transformation:

```
q2lm_t <- lm(log(q2data$visc) ~ q2data$temp + exp(q2data$temp))
summary(q2lm_t)
plot(q2lm_t$fitted.values, resid(q2lm_t), main="[Transformed] Residuals vs Fitted", xlab="Fitted", ylab=
plot(q2data$temp, resid(q2lm_t), main="[Transformed] Residuals vs Temp", xlab="Temp", ylab="Residuals")
plot(exp(q2data$temp), resid(q2lm_t), main="[Transformed] Residuals vs e^(temp)", xlab="e^(temp)", ylab=
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.4036	0.0192	20.99	0.0000
q2data$temp	-0.0121	0.0003	-37.02	0.0000
exp(q2data$temp)	0.0000	0.0000	3.23	0.0232

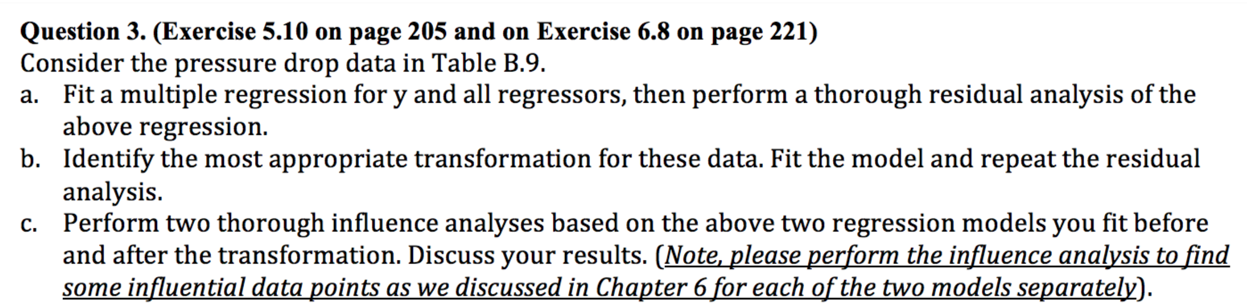

[Transformed] Residuals vs Fitted


[Transformed] Residuals vs Temp


[Transformed] Residuals vs e^(temp)

Looks like we still have non-constant variance, but the p values and R^2 got better after both the additional β and the log transformation.

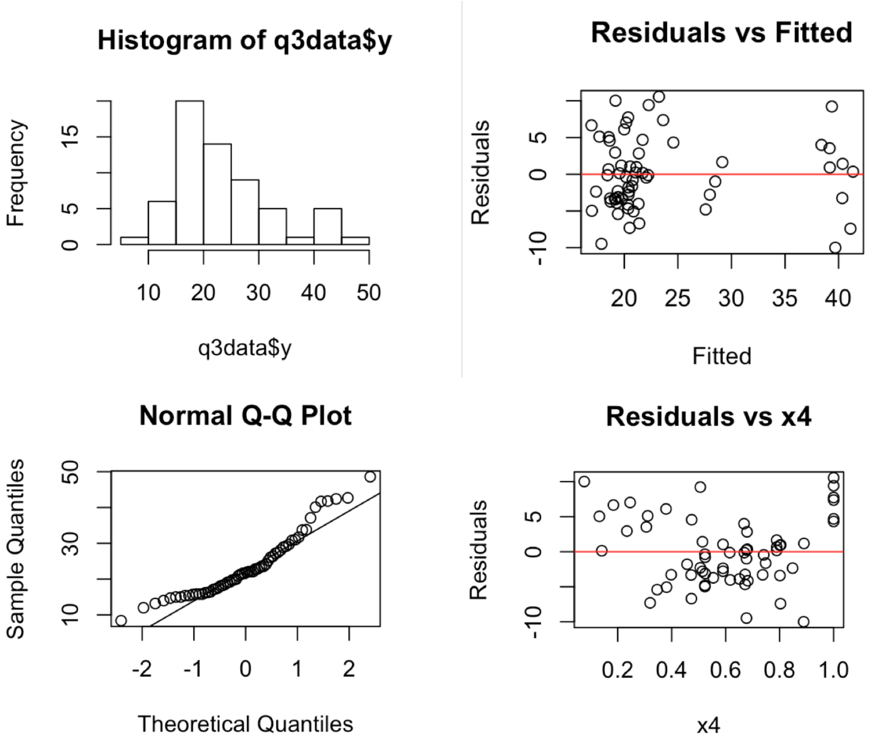
Question 3



Question 3. (Exercise 5.10 on page 205 and on Exercise 6.8 on page 221)
Consider the pressure drop data in Table B.9.

a. Fit a multiple regression for y and all regressors, then perform a thorough residual analysis of the above regression.
b. Identify the most appropriate transformation for these data. Fit the model and repeat the residual analysis.
c. Perform two thorough influence analyses based on the above two regression models you fit before and after the transformation. Discuss your results. (*Note, please perform the influence analysis to find some influential data points as we discussed in Chapter 6 for each of the two models separately*).

Q3 Part (a)



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.8945	4.3251	1.36	0.1783
x1	-0.4779	0.3400	-1.41	0.1653
x2	0.1827	0.0172	10.63	0.0000
x3	35.4028	11.0996	3.19	0.0023
x4	5.8439	2.9098	2.01	0.0494

```
Multiple R-squared:  0.6914,Adjusted R-squared:  0.6697
```

The five key assumptions:

a. Normality - **Looks somewhat normal, maybe with a long right tail.**

b. Independence - **Since there doesn't appear to be any pattern in the residuals vs fitted values, I would say the errors are independent.**

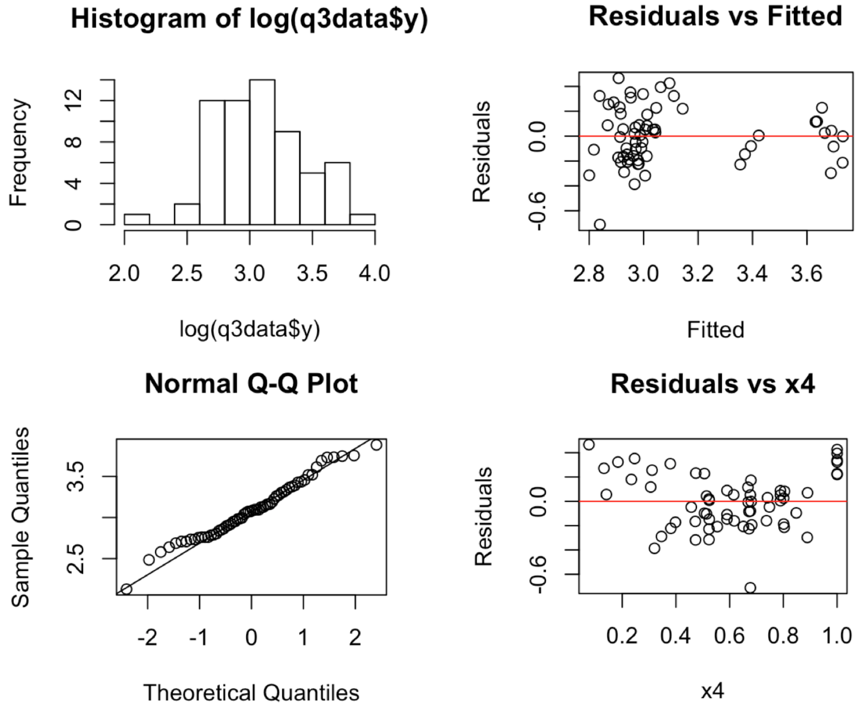
c. Constant Variance - **It looks okay, but there might be a slight quadratic pattern.**

d. E[ϵ] $= 0$ - **This is assumed since that's the way we build our model (i.e. via least squares).**

e. Linearity - Multiple R-squared: 0.6914 Adjusted R-squared: 0.6697

x_2 (***), x_3 (**), and x_4 (*) are all significant, so if H_0 is that there is no linear relationship between any of our regressors and our response variable, then we reject H_0.

Q3 Part (b)

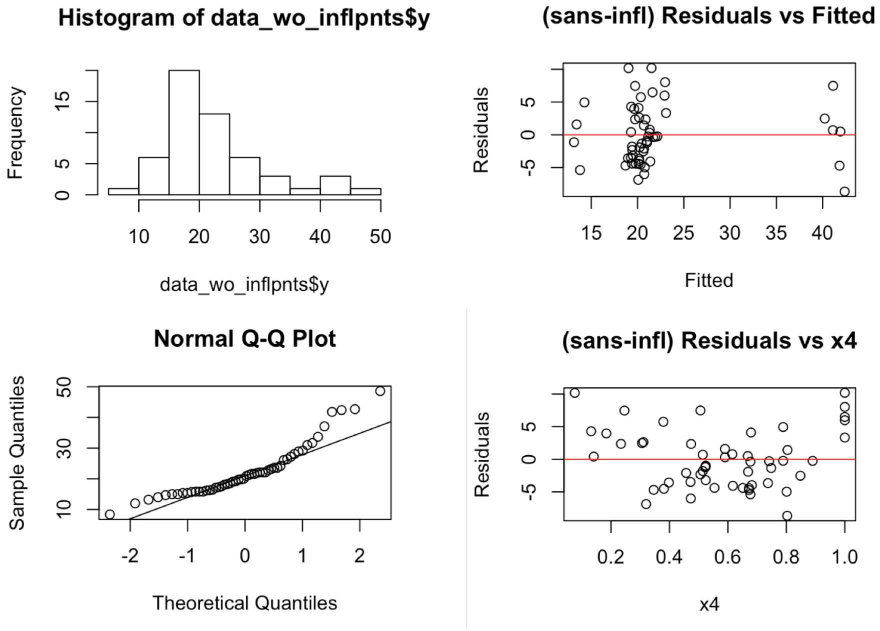


	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.2285	0.2033	10.96	0.0000
x1	-0.0162	0.0160	-1.01	0.3159
x2	0.0066	0.0008	8.13	0.0000
x3	1.8502	0.5218	3.55	0.0008
x4	0.2548	0.1368	1.86	0.0677

Multiple R-squared: 0.5894 Adjusted R-squared: 0.5606

Oddly enough, it looks like we improved some aspects of our model (our y looks more normal and our residuals look a little better, but our R^2 went down significantly.

Q3 Part (c): Pre-Log Transformation, Influential Points Removed



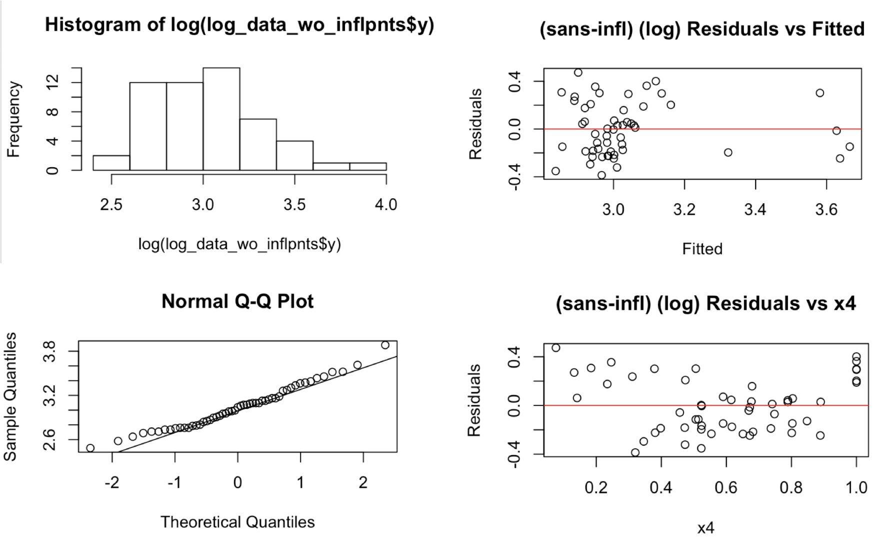
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-13.1146	9.4792	-1.38	0.1728
x1	0.0268	0.3511	0.08	0.9395
x2	0.1984	0.0192	10.35	0.0000
x3	87.5725	27.1116	3.23	0.0022
x4	4.2155	2.8757	1.47	0.1491

```
Multiple R-squared:  0.7178, Adjusted R-squared:  0.6947
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	20.96	20.96	0.97	0.3291
x2	1	2402.11	2402.11	111.37	0.0000
x3	1	218.45	218.45	10.13	0.0025
x4	1	46.35	46.35	2.15	0.1491
Residuals	49	1056.90	21.57		

Influential Points (by data sample index):
22 50 51 52 53 54 56 59

Q3 Part (c): Post-Log Transformation, Influential Points Removed



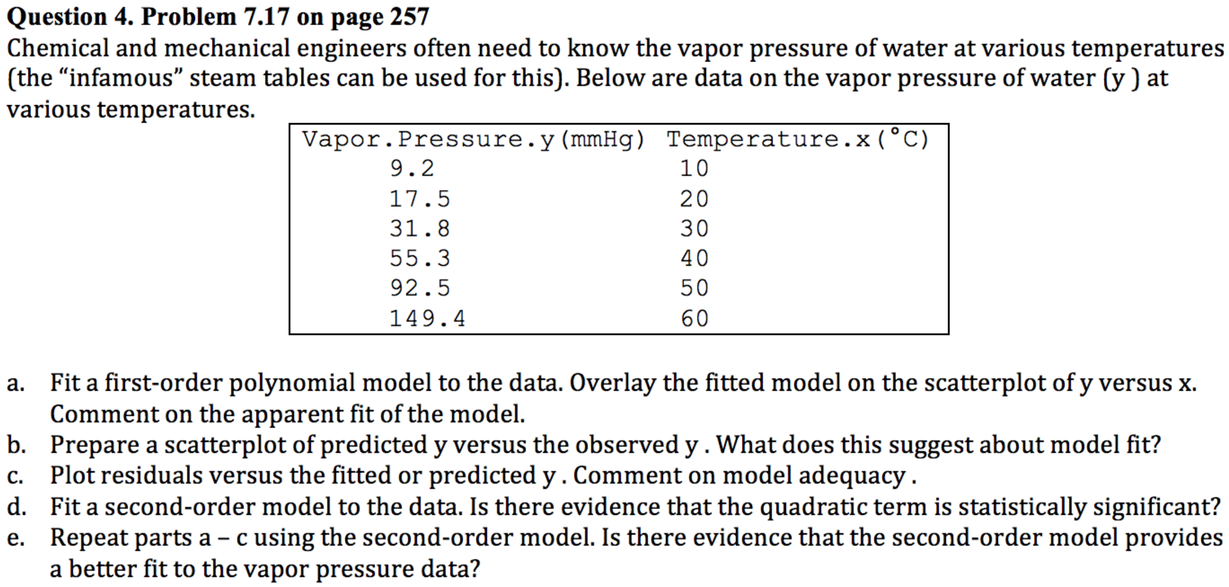
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.2988	0.3211	7.16	0.0000
x1	-0.0128	0.0172	-0.75	0.4591
x2	0.0057	0.0011	5.04	0.0000
x3	1.6190	0.8849	1.83	0.0735
x4	0.2822	0.1371	2.06	0.0450

```
Multiple R-squared:  0.4198,  Adjusted R-squared:  0.3715
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	0.00	0.00	0.00	0.9613
x2	1	1.47	1.47	27.53	0.0000
x3	1	0.16	0.16	2.96	0.0918
x4	1	0.23	0.23	4.24	0.0450
Residuals	48	2.56	0.05		

Influential Points (by data sample index):
35 50 51 52 54 57 59 60 62
Influential Points in both the pre and post transformed models:
50 51 52 54 59
Conclusion:
Transforming the data via log(y) actually does worse in both cases (with and without influential points). It turns out that the transformed model without influential points had the worst R^2 out of the four, but the original model without influential points had the best R^2 out of the four. Transforming the data overall tends to improve how normally distributed the predictor is (and slightly improve the consistency of the variance of the error term), but still makes the model worse in terms of R^2.

Question 4



Question 4. Problem 7.17 on page 257

Chemical and mechanical engineers often need to know the vapor pressure of water at various temperatures (the "infamous" steam tables can be used for this). Below are data on the vapor pressure of water (y) at various temperatures.

Vapor.Pressure.y(mmHg)	Temperature.x(°C)
9.2	10
17.5	20
31.8	30
55.3	40
92.5	50
149.4	60

a. Fit a first-order polynomial model to the data. Overlay the fitted model on the scatterplot of y versus x. Comment on the apparent fit of the model.

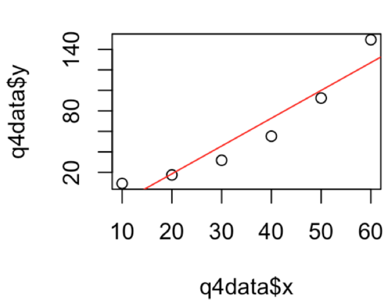
b. Prepare a scatterplot of predicted y versus the observed y . What does this suggest about model fit?

c. Plot residuals versus the fitted or predicted y . Comment on model adequacy .

d. Fit a second-order model to the data. Is there evidence that the quadratic term is statistically significant?

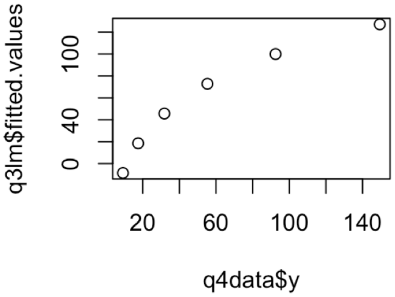
e. Repeat parts a – c using the second-order model. Is there evidence that the second-order model provides a better fit to the vapor pressure data?

Q4 Part (a)



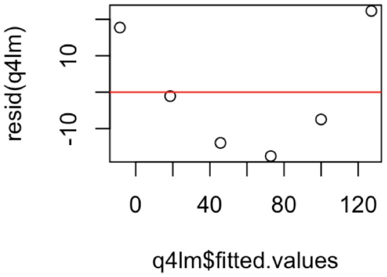
It fits okay. Could be better. $\mathbf{R^2} = \mathbf{0.9038}$

Q4 Part (b)



It suggests that, since there's a slight curve, perhaps y isn't linear.

Q4 Part (c)



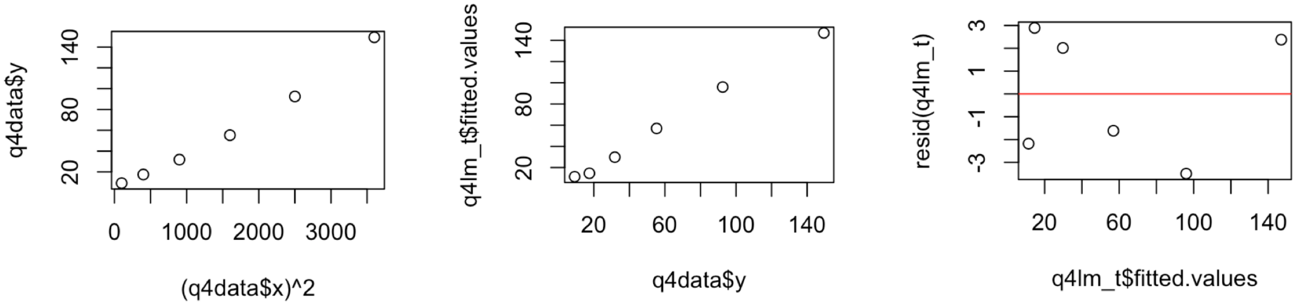
It looks like we have non constant variance (due to the quadratic shape).

Q4 Part (d)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	20.1000	6.3360	3.17	0.0504
x	-1.4696	0.4145	-3.55	0.0382
I(x^2)	0.0598	0.0058	10.31	0.0019

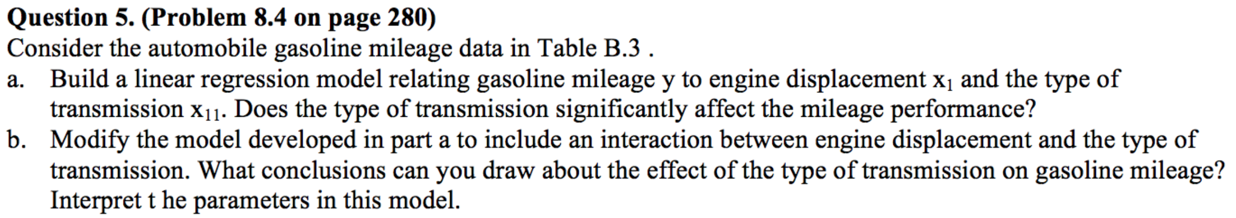
Yes, the x^2 term is statistically significant (**).

Q4 Part (e)



Yes, there is evidence it is a better fit. The non-constant variance problem seems to be reduced (although you could argue that there is still a slight quadratic shape, but there's not enough data to really say either way). The R^2 is also a lot better: 0.9974

Question 5



Question 5. (Problem 8.4 on page 280)

Consider the automobile gasoline mileage data in Table B.3 .

a. Build a linear regression model relating gasoline mileage y to engine displacement x_1 and the type of transmission x_{11}. Does the type of transmission significantly affect the mileage performance?

b. Modify the model developed in part a to include an interaction between engine displacement and the type of transmission. What conclusions can you draw about the effect of the type of transmission on gasoline mileage? Interpret t he parameters in this model.

Q5 Part (a)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	33.6184	1.5395	21.84	0.0000
x1	-0.0457	0.0087	-5.27	0.0000
as.factor(x11)1	-0.4987	2.2282	-0.22	0.8245

The p value of x11 in the model is 0.8245, so there isn't significant evidence to conclude that there is a linear relationship between y and x11.

Q5 Part (b)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	42.9196	2.7349	15.69	0.0000
x1	-0.1168	0.0198	-5.89	0.0000
as.factor(x11)1	-13.4637	3.8441	-3.50	0.0016
x1:as.factor(x11)1	0.0816	0.0213	3.84	0.0006

x1 (***), x11 (**), and x1:x11(***) are significant. It looks like the type of transmission (x11) has a significant effect on the gasoline mileage (y) when you account for the interaction between the engine displacement (x1) and type of transmission (x11).