Homework 6

Question 1

```
x = c(110.5, 105.4, 118.1, 104.5, 93.6, 84.1, 77.8, 75.6)
y = c(5.755, 5.939, 6.010, 6.545, 6.730, 6.750, 6.899, 7.862)
```

1a. Give the least squares estimate of the slope. Give a brief interpretation.

```
plant_data <- data.frame(x, y)
pd.fit <- lm(formula = y ~ x, data = plant_data)
pd.fit
```

```
##
## Call:
## lm(formula = y ~ x, data = plant_data)
##
## Coefficients:
## (Intercept)            x
##    10.13746     -0.03717
```

The slope of the line is -0.03717. The negative slope suggests that for each unit increase in plant height there is an expected 0.3717 unit decrease in grain yield.

1b. Perform a test for null hypothesis vs Ha B =/ 0 using an F test first and then a T test. Your conclusion?

```
### F-test (Anova)
anova(pd.fit)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## x          1 2.42357 2.42357  18.455 0.005116 **
## Residuals  6 0.78794 0.13132
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the anova we see that we can reject the null hypothesis at the 0.01 level.

```
### T Test
summary(pd.fit)
```

```
##
## Call:
```

```
## lm(formula = y ~ x, data = plant_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34626 -0.27605 -0.09448  0.27023  0.53495
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.137455   0.842265  12.036    2e-05 ***
## x           -0.037175   0.008653  -4.296  0.00512 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3624 on 6 degrees of freedom
## Multiple R-squared:  0.7547, Adjusted R-squared:  0.7138
## F-statistic: 18.46 on 1 and 6 DF,  p-value: 0.005116
```

Again, we see that we can reject the null hypothesis at the 0.01 level, so the t test and the anova have the same result. Overall, this indicates the regression model is well supported.

1c. Construct a 95% confidence interval for the intercept by hand using the equation from the lecture, compare your results with those from R and briefly interpret the 95% confidence interval. You can get using R code qt(alpha/2, n-2) where alpha is 0.05 here.

```
t <- qt(0.05/2, 8-2)
B <- 10.137455
SE <- 0.842265

### B +/- (T x SE)
Int1 <- (B - (t*SE))
Int2 <- (B+(t*SE))
Int1
```

```
## [1] 12.1984
```

```
Int2
```

```
## [1] 8.076507
```

```
##Now have R calculate
pd.int <- confint(pd.fit, level = 0.95)
pd.int
```

```
##                   2.5 %      97.5 %
## (Intercept)  8.07650745 12.19840320
## x           -0.05834895 -0.01600043
```

The results of my hand calculations and the R formula are consistent. I think this means that there is a 95% chance that repeating the study would result in a intercept between those two numbers.

(1 point) Give the fitted regression line (as a equation that looks like y =a+bx) and the raw residuals.

```r
## get intercept and slope
B_int <- coef(pd.fit)[1]
slope <- coef(pd.fit)["x"]
B_int
```

```
## (Intercept)
##    10.13746
```

```r
slope
```

```
##           x
## -0.03717469
```

y = -0.3717x + 10.1375

Residuals: 0.78794 0.13132 (taken from the analysis of variance table from the anova)

1e. Give an estimate of the error variance.

```r
pd.mse <- mean(pd.fit$residuals)^2
pd.mse
```

```
## [1] 7.70372e-34
```

taking the square of the mean of the residuals

1f. Estimate the expected yield of a rice variety that has height x =100 and provide a 95% confidence interval.

```r
conf_yield <- predict(pd.fit, newdata = data.frame(x = 100), interval = "confidence")
conf_yield
```

```
##        fit      lwr      upr
## 1 6.419986 6.096321 6.743651
```

This indicates there is a 95% confidence level the predicted yield is between 6.09 and 6.74.

1g. Predict the yield of a new rice variety that has height x =100 and provide a 95% prediction interval. Compare the results with those from (f), which one is wider?

```r
pred_yield <- predict(pd.fit, newdata = data.frame(x = 100), interval = "prediction")
pred_yield
```

```
##        fit      lwr      upr
## 1 6.419986 5.476038 7.363934
```

These results have a wider interval than the results given using a 95% confidence interval.

1h. Compute the coefficient of determination R^2 and briefly interpret what does it mean.

```
sum.fit <- summary(pd.fit)
sum.fit$r.squared
```

```
## [1] 0.7546518
```

The r^2 gives the percent of the variation in the dependent variable that is predictable using the independent variable. So in this case, 75% of the variation in grain yield can be predicted by the rice plant height.

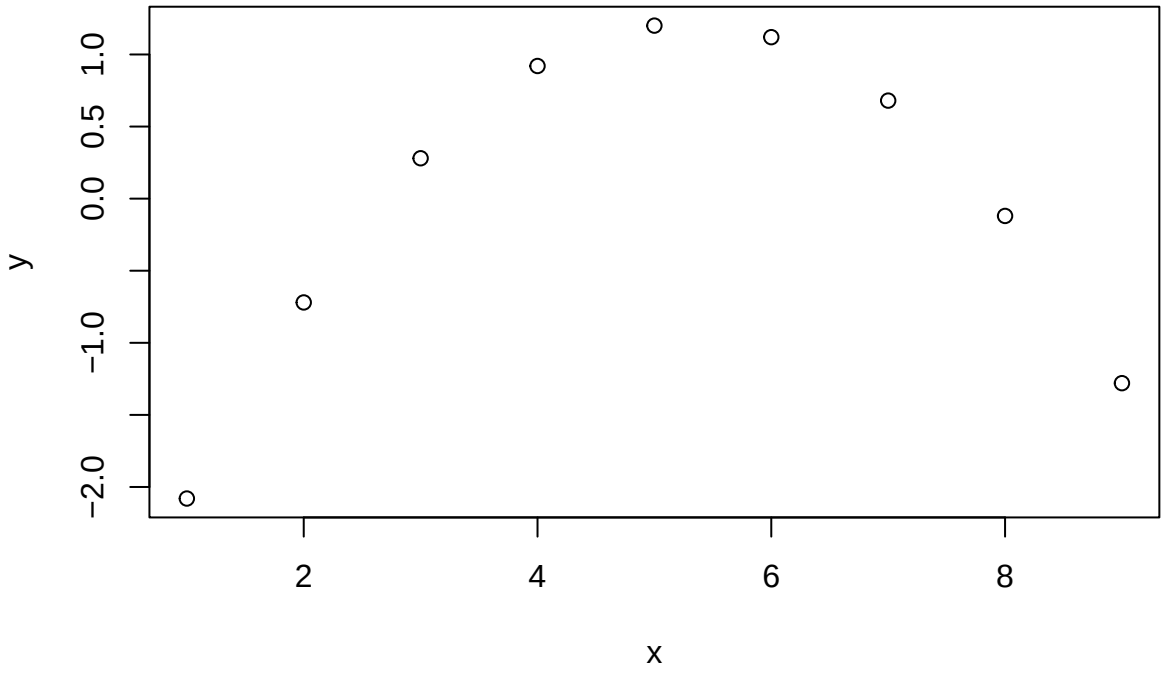
Question 2

This problem is designed to demonstrate why residuals are plotted against y (instead of yy). Consider the following (artificial) data set that was constructed so that the relationship between yy and xx is quadratic. It is immediately evident that a linear fit is not appropriate. However, we adopt the point of view that the residual plot will provide diagnostic information on the lack of fit.

```
x = c(1, 2, 3, 4, 5, 6, 7, 8, 9)
y = c(-2.08, -0.72, 0.28, 0.92, 1.20, 1.12, 0.68, -0.12, -1.28)
q2 <- data.frame(x, y)
```

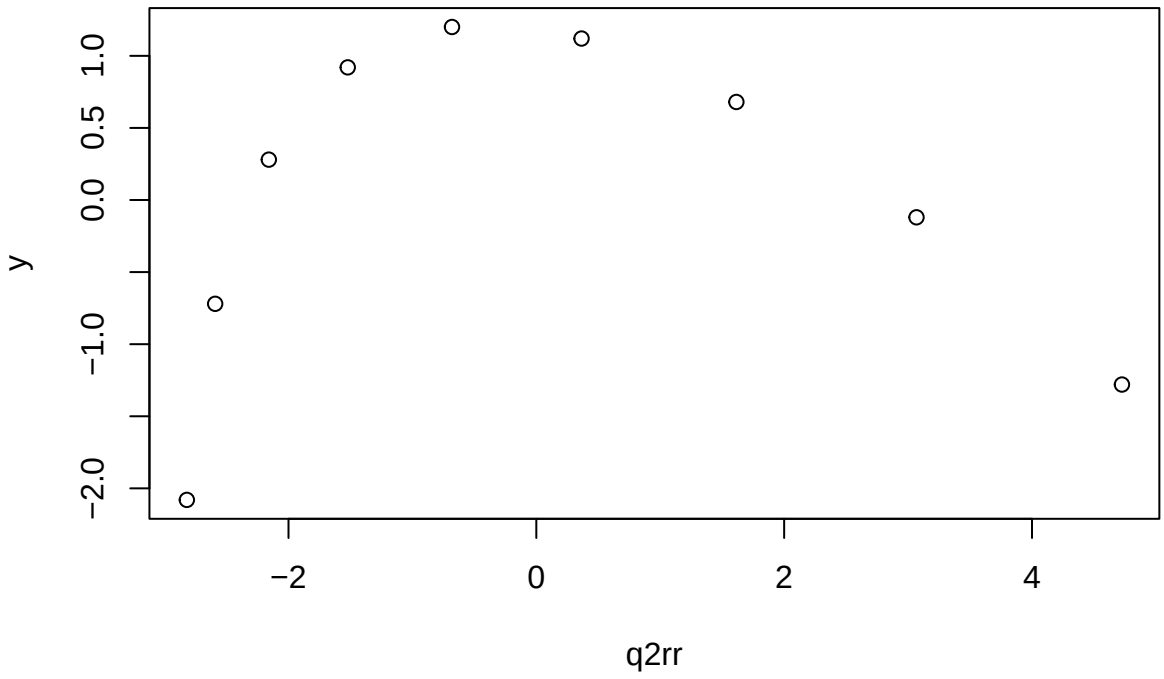
2a. Plot y vs. x.

```
q2lm <- lm(x~y, data = q2)
plot(x, y)
```



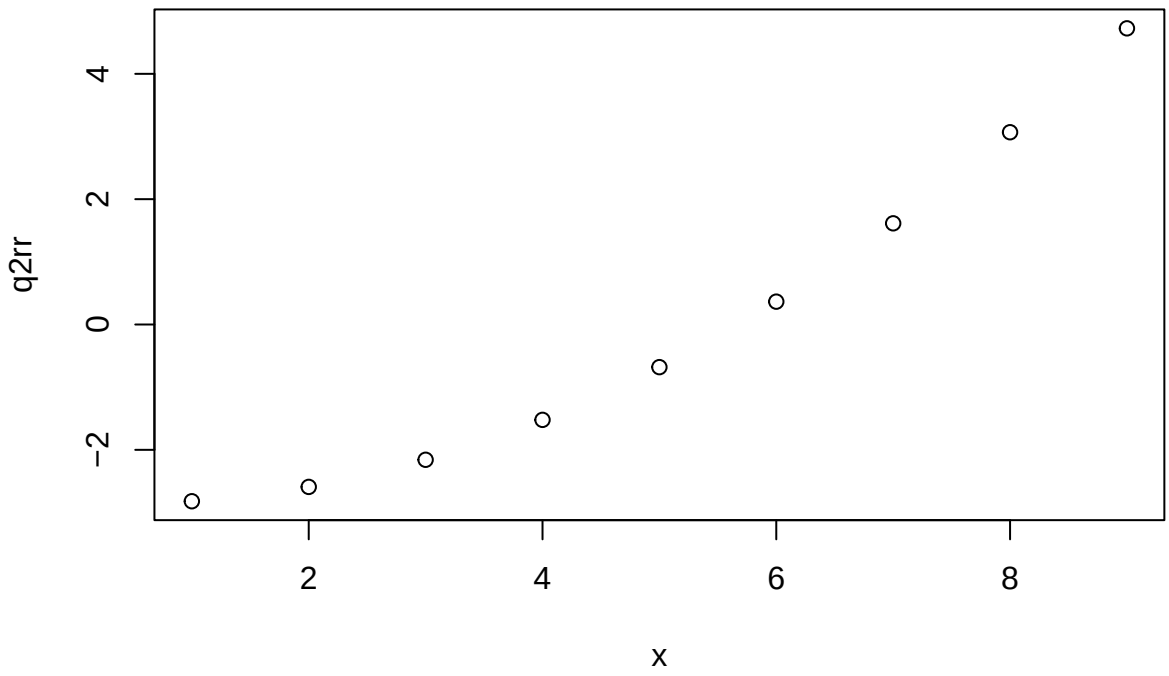
2b. Plot the raw residuals vs. y.

```
q2rr <- q2lm$residuals
plot(q2rr, y)
```



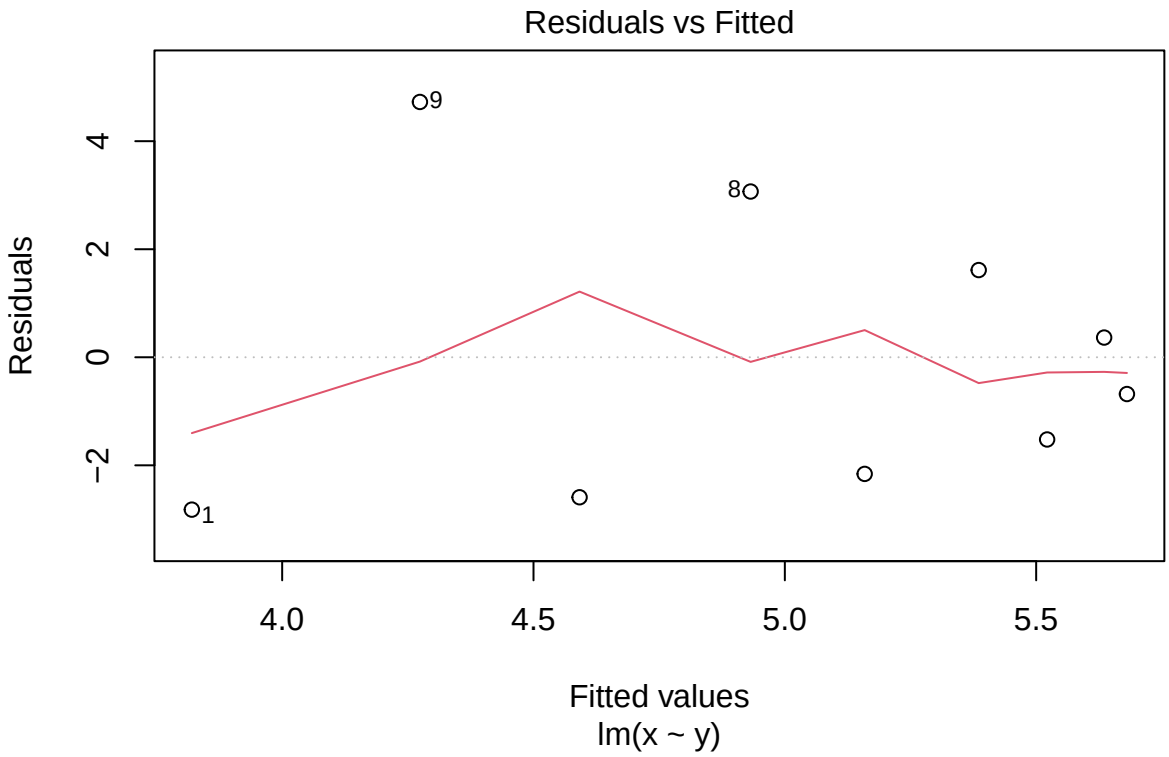
2c. Plot the raw residuals vs. x.

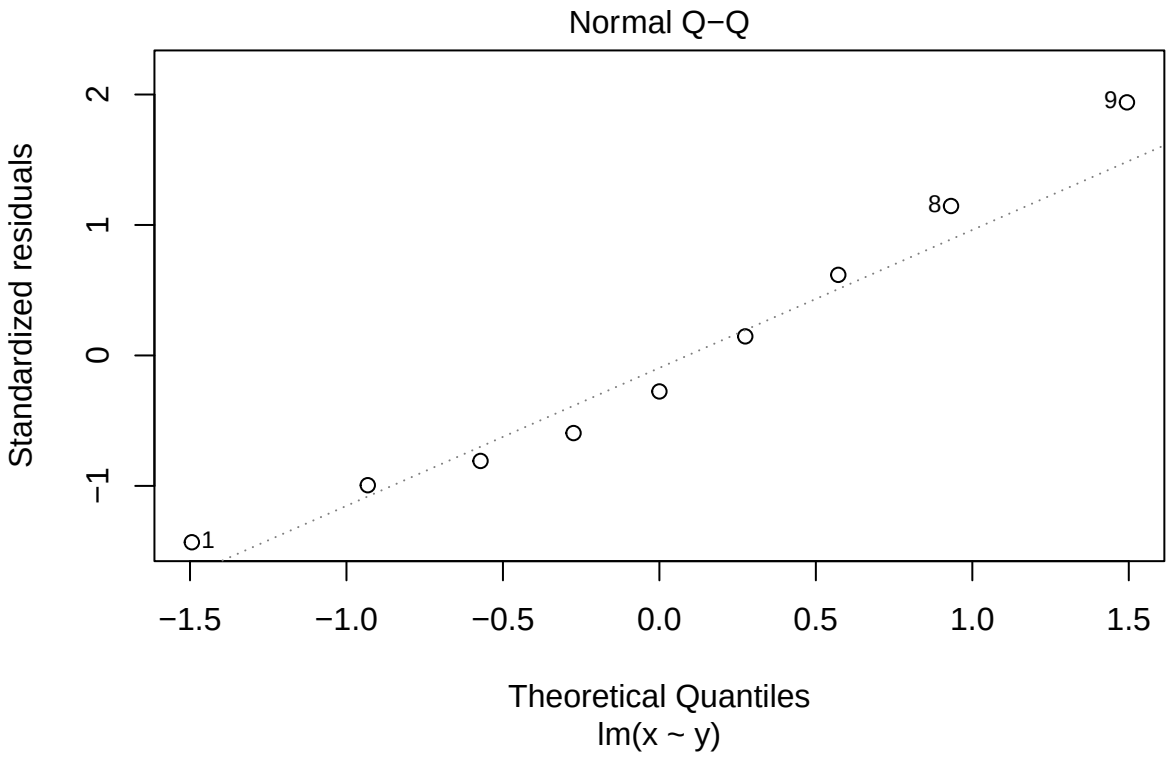
```
plot(x, q2rr)
```

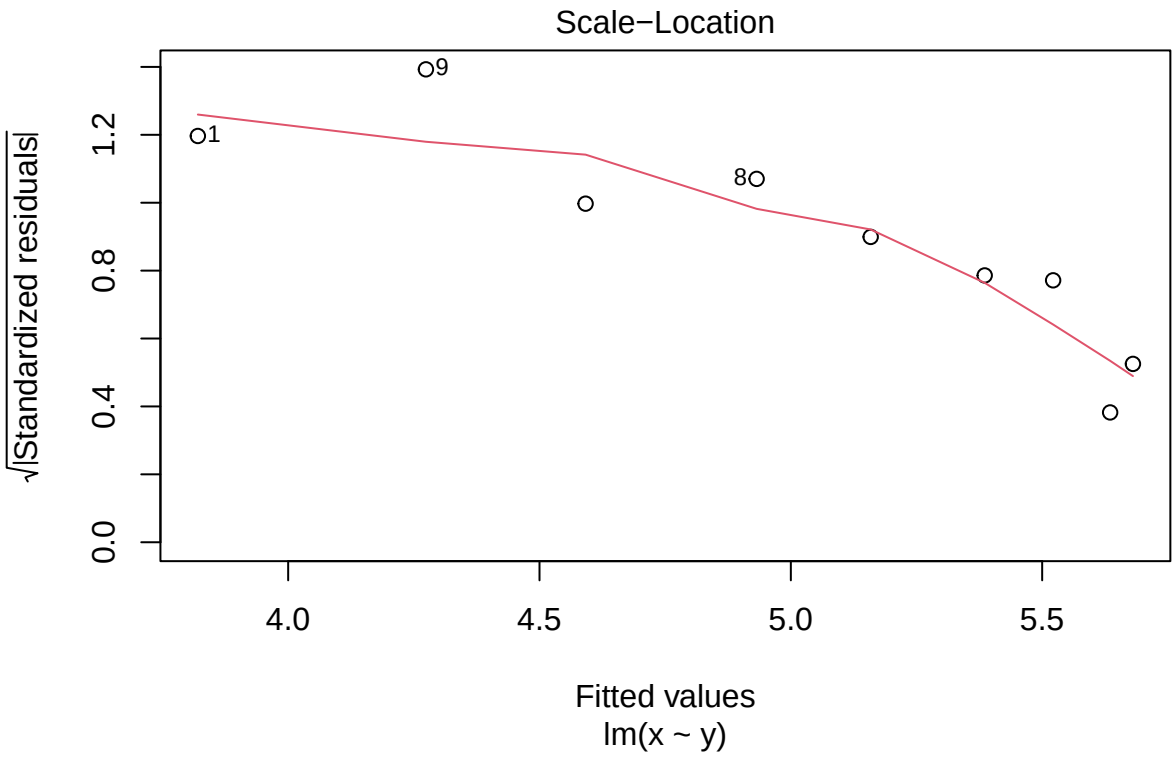


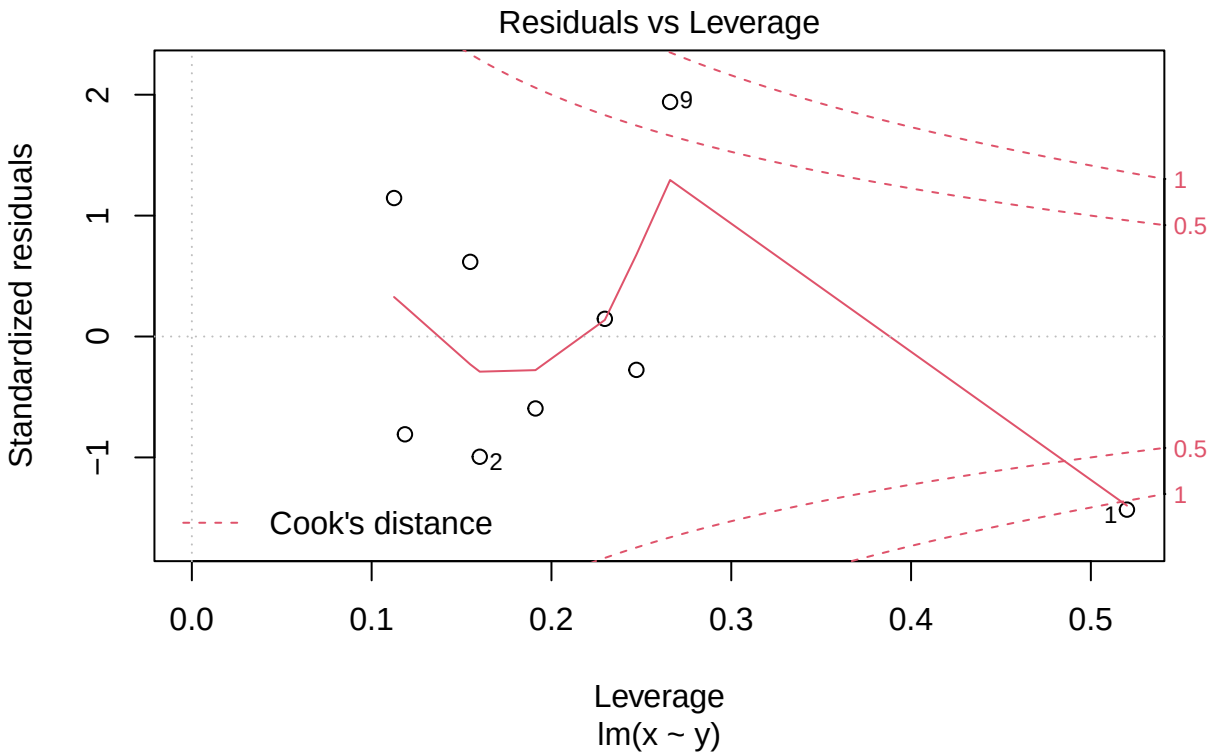
2d. Plot the raw residuals vs. y.

```
### residuals vs fitted values from the below plots
plot(q2lm)
```


Residuals vs Fitted
Residuals
4
2
0
−2
9
8
1
4.0
4.5
5.0
5.5
Fitted values
lm(x ~ y)


Normal Q−Q
Standardized residuals
2
1
0
−1
9
8
1
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles
lm(x ~ y)


Scale−Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
9
1
8
4.0
4.5
5.0
5.5
Fitted values
lm(x ~ y)



2e. Compare the plots from (b), (c), and (d). Is there a meaningful difference between (c) and (d)? Explain. Which of the plots (b) or (d) gives a better indication of the lack of fit? Explain.

We see a better indication of lack of fit from the plot in 2d because we see the fan shape of the data, indicating unequal variance. In 2c, we are plotting the independent variable which doesn't tell us much about the variance, and we don't see the same fan shape as 2d. In 2b, we curvature indicating the data is nonlinear instead of the fan shape we see in 2d.