Regression Analysis Examples in R

Chapter 2 - Simple Linear Regression

Example. Airfreight Data

	1	2	3	4	5	6	7	8	9	10
Shipment Route (x)	1	0	2	0	3	1	0	1	2	0
Airfreight Breakage (y)	16	9	17	12	22	13	8	15	19	11

a) Compute the ANOVA table
b) Compute the confidence intervals for the parameters
c) Compute the confidence interval on the average (mean) response when $X = 1$.
d) What is the total variability in y explained by this model?

Solution.

Part a.

We can compute the anova table manually as follows,

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 20 - \frac{1}{10}(100) = 10$$

$$S_{xy} = \sum_{i=1}^{n} y_i(x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i = 182 - \frac{1}{10}(10)(142) = 40$$

Therefore,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{40}{10} = 4$$

Then, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$, so

$$\hat{\beta}_0 = \frac{1}{10}(142) - 4 \cdot \frac{1}{10}(10) = 10.2$$

This gives us our linear model

$$\hat{y} = 10.2 + 4x$$

The sum of squares for regression is

$$SSR = \hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 = \hat{\beta}_1^2 S_{xx} = 16 \cdot 10 = 160$$

The total sum of squares is

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = 2194 - 10(14.2)^2 = 177.6$$

Then, the residual sum of squares is

$$SSE = SST - SSR = 177.6 - 160 = 17.6$$

Now we can construct the anova table

Source	Sum of Squares	DF	MS=SS/df	F = MSR/MSE
Regression	160	1	160	72.727
Error	17.6	8	2.2	
Total	177.6			

We conclude tbat the regression is highly significant since the F value is very large. We can also do this in R

```
x <- c(1,0,2,0,3,1,0,1,2,0)
y <- c(16,9,17,12,22,13,8,15,19,11)
model <- lm(formula = y ~ x)
print(model)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##        10.2          4.0
```

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1  160.0   160.0  72.727 2.749e-05 ***
## Residuals  8   17.6     2.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that we get the same results and reach the same conclusion, and we also get the p-value which is very small and we can conclude from there as well that the regression is highly significant.

Part b.

We can construct confidence intervals, first we need to compute $se(\hat{\beta}_1)$ and $se(\hat{\beta}_0)$.

$$se^2(\hat{\beta}_0) = MSE\left(\frac{1}{n} + \frac{\bar{x}}{S_{xx}}\right) = 2.2\left(\frac{1}{10} + \frac{1}{10}\right) = 0.44 \implies se(\hat{\beta}_0) = \sqrt{0.44} = 0.6633$$

$$se^2(\hat{\beta}_1) = \frac{MSE}{S_{xx}} = \frac{2.2}{10} = 0.22 \implies se(\hat{\beta}_1) = \sqrt{0.22} = 0.490$$

Then, we have to compute $t_{\alpha/2,n-2} = t_{0.025,8}$, we either use a t look up table or in R,

```r
qt(0.025, 8, lower.tail=FALSE)
```

```
## [1] 2.306004
```

Thus, our confidence intervals are

$$\hat{\beta}_0 - t_{\alpha/2,n-2}se(\hat{\beta}_0) \leq \hat{\beta}_0 \leq \hat{\beta}_0 + t_{\alpha/2,n-2}se(\hat{\beta}_0) \rightarrow 10.2 \pm 2.306(0.6633) = (8.6704, 11.7296)$$

$$\hat{\beta}_1 - t_{\alpha/2,n-2}se(\hat{\beta}_1) \leq \hat{\beta}_1 \leq \hat{\beta}_1 + t_{\alpha/2,n-2}se(\hat{\beta}_1) \rightarrow 4 \pm 2.306(0.490) = (2.9392, 5.0608)$$

We can compute these confidence intervals in R as well

```r
confint(model, level=0.95)
```

```
##                2.5 %    97.5 %
## (Intercept) 8.670370 11.729630
## x           2.918388  5.081612
```

Part c.

We want to compute first $E(y|x_0)$, where $x_0 = 1$. An unbiased estimator for $E(y|x_0)$ is

$$\widehat{E(y|x_0)} = \hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 10.2 + 4(1) = 14.2$$

Then, the confidence interval is

$$\left[\hat{\mu}_{y|x_0} \pm t_{\alpha/2,n-2}\sqrt{MSE\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}\right] = \left[14.2 \pm 2.306\sqrt{2.2\left(\frac{1}{10} + \frac{(1-1)^2}{S_{xx}}\right)}\right] = (13.11839, 15.28161)$$

We can do this in R with

```r
predict(model, newdata = data.frame(x=1), interval = 'confidence', level=0.95)
```

```
##    fit      lwr      upr
## 1 14.2 13.11839 15.28161
```

Part d.

The total variability in y explained by the regressor x is measured by the coefficient of determination

$$R^2 = \frac{SSR}{SST} = \frac{160}{177.6} = 0.9009$$

We can also see this in the summary of the model in R

```r
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

```
##   -2.2   -1.2    0.3    0.8    1.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.2000     0.6633  15.377 3.18e-07 ***
## x             4.0000     0.4690   8.528 2.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.483 on 8 degrees of freedom
## Multiple R-squared:  0.9009, Adjusted R-squared:  0.8885
## F-statistic: 72.73 on 1 and 8 DF,  p-value: 2.749e-05
```

The R^2 value is 0.9009.

Chatper 3 - Multiple Linear Regression

Question 3.1

Consider the National Football League data in Table B.1

a. Fit a multiple linear regression model relating the number of games won to the team's passing yardage (x_2), the percentage of rushing plays (x_7), and the opponents' yards rushing (x_8).
b. Construct the analysis-of-variance table and test for significance of regression.
c. Calculate t statistics for the hypotheses $H_0 : \beta_2 = 0$, $H_0 : \beta_7 = 0$, and $H_0 : \beta_8 = 0$. What conclusions can you draw about the roles of variables in x_2, x_7, and x_8 play in the model?
d. Calculate R^2 and R^2_{Adj} for this model.
e. Using the partial F test, determine the contribution of x_7 to the model. How is the partial F statistic related to the t test for β_7 calculated in part c above?

Question 3.3

Refer to problem 3.1

a. Find a 95% CI for β_7.

b. Find a 95% CI on the mean number of games won by a team when $x_2 = 2300$, $x_7 = 56$, and $x_8 = 2100$.

Question 3.4

Reconsider the National Football League data from Problem 3.1. Fit a model to these data using only x_7 and x_8 as regressors.

a. Test for significance of regression.

b. Calculate R^2 and R^2_{Adj}. How do these quantities compare to the value computed for the model in Problem 3.1, which included an additional regressor (x_2)?

c. Calculate a 95% CI on β_7. Also find a 95% CI on the mean number of games won by a team when $x_7 = 56$, and $x_8 = 2100$. Compare the lengths of these CIs to the lengths of the corresponding CIs from Problem 3.3.

d. What conclusions can you draw from this problem about the consequences of omitting an important regressor from a model?

Solutions.

Question 3.1

Part a.

We can fit a linear model using the same R function,

```
# Table b1 was loaded ahead of time.
model <- lm(formula = y ~ x2 + x7 + x8, tableb1)
model
```

```
##
## Call:
## lm(formula = y ~ x2 + x7 + x8, data = tableb1)
##
## Coefficients:
## (Intercept)           x2           x7           x8
##   -1.808372     0.003598     0.193960    -0.004815
```

This gives us our linear model with the estimates $\hat{\beta}_0 = -1.808$, $\hat{\beta}_2 = 0.00360$, $\hat{\beta}_7 = 0.194$, and $\hat{\beta}_8 = -0.00482$.

$$y = -1.808 + 0.00360x_2 + 0.194x_7 - 0.00482x_8$$

Part b.

We test for significance of regession using the hypotheses

$$H_0 : \beta_2 = \beta_7 = \beta_8 = 0, \;\; H_1 : \beta_j \neq 0, j = 2, 7, 8$$

We use the F-statistic

$$F_0 = \frac{SSR/k}{SSE/(n-p)} = \frac{MSR}{MSE} \sim F_{k,n-p}$$

We reject the null hypothesis when $F > F_{\alpha,k,n-k-1}$, we compute these values in R. In this case, we have $k = 3$ regressors and coefficients, so $p = k + 1 = 4$, thus

```
n <- nrow(tableb1)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x2         1  76.193  76.193  26.172 3.100e-05 ***
## x7         1 139.501 139.501  47.918 3.698e-07 ***
## x8         1  41.400  41.400  14.221 0.0009378 ***
## Residuals 24  69.870   2.911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
qf(0.05, 3, n-4, lower.tail=FALSE)
```

```
## [1] 3.008787
```

Source	Sum of Squares	DF	MS	F	P
x_2	76.193	1	76.193	26.172	$3.1 \cdot 10^{-5}$
x_7	139.501	1	139.501	47.918	$3.698 \cdot 10^{-7}$
x_8	41.400	1	41.400	14.221	0.0009378
Residuals	69.8790	24	2.911		

$$F_0 = \frac{SSR/k}{SSE/(n-p)} = \frac{(76.193 + 139.501 + 41.4)/3}{69.870/24} = 29.439$$

We can also obtain the F-statistic from the summary of the model,

```r
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x2 + x7 + x8, data = tableb1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0370 -0.7129 -0.2043  1.1101  3.7049
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.808372   7.900859  -0.229 0.820899
## x2           0.003598   0.000695   5.177 2.66e-05 ***
## x7           0.193960   0.088233   2.198 0.037815 *
## x8          -0.004816   0.001277  -3.771 0.000938 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.706 on 24 degrees of freedom
## Multiple R-squared:  0.7863, Adjusted R-squared:  0.7596
## F-statistic: 29.44 on 3 and 24 DF,  p-value: 3.273e-08
```

Therefore, we reject the null hypothesis $H_0 : \beta_2 = \beta_7 = \beta_8 = 0$, and conclude our regression is significant.

Part c.

We want to conduct tests on the indivudal coefficients, with the hypotheses $H_0 : \beta_2 = 0$, $H_0 : \beta_7 = 0$, and $H_0 : \beta_8 = 0$. We need the t-statistic

$$t_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

We reject the null hypothesis when $|t_0| > t_{\alpha/2, n-p}$,

```r
qt(0.025, n-4, lower.tail=FALSE)
```

```
## [1] 2.063899
```

We have all the information we need however in the summary of the model, we can see the estimate and the standard error for each coefficient, which tells us the t-value, but also the t-value is included. We can see that for all coefficents and their respective t-values, $|t_0| > t_{\alpha/2,n-p}$ so we reject all 3 of the null hypotheses.

Part d.

The R^2 value can be computed with

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{76.193 + 139.501 + 41.4}{76.193 + 139.501 + 41.4 + 69.8790} = 0.7863$$

We get these values from the ANOVA table above and also in the summaryof our model, we have $R^2 = 0.7863$, and the adjusted R^2 value is

$$R^2_{Adj} = 1 - \frac{SSE/(n-k)}{SST/(n-1)} = 1 - \frac{68.8790/24}{326.973/27} = 0.7596$$

This value is also in the summary R output above.

Part e.

To conduct the partial F test to determine the contribution of x_7, we want to tes tthe hypotheses

$$H_0 : \beta_7 = 0, \ H_1 : \beta_7 \neq 0$$

We fit the model assuming the null hypothesis is true to get the reduced model and obtain the anova table,

```r
reduced_model <- lm(formula = y ~ x8 + x2, data=tableb1)
anova(reduced_model)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x8         1 178.092 178.092  53.043 1.245e-07 ***
## x2         1  64.934  64.934  19.340 0.0001775 ***
## Residuals 25  83.938   3.358
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Then, we want to use the F statistic to test the hypotheses

$$F_0 = \frac{SSR(\beta_7|\beta_2, \beta_0)/r}{MSE}$$

So, we have

$$SRR(\beta_7|\beta_2, \beta_8) = SSR(\beta_7, \beta_2, \beta_8) - SSR(\beta_2, \beta_8) = 257.094 - (178.092 + 64.934) = 14.064$$

Therefore,

$$F_0 = \frac{SSR(\beta_7|\beta_2, \beta_8)}{MSE} = \frac{14.064}{2.911} \approx 4.831$$

we reject the null hypothesis if $F_0 > F_{\alpha,r,n-p}$,

```
qf(0.05, 1, 24, lower.tail=FALSE)
```

```
## [1] 4.259677
```

We conclude that x_7 contributed significantly to this model. We also notice that the F statistic is the square of the t test used in part c.

Question 3.3

Part a.

To consturct a confidence interval on β_7, we need to compute

$$\hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \leq \beta_j \leq \hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}}$$

We have the standard error for β_7 from the previous summary output from R, so $se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}} = 0.088233$. Then, the t-value was also obtained earlier and we have $t_{\alpha/2,n-p} = 2.063899$, and the coefficient estimator $\hat{\beta}_7 = 0.193960$. Therefore, the confidence interval is

$$0.193960 \pm 2.063899 \cdot (0.088233) = (0.011856, 0.376064)$$

This can be also obtained in R

```
confint(model, "x7")
```

```
##         2.5 %    97.5 %
## x7 0.01185532 0.3760651
```

Part b.

The mean response at $x_2 = 2300$, $x_7 = 56$ and $x_8 = 2100$ is

$$y_0 = \beta_0 + \beta_2(2300) + \beta_7(56) + \beta_8(2100) = 7.215188$$

Then, the confidence interval on the mean response is

$$\left[\hat{y}_0 \pm t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 x_0'(X'X)^{-1}x_0}\right]$$

This can be calculated in R with

```
predict(model, newdata=data.frame(x2 = 2300, x7 = 56, x8 = 2100),
        interval = 'confidence', level=0.95)
```

```
##        fit      lwr      upr
## 1 7.216424 6.436203 7.996645
```

Question 3.4

Part a

We want to test the hypotheses for significance,

$$H_0 : \beta_7 = \beta_8 = 0,\ H_1 : \beta_j \neq 0,\ j = 7, 8$$

We can fit the model and get the anova table,

```
model <- lm(formula = y ~ x7 + x8, tableb1)
anova(model)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## x7         1  97.238  97.238  16.437 0.000431 ***
## x8         1  81.828  81.828  13.832 0.001015 **
## Residuals 25 147.898   5.916
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see from the anova table that the regression is highly significant, but we can also use the F test statistic,

$$F_0 = \frac{SSR/k}{SSE/(n-k-1)} = \frac{(97.238 + 81.828)/2}{147.898/25} = \frac{89.533}{5.916} = 15.134$$

Then, we can compute $F_{\alpha,k,n-k-1}$,

```
qf(0.05, 3, 25, lower.tail=FALSE)
```

```
## [1] 2.991241
```

We can see that $F_0 > F_{\alpha,k,n-k-1}$ so we reject the null hypotheses that $H_0 : \beta_7 = \beta_8 = 0$, and conclude that the regression is significant.

Part b.

We can obtain a summary for the model and look at the R^2 and R^2_{Adj} values similar to the previous questions.

```
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x7 + x8, data = tableb1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7985 -1.5166 -0.5792  1.9927  4.5248
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.944319   9.862484   1.819  0.08084 .
## x7           0.048371   0.119219   0.406  0.68839
## x8          -0.006537   0.001758  -3.719  0.00102 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.432 on 25 degrees of freedom
## Multiple R-squared:  0.5477, Adjusted R-squared:  0.5115
## F-statistic: 15.13 on 2 and 25 DF,  p-value: 4.935e-05
```

We get an R-squared value $R^2 = 0.5477$ and the adjusted R-squared is $R^2_{Adj} = 0.5115$, we can see that these values are lower than when we had x_2 in the model. So, the model with x_2 was able to better explain the variablity in y and this suggests x_2 may have been contributing significantly to the model.

Part c.

```
confint(model, "x7")
```

```
##         2.5 %   97.5 %
## x7 -0.1971643 0.293906
```

A 95% confidence interval on β_7 is

$$(-0.1971643, 0.293906)$$

```
new_data <- data.frame(x7 = 56, x8=2100)
predict(model, newdata=new_data, interval='confidence', level=0.95)
```

```
##        fit      lwr      upr
## 1 6.926243 5.828643 8.023842
```

Our 95% confidence interval on the mean number of games one when $x_7 = 56$ and $x_8 = 2100$ is

$$(5.828643, 8.023842)$$

We can see that the length of both confidence intervals are greater than when x_2 was included in the model. This suggests we were more confident with our estimates when x_2 was included.

d.

We can conclude that omitting an important regressor (x_2) affected our estimates and standard error of coeffficients, resulting in larger lenghts in the confidence intervals and lower values for R^2.

Chapter 4 - Model Adequecy

Question 4.1

Consider the simple regression model fit to the National Football League team performance data in Problem 2.1. (Same data as previous questions, with the model y $x8$.

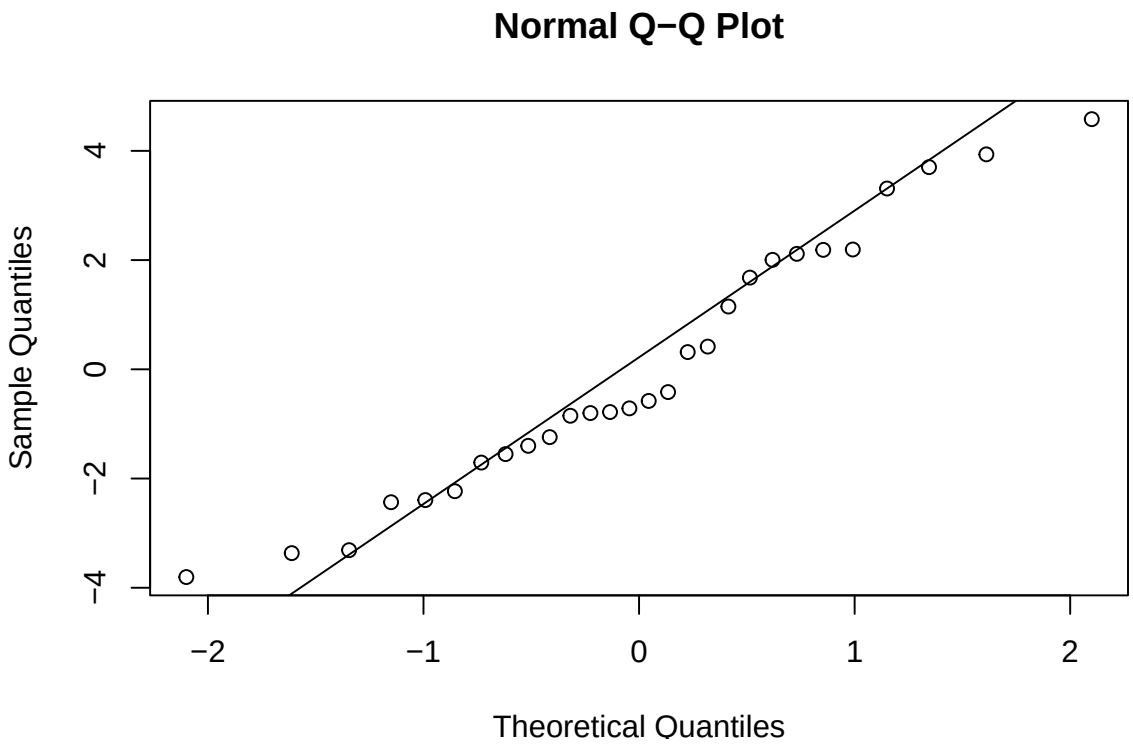
a. Construct a normal probability plot of the residuals. Does there seem to be any problem with the normality assumption?

b. Construct an interpret a plot of the residuals versus the predicted repsonse.

c. Plot the residuals versus the team passing yardage, x_2. DOes this plot indicate that the model will be improved by adding x_2 to the model?

Part a.

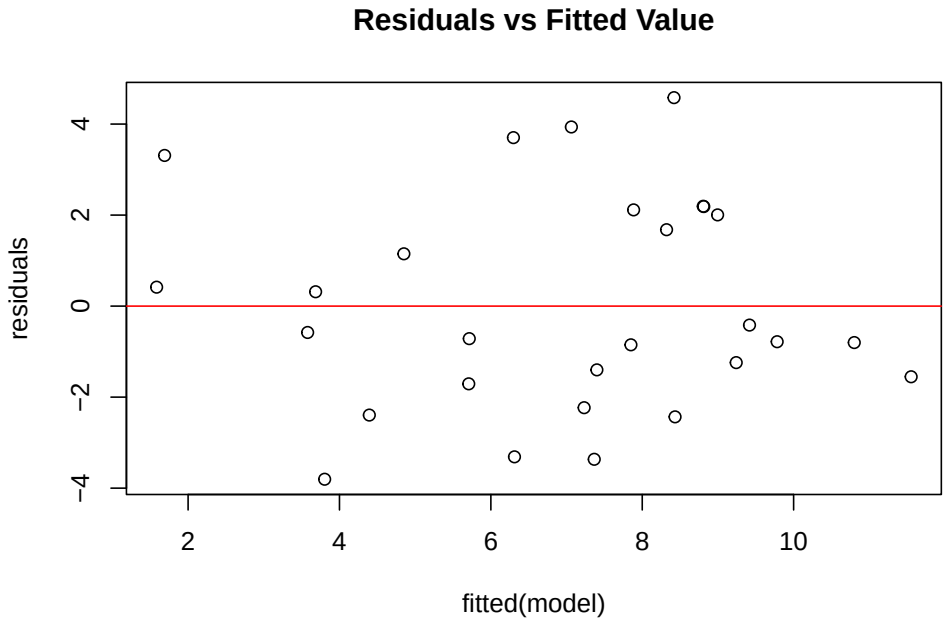
```
library(ggplot2)
model <- lm(formula = y ~ x8, tableb1)
residuals <- residuals(model)
qqnorm(residuals)
qqline(residuals)
```



It appears that the normality assumption is fine since the standardized residuals closely follow the theoretical quantiles for a normal distribution as observered in the quantile-quantile plot.

Part b.

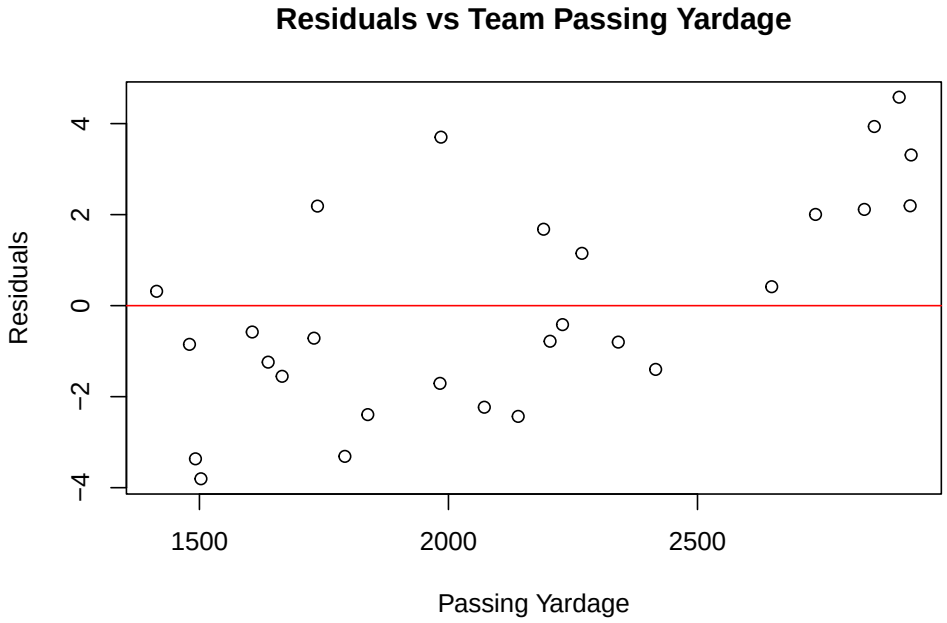
```
plot(fitted(model), residuals, main = "Residuals vs Fitted Value")
abline(h=0, col = 'red')
```



The model appears to be adequate since the residuals vs fitted values appear to be randomly distributed around 0, showing no patterns.

Part c.

```
residuals <- residuals(model)
plot(tableb1$x2, residuals, xlab="Passing Yardage",
     ylab="Residuals", main="Residuals vs Team Passing Yardage")
abline(h=0, col="red")
```



The model appears to be improved when adding x_2 since the residuals appear to be more close to 0.

Lack of Fit Example

We are going to use this data to conduct a lack of fit test.

x	1	1	2	3.3	3.3	4	4	4	4.7	5
y	10.84	9.30	16.35	22.88	24.35	24.56	25.86	29.16	24.59	22.25
x	5.6	5.6	5.6	6.0	6.0	6.5	6			
y	25.90	27.20	25.61	25.45	26.56	21.03	21.46			

We can count that there are $m = 10$ levels, and $n = 17$ observations, so

$$\sum_{i=1}^{n} n_i = 2 + 1 + 3 + \cdots = 17$$

We can compute then fit a model with this data in R,

```
x <- c(1,1,2,3.3,3.3,4,4,4,4.7,5,5.6,5.6,5.6,6,6,6.5,6)
y <- c(10.84 , 9.30 , 16.35 ,22.88  ,24.35 , 24.56, 25.86 ,29.16
      ,24.59, 22.25,25.90 , 27.20 , 25.61 , 25.45 , 26.56 , 21.03 ,21.46)
model <- lm(formula = y ~ x)
print(model)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##      12.523        2.316
```

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x          1 260.44 260.439  17.196 0.0008606 ***
## Residuals 15 227.17  15.145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We get the linear model

$$\hat{y}_i = 12.5323 + 2.316x_i$$

Then, we can compute lack of fit and pure error with

$$SS_{PE} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2, SS_{LOF} = SSE - SS_{PE}$$

In R, we get

```r
library(EnvStats)
anovaPE(model)
```

```
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## x              1 260.439 260.439 71.0262 2.996e-05 ***
## Lack of Fit    7 197.840  28.263  7.7078  0.004971 **
## Pure Error     8  29.334   3.667
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our F-statistic is 7.70, and the critical value is $F_{\alpha,m-2,n-m}$, which we can compute in R or look at a table.

```r
qf(0.05, 8, 7, lower.tail=FALSE)
```

```
## [1] 3.725725
```

Therefore, we reject the null hypothesis $H_0 : E(y_i) = \beta_0 + \beta_1 x_i$.

Interprolating Values From Look-up Table

Say we want to find the t-value for t_{α_0} where α_0 is not located on the look up table, we then choose the 2 closest α values, call them α_1, α_2, so that $\alpha_1 < \alpha_0 < \alpha_2$. Then, we use the following formula to interprolate the t-value as

$$t_{\alpha_0} \approx t_{\alpha_1} + \frac{(\alpha_0 - \alpha_1)(t_{\alpha_2} - t_{\alpha_1})}{\alpha_2 - \alpha_1}$$

Example.

Suppose you have the critical value $\alpha_0 = 0.015$, the 2 closest values would be $\alpha_1 = 0.01$, and $\alpha_2 = 0.025$. So, we get

$$t_{0.015} \approx \frac{(0.015 - 0.01)(t_{0.025} - t_{0.01})}{0.025 - 0.01}$$

```r
noint <- lm(formula = y~ 0)
print(noint)
```

```
##
## Call:
## lm(formula = y ~ 0)
##
## No coefficients
```

```r
anova(noint)
```

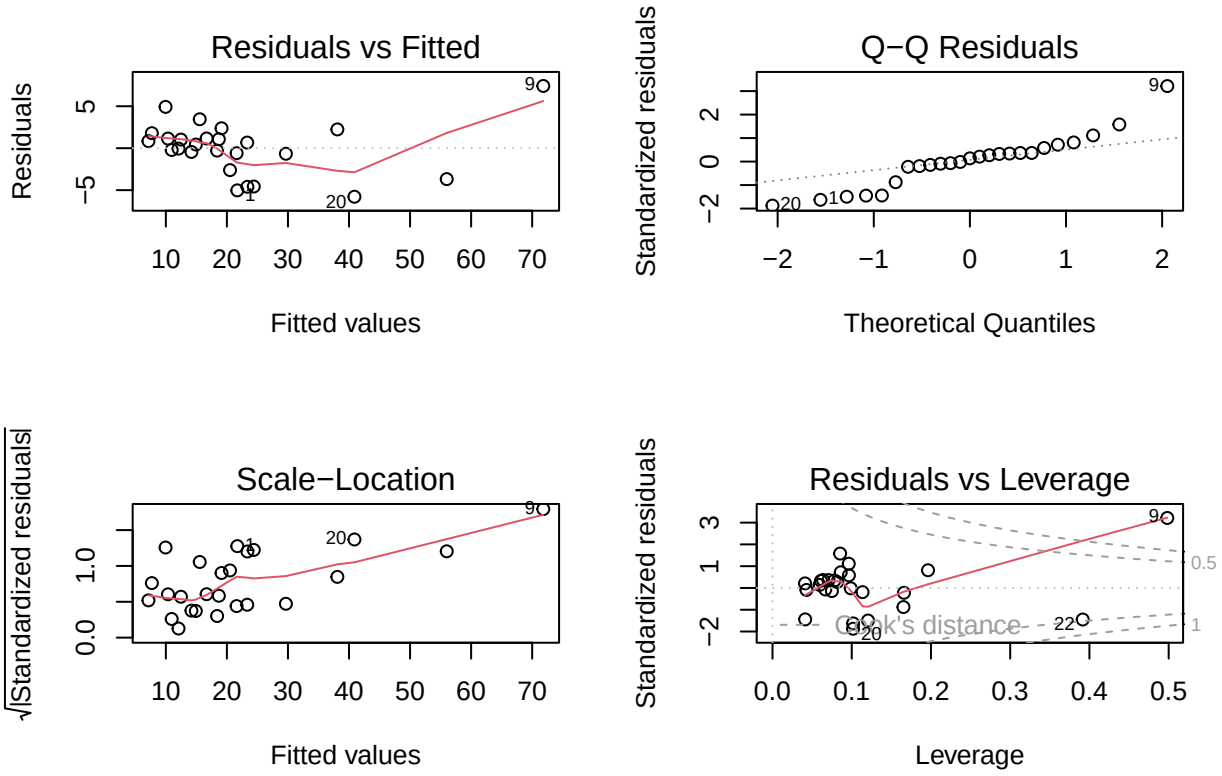
```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value Pr(>F)
## Residuals 17 9132.2  537.19
```

Chapter 5 - Weighted Least Squared and Transformations

Example of checking for model adequecy with BP test

Using the delivery time data, we can examine the model adequecy of

```
delivery <- read.table("./Data/delivery.txt",header=TRUE, sep='\t')
X1 = delivery$Number.of.Cases
X2 = delivery$Distance
Y = delivery$Delivery.Time
fit = lm(Y~X1+X2, data=delivery)
par(mfrow=c(2,2))
plot(fit)
```



We can see in the qq-plot that the residuals do not follow the theoretical quantiles of a normal distribution, so there is likely an issue with the normality assumption of the model. The residuals vs fitted appear to have some pattern resembling a parabola.

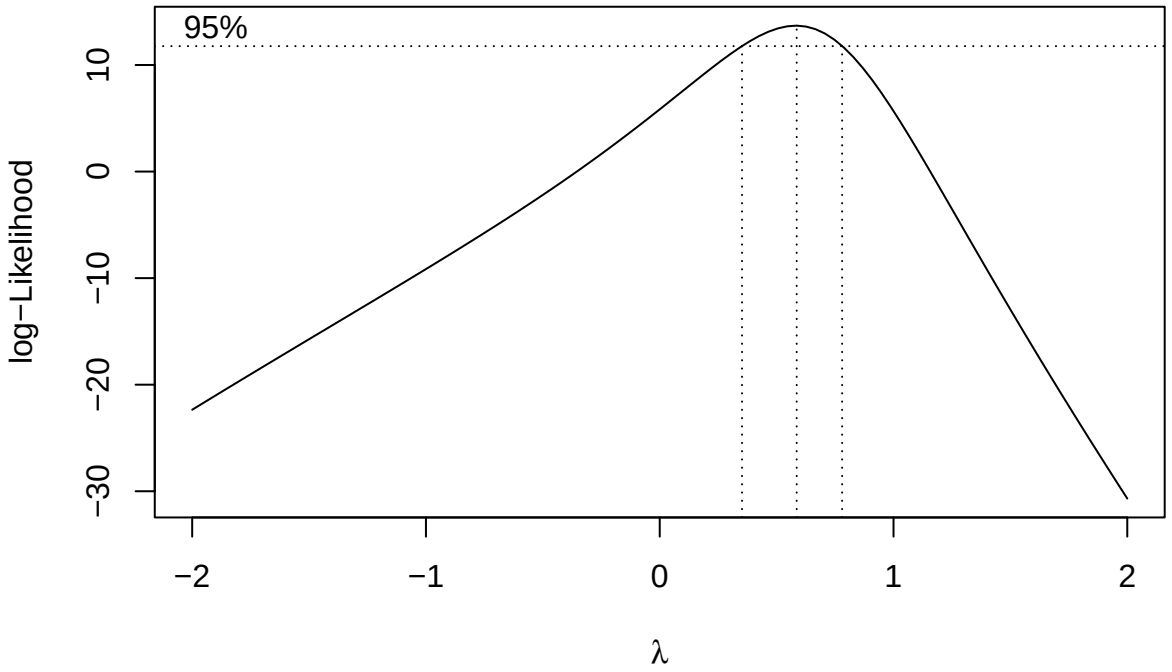
```
library(lmtest)
bptest(fit)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit
## BP = 11.988, df = 2, p-value = 0.002493
```

We can see that the p-value that it is low, so we reject the null hypothesis that the variance is constant.

We can attempt to remede the issue with the residuals by applying a boxcox transformation.

```
library(MASS)
boxcox(fit)
b <- boxcox(fit)
```

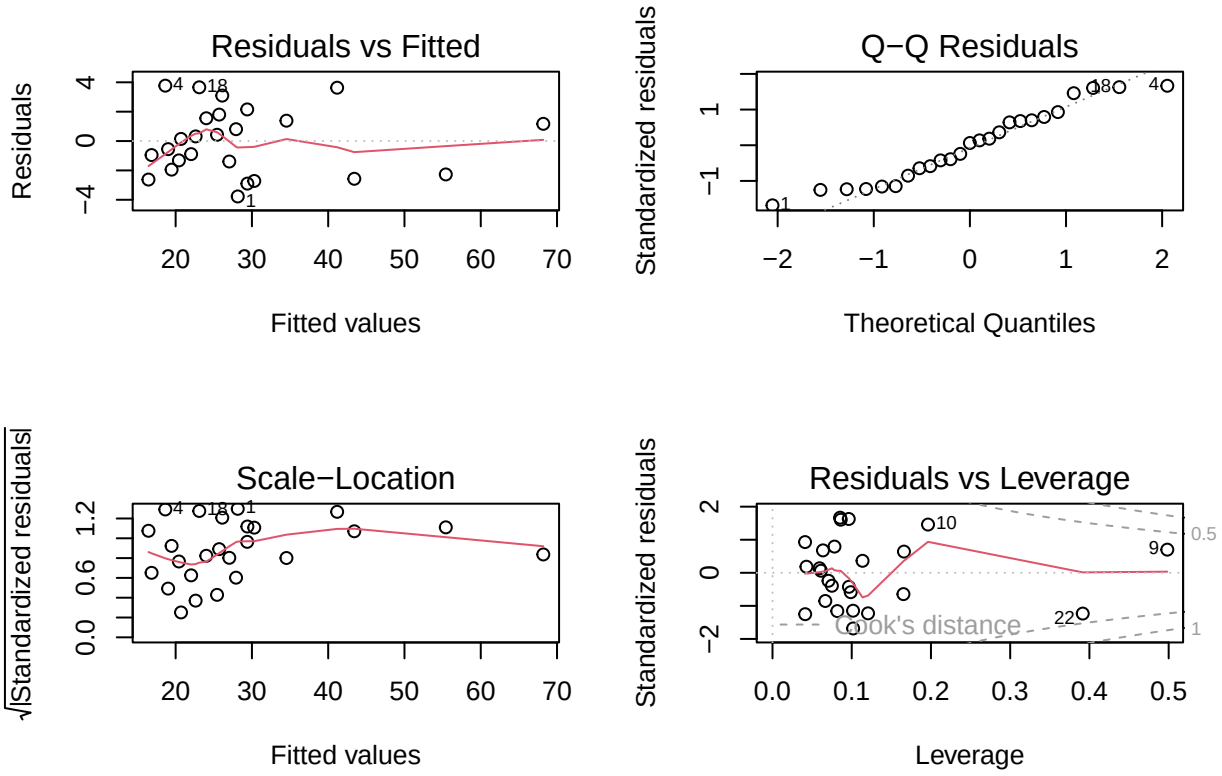



```
lambda <- b$x[which.max(boxcox(fit)$y)]
lambda
```

```
## [1] 0.5858586
```

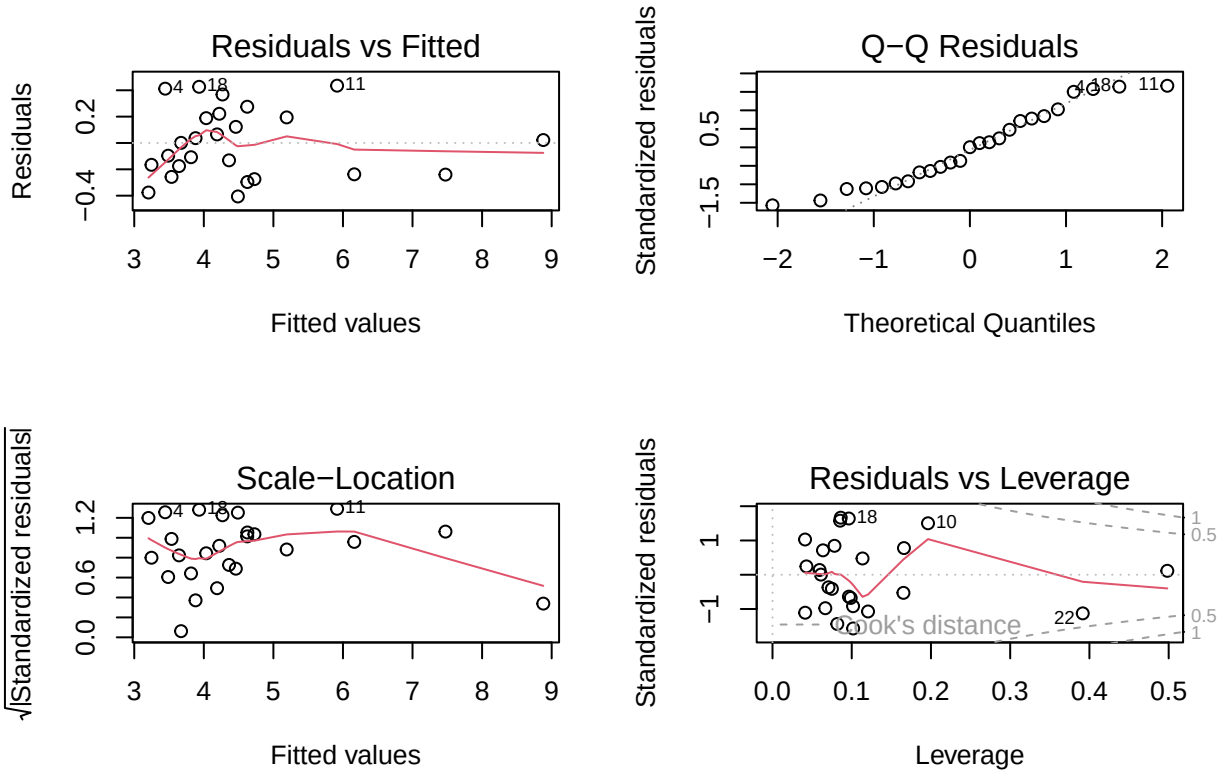
Now using our lambda value, we can transform y and refit the model,

```
n = length(Y)

y_dot <- exp(1/n * sum(log(Y)))

Y_transformed <- (Y^lambda - 1) / (lambda * y_dot^(lambda - 1))

new_model <- lm(Y_transformed ~ X1 + X2, data=delivery)

par(mfrow=c(2,2))

plot(new_model)
```



We can see that the qq-plot looks better now and the residuals appear fairly normally distributed, as well as the residuals vs fitted look more random, so we also can conclude that there is no longer an issue with the constant variance assumption. Another approach to this problem is to use another transformation on y. Since the reponse variable y is the time, which is count, the simplest probabilistic model for count data is the Poisson distribution, thus we transform $y' = \sqrt{y}$, and we plot this model.

```
y_prime <- sqrt(Y)

sqrt_y_model <- lm(y_prime ~ X1 + X2, data=delivery)

par(mfrow=c(2,2))

plot(sqrt_y_model)
```



We see that we get fairly similar results to the boxcox transformation, and can conlude the assumption of constant variance is no longer violated.

Example of Weighted Least Squares

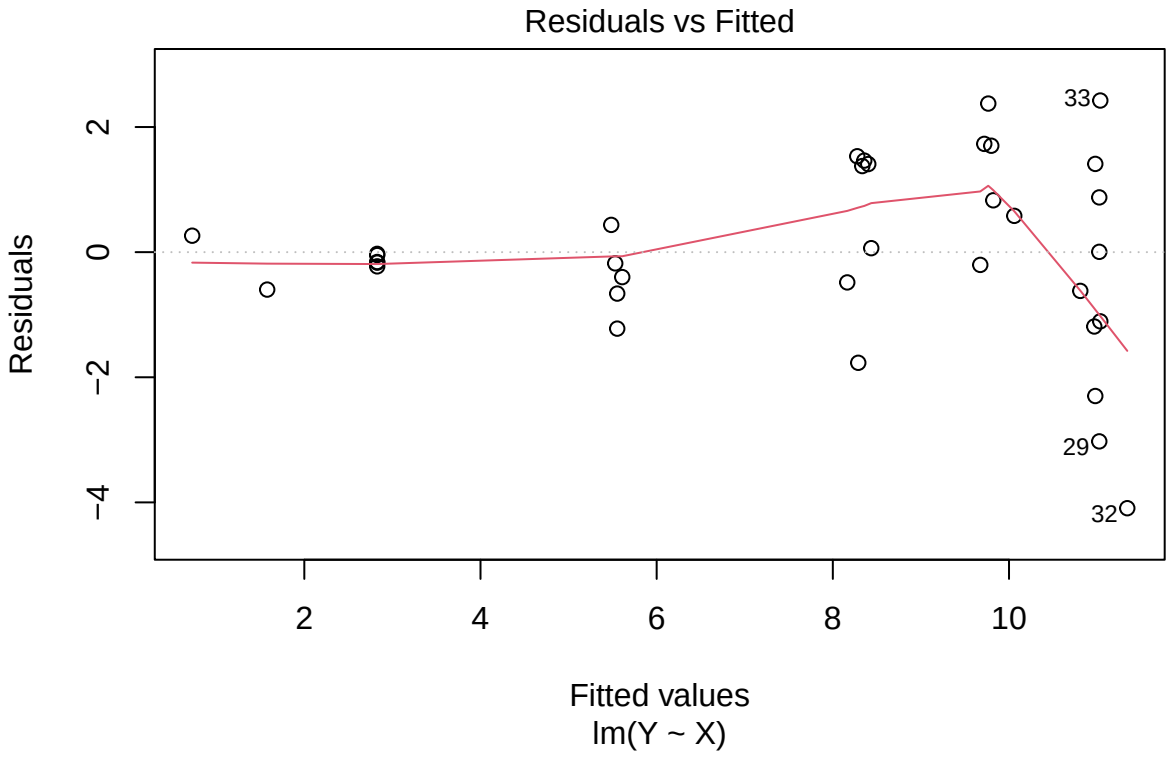
Using the weighted Turkey data, we can fit a model an examine the residuals.

```
turkey <- read.table("./Data/weighted.txt",header=TRUE, sep='\t')



fit <- lm(Y ~ X, data=turkey)

summary(fit)
```

```
##
## Call:
## lm(formula = Y ~ X, data = turkey)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0928 -0.6087 -0.0473  1.1256  2.4238
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.57895    0.67919  -0.852      0.4
## X            1.13540    0.08622  13.169 1.09e-14 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.457 on 33 degrees of freedom
## Multiple R-squared:  0.8401, Adjusted R-squared:  0.8353
## F-statistic: 173.4 on 1 and 33 DF,  p-value: 1.089e-14
```

```
plot(fit,1)
```



As we can see, there appears to be a telescoping effect. One way to proceed is to perform the usual regression. Then, group the data using the X variable. Estimate the variances s_i^2 of the Y_i for each group. Then fit the variances against the averages of the X_i of the groups. Next we computed averages and variances for subsets of the data and then fitted the variances against the averages.

```
turkey
```

```
##         X     Y
## 1    1.15  0.99
## 2    1.90  0.98
## 3    3.00  2.60
## 4    3.00  2.67
## 5    3.00  2.66
## 6    3.00  2.78
## 7    3.00  2.80
## 8    5.34  5.92
## 9    5.38  5.35
## 10   5.40  4.33
## 11   5.40  4.89
```

```
## 12  5.45  5.21
## 13  7.70  7.68
## 14  7.80  9.81
## 15  7.81  6.52
## 16  7.85  9.71
## 17  7.87  9.82
## 18  7.91  9.81
## 19  7.94  8.50
## 20  9.03  9.47
## 21  9.07 11.45
## 22  9.11 12.14
## 23  9.14 11.50
## 24  9.16 10.65
## 25  9.37 10.64
## 26 10.17  9.78
## 27 10.18 12.39
## 28 10.22 11.03
## 29 10.22  8.00
## 30 10.22 11.90
## 31 10.18  8.68
## 32 10.50  7.25
## 33 10.23 13.46
## 34 10.03 10.19
## 35 10.23  9.93
```

We can see that we have many data points at 3, 5.4, 7.8, 9.1, and 10.2. These aren't perfect groupings but there are many points at this X value or very close to it, so we will use these as our groups. Now we can compute the variances at each of these points.

```
s1 <- round(var(turkey[turkey$X == 3, ]$Y),4)
s2 <- round(var(subset(turkey, X >= 5.34 & X <= 5.45)$Y),4)
s3 <- round(var(subset(turkey, X >= 7.7 & X <= 7.94)$Y),4)
s4 <- round(var(subset(turkey, X >= 9.03 & X <= 9.37)$Y),4)
s5 <- round(var(subset(turkey, X >= 10.03 & X <= 10.5)$Y),4)

s <- c(s1, s2, s3, s4, s5)

df <- data.frame(X = c(3, 5.4, 7.8, 9.1, 10.2), s = s)
df
```

```
##      X      s
## 1  3.0 0.0072
## 2  5.4 0.3440
## 3  7.8 1.7404
## 4  9.1 0.8683
## 5 10.2 3.8964
```

So now that we have our dataframe, we can fit a model with s as the response and the averages that we picked,

```r
fit2 <- lm(s ~ X + I(X^2), data=df)
summary(fit2)
```

```
##
## Call:
## lm(formula = s ~ X + I(X^2), data = df)
##
## Residuals:
##       1       2       3       4       5
## -0.1198  0.1980  0.5586 -1.2990  0.6621
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.53291    3.78395   0.405    0.725
## X           -0.73343    1.28494  -0.571    0.626
## I(X^2)       0.08826    0.09666   0.913    0.458
##
## Residual standard error: 1.116 on 2 degrees of freedom
## Multiple R-squared:  0.7427, Adjusted R-squared:  0.4853
## F-statistic: 2.886 on 2 and 2 DF,  p-value: 0.2573
```

Now, we have the regression model

$$\hat{s}^2 = 1.5329 - 0.7334\bar{X} + 0.0883\bar{X}^2$$

The weights are then computes as inverses of the predicted variances,

```r
weights <- 1/predict(fit2, newdata = data.frame(X = turkey$X))

weighted_model <- lm(Y ~ X, data=turkey, weights=weights)
summary(weighted_model)
```

```
##
## Call:
## lm(formula = Y ~ X, data = turkey, weights = weights)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8010 -0.5572  0.1544  0.9843  1.6397
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.88908    0.30058  -2.958  0.00569 **
## X            1.16469    0.05945  19.590  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.139 on 33 degrees of freedom
```

```
## Multiple R-squared:  0.9208, Adjusted R-squared:  0.9184
## F-statistic: 383.8 on 1 and 33 DF,  p-value: < 2.2e-16
```

The original model was

$$\hat{Y} = -0.579 + 1.14X$$

The new weighted model becomes

$$\hat{Y} = -0.89 + 1.16X$$

Chapter 6 - Regression Diagnostics and Measures of Influence

We will examine the bank dataset and fit a model to the data. Then examine the residuals and look for any outliers or influential points.

```
library(olsrr)
```

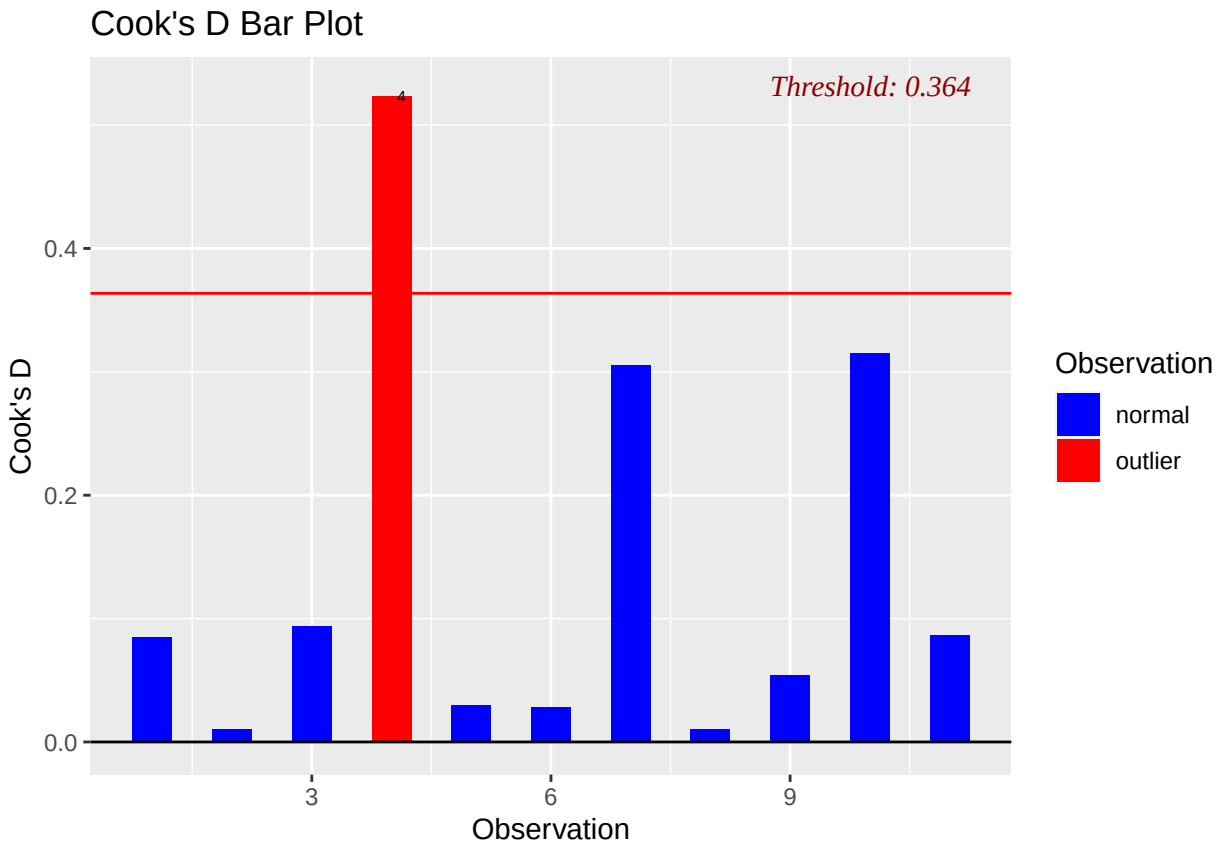
```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:MASS':
##
##     cement
```

```
## The following object is masked from 'package:MPV':
##
##     cement
```

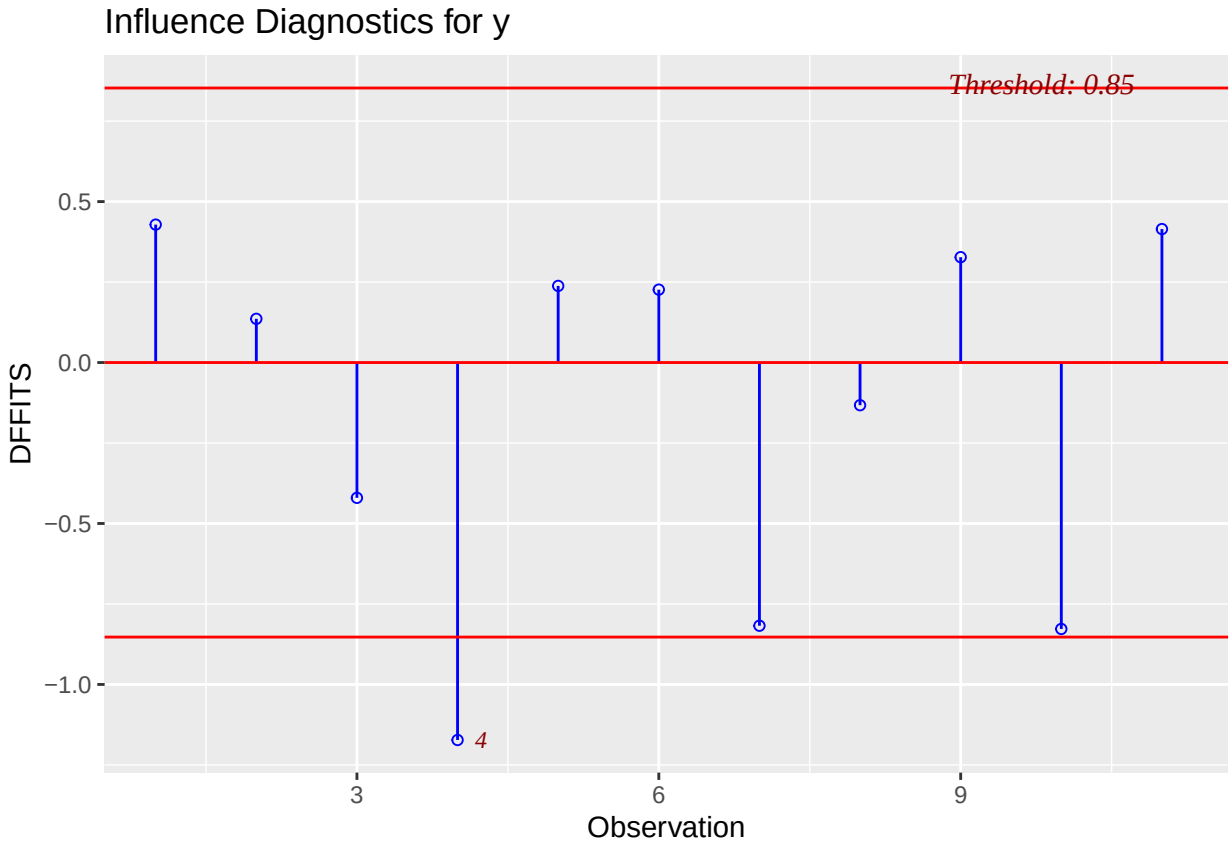
```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
bank <- read.table("./Data/bank.txt",header=TRUE, sep='\t')

y <- bank$Number.New.accounts
x <- bank$Minimum.Deposit
fit <- lm(y~x)
# Cook's Distance vs. Observations
ols_plot_cooksd_bar(fit)
```



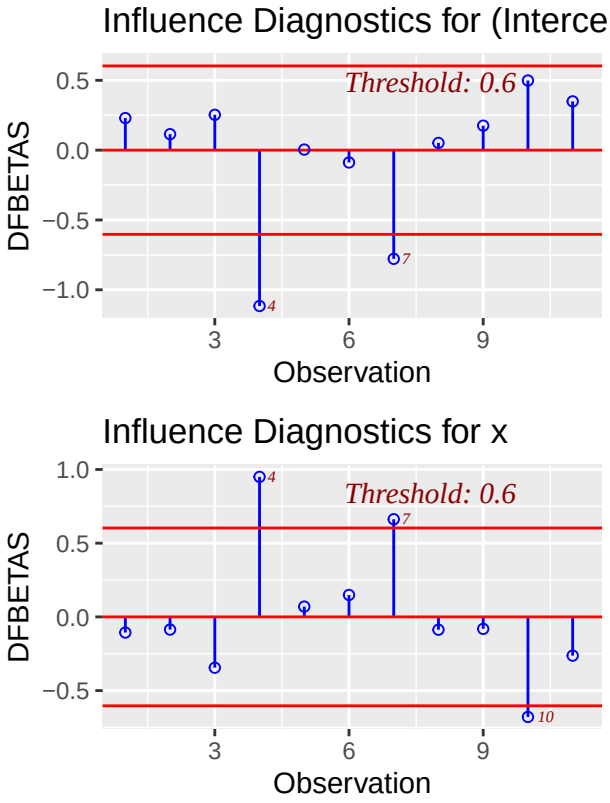
From the Cook's distant plot, we see that the 4th observation is influential on the fitted values at all X values. We can look at the plot for DFFITS,

```
ols_plot_dffits(fit)
```



We see again that the 4th observation is influential on the 4th fitted value of the model. We can also look at the DFBETAS plot,

```
ols_plot_dfbetas(fit)
```



From the DFBETAS vs Observation plots, we see that the 4th, 7th, and 10th observations are influential on the slope, and the 4th and 7th oberservations are influential on the intercept. To summarize, we can examine the measures of influence with

```
influence.measures(fit)
```

```
## Influence measures of
##   lm(formula = y ~ x) :
##
##      dfb.1_   dfb.x  dffit cov.r  cook.d    hat inf
## 1   0.22952 -0.1062  0.428 0.951 0.08506 0.0969
## 2   0.11433 -0.0860  0.136 1.454 0.01024 0.1518
## 3   0.25317 -0.3446 -0.420 1.566 0.09395 0.2775
## 4  -1.11603  0.9498 -1.173 0.787 0.52318 0.2644   *
## 5   0.00406  0.0698  0.238 1.241 0.02993 0.0995
## 6  -0.08838  0.1486  0.226 1.409 0.02790 0.1597
## 7  -0.77818  0.6623 -0.818 1.133 0.30507 0.2644
## 8   0.05173 -0.0870 -0.133 1.472 0.00977 0.1597
## 9   0.17533 -0.0811  0.327 1.108 0.05356 0.0969
## 10  0.49853 -0.6786 -0.828 1.171 0.31504 0.2775
## 11  0.34910 -0.2626  0.415 1.189 0.08634 0.1518
```

Chapter 7 - Polynomial and Indicator Regression

Example using Indicator Variables

We will use the turkey dataset

```
turkey <- read.table("./Data/turkey.txt",header=TRUE, sep='\t')
turkey
```

```
##    Age Weight Origin Z1 Z2
## 1   28   13.3      G  1  0
## 2   20    8.9      G  1  0
## 3   32   15.1      G  1  0
## 4   22   10.4      G  1  0
## 5   29   13.1      V  0  1
## 6   27   12.4      V  0  1
## 7   28   13.2      V  0  1
## 8   26   11.8      V  0  1
## 9   21   11.5      W  0  0
## 10  27   14.2      W  0  0
## 11  29   15.4      W  0  0
## 12  23   13.1      W  0  0
## 13  25   13.8      W  0  0
```

We can see that we have a catagorical variable "origin", which has 3 levels, G, V, and W. So, we create 2 dummy variables Z_1 and Z_2 to represent the levels. In this case, it was done for us with $(Z_1, Z_2) = (0,0) \implies W$, $(Z_1, Z_2) = (0,1) \implies V$, and $(Z_1, Z_2) = (1,0) \implies G$. 1 dummy variables allows for $2^1 = 2$ levels, 2 dummy variables gives us $2^2 = 4$ levels, and so on.

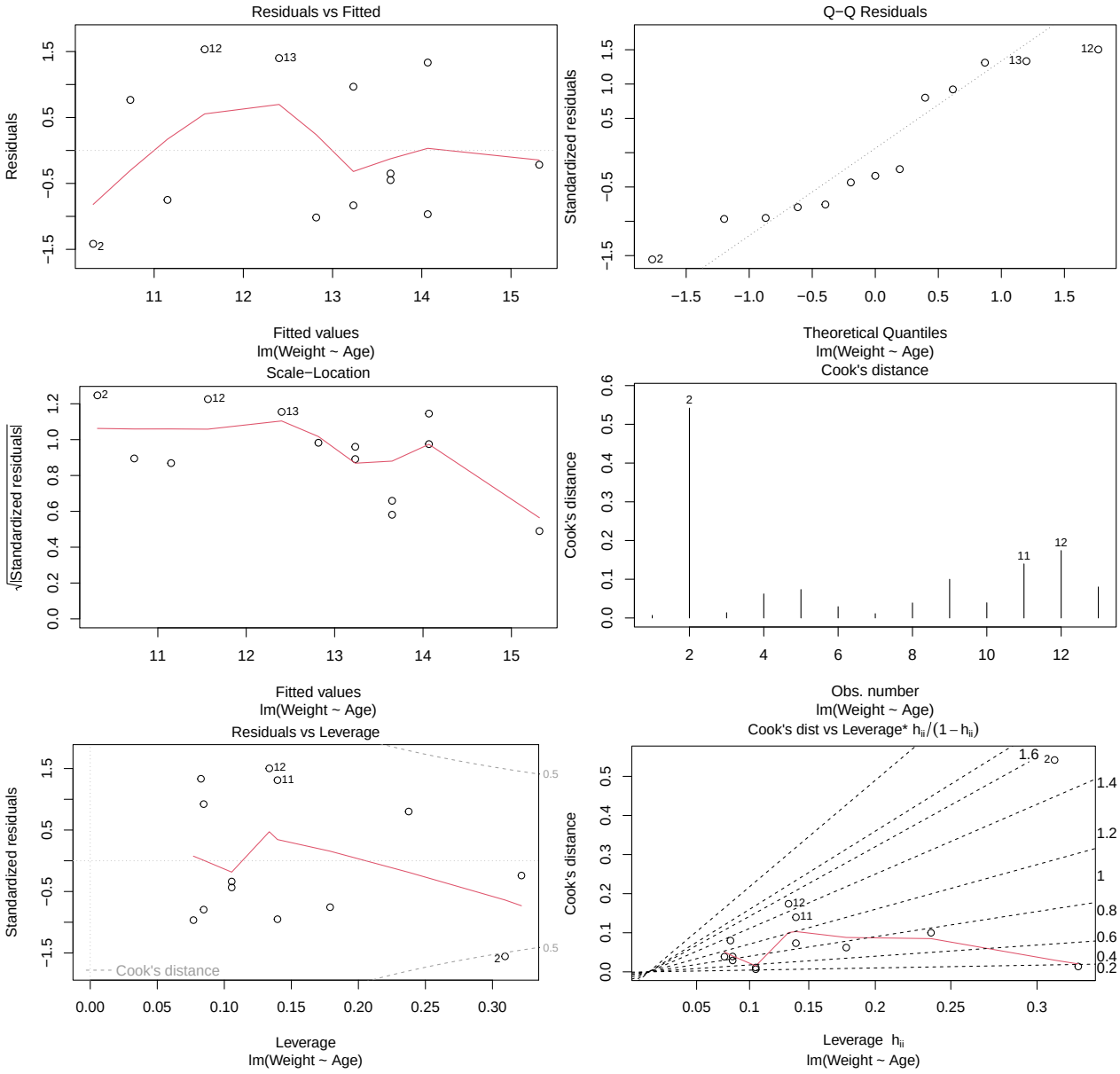
We can plot a model using just age and weight,

```
model1 <- lm(Weight ~ Age, data=turkey)
summary(model1)
```

```
##
## Call:
## lm(formula = Weight ~ Age, data = turkey)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4167 -0.8333 -0.3500  0.9667  1.5333
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.98333    2.33273    0.85 0.413327
## Age          0.41667    0.08922    4.67 0.000682 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.096 on 11 degrees of freedom
## Multiple R-squared:  0.6647, Adjusted R-squared:  0.6343
```

```
## F-statistic: 21.81 on 1 and 11 DF,  p-value: 0.0006824
```

We see that there is a significant relationship between age and weight, the R^2 value is a bit low at 0.66, we can examine the model adequency.



We can see from the qq-plot, that the normally assumption of the residuals may not be reasonable. We can also see from the residuals vs fitted plot that there is a curve in the data, which tells us that our linear model may not be a good fit. Using the criteria that an observation is influential when Cook's distance $D_i > 1$, we see there are no influential points.

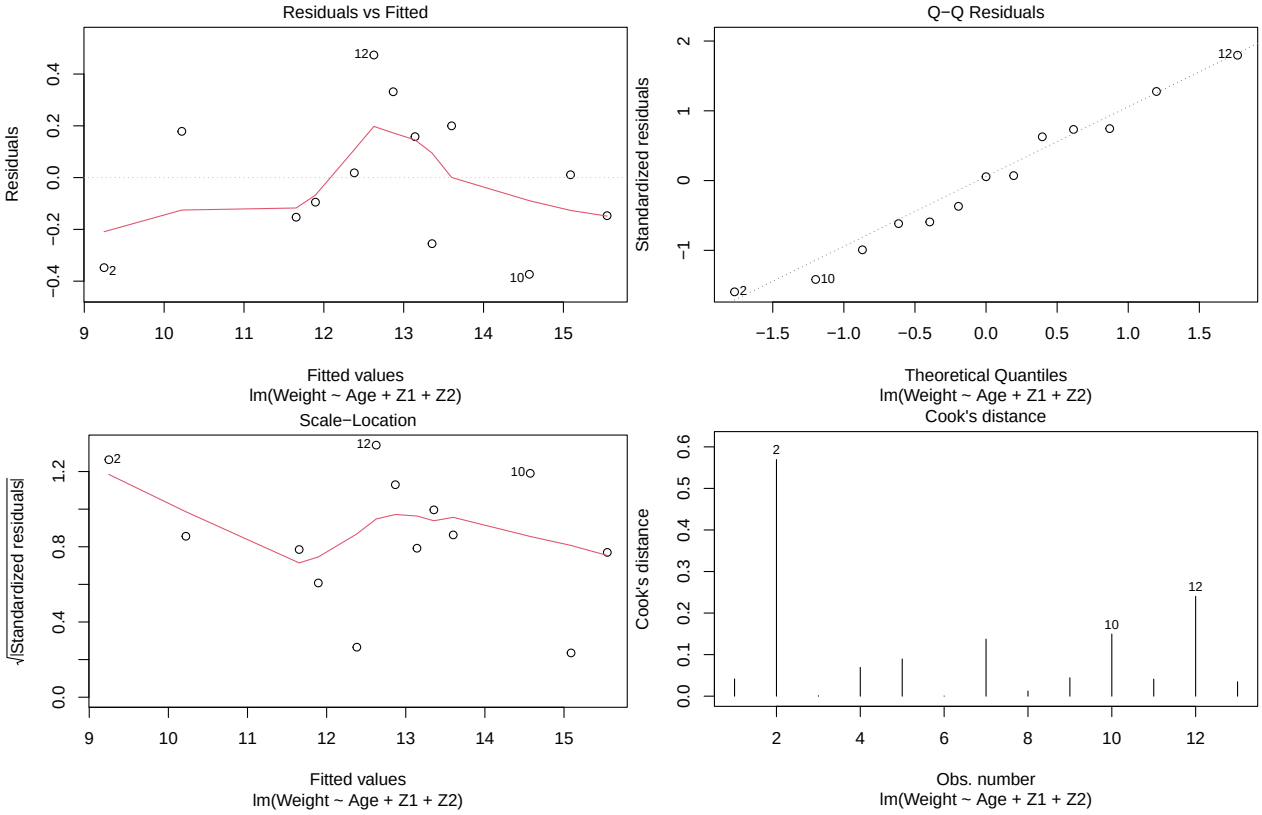
Now we can try fitting the model with our 2 dummy variables to see if it improves,

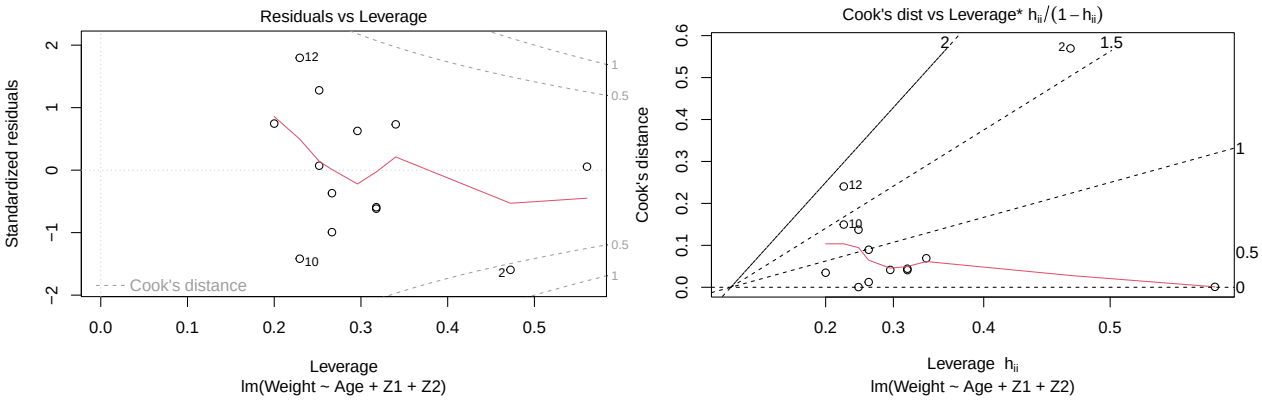
```
model2 <- lm(Weight ~ Age + Z1 + Z2, data=turkey)
summary(model2)
```

```
##
```

```
## Call:
## lm(formula = Weight ~ Age + Z1 + Z2, data = turkey)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37353 -0.15294  0.01103  0.17868  0.47353
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.43088    0.65744   2.176   0.0575 .
## Age          0.48676    0.02574  18.908 1.49e-08 ***
## Z1          -1.91838    0.20180  -9.506 5.45e-06 ***
## Z2          -2.19191    0.21143 -10.367 2.65e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3002 on 9 degrees of freedom
## Multiple R-squared:  0.9794, Adjusted R-squared:  0.9726
## F-statistic: 142.8 on 3 and 9 DF,  p-value: 6.6e-08
```

We see that all the predictors are significant, and our R^2 value has increased significantly. We can examine the model adequency again,





The qq-plot shows the assumption of normally distributed residuals is more reasonable now, and the residuals vs fitted is looking more random, however there still is a slight curve which could require more adjustments.