Linear Regression Practice Solutions

Part 1, Predict Water Use

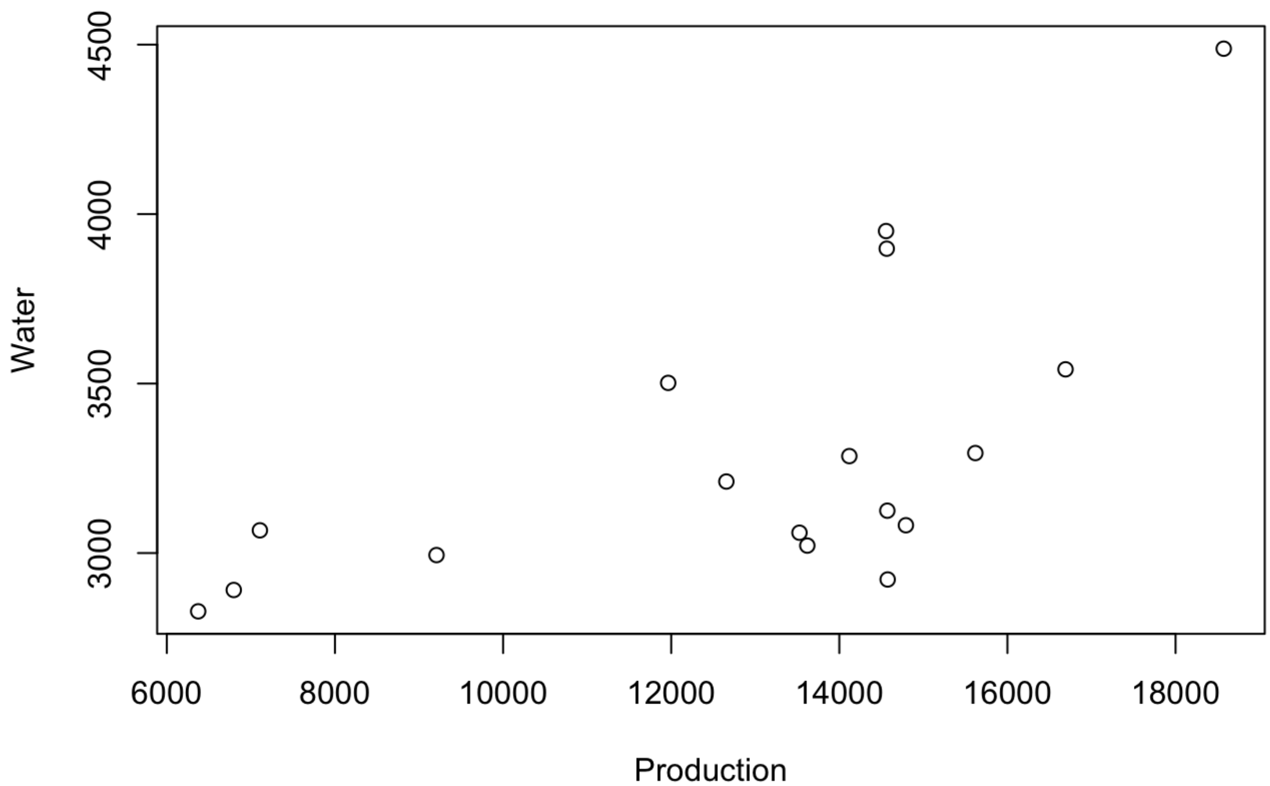
```
mydata <- read_csv("water.csv")
```

```
Parsed with column specification:
cols(
  Production = col_double(),
  Water = col_double()
)
```

```
attach(mydata)
head(mydata)
```

```
# A tibble: 6 x 2
  Production Water
       <dbl> <dbl>
1       7107  3067
2       6373  2828
3       6796  2891
4       9208  2994
5      14792  3082
6      14564  3898
```

```
plot(mydata)
```



I see some linear association in the plot.

Create the model

```
mydata_lm <- lm(Water ~ Production)
summary(mydata_lm)
```

```

Call:
lm(formula = Water ~ Production)

Residuals:
    Min      1Q  Median      3Q     Max 
-515.48 -293.68  -64.53  226.13  731.12 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.273e+03  3.387e+02   6.711 6.97e-06 ***
Production  7.989e-02  2.538e-02   3.148  0.00663 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 358 on 15 degrees of freedom
Multiple R-squared:  0.3978,    Adjusted R-squared:  0.3577 
F-statistic: 9.911 on 1 and 15 DF,  p-value: 0.006632
```

Assess the Model

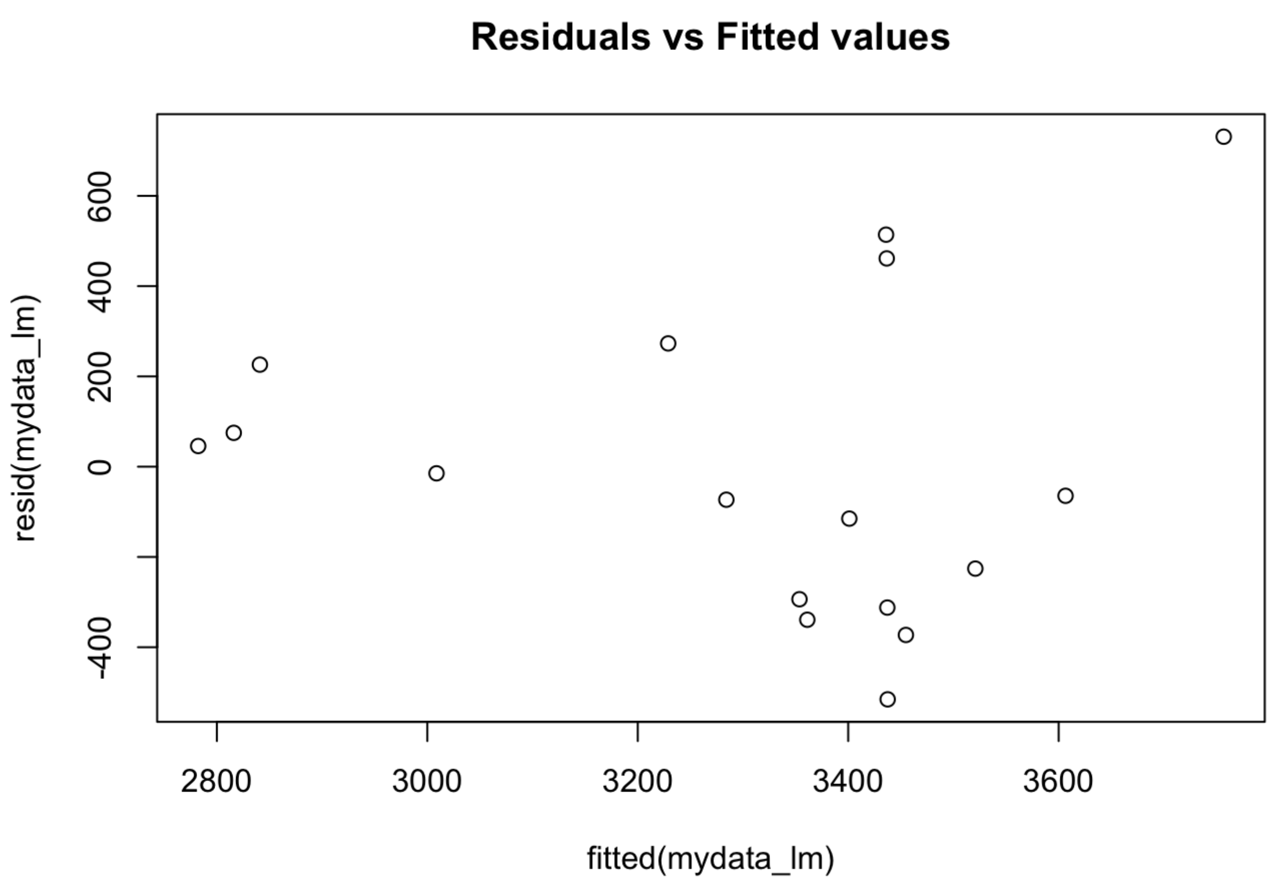
The F-statistic has a very small p-value indicating a significant relationship. The R^2 value is only moderate indicating about 40% of variability in water use is explained by this model. There are likely other variables that should be considered.

Now we need to check:

1. The relationship is linear
2. The errors are independent
3. The errors at each predictor value are normally distributed
4. The errors have equal variance across predictors (homoscedasticity)

We'll start with the residual vs fitted plot

```
plot(fitted(mydata_lm), resid(mydata_lm), main =  "Residuals vs Fitted values")
```



There is no pattern, so #1 is ok. When considering #4, I notice that the variability seems to change slightly across the plot, however there are not that many observations, so I don't think there is enough here to be concerned about violating #4. I will proceed cautiously.

For #2, we are told the manager randomly selected days, so it is reasonable to conclude the data will be independent.

Finally a normal quantile plot for #3:

```
library(car)
```

```
Loading required package: carData
```

```


Attaching package: 'car'
```

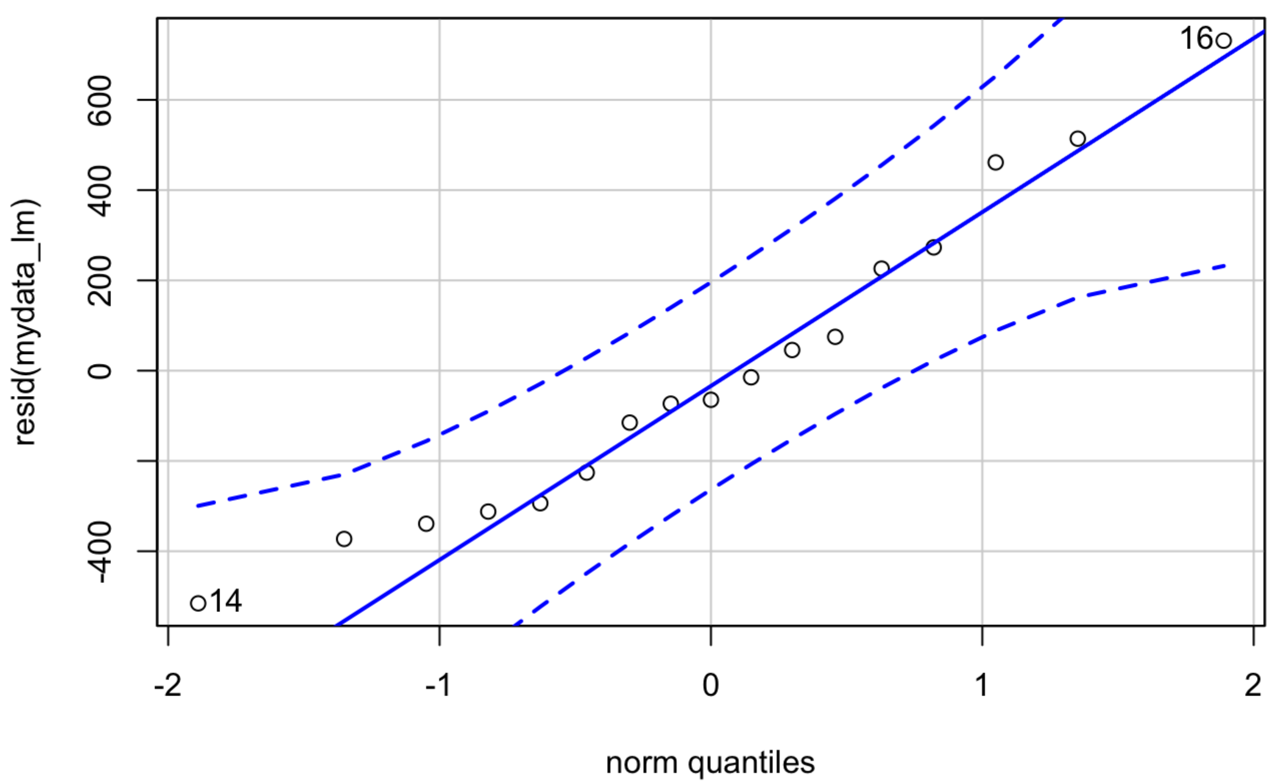
```
The following object is masked from 'package:dplyr':

    recode
```

```
The following object is masked from 'package:purrr':

    some
```

```
qqPlot(resid(mydata_lm))
```




```
[1] 16 14
```

This plot indicates some skewing at the low end, but overall looks quite good, so #3 is satisfied.

Summary

The model satisfies all assumptions, though we are slightly concerned about heteroscedacity. The F-statistic indicates this model is significant. The model is:

$\texttt{Water} = 2{,}273 + 0.07989\ \texttt{Production}$

Part 2, LinearReg1 Data Set

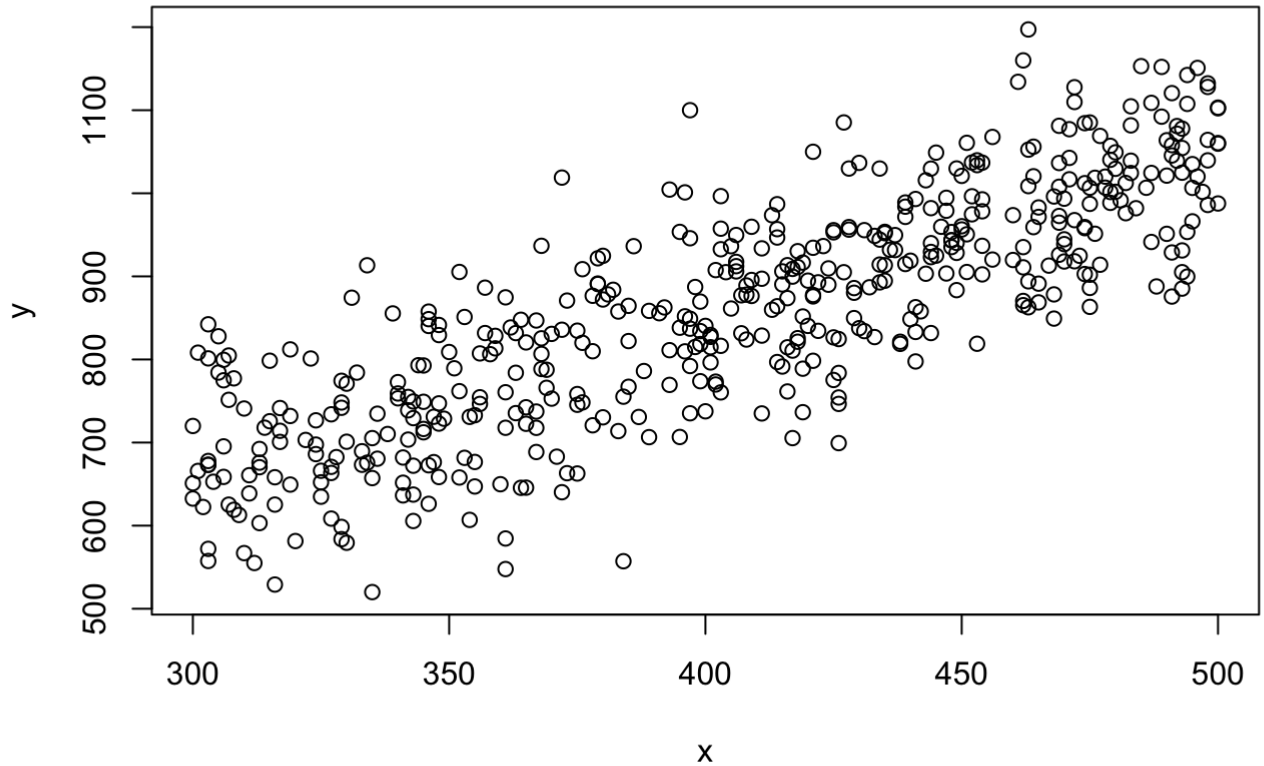
```
mydata <- read_csv("LinearReg1.csv")
```

```
Parsed with column specification:
cols(
  x = col_double(),
  y = col_double()
)
```

```
attach(mydata)
head(mydata)
```

```
# A tibble: 6 x 2
      x     y
  <dbl> <dbl>
1   468  878.
2   487 1024.
3   498  986.
4   301  666.
5   342  755.
6   402  773.
```

```
plot(mydata)
```



Create the model

```
mydata_lm <- lm(y ~ x)
summary(mydata_lm)
```

```


Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-260.267  -53.773    1.368   51.476  257.223 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 69.19987   24.83291   2.787  0.00553 ** 
x            1.94857    0.06064  32.135  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 78.71 on 498 degrees of freedom
Multiple R-squared:  0.6746,	Adjusted R-squared:  0.674 
F-statistic:  1033 on 1 and 498 DF,  p-value: < 2.2e-16
```

Assess the Model

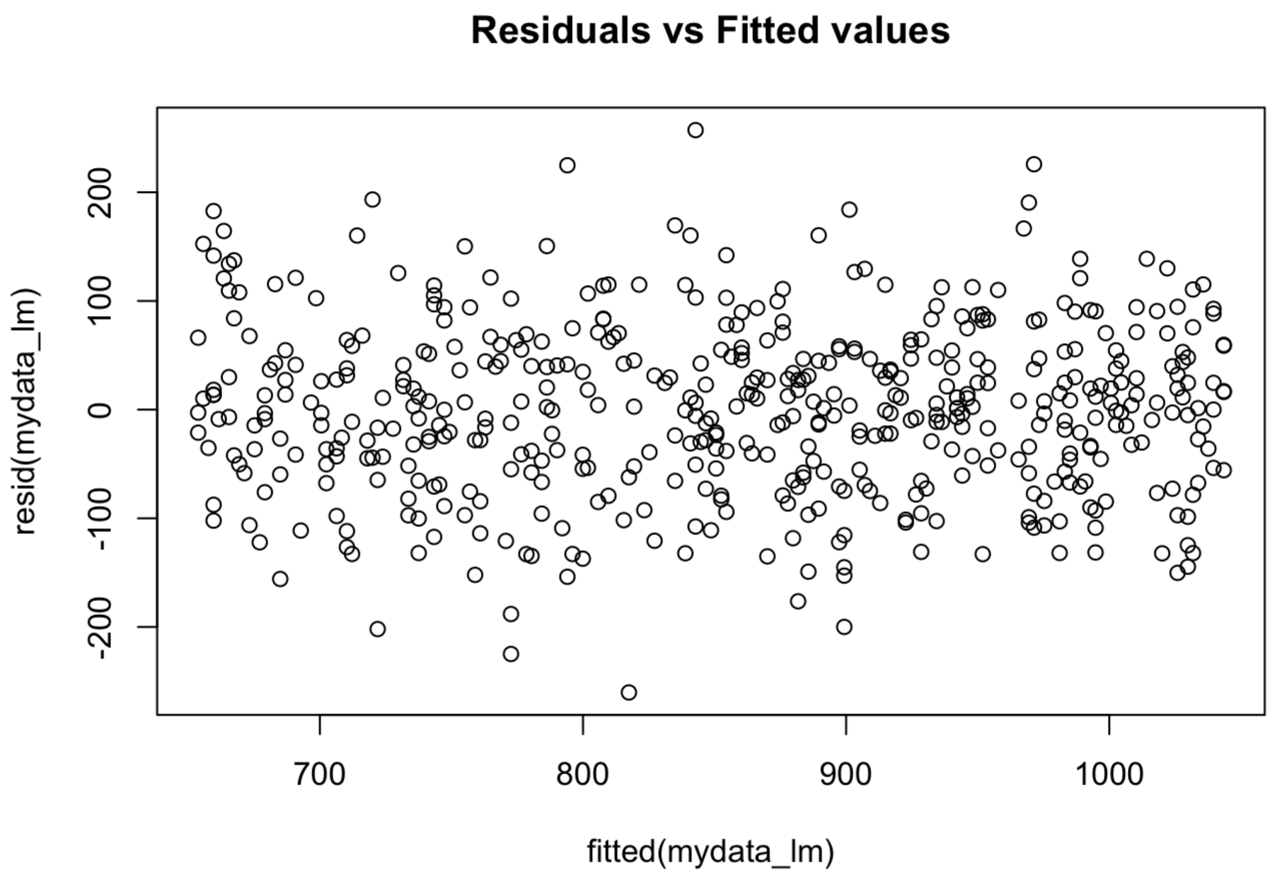
The F-statistic has a very small p-value indicating a significant relationship. The R^2 value is fairly high indicating 67% of variability in y is explained by this model.

We need to check:

1. The relationship is linear
2. The errors are independent
3. The errors at each predictor value are normally distributed
4. The errors have equal variance across predictors (homoscedasticity)

We'll start with the residual vs fitted plot

```
plot(fitted(mydata_lm), resid(mydata_lm), main =  "Residuals vs Fitted values")
```

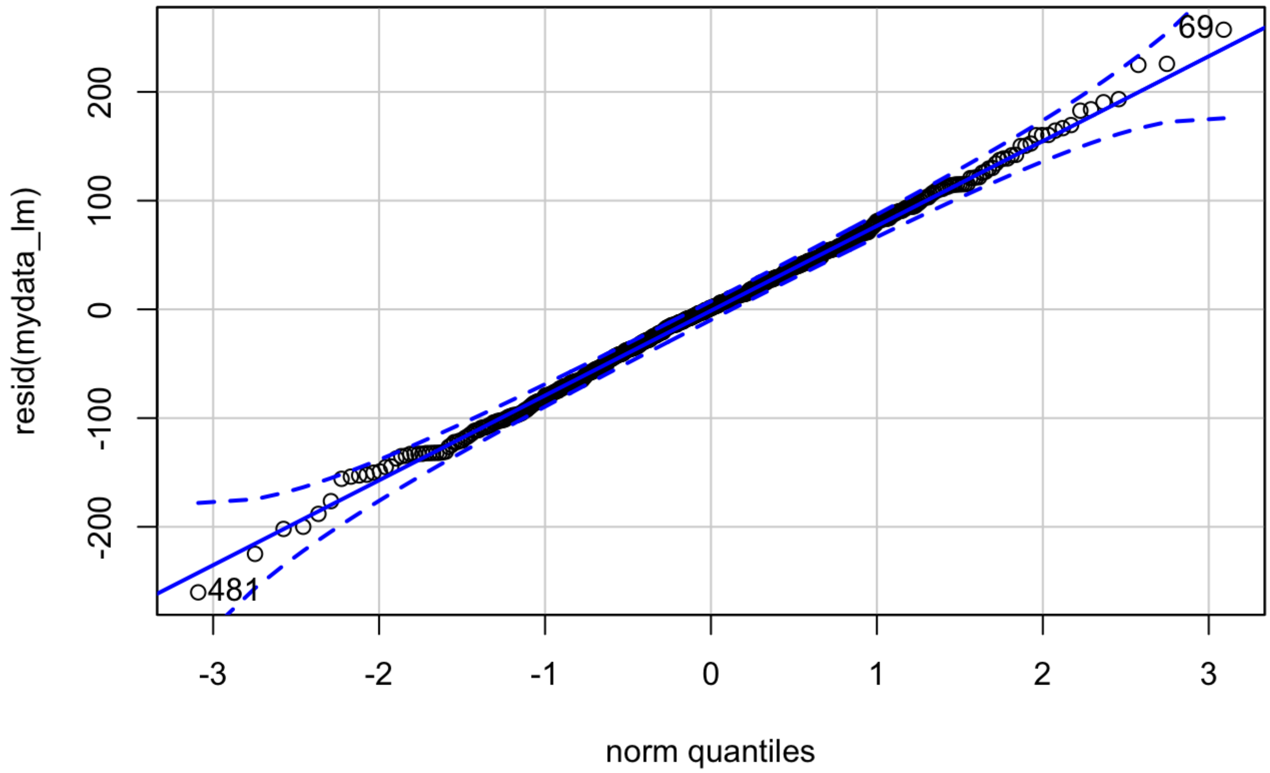


There is no pattern, so #1 is ok, and the variability looks uniform so #4 is ok.

We are not really told anything about this data, so we cannot check #2. Since this is an artificial situation we will need to assume the data were obtained randomly and are independent.

Finally a normal quantile plot for #3:

```
qqPlot(resid(mydata_lm))
```




```
[1] 481  69
```

This looks quite good, so #3 is satisfied.

Summary

The model satisfies all assumptions and the F-statistic indicates it is significant. The model is:

y = 69.19987 + 1.94857 x

Part 3, LinReg2 data set

```
mydata <- read_csv("LinearReg2.csv")
```

```
Parsed with column specification:
cols(
  x = col_double(),
  y = col_double()
)
```

```
attach(mydata)
```

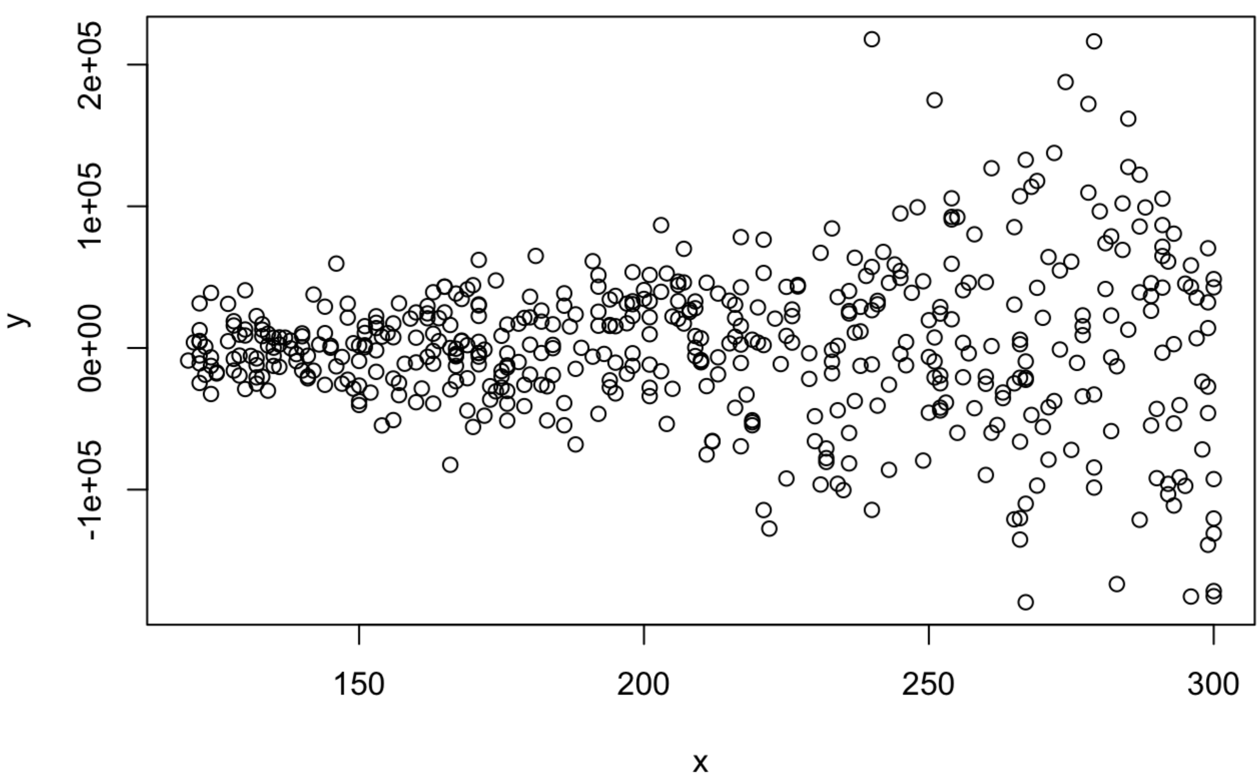
```
The following objects are masked from mydata (pos = 3):

    x, y
```

```
head(mydata)
```

```
# A tibble: 6 x 2
      x       y
  <dbl>   <dbl>
1   242  67809.
2   166 -82457.
3   254  90713.
4   256 -20676.
5   205  22262.
6   132  22533.
```

```
plot(mydata)
```



This dot plot does not look linear. There really isn’t much of a pattern at all.

Create the model

```
mydata_lm <- lm(y ~ x)
summary(mydata_lm)
```

```

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-181951  -26832     907   29139  215913 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1882.46    9882.44  -0.190    0.849
x              16.31      45.71   0.357    0.721

Residual standard error: 54090 on 498 degrees of freedom
Multiple R-squared:  0.0002555, Adjusted R-squared:  -0.001752 
F-statistic: 0.1273 on 1 and 498 DF,  p-value: 0.7214
```

Assess the Model

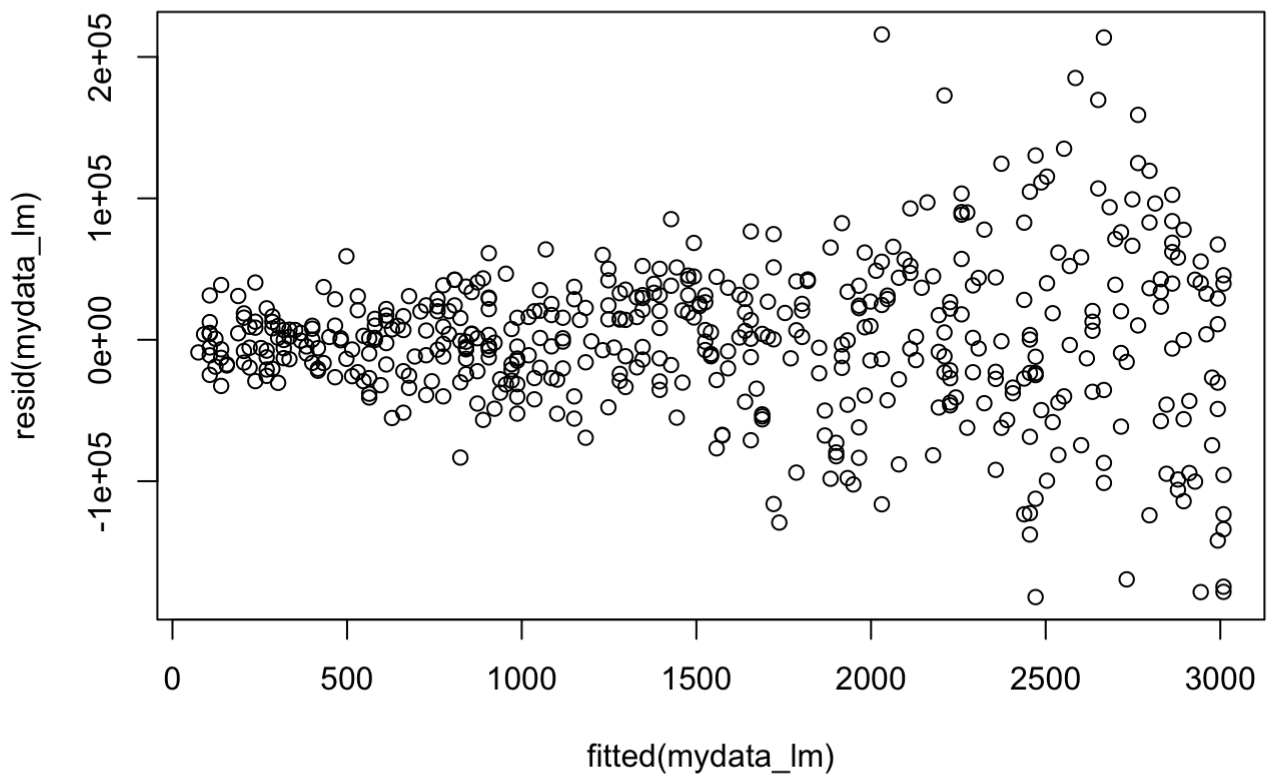
The F-statistic has a large p-value indicating this model is not significant. The R^2 value is very low indicating essentially none of variability in y is explained by this model.

We need to check:

1. The relationship is linear
2. The errors are independent
3. The errors at each predictor value are normally distributed
4. The errors have equal variance across predictors (homoscedasticity)

We'll start with the residual plot.

```
plot(fitted(mydata_lm), resid(mydata_lm))
```

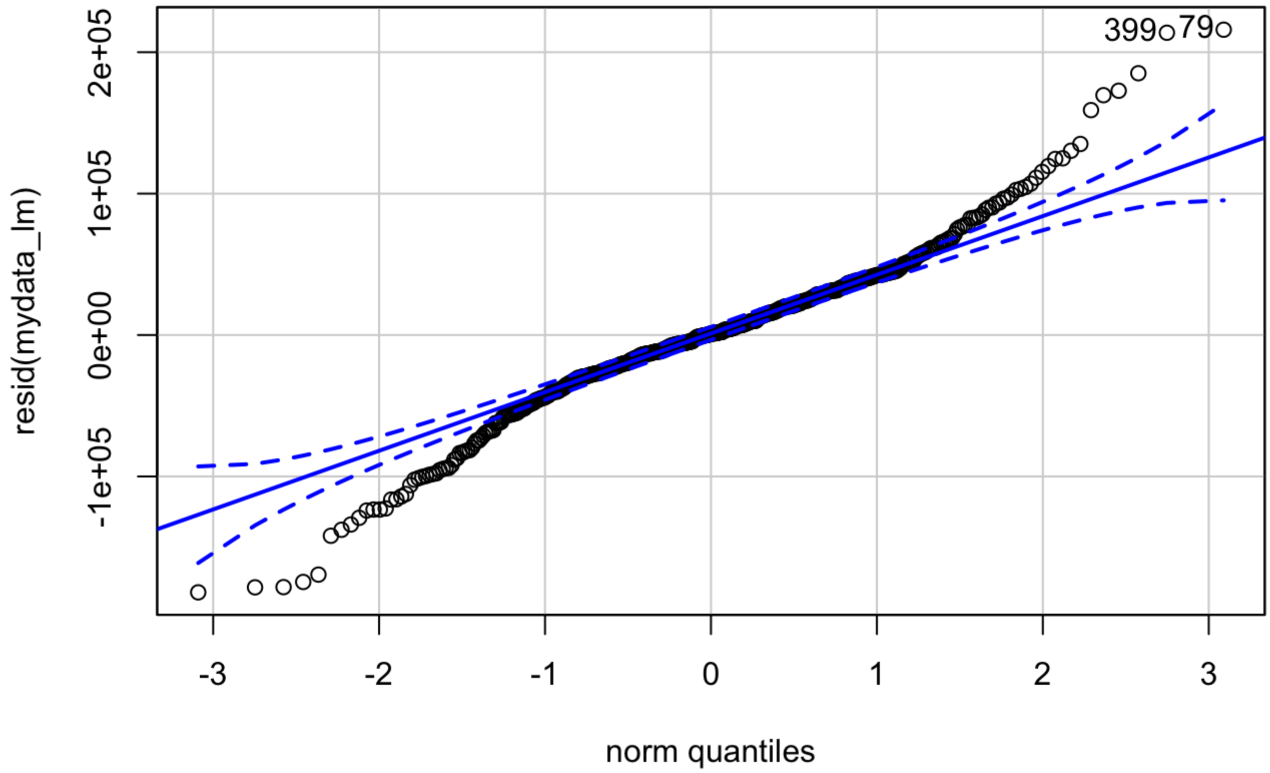


There is no real pattern, so #1 is ok, however the variability changes drastically as we move left to right, so #4 is not satisfied.

We are not really told anything about this data, so we cannot check #2.

Finally a normal quantile plot for #3:

```
qqPlot(resid(mydata_lm))
```




```
[1]  79 399
```

This is not a good normal quantile plot. The points leave the confidence bands significantly at either end, so the assumption of normally distributed errors is in doubt.

Summary

We cannot use this model. It does not satisfy all assumptions of the linear regression procedure, and even if it did, the F-statistic indicates the relationship is not significant. If there is a relationship between these variables it is likely not linear.