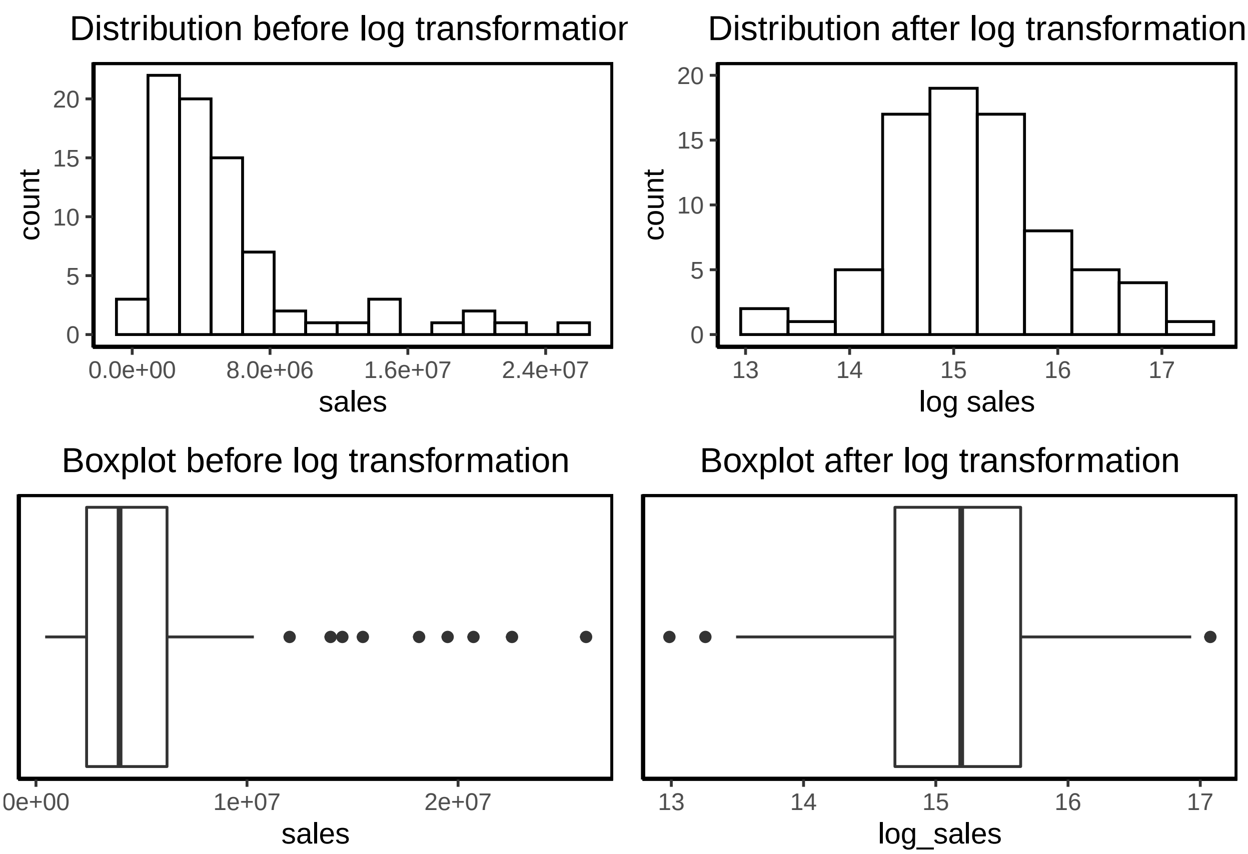
# Individual assignment on linear regression

LIU Yuzhu 1155091887

## Problem 1

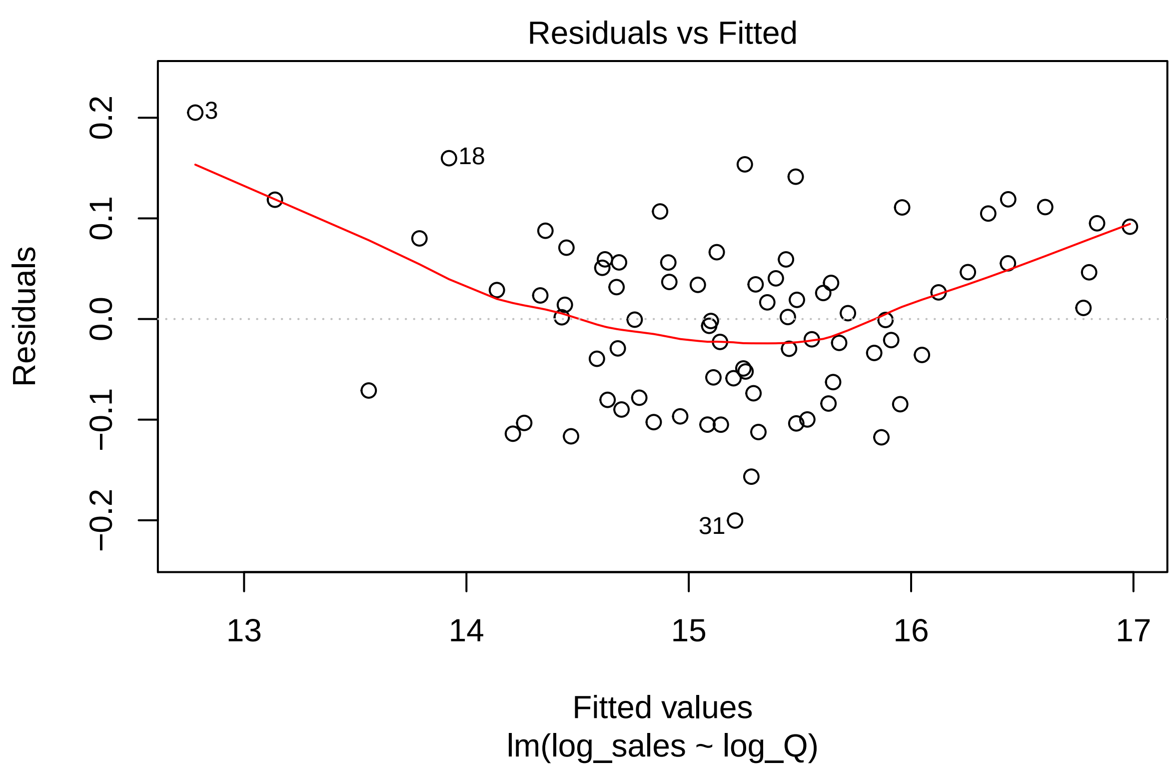
```
## Summary statistic of sales before log transformation
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   436062  2399798  3961997  5583184  6210331 26064575
##
## Summary statistic of sales after log transformation
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.99   14.69   15.19   15.21   15.64   17.08
```

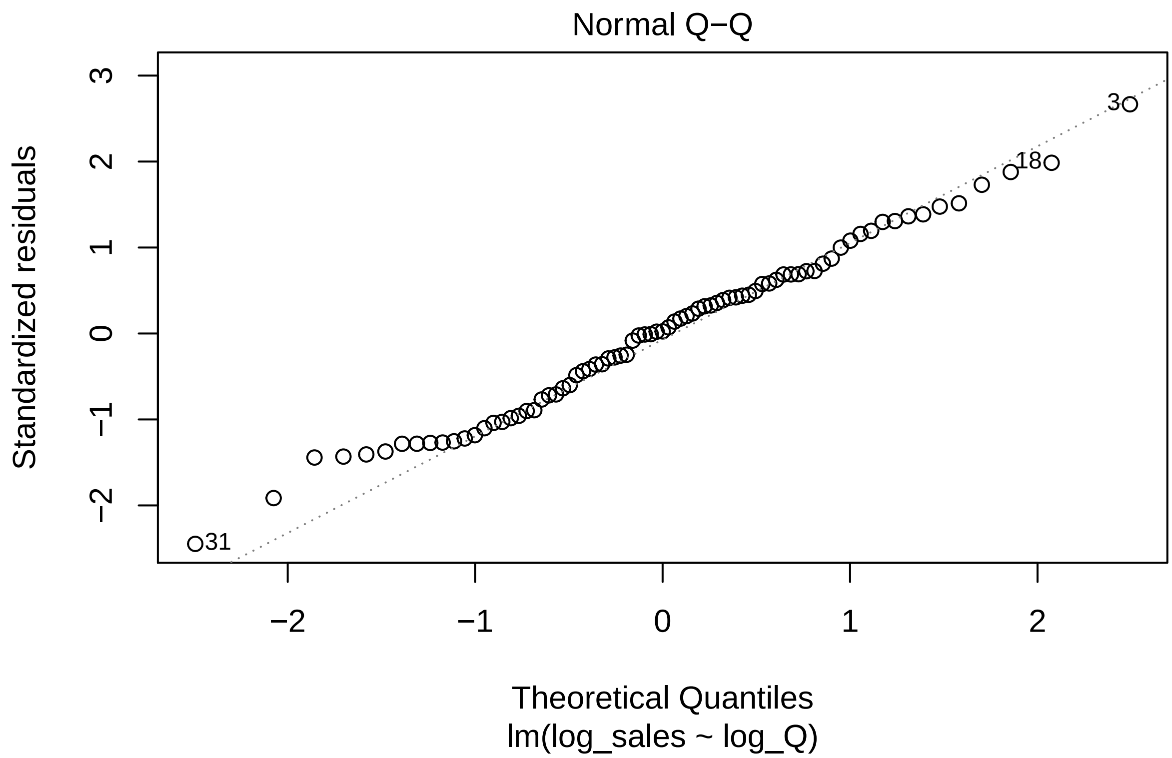


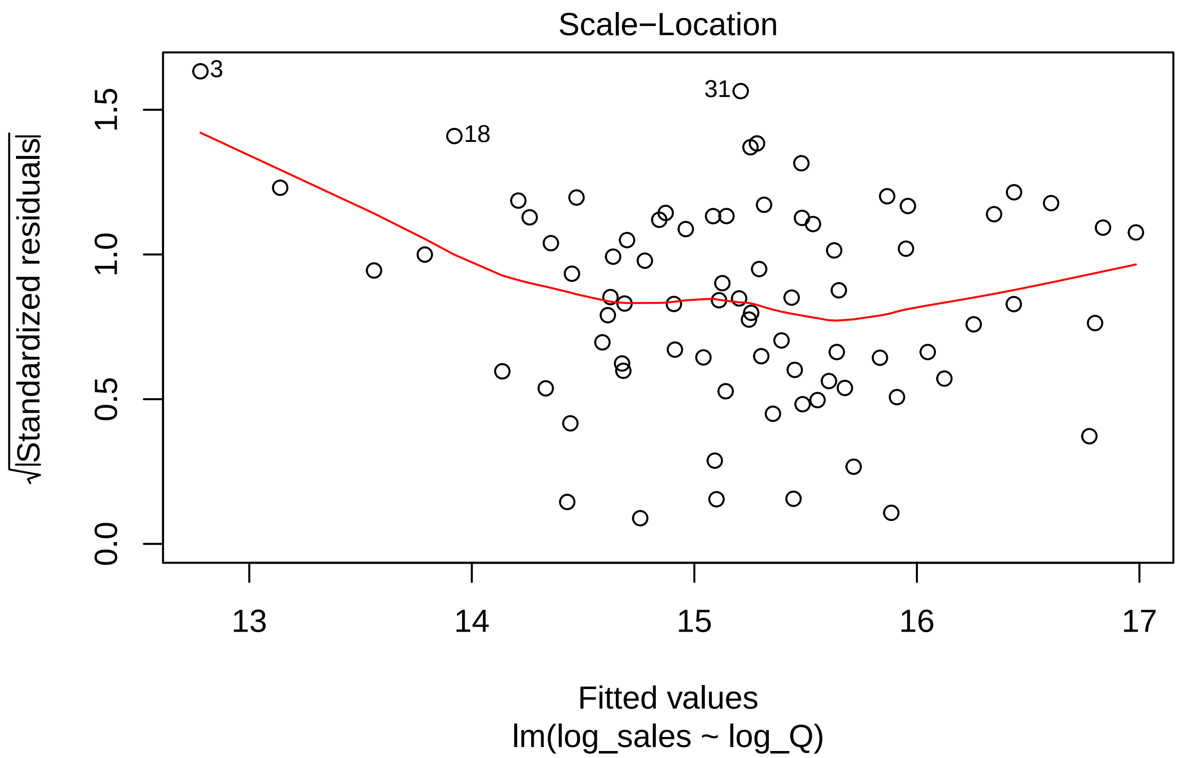

As shown in the above histogram chart, before log transformation, sales data is right-skewed with a long tail on the upper side of the distribution. Its mean (5583184) is much larger than its median (3961997). After log transformation, due to concaveness of logarithm function, the distribution becomes more symmetric, and its mean (15.21) is very close to its median (15.19). Furthermore, the above boxplots show that before log transformation, many outliers are more than 1.5 IQR above Q3, while after log transforamtion, the number of outliers are reduced dramatically.

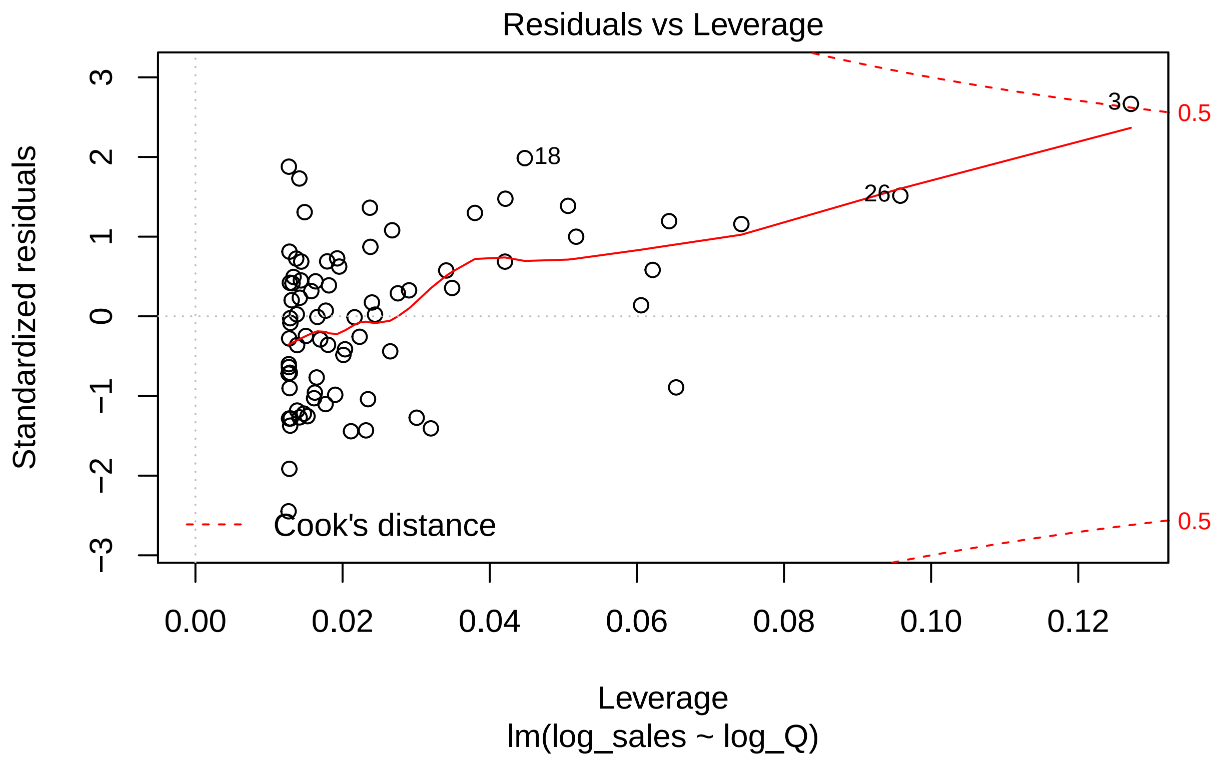
```
##
## Call:
## lm(formula = log_sales ~ log_Q, data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.200244 -0.066863  0.001987  0.055742  0.205113
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.0071     0.0718  125.44   <2e-16 ***
## log_Q         1.0792     0.0124   87.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08234 on 77 degrees of freedom
## Multiple R-squared:  0.9899, Adjusted R-squared:  0.9898
## F-statistic:  7579 on 1 and 77 DF,  p-value: < 2.2e-16
```

The estimated slope coefficient is 1.0792, which means one unit increase in the logarithm of Q would result in on average 1.0792 units of increase in logarithm of sales.



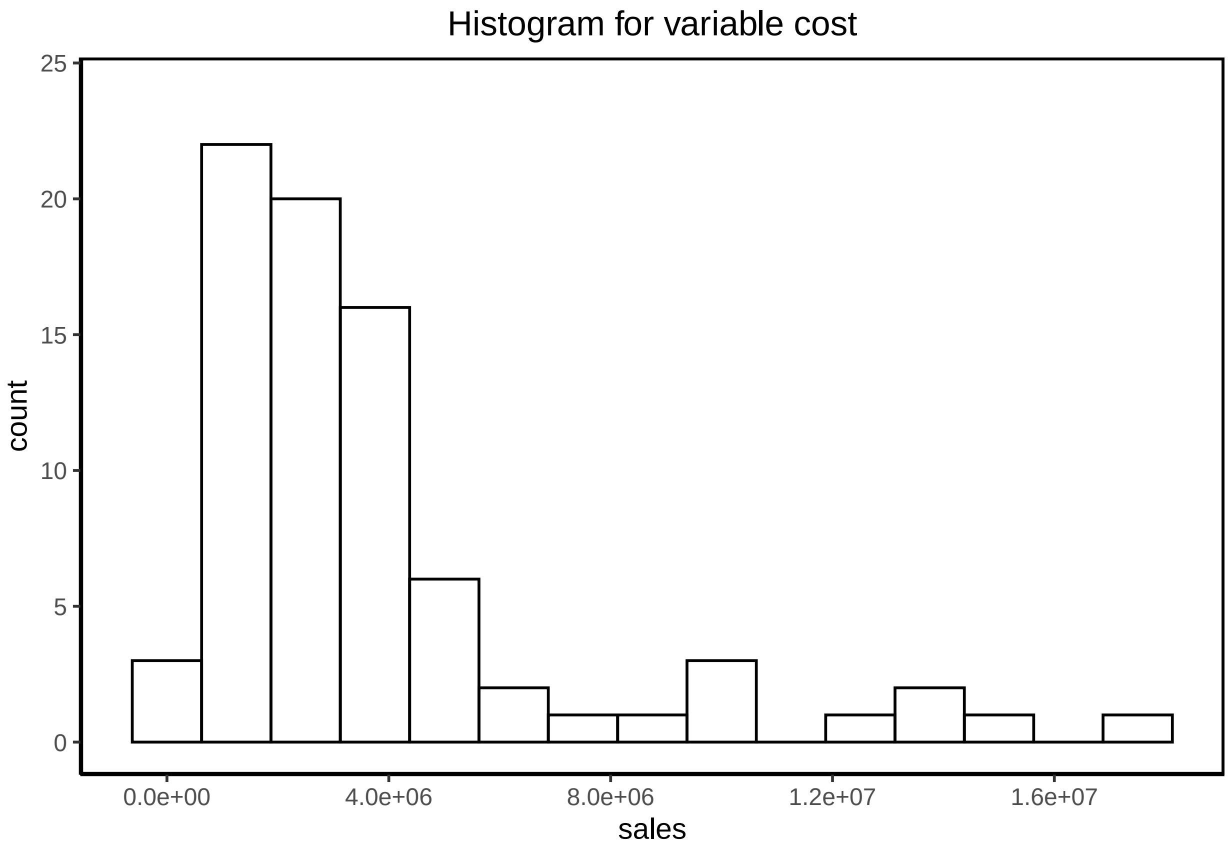

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log_sales ~ log_Q)
3
18
31
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
3
18
31
13
14
15
16
17
Fitted values
lm(log_sales ~ log_Q)



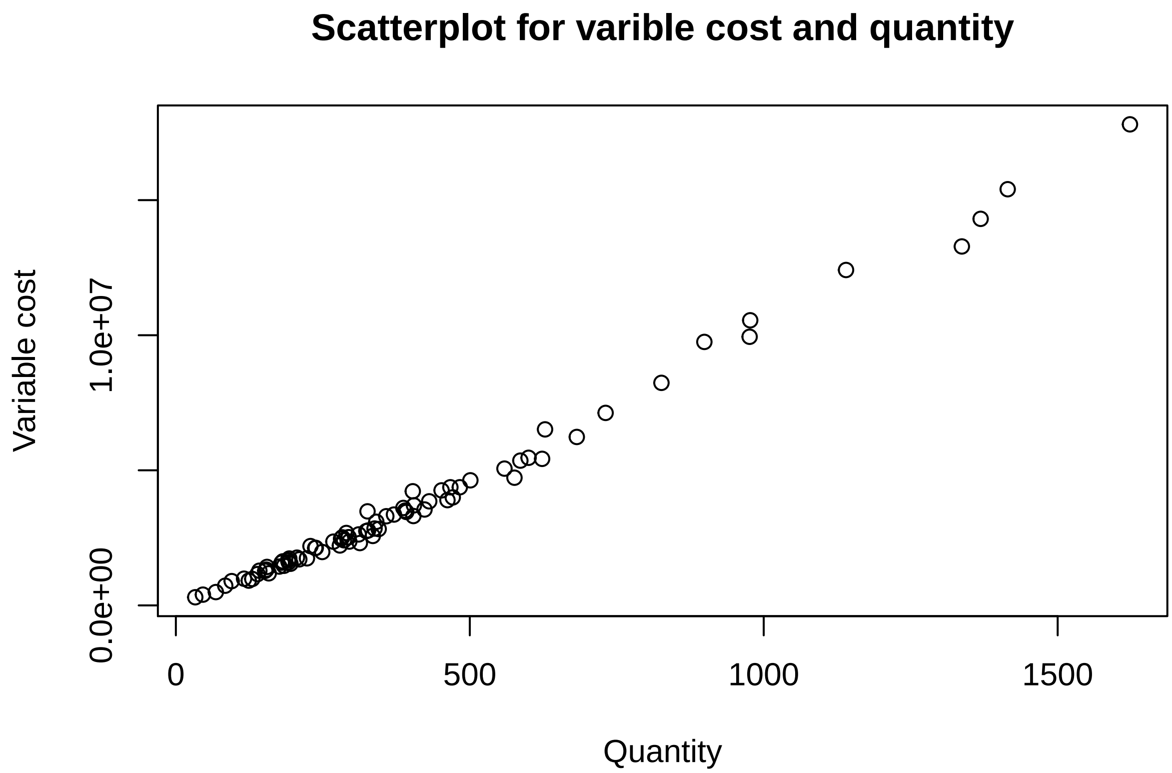

Compared with the old sales_model, the sales_model after log transformation improves a bit in goodness of fit, indicated by the less significant curvature of the plot of residuals and fitted values. However, the normal Q-Q plot still deviates from straight line in the lower part of the data, which alarms normal assumption about residuals may not be valid.

**Problem 2**




```
## Summary statistics for variable cost
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   301688  1650542  2680286  3841375  4314705 17804291
```

The histogram of variable cost is significantly right-skewed with a long tail on the upper side. The variable's mean (3841375) is much larger than its median (2680286), which indicates more data lies below the mean rather than above the mean.




As shown in the scatterplot, points lie roughly around a line, so there is a strong linear association between variable cost and quantity.

```
##
## Call:
## lm(formula = varCost ~ Q, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -927645 -209737    4054  226612  828009
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -575210.4    63426.9  -9.069 8.83e-14 ***
## Q             10814.2      121.5  88.973  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 350900 on 77 degrees of freedom
## Multiple R-squared:  0.9904, Adjusted R-squared:  0.9902
## F-statistic:  7916 on 1 and 77 DF,  p-value: < 2.2e-16
```

The fitted model is varCost = -575210.4 + 10814.2 * Q + error.

Null hypothesis that population coefficient of Q is 0 is rejected since p-value $<$ 2e-16 and is lower than 0.05, the commonly-used significance level. The slope coefficient of 10814.2 means that one unit of increase in quantity would result in on average 10814.2 units of increase in variable cost.

Multiple R-squared is 0.9904, which means more than 99% of variation in variable costs can be explained by the model. After adjusted for the number of variables in the model, the resulting adjusted R-squared is still more than 0.9, indicating high goodness of fit. P-value for F-test is lower than 2.2e-16 and 0.05, so the overall model is useful in predicting variable cost.