Simple Regression

Patrick Kelly

2/12/2020

Source

This exercise was inspired by an article on medium.com.

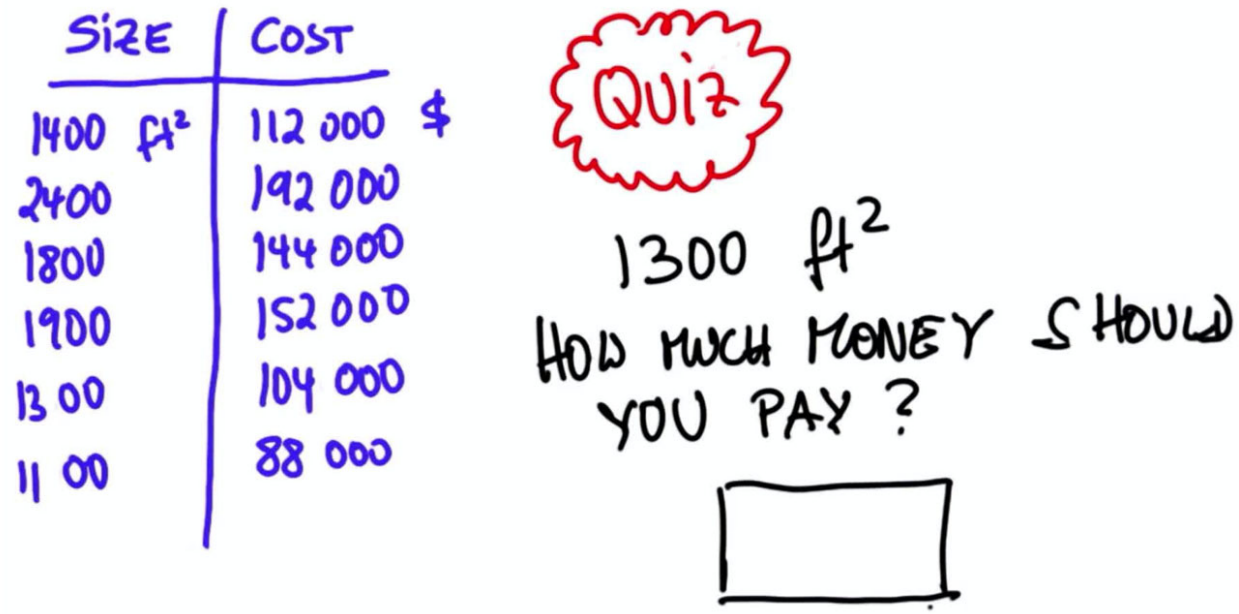
Since the tutorial was written for python users, I decidwd to do an equivalent analysis in R.

Click Here"Intro to Statistics"

Interview for a data scientist job

Here is a challenge. Given the data, create a model that predicts the cost of a 1300 square feet house.

If you are given 2 days to find an answer, using R, could you do it?

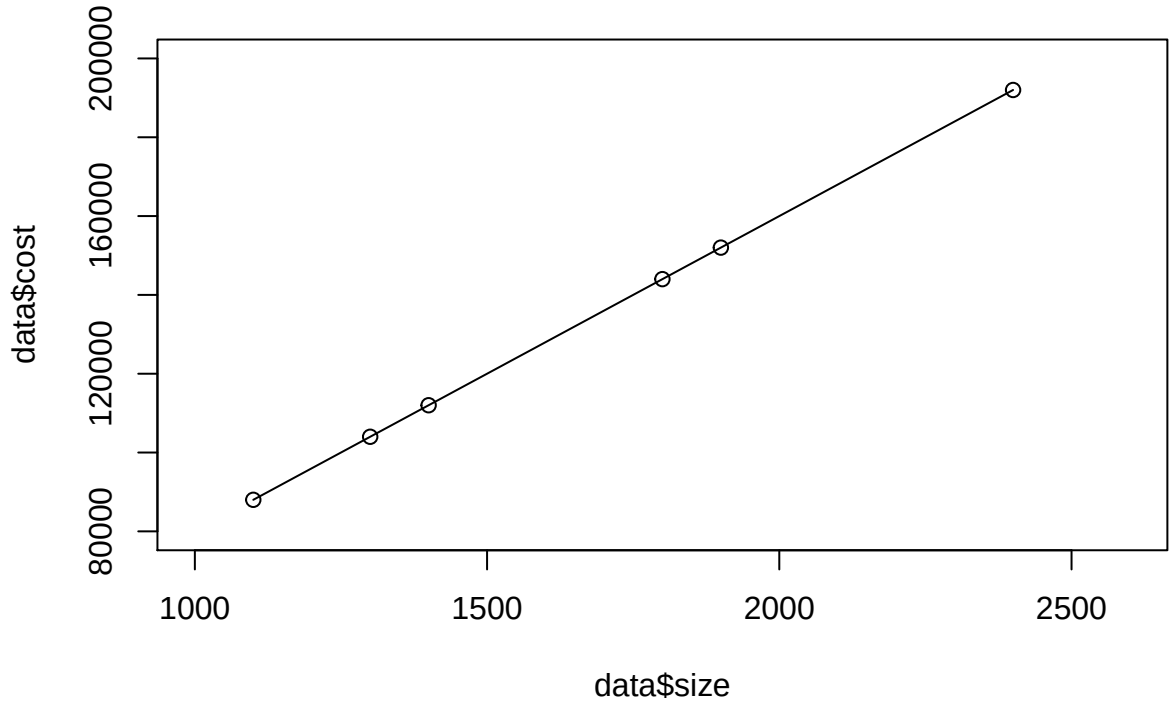


Create a dataframe with the data

```
size <- c(1400,2400,1800,1900,1300,1100)
cost <- c(112000,192000,144000,152000,104000,88000)
data <- data.frame(size,cost)
```

Create a scatterplot

```
scatter.smooth(data$size,data$cost,
     xlim = c(1000,2600),
     ylim = c(80000,200000))
```

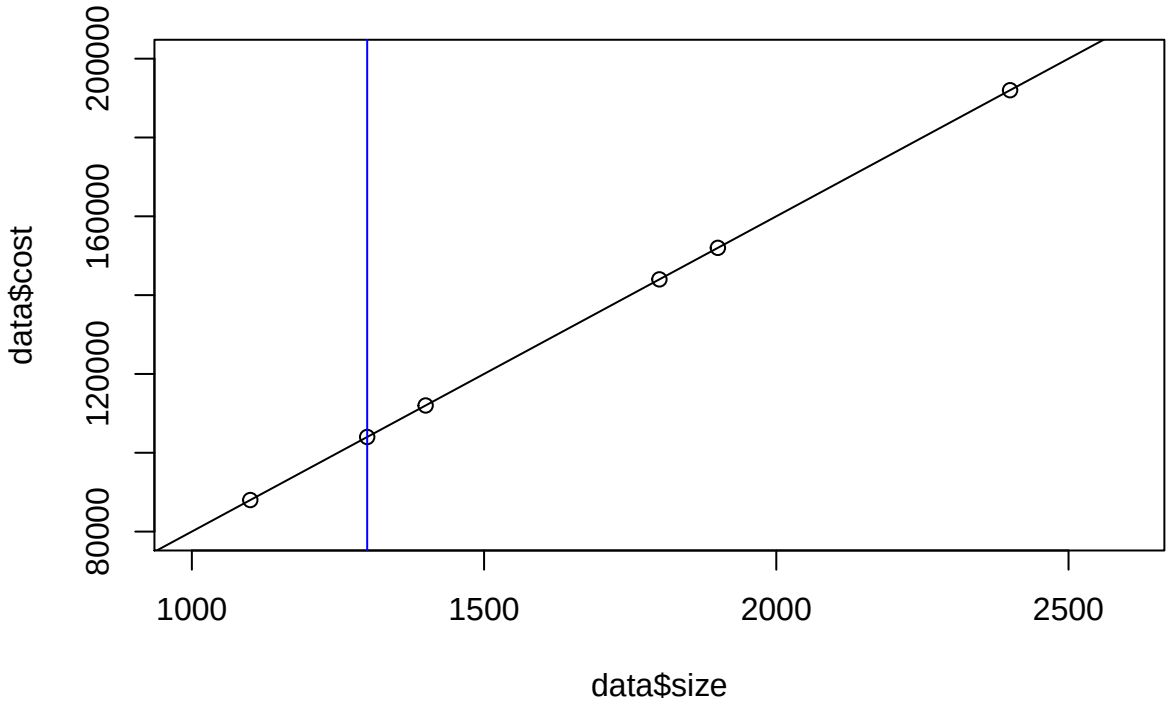


What model to try?

The scatteplot shows that the data points lie on a straight line, so simple linear regression will be an excellent model to try.

In statistics, linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression.

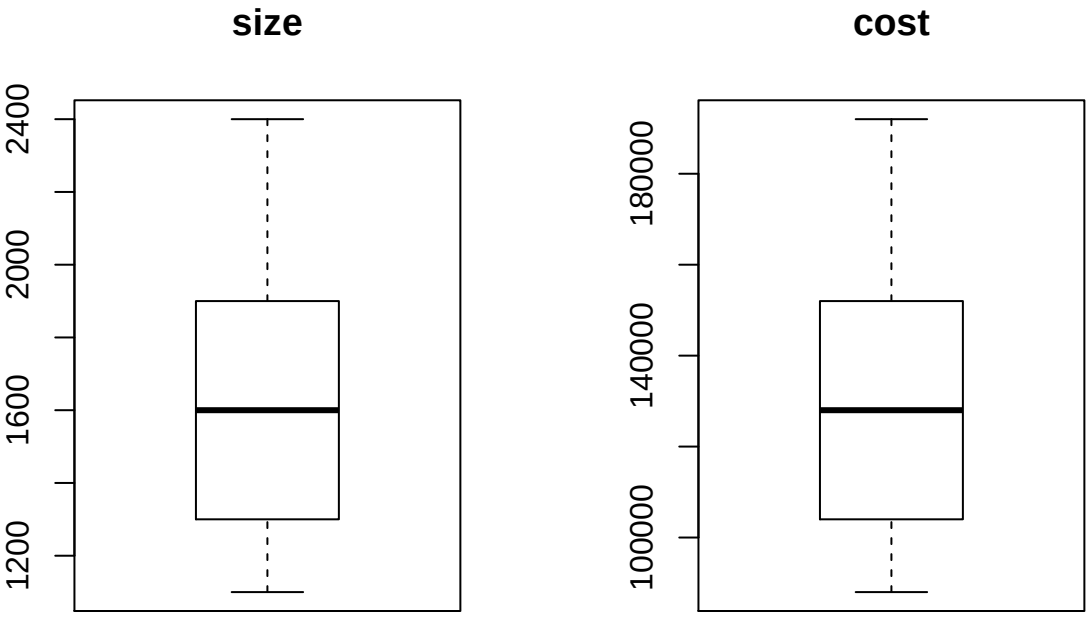
```
model <- lm(cost ~ size,data=data)
plot(data$size,data$cost,
     xlim = c(1000,2600),
     ylim = c(80000,200000))
abline(lm(cost ~ size,data=data))
abline(v=1300, col = "blue")
```



#

Check data for outliers

```
par(mfrow=c(1, 2))
boxplot(data$size, main = "size")
boxplot(data$cost,main = "cost")
```




```
# There are no outliers
```

Model details

```
model$coefficients
```

```
##  (Intercept)         size
## 7.128953e-11 8.000000e+01
```

Predicted price for house size = 1300 square feet

```
Pred_1300 <- model$coefficients[1]+
  (model$coefficients[2] * 1300)
Pred_1300
```

```
## (Intercept)
##      104000
```

Model interpretation

```
summary(model)
```

```
## Warning in summary.lm(model): essentially perfect fit: summary may be unreliable
```

```
##
## Call:
## lm(formula = cost ~ size, data = data)
##
## Residuals:
##          1          2          3          4          5          6
##  1.565e-11 -9.950e-12  6.468e-12  5.701e-12 -4.246e-12 -1.363e-11
##
## Coefficients:
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 7.129e-11  1.997e-11 3.569e+00   0.0234 *
## size        8.000e+01  1.171e-14 6.835e+15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.247e-11 on 4 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 4.671e+31 on 1 and 4 DF,  p-value: < 2.2e-16
```

Is this exercise too simple to be useful?

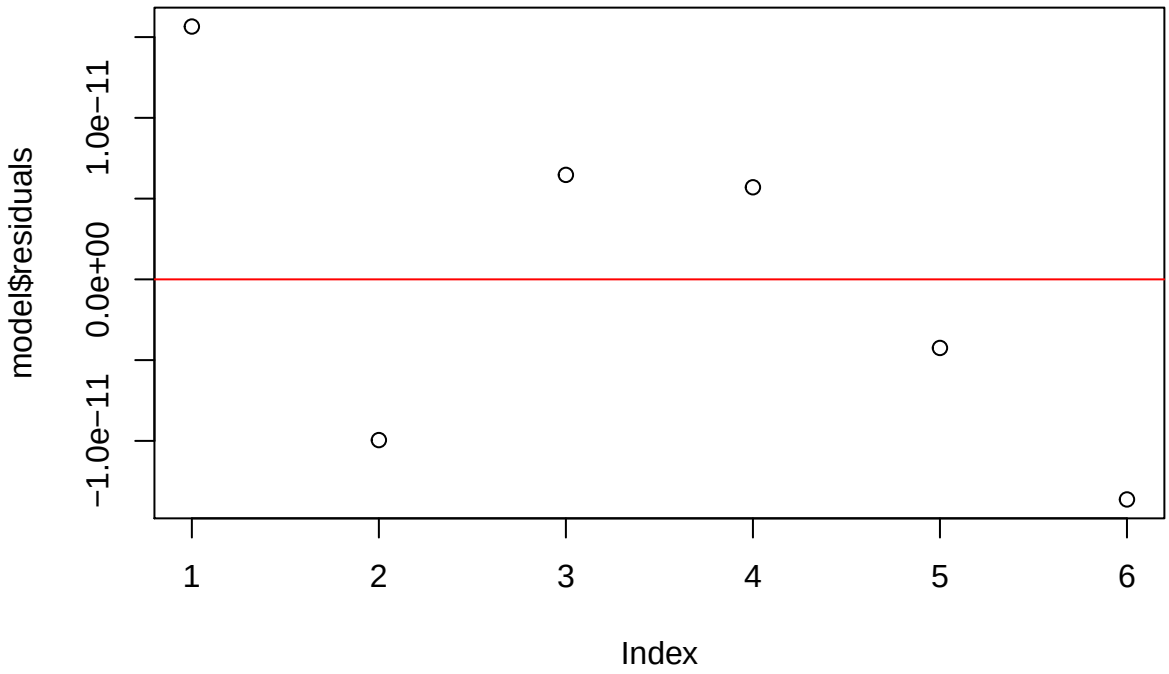
1. Who would do a regression on a sample size of only 6?
2. The model predicted values are perfect, with no error. (All the data points lie exactly on the prediction line). Thus the R sqrared value = 1. This never happens with real data. It has been said that there are no good models, only useful ones.

Here are assumptions that should be chekced for serious analysis and to respond to your critics:

1. The regression model is linear in parameters.
2. The mean of residuals is zero.
3. Homoscedasticity of residuals or equal variance.
4. No autocorrelation of residuals.
5. The X variables and residuals are uncorrelated.
6. The number of observations must be greater than number of Xs.
7. The variability in X values is positive.
8. The regression model is correctly specified/
9. No perfect multicollinearity.
10. Normality of residuals.
11. The data points are independent.

Model Diagnostics

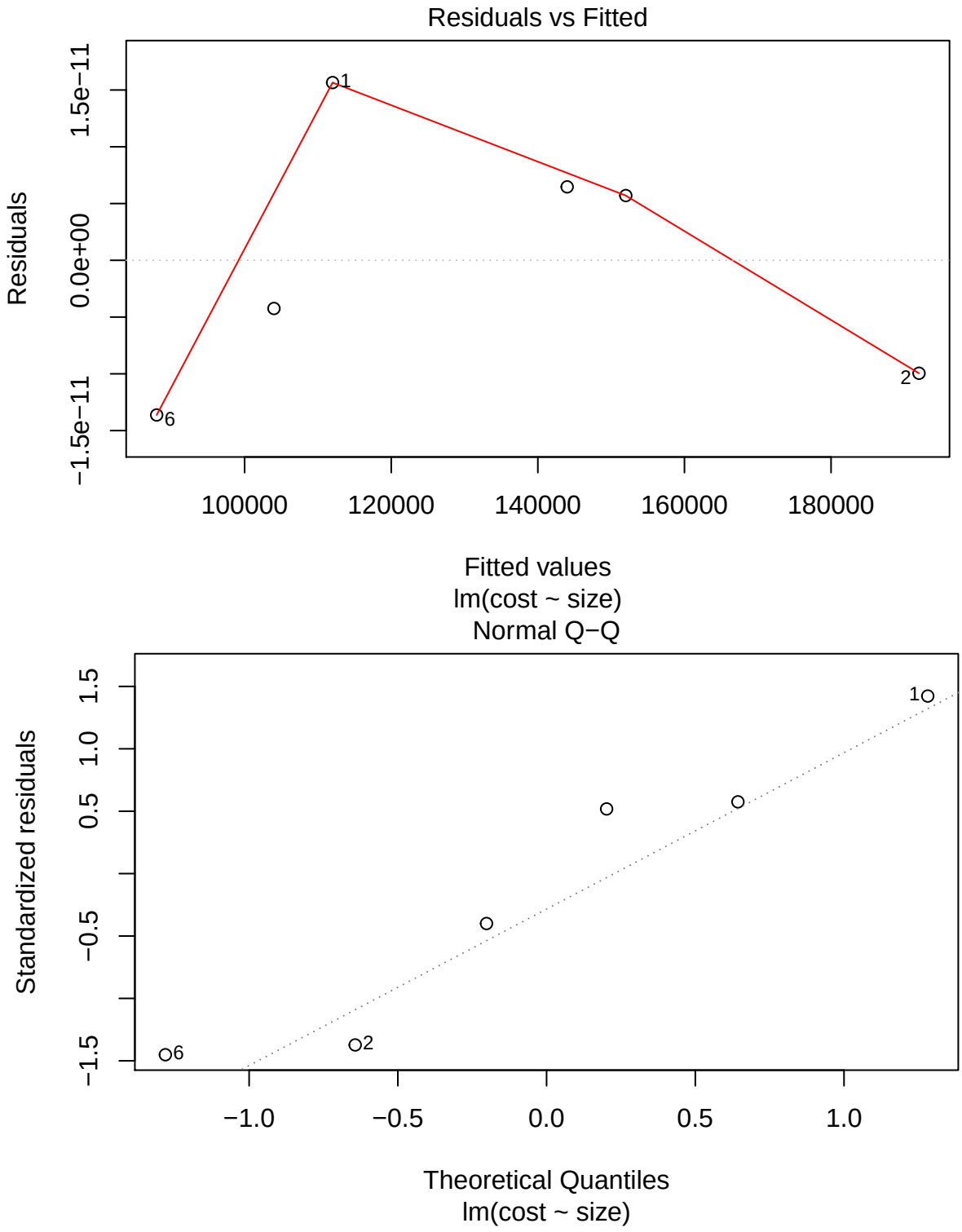
```
plot(model$residuals)
abline(h=0, col = "red")
```

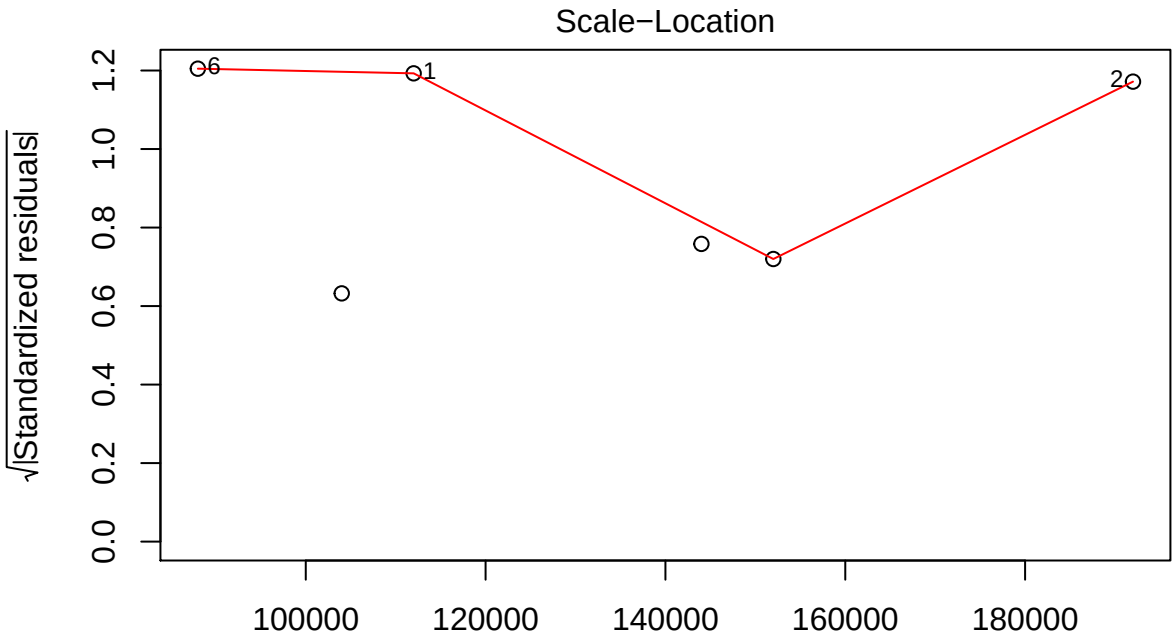


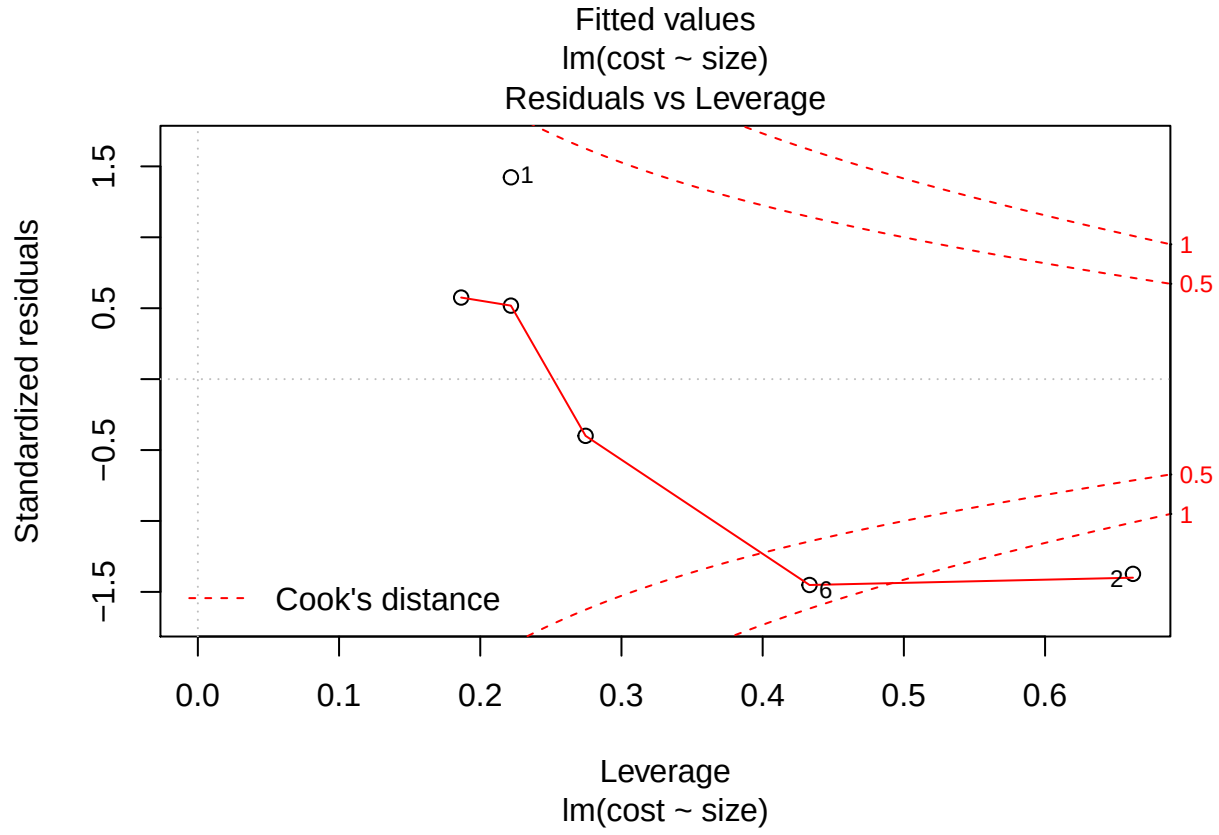

```
round(mean(model$residuals))
```

```
## [1] 0
```

```
plot(model)
```


Residuals vs Fitted
Residuals
1.5e−11
0.0e+00
−1.5e−11
1
6
2
100000
120000
140000
160000
180000
Fitted values
lm(cost ~ size)
Normal Q−Q
Standardized residuals
1.5
1.0
0.5
−0.5
−1.5
1
6
2
−1.0
−0.5
0.0
0.5
1.0
Theoretical Quantiles
lm(cost ~ size)






```
cor.test(data$size, model$residuals)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$size and model$residuals
```

```
## t = -2.0977e-16, df = 4, p-value = 1
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8115613  0.8115613
## sample estimates:
##           cor
## -1.048848e-16
```

```
# p-value is high, so null hypothesis that true correlation is 0 can't be rejected
var(data$size)
```

```
## [1] 227000
```

```
# The variance in the X variable is much larger than 0. So, this assumption is satisfied.
influence.measures(model)
```

```
## Influence measures of
##   lm(formula = cost ~ size, data = data) :
##
##    dfb.1_ dfb.size  dffit cov.r cook.d   hat inf
## 1  0.6578  -0.4664  0.936 0.557 0.2884 0.222
## 2  1.6226  -1.9808 -2.290 1.472 1.8481 0.662   *
## 3 -0.0185   0.0812  0.249 1.839 0.0379 0.186
## 4 -0.0647   0.1236  0.248 1.988 0.0383 0.222
## 5 -0.1750   0.1363 -0.217 2.259 0.0303 0.275
## 6 -1.4636   1.2526 -1.597 0.703 0.8048 0.433   *
```

AIC and BIC

The Akaike's information criterion - AIC (Akaike, 1974) and the Bayesian information criterion - BIC (Schwarz, 1978) are measures of the goodness of fit of an estimated statistical model and can also be used for model selection. Both criteria depend on the maximized value of the likelihood function L for the estimated model. For model comparison, the model with the lowest AIC and BIC score is preferred.

```
AIC(model)
```

```
## [1] -280.6975
```

```
BIC(model)
```

```
## [1] -281.3222
```

Package for diagnostics = gvlma

```
# install.packages("gvlma")
# Global validations of linear model assumptions
library(gvlma)
gvlma::gvlma(model)
```

```
##
## Call:
## lm(formula = cost ~ size, data = data)
##
## Coefficients:
## (Intercept)         size
##   7.129e-11    8.000e+01
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma::gvlma(x = model)
##
##                     Value p-value                   Decision
## Global Stat        4.8756  0.3003    Assumptions acceptable.
## Skewness           0.0122  0.9120    Assumptions acceptable.
## Kurtosis           0.4450  0.5047    Assumptions acceptable.
## Link Function      4.2181  0.0400 Assumptions NOT satisfied!
## Heteroscedasticity 0.2003  0.6545    Assumptions acceptable.
```

Comment

Ordinary Least Squares (OLS) also known as Simple Linear Regression is an important foundation upon which data science is built.

It certainly does not appear to be as simple today as it was almost 50 years ago, when I first learned about it.

The internet is a great resource for finding answers to questions and uncovering even more questions.

How does one become a world-class tennis player? Practice... practice... practice. The same goes for data science.

Keys to success:

1. Don o don, tolo be ta kalanso.
2. Woshi ji.

Click Here"Assumptions of Linear Regression"

Click Here"Linear Regression"