Finding LS estimates in R

Dr. Lasanthi Watagoda

August 24th 2021

Read data from an URL

```
toluca <- read.table ("http://www.cnachtsheim-text.csom.umn.edu/Kutner/Chapter%20%201%20Data%20Sets/CH0
toluca
```

```
    V1  V2
1   80 399
2   30 121
3   50 221
4   90 376
5   70 361
6   60 224
7  120 546
8   80 352
9  100 353
10  50 157
11  40 160
12  70 252
13  90 389
14  20 113
15 110 435
16 100 420
17  30 212
18  50 268
19  90 377
20 110 421
21  30 273
22  90 468
23  40 244
24  80 342
25  70 323
```

```
toluca <- read.table("http://www.cnachtsheim-text.csom.umn.edu/Kutner/Chapter%20%201%20Data%20Sets/CH01


#Look at the first 6 entries
head(toluca)
```

```
  V1  V2
1 80 399
```

```
2 30 121
3 50 221
4 90 376
5 70 361
6 60 224
```

Rename columns

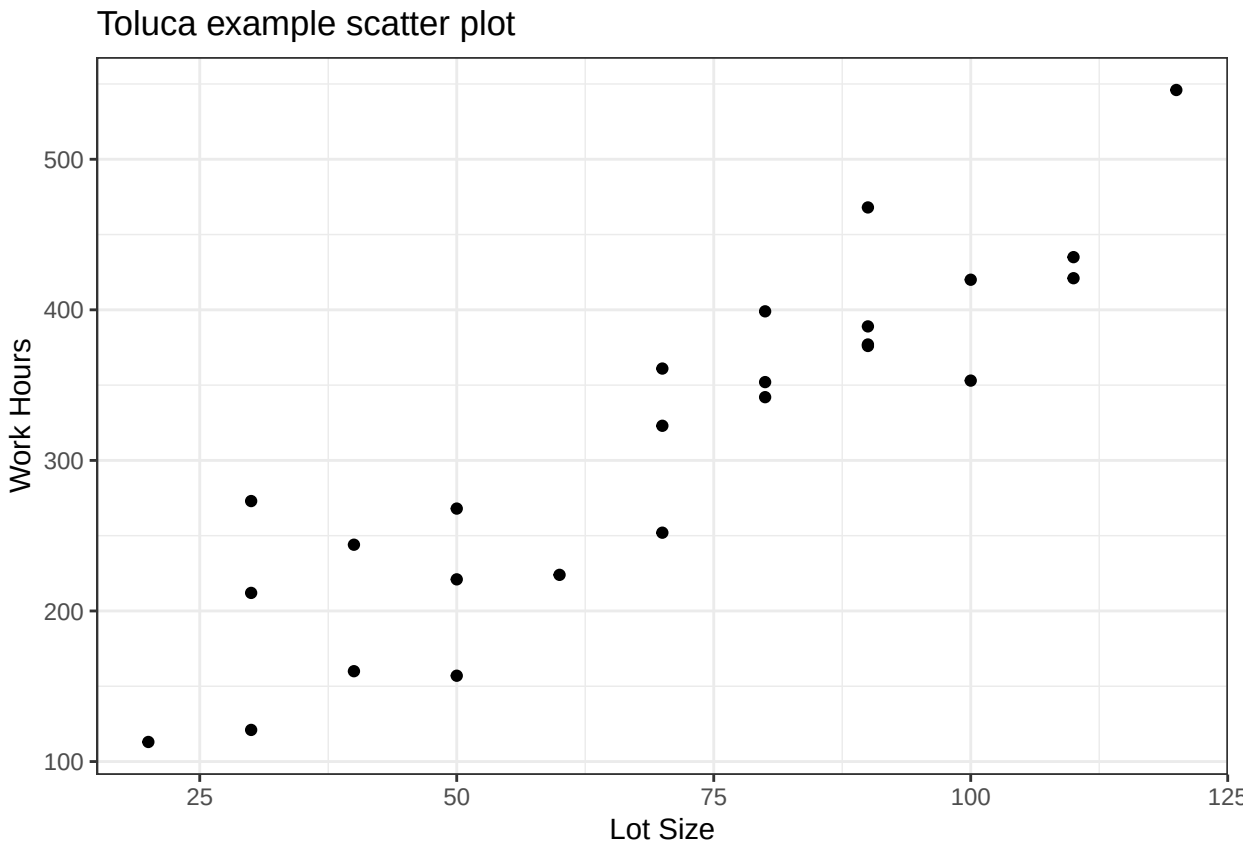
```
colnames(toluca) <- c("lotSize", "hours")

#Look at the first 6 entries
head(toluca)
```

```
  lotSize hours
1      80   399
2      30   121
3      50   221
4      90   376
5      70   361
6      60   224
```

Creating a scatter plot

```
library(ggplot2)
ggplot(toluca, aes(x = lotSize, y = hours)) +
  geom_point() +
  labs(x = "Lot Size", y = "Work Hours", title = "Toluca example scatter plot") +
  theme_bw()
```

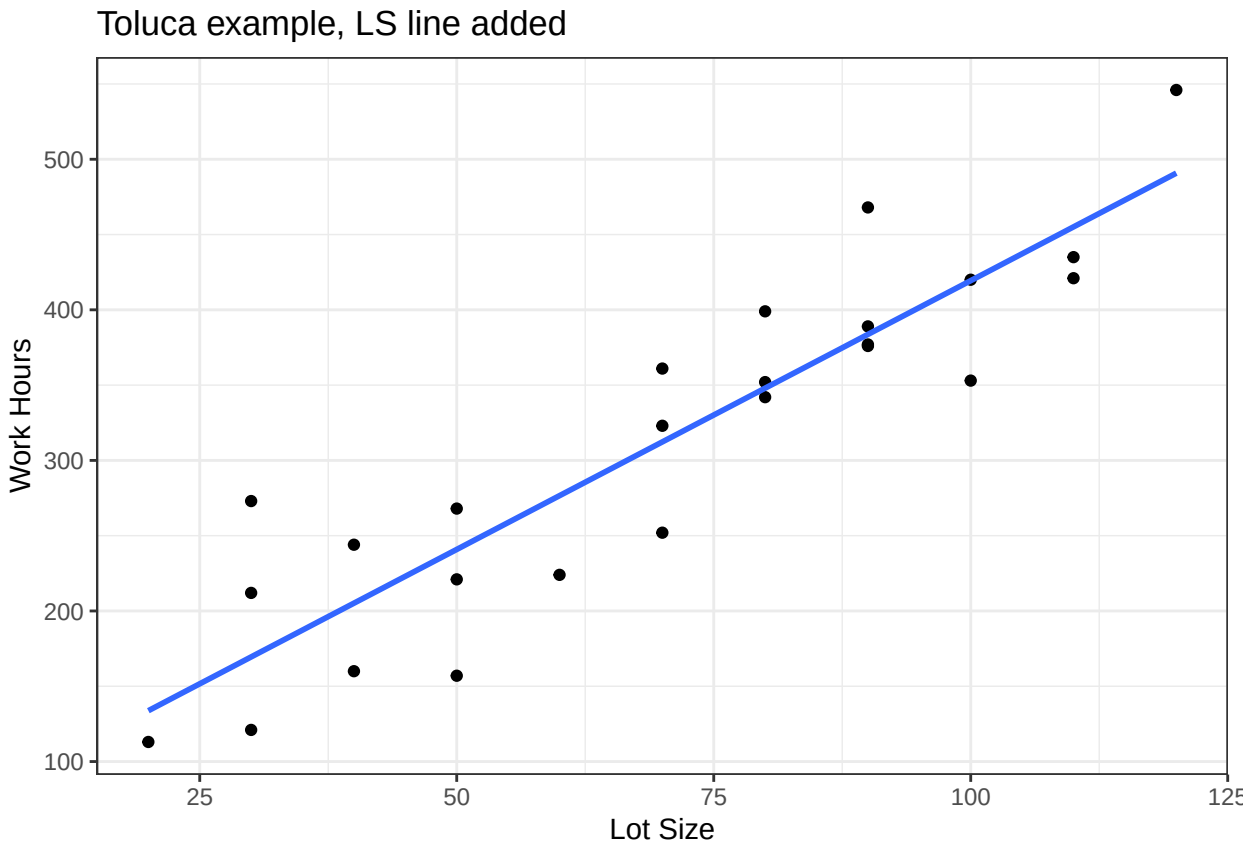



```
# within the ggplot function, first give the name of the data set then... since we need a scatterplot u
```

Note: Lot Size and Work hours has a strong, linear, positive association

Creating a scatter plot, LS line added

```
ggplot(toluca, aes(x = lotSize, y = hours)) +
  geom_point() +
  labs(x = "Lot Size", y = "Work Hours", title = "Toluca example, LS line added") +
  geom_smooth(method = "lm", se = FALSE) +
  theme_bw()
```



Finding the LS estimates

```
toluca_LS_model <- lm(hours ~ lotSize, data = toluca)
summary(toluca_LS_model)
```

```
Call:
lm(formula = hours ~ lotSize, data = toluca)

Residuals:
    Min      1Q  Median      3Q     Max
-83.876 -34.088  -5.982  38.826 103.528

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   62.366     26.177   2.382   0.0259 *
lotSize        3.570      0.347  10.290 4.45e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 48.82 on 23 degrees of freedom
Multiple R-squared:  0.8215,    Adjusted R-squared:  0.8138
F-statistic: 105.9 on 1 and 23 DF,  p-value: 4.449e-10
```

Finding fitted values $\hat{y}_i$, and residuals $e_i = (y_i - \hat{y}_i)$

```
library(moderndive)
Fittedandresiduals <-get_regression_points(toluca_LS_model)
Fittedandresiduals
```

```
# A tibble: 25 x 5
      ID hours lotSize hours_hat residual
   <int> <int>   <int>     <dbl>    <dbl>
 1     1   399      80      348.    51.0
 2     2   121      30      169.   -48.5
 3     3   221      50      241.   -19.9
 4     4   376      90      384.    -7.68
 5     5   361      70      312.    48.7
 6     6   224      60      277.   -52.6
 7     7   546     120      491.    55.2
 8     8   352      80      348.     4.02
 9     9   353     100      419.   -66.4
10    10   157      50      241.   -83.9
# ... with 15 more rows
```

Calculating $SSE = \sum(y_i - \hat{y}_i)^2$

```
sum_of_square_of_residuals <- sum(Fittedandresiduals$residual^2)
sum_of_square_of_residuals
```

```
[1] 54825.46
```

Calculating $MSE = SSE/(n-2)$

```
Mean_Square_Error <- sum_of_square_of_residuals/(nrow(toluca) -2)
Mean_Square_Error
```

```
[1] 2383.716
```

Calculating Residual Standard Error (estimator of standard deviation σ) $s = \sqrt{(MSE)}$

```
s <- sqrt(Mean_Square_Error)
s
```

```
## [1] 48.82331
```