Lab 3 Worksheet

Saksham Goel

January 30, 2018

The Timing of Production Runs.

We have talked about the study of production runs during the lecture. Now it is time for you to explore the data firsthand.

Download data

Download the data **production.txt** from the textbook website <www.stat.tamu.edu/~shether/book>

Read data into R

Read the data into R by using function "read.table( )" and save the data in a variable called "prod". Prod will be a type of R object called data frame.

You can provide a complete address for the file or set the working directory to the folder where you have saved the data:

prod = read.table('/myComputer/myFolder1/myFolder2/production.txt', header = TRUE)

Or

setwd('/myComputer/myFolder1/myFolder2')

prod = read.table('production.txt', header = TRUE)

```
# read data into R.
prod = read.table("production.txt", header= TRUE)
# Use function View(prod) or head(prod) to see if the data has been imported successfully.
#View(prod)
head(prod)
```

```
##   Case RunTime RunSize
## 1    1     195     175
## 2    2     215     189
## 3    3     243     344
## 4    4     162      88
## 5    5     185     114
## 6    6     231     338
```

```
# Advanced: can you try importing data using "read.csv( )" instead of "read.table( )":
csv_type = read.csv("production.txt", header = TRUE, sep = "\t")
#View(csv_type)
head(csv_type)
```

```
##   Case RunTime RunSize
## 1    1     195     175
## 2    2     215     189
## 3    3     243     344
```

```
## 4     4      162       88
## 5     5      185      114
## 6     6      231      338
```

Explore data

summary( ) is a very important function. See what happens when you apply it to a dataframe.

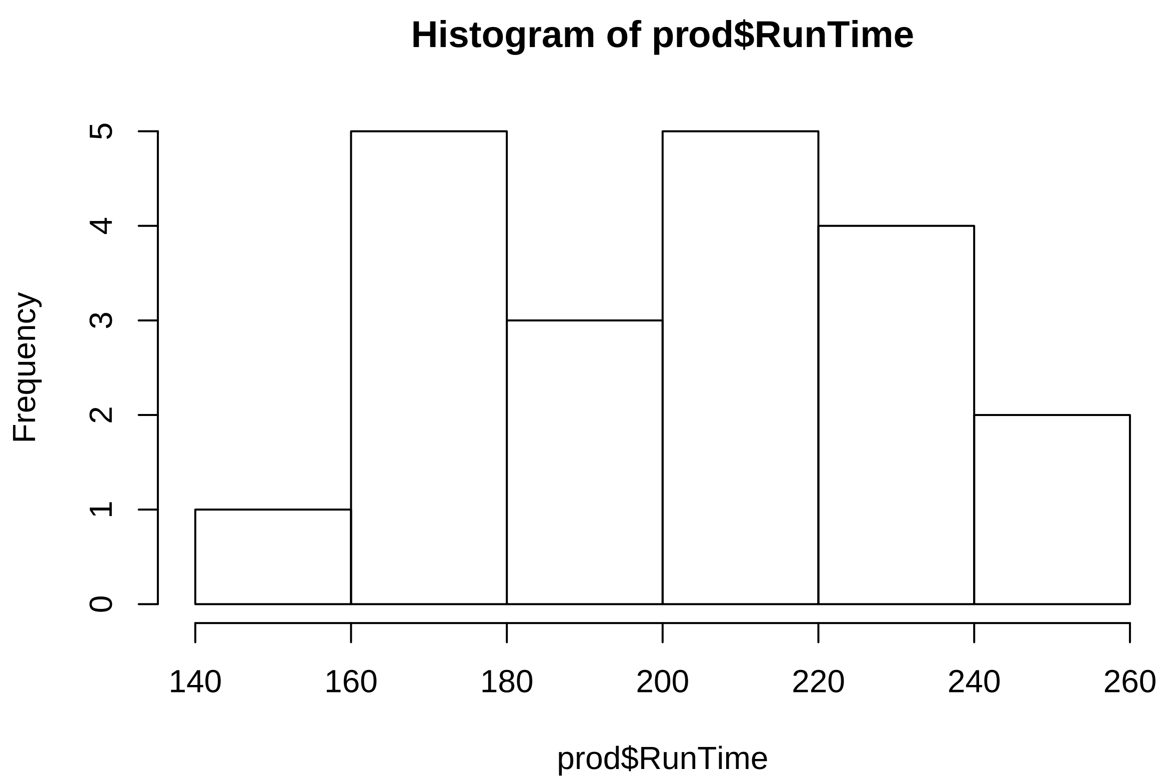
```
# explore data using summary()
summary(prod)
```

```
##       Case          RunTime         RunSize     
##  Min.   : 1.00   Min.   :147.0   Min.   : 58.0  
##  1st Qu.: 5.75   1st Qu.:171.2   1st Qu.:120.8  
##  Median :10.50   Median :207.5   Median :182.0  
##  Mean   :10.50   Mean   :202.1   Mean   :201.8  
##  3rd Qu.:15.25   3rd Qu.:226.2   3rd Qu.:278.8  
##  Max.   :20.00   Max.   :253.0   Max.   :344.0  
```

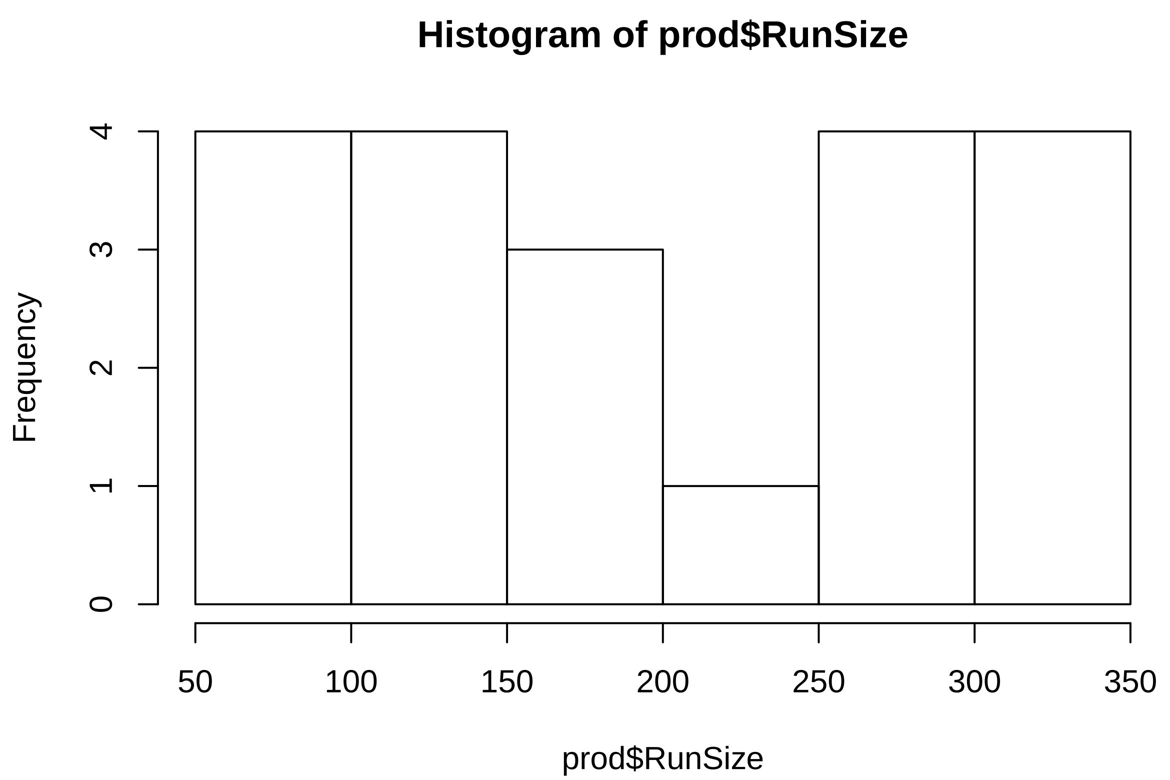
Now let us draw some exploratory plots.

For each variable in the data frame, we can draw a histogram. A histogram can give us an idea of the distribution of a variable.

```
# explore the data with histograms
# Hint: use hist() for histograms
hist(prod$RunTime)
```

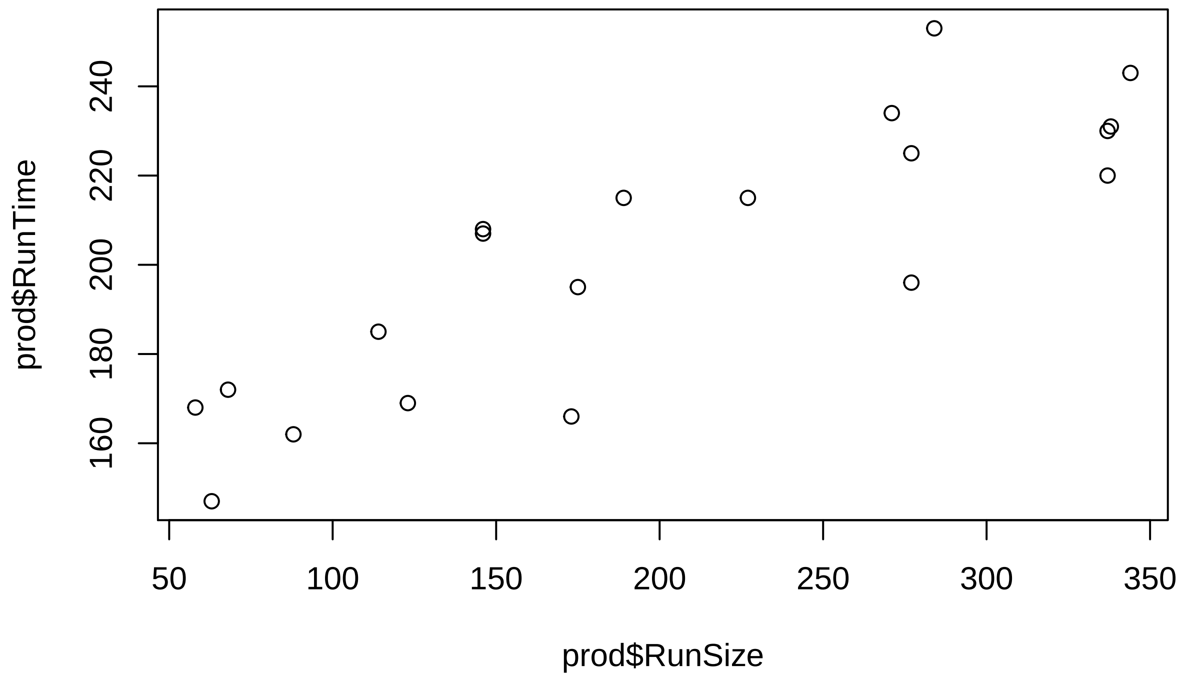


```r
hist(prod$RunSize)
```



Next we draw a scatterplot with one variable (RunSize) on the x axis and another variable (RunTime) on the y axis.

```r
# explore data with scatterplot
# Hint: use plot( )
plot(x = prod$RunSize, y = prod$RunTime)
```



Fit a regression model

The linear trend in the scatterplot seems strong, which means that a linear regression model is appropriate.

```r
# fit a linear regression model
# Hint: use lm( ). Don't forget to save the model to a variable called "mod"
mod = lm( prod$RunTime ~ prod$RunSize )
```

We have applied the function summary( ) to a data frame. This function can also be applied to a model.

```r
# apply summary( ) to your model
summary(mod)
```

```
##
## Call:
## lm(formula = prod$RunTime ~ prod$RunSize)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.597 -11.079   3.329   8.302  29.627
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  149.74770    8.32815   17.98 6.00e-13 ***
## prod$RunSize   0.25924    0.03714    6.98 1.61e-06 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.25 on 18 degrees of freedom
## Multiple R-squared:  0.7302, Adjusted R-squared:  0.7152
## F-statistic: 48.72 on 1 and 18 DF,  p-value: 1.615e-06
```

Please complete the following formula of the model:

Fitted RunTime = 149.74770 + 0.25924 * RunSize

Or we can use mathematical symbols:

Let X = RunSize, Y = RunTime. $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

Compare the observed RunTime and the fitted RunTime

```r
# Print out the observed RunTime
# prod$RunTime
# Print out the observed RunTime
# Hint: Use mod$fitted.values OR use predict(mod)
prod$RunTime
```

```
##  [1] 195 215 243 162 185 231 234 166 253 196 220 168 207 225 169 215 147
## [18] 230 208 172
```

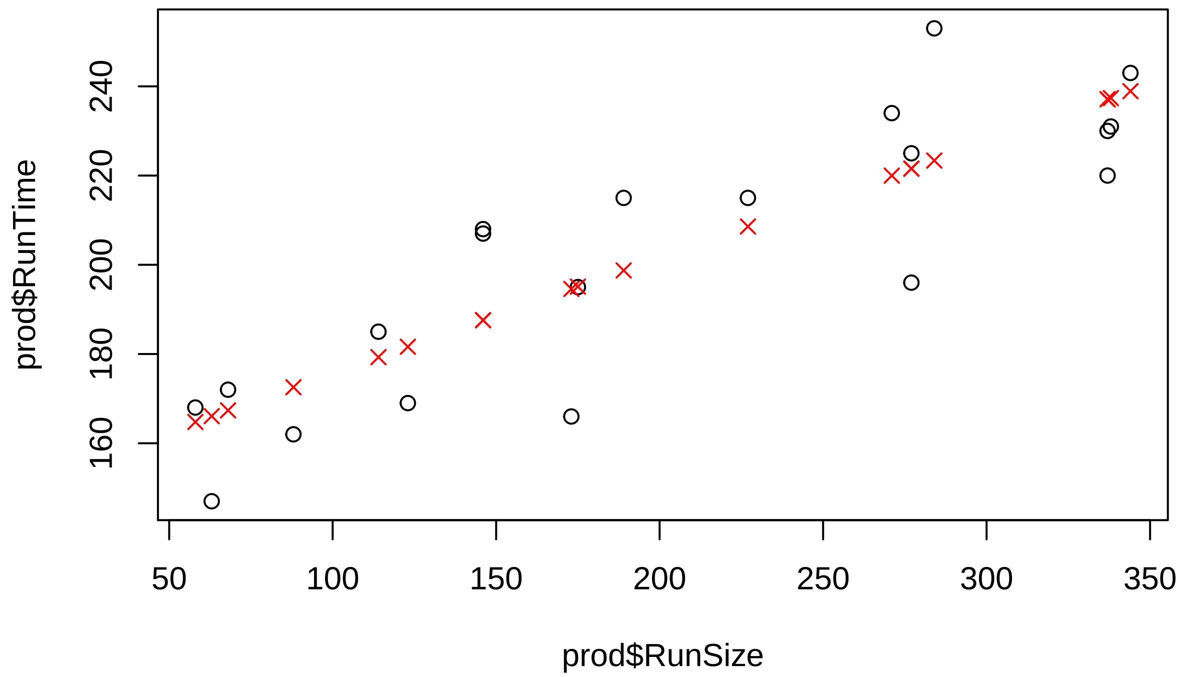
```r
predict(mod)
```

```
##        1        2        3        4        5        6        7        8
## 195.1152 198.7447 238.9273 172.5611 179.3014 237.3719 220.0026 194.5968
##        9       10       11       12       13       14       15       16
## 223.3727 221.5580 237.1126 164.7838 187.5972 221.5580 181.6346 208.5959
##       17       18       19       20
## 166.0800 237.1126 187.5972 167.3762
```

The observed RunTime is different from the fitted RunTime.

visualize the observed RunTime and the fitted RunTime

```r
# Hint: first draw a scatterplot of the data
plot(x = prod$RunSize, y = prod$RunTime)
# Next add the fitted values
points(predict(mod) ~ prod$RunSize, col = 'red', pch = 4)
```



Calculating the estimated coefficients using formulas

We learnt that the following formulas during the lecture:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

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

Let's use these two formulas to estimate the coefficients.

```
x = prod$RunSize
y = prod$RunTime
xbar = mean(x)
ybar = mean(y)

beta1_hat = sum((x-xbar)*(y-ybar)) / sum((x-xbar)^2)
beta1_hat
```

```
## [1] 0.2592431
```

```
beta0_hat = ybar - beta1_hat*xbar
beta0_hat
```

```
## [1] 149.7477
```

Compare beta1_hat and beta2_hat with the estimated coefficients obtained from the linear regression model. They should be identical!

\ *Congratulations! You've fitted a simple linear regression model. Please generate the pdf and wait for further instruction.*