lab3_challenge

Saksham Goel

January 30, 2018

The Timing of Processing Invoices.

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

```
##   Day Invoices Time
## 1   1      149  2.1
## 2   2       60  1.8
## 3   3      188  2.3
## 4   4       23  0.8
## 5   5      201  2.7
## 6   6       58  1.0
```

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

```
##   Day Invoices Time
## 1   1      149  2.1
## 2   2       60  1.8
## 3   3      188  2.3
```

```
## 4   4        23  0.8
## 5   5       201  2.7
## 6   6        58  1.0
```

Explore data

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

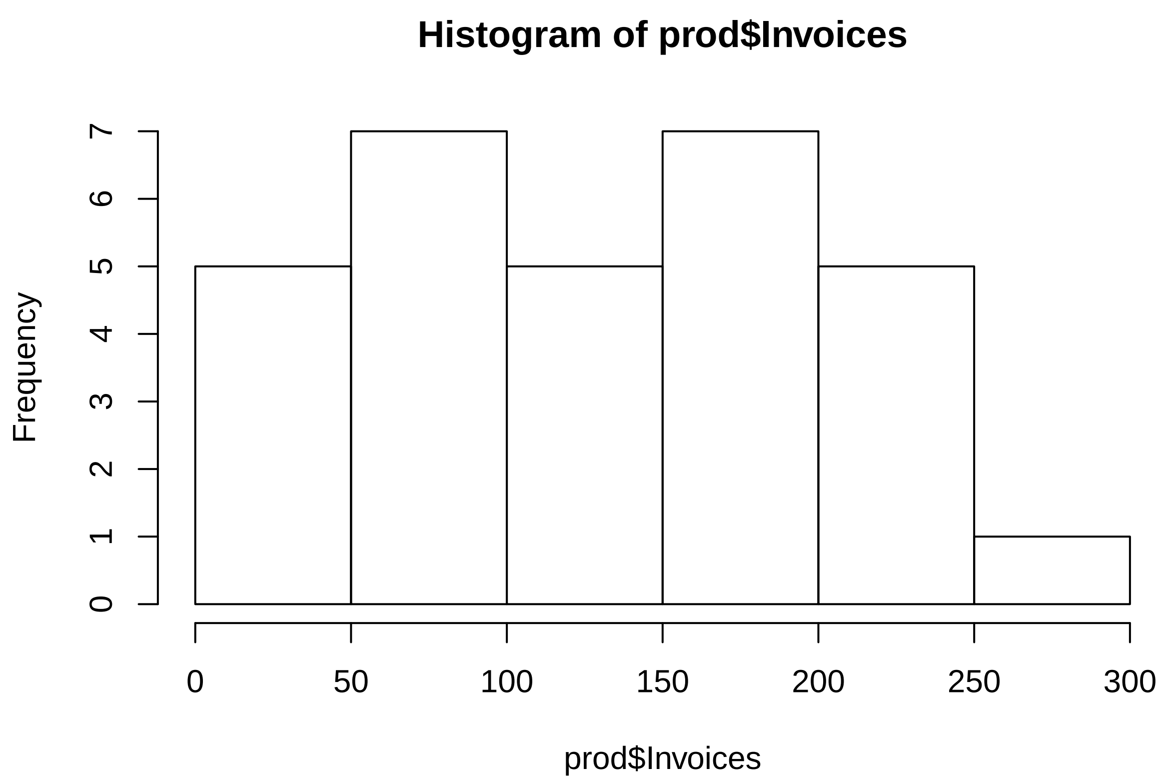
```
# explore data using summary()
summary(prod)
```

```
##       Day           Invoices          Time      
##  Min.   : 1.00   Min.   : 23.0   Min.   :0.800  
##  1st Qu.: 8.25   1st Qu.: 62.5   1st Qu.:1.500  
##  Median :15.50   Median :127.5   Median :2.000  
##  Mean   :15.50   Mean   :130.0   Mean   :2.110  
##  3rd Qu.:22.75   3rd Qu.:189.5   3rd Qu.:2.775  
##  Max.   :30.00   Max.   :289.0   Max.   :4.100
```

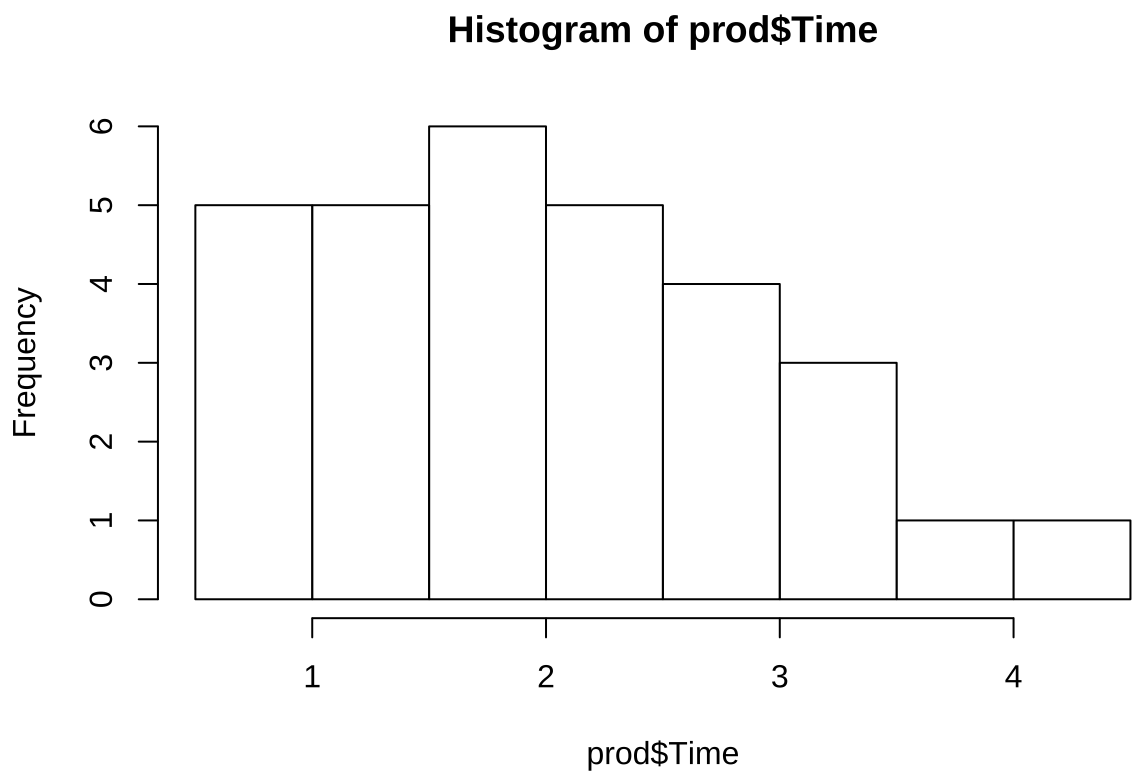
Now let us draw some exploratory plots.

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$Invoices)
```

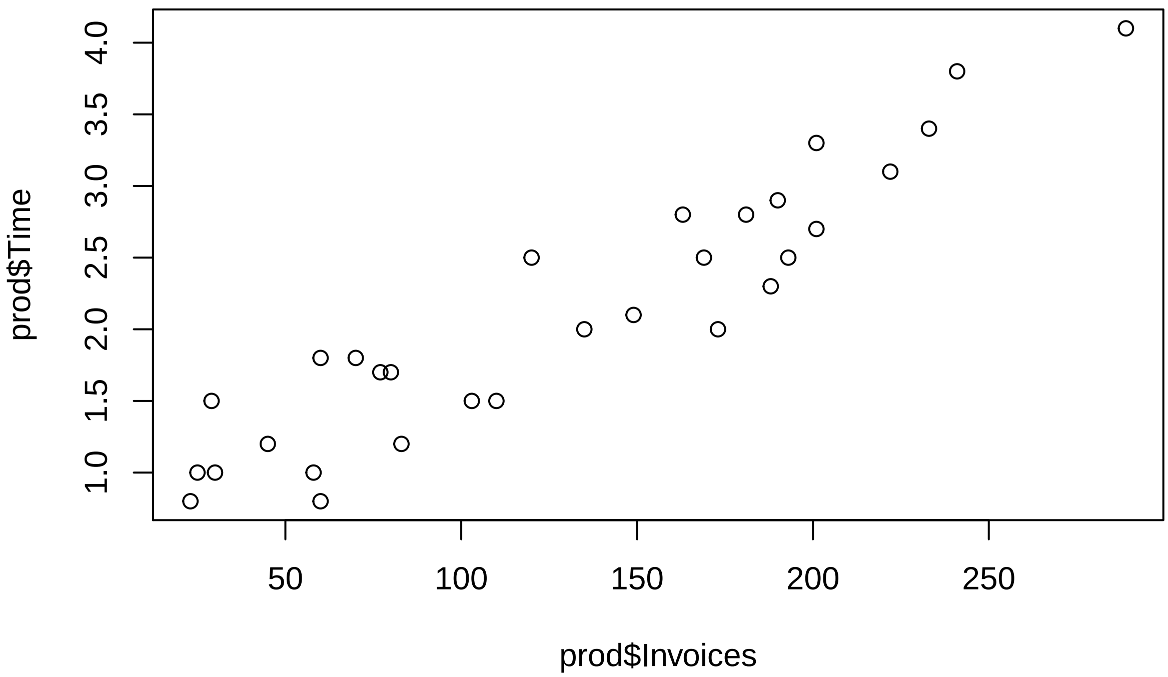


```r
hist(prod$Time)
```



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$Invoices, y = prod$Time)
```



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$Time ~ prod$Invoices )
```

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$Time ~ prod$Invoices)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59516 -0.27851  0.03485  0.19346  0.53083
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.6417099  0.1222707   5.248 1.41e-05 ***
## prod$Invoices 0.0112916  0.0008184  13.797 5.17e-14 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3298 on 28 degrees of freedom
## Multiple R-squared:  0.8718, Adjusted R-squared:  0.8672
## F-statistic: 190.4 on 1 and 28 DF,  p-value: 5.175e-14
```

Please complete the following formula of the model:

Fitted Time = 0.6417099 + 0.0112916 * Invoices

Or we can use mathematical symbols:

Let X = Invoices, Y = Time. $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

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$Time
```

```
##  [1] 2.1 1.8 2.3 0.8 2.7 1.0 1.7 3.1 2.8 1.0 1.5 1.2 0.8 1.0 2.0 2.5 2.9
## [18] 3.4 4.1 1.2 2.5 1.8 3.8 1.5 2.8 2.5 3.3 2.0 1.7 1.5
```

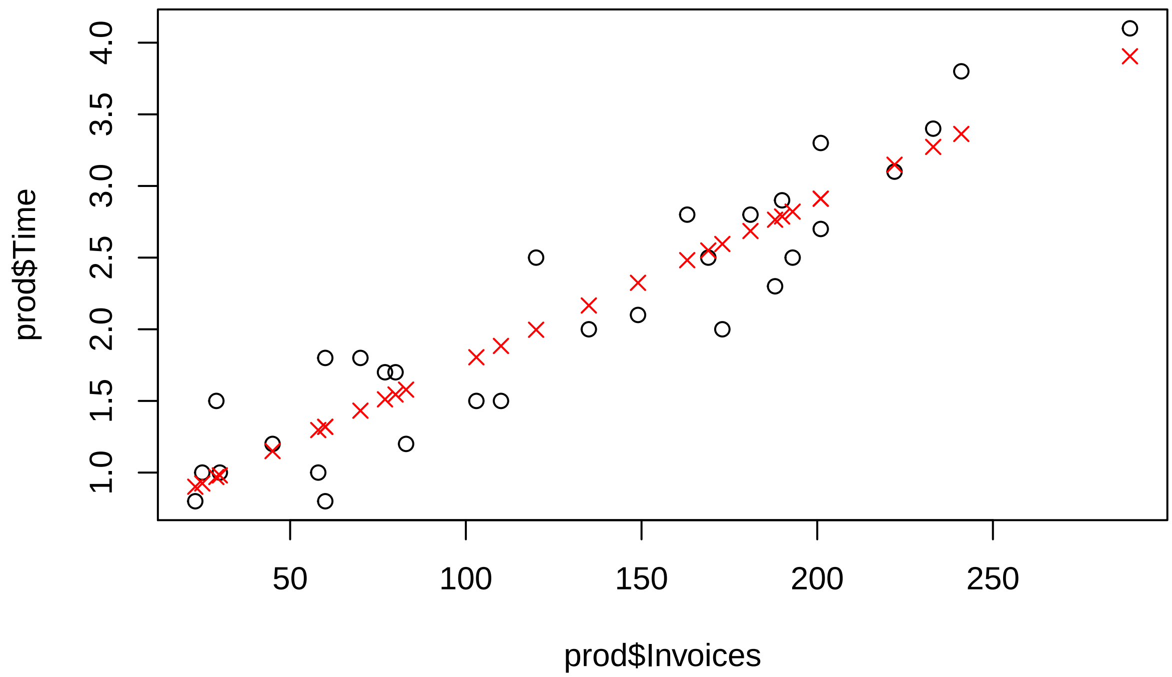
```r
predict(mod)
```

```
##         1         2         3         4         5         6         7
## 2.3241648 1.3192085 2.7645390 0.9014177 2.9113303 1.2966252 1.5111665
##         8         9        10        11        12        13        14
## 3.1484549 2.6854975 0.9804592 1.8837907 1.5789163 1.3192085 0.9240010
##        15        16        17        18        19        20        21
## 2.5951643 2.5499977 2.7871223 3.2726630 3.9049950 1.1498339 2.8209972
##        22        23        24        25        26        27        28
## 1.4321250 3.3629961 1.8047492 2.4822479 1.9967072 2.9113303 2.1660818
##        29        30
## 1.5450414 0.9691676
```

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$Invoices, y = prod$Time)
# Next add the fitted values
points(predict(mod) ~ prod$Invoices, 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$Invoices
y = prod$Time
xbar = mean(x)
ybar = mean(y)

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

```
## [1] 0.01129164
```

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

```
## [1] 0.6417099
```

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