PSTAT 126 - Lab 1

Fall 2022

1 R Markdown

You will use R Markdown for assignments. Refer to the excellent online book, R Markdown Cookbook, for documentation and exploring R Markdown format's rich set of features.

Following section is taken from Introduction to R Markdown

1.1 Introduction to R Markdown

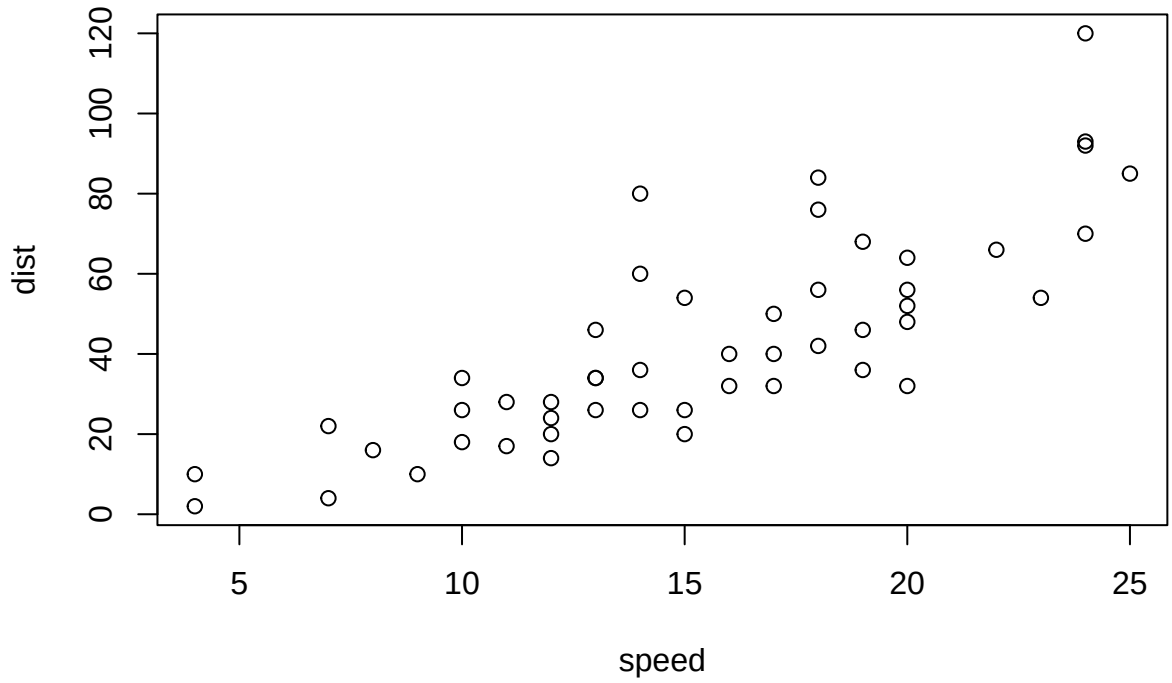
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see .

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot. View the chapter on chunk options.

Markdown can also display LaTeX equations:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

View the chapter on LaTex output.

2 Simple Linear Regression

In this section, we will manually compute elements of simple linear regression (SLR) model.

Data: Applied Linear Statistical Models (Fifth Edition) by Michael H. Kutner

2.1 Reading Data into R

Confirm that the data file, `copier.txt`, exists. The command to read the file is

```
copier <- read.table("copier.txt", header = FALSE)
```

The argument `header = FALSE` because `copier.txt` does not have a header row. Now view first few lines of `copier`:

```
head(copier)
```

```
##    V1 V2
## 1  20  2
```

```
## 2  60  4
## 3  46  3
## 4  41  2
## 5  12  1
## 6 137 10
```

The response variable, Y, is the total number of minutes (column `V1`) spent by the service time for maintaining a number copiers (column `V2`), which is the independent variable X. Let's set suitable column names.

```
colnames(copier) <- c("service_time", "copiers")
```

Execute the summary function to view summary statistics for each column:

```
summary(copier)
```

```
##   service_time       copiers
##  Min.   :  3.00   Min.   : 1.000
##  1st Qu.: 36.00   1st Qu.: 2.000
##  Median : 74.00   Median : 5.000
##  Mean   : 76.27   Mean   : 5.111
##  3rd Qu.:111.00   3rd Qu.: 7.000
##  Max.   :156.00   Max.   :10.000
```

2.2 OLS Solution: Coefficient Estimates

The OLS estimators that minimize mean squared error (MSE) were derived in lecture:

$$\beta_0^* = \mu_Y - b_1^* \mu_X, \quad \beta_1^* = \frac{\text{Cov}(X,Y)}{\sigma_X^2},$$

where the regression function is

$$E(Y_i) = \beta_0 + \beta_1 X_i.$$

Since we do not know the true values μ_X, μ_Y, σ_X^2, σ_Y^2, and $\text{Cov}(X,Y)$, we compute estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ using quantities estimated data:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 \quad = \frac{\text{SXY}}{\text{SXX}},$$

where sample estimates are computed as defined in the following table:

True quantity	Sample quantity	Formula for sample estimate	Description
$E(X)$	$\bar{x}$	$\sum x_i/n$	Sample average of x
$E(Y)$	$\bar{y}$	$\sum y_i/n$	Sample average of y
	SXX	$\sum (x_i - \bar{x})^2 = \sum (x_i - \bar{x})\, x_i$	Sum of squares for the x's
$\text{Var}(X) = \sigma_X^2$	SD_x^2	$\text{SXX}/(n-1)$	Sample variance of the x's
σ_X	SD_x	$\sqrt{\text{SXX}/(n-1)}$	Sample standard deviation of the x's
	SYY	$\sum (y_i - \bar{y})^2 = \sum (y_i - \bar{y})\, y_i$	Sum of squares for the y's
$\text{Var}(Y) = \sigma_Y^2$	SD_y^2	$\text{SYY}/(n-1)$	Sample variance of the y's
σ_Y	SD_y	$\sqrt{\text{SYY}/(n-1)}$	Sample standard deviation of the y's
	SXY	$\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i - \bar{x})\, y_i$	Sum of cross-products
$\text{Cov}(X,Y)$	s_{xy}	$\text{SXY}/(n-1)$	Sample covariance
$\text{Corr}(X,Y)$	r_{xy}	$s_{xy}/(\text{SD}_x \text{SD}_y)$	Sample correlation

```
n = nrow(copier)
mx = mean(copier$copiers)
my = mean(copier$service_time)
SXX = sum((copier$copiers - mx)^2)
SYY = sum((copier$service_time - my)^2)
SXY = sum((copier$copiers - mx)*(copier$service_time - my))
beta1 = SXY/SXX
beta0 = my - beta1*mx
```

Estimated regression function is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = -0.5801567 + 15.035248 x_i$$

2.3 OLS Solution: Residual and Coefficient Variances

Recall the variance of error term is $\text{Var}(\epsilon_i) = \sigma^2$.

The residual $e_i = y_i - \hat{y}_i$ is used to compute $\text{RSS} = \sum_{i=1}^{n} e_i^2$, which is then used to estimate the error variance and the coefficient variance:

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n-2}\text{RSS} \\
\widehat{\text{Var}\left(\hat{\beta}_1\right)} &= \frac{\hat{\sigma}^2}{\text{SXX}} \\
\widehat{\text{Var}\left(\hat{\beta}_0\right)} &= \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\text{SXX}}\right)
\end{aligned}$$

```
e = copier$service_time - (beta0 + beta1*copier$copiers)
RSS = sum(e^2)
MSE = RSS/(n-2)
var_beta1 = MSE/SXX
var_beta0 = MSE*(1/n + (mx^2)/SXX)
```

Following are the computed estimates:

$$\begin{aligned}
\hat{\sigma}^2 &= 79.4506285 \\
\widehat{\text{Var}\left(\hat{\beta}_1\right)} &= 0.2333733 = 0.4830872^2 \\
\widehat{\text{Var}\left(\hat{\beta}_0\right)} &= 7.8620857 = 2.8039411^2
\end{aligned}$$

2.4 OLS Solution: Built-in `lm` function

Run the `lm` function in R:

```
lm_copier = lm(service_time ~ copiers, data = copier)
summary(lm_copier)
```

```
##
## Call:
## lm(formula = service_time ~ copiers, data = copier)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -22.7723  -3.7371   0.3334   6.3334  15.4039
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.5802     2.8039  -0.207    0.837
## copiers      15.0352     0.4831  31.123   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.914 on 43 degrees of freedom
## Multiple R-squared:  0.9575, Adjusted R-squared:  0.9565
## F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

where `(Intercept)` is β_0 and `copiers` is β_1.

In the summary output, can you locate the quantities we computed manually?

We can also print just the coefficient estimates:

```
coef(lm_copier)
```

```
## (Intercept)     copiers
##  -0.5801567  15.0352480
```

Print variance-covariance estimates of the regression coefficients:

```
vcov(lm_copier)
```

```
##             (Intercept)    copiers
## (Intercept)    7.862086 -1.1927966
## copiers       -1.192797  0.2333733
```

Locate the two variances of regression coefficients in the output. What does the third quantity represent?