Linear Least Squares Example

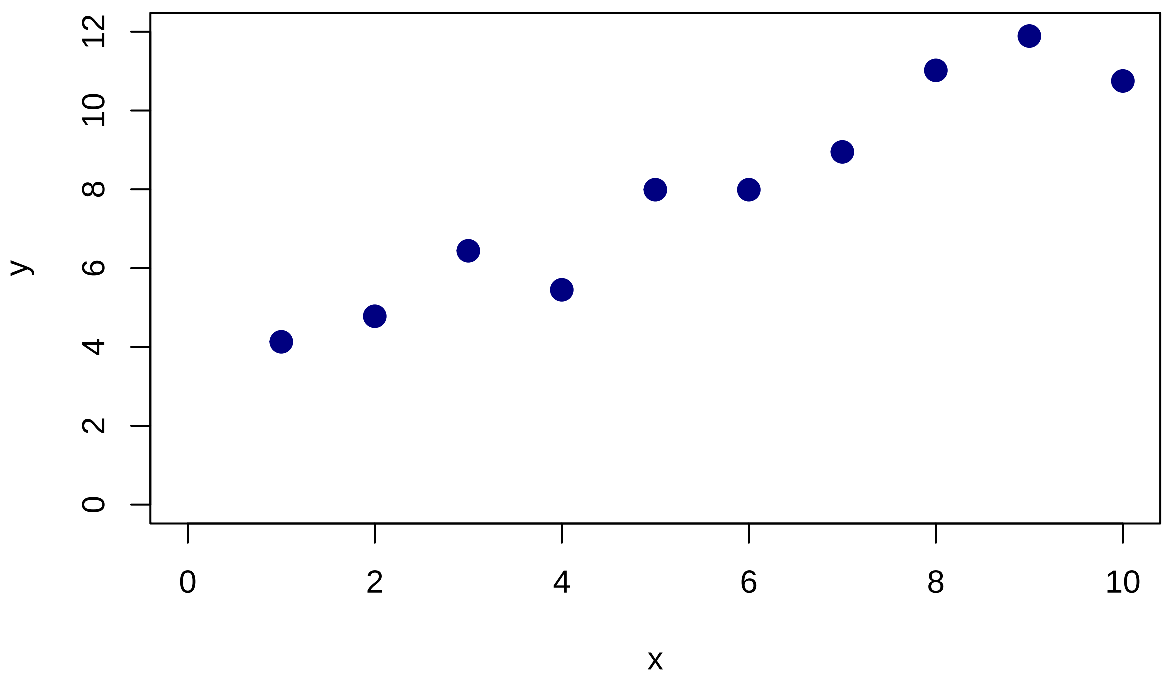
Sampling the data

Consider the following observations $Y_1, Y_2, \ldots, Y_n$ following linear model $E(Y) = \beta_0 + \beta_1 x$:

x	1	2	3	4	5	6	7	8	9	10
y	4.13	4.78	6.44	5.45	7.99	7.99	8.95	11.02	11.89	10.75

The following code plots each point (x, y) in the plane.

```
x <- 1:10
y <-  c(4.13,  4.78,  6.44,  5.45,  7.99,  7.99,  8.95, 11.02, 11.89, 10.75)
plot(x, y, xlim = c(0,10), ylim = c(0,12), col = "navy", pch = 19, cex = 1.5)
```



scatterplot-1.pdf

Estimating the Linear Model

We can calculate the terms S_{xy} and S_{xx} in the formula for $\hat{\beta}_1$ using the identities:

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i$$
$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2$$

The following code calculates the necessary intermediate values.

```
# Number of observations.
n <- length(x)

# Sums.
sumx <- sum(x = x)
sumy <- sum(x = y)
sumxy <- sum(x = x*y)
sumxx <- sum(x = x^2)
```

This yields:

$$\sum_{i=1}^n x_i = 55 \qquad \sum_{i=1}^n y_i = 79.39 \qquad \sum_{i=1}^n x_i y_i = 508.02 \qquad \sum_{i=1}^n x_i^2 = 385$$

The following code uses these sums to calculate S_{xx}, S_{xy}, and $\hat{\beta}_1$.

```
# Calculate Sxy and Sxx
Sxy <- sumxy - 1/n*sumx*sumy
Sxx <- sumxx - 1/n*sumx^2

# Calculate hat beta_1
hbeta1 <- Sxy/Sxx
```

This yields

$$S_{xy} = 71.375 \qquad S_{xx} = 82.5$$

and, therefore,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = 0.8651515.$$

On the other hand, we can calculate $\hat{\beta}_0$ using the identity

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{1}{n} \left(\sum_{i=1}^n y_i - \hat{\beta}_1 \sum_{i=1}^n x_i \right),$$

which can be evaluated as follows.

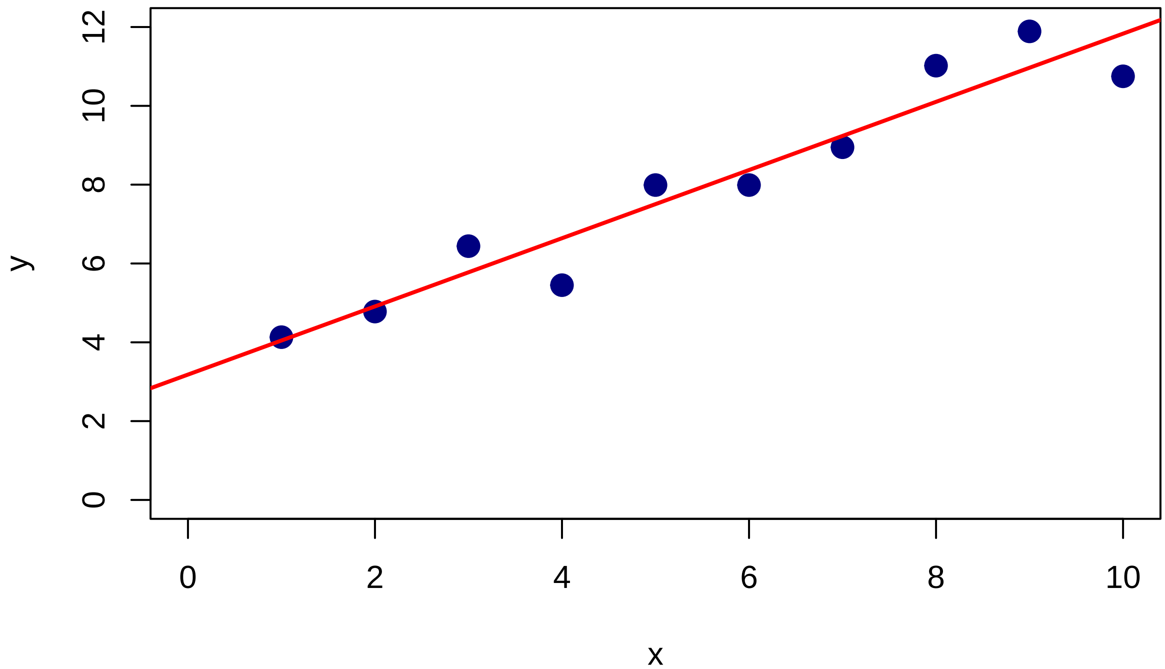
```
hbeta0 <- (sumy - hbeta1*sumx)/n
```

This yields $\hat{\beta}_0 = 3.1806667$. The fitted line is

$$\hat{y} = 3.1806667 + (0.8651515)x.$$

We can plot the fitted line, as well as the observed data using the **abline** function, which plots a line with given slope and intercept.

```
plot(x, y, xlim = c(0,10), ylim = c(0,12), col = "navy", pch = 19, cex = 1.5)
abline(hbeta0, hbeta1, col = "red", lwd = 2)
```



fitted line-1.pdf

Finding the Variances of the Estimators

We can calculate the variance of the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ using the identities

$$V(\hat{\beta}_0) = \frac{\sigma^2 \sum x_i^2}{nS_{xx}}, \qquad V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}.$$

For this particular example, we have $S_{xx} = 82.5$ and $\sum x_i^2 = 385$, which, in turn, yields

$$V(\hat{\beta}_0) = \frac{385}{(10)(82.5)}\sigma^2 = 0.4666667\sigma^2, \qquad V(\hat{\beta}_1) = \frac{\sigma^2}{82.5}.$$

We can also calculate the covariance between the two estimators $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$ using the formula

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\bar{x}\sigma^2}{S_{xx}} = -\frac{\sum x}{nS_{xx}}\sigma^2.$$

We can calculate the coefficient of σ^2 in this expression using the following code.

```
covFactor <- -(sumx)/(n*Sxx)
```

This gives covariance $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -0.0666667\sigma^2$.

Estimating the Population Variance

In order to estimate σ^2, we need to calculate the sum of squared errors (SSE), which can be accomplished using the formula

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 - \hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = S_{yy} - \hat{\beta}_1 S_{xy}.$$

To evaluate this formula, we can need to calculate

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2,$$

which can be accomplished using the following code.

```
# Calculate sum of squared ys.
sumyy <- sum(y^2)

# Calculate Syy
Syy = sumyy - sumy^2/n
```

Using the calculated values, we obtain

$$SSE = S_{yy} - \hat{\beta}_1 S_{xy} = 66.96189 - (0.8651515)(71.375) = 5.2117006,$$

which could also be evaluated using the code.

```
SSE <- Syy - hbeta1*Sxy
```

The estimate of the variance σ^2 given by $SSE/(n-2)$ can be evaluated using the following code.

```
Ssquared <- SSE/(n-2)
```

This gives $s^2 = 0.6514626$.

Using the LM function

We can also calculate the linear model using the R function **lm**. The following code saves the output of the linear regression function to the list **mod**.

```
mod <- lm(y~x)
```

We can view the estimated coefficients of the linear model, i.e., $\hat{\beta}_0$ and $\hat{\beta}_1$, calculated by **lm** using the following code:

```
mod$coefficients
```

```
## (Intercept)           x
##   3.1806667   0.8651515
```

Note that the intercept and slope terms calculated by **lm** agree with the values of $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively, calculated above.