Working Lab 1

Lecture 5

Joaquin Cavieres

2023-05-17

Exercise

Consider the following set of simulated data:

```
set.seed(123)
n <- 500
x <- rnorm(n, mean = 5, sd=2)
beta_0 <- 2
beta_1 <- 0.5
epsilon <- rnorm(n, mean=0, sd = sqrt(3))
y <- beta_0 + x * beta_1 + epsilon
```

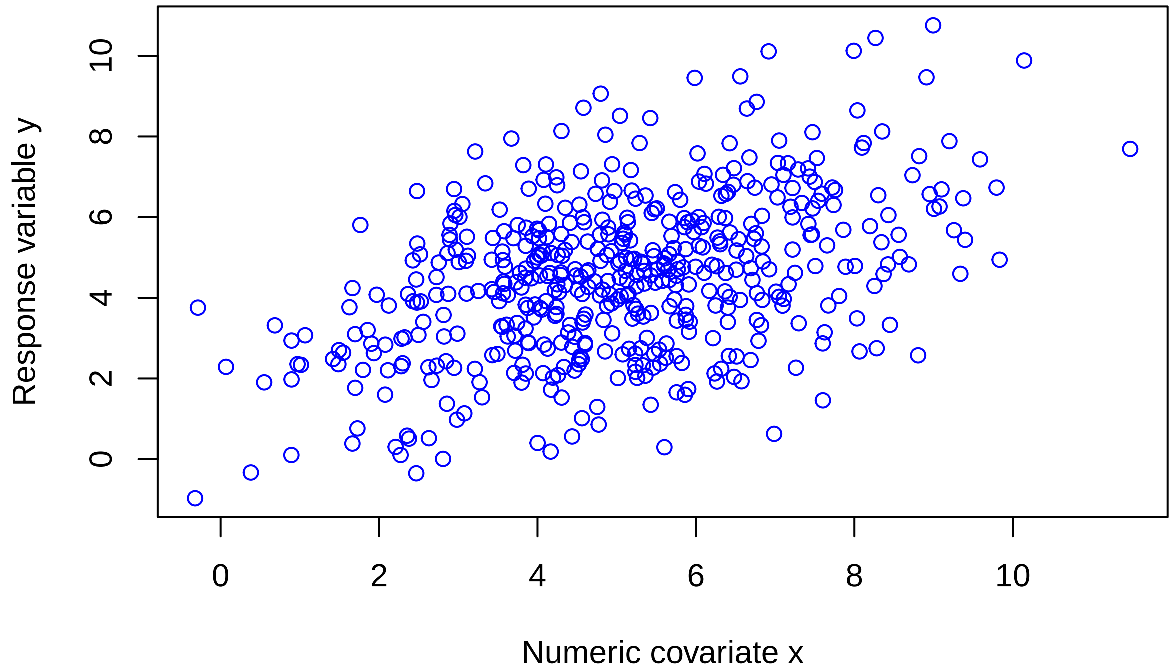
From the simulated data we are interested in estimate the parameters (coefficients) β_0 and β_1 in a simple linear regression. Considering the above;

1) Make a plot to show the relationship between the response variable and the independent covariate.
2) Using the equations of the OLS method, compute $\hat{\beta}_0$ and $\hat{\beta}_1$.
3) Calculate the residuals of the model.
4) Using the `lm()` function of R, calculate $\hat{\beta}_0$ and $\hat{\beta}_1$ and compare them with the obtained in the point 2.
5) How can you know if the model has a good fit?

Results

1) Make a plot to show the relationship between the response variable and the independent covariate.

```
plot(x, y, type = "p", col = "blue", xlab = "Numeric covariate x", ylab = "Response variable y")
```



2) Using the equations of the OLS method, compute $\hat{\beta}_0$ and $\hat{\beta}_1$.

For $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}, \tag{1}$$

$$\tag{2}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}. \tag{3}$$

The predicted (fitted) values $\hat{Y}_i$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \tag{4}$$

and the residuals $\hat{\epsilon}_i$

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i. \tag{5}$$

```
sxy <- sum((x - mean(x)) * (y - mean(y)))
sxx <- sum((x - mean(x))^2)
hat_beta1 <- sxy / sxx
hat_beta0 <- mean(y) - hat_beta1 * mean(x)

hat_beta0
```

```
## [1] 2.232899
```

```
hat_beta1
```

```
## [1] 0.4532581
```

3) Calculate the residuals of the model.

```
yhat <- hat_beta0  + hat_beta1*x
resid_model <- y - yhat
head(resid_model, 10)  # only the 10 firts values
```

```
##  [1] -1.094094189 -1.741844116  1.924968076  1.308277928 -2.601056610
##  [6] -0.003659211 -1.507928748 -3.704098608  0.196616273 -0.178050684
```

4) Using the `lm()` function of R, calculate $\hat{\beta}_0$ and $\hat{\beta}_1$ and compare them with the obtained in the point 3.

```
linear_fit <- lm(y ~ x)
coef(linear_fit)
```

```
## (Intercept)           x
##   2.2328993   0.4532581
```

5) How can you know if the model has a good fit?

Using the diagnostics of the mode, and some of them are:

- The R^2
- Considering the assumptions of the model, check the normality of the residuals.

```
summary(linear_fit)$r.squared     # R2
```

```
## [1] 0.2027639
```

```
hist(linear_fit$residuals)        # Are normallly distribuited?
```