Lab Week 03 - 01-24-2022

Pouria

1/21/2022

Lab 1

In this first exercise, we will simulate different datasets

$$y = 2 + 3 * X + \sigma,$$

with X randomly generated via a uniform distribution from 0 to 10, and σ generated via a normal (Gaussian) distribution of 0 mean and variance 1 (which we note, $\sigma \sim N(0, 1)$)

Lab 1(a)

Generate 10 “experiments” with 5 observations each. Compute the slopes and the intercepts using the formula of Equation 3.4, check that the values are the same as given by the lm function in R , and plot the 10 different lines. Also plot in bold the “true” line.

Coefficient Estimation Function

Better define your computational algorithms as functions

Let's define the equations of the least squares approach to obtain values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the **residual sum of squares** (RSS).

```
Eq3.4 <- function(X, y){

  # Equations 3.4 to estimate beta0 and beta1
  beta1 = sum((X-mean(X))*(y-mean(y))) / sum((X-mean(X))^2)
  beta0 = mean(y) - beta1*mean(X)

  return(list(beta0, beta1))
}
```

The Loop of Experiments Function

Define the experiments in a loop as a function

```
Experiment <- function(N.Obs, N.Exp, X.min, X.max, sigma.mu

  beta0 <- beta1 <- 0

  for (i in 1:N.Exp){
    X = runif(N.Obs, min = X.min, max = X.max)
    sigma = rnorm(N.Obs, mean = sigma.mu, sd = sigma.sd)
    y = 2 + 3*X + sigma

    coeff <- Eq3.4(X, y)
    beta0[i] <- coeff[[1]]
    beta1[i] <- coeff[[2]]

    fit.lm <- lm(y~X)
    beta0_lm <- fit.lm$coefficients[1]
    beta1_lm <- fit.lm$coefficients[2]
```

Visualization function

Plot the 10 different lines and the “true” line

```
Visualize <- function(Coeff.df){
  ggplot() +
    geom_abline(data = Coeff.df, aes(slope=beta1 , intercep
    geom_abline(aes(slope = 3, intercept = 2), size=1.5, co
    scale_x_continuous(name="X", limits=c(-2,2)) +
    scale_y_continuous(name="y", limits=c(-10,10))
}
```

Solution 1(a)

Always define your parameters first.

Let's assign values to the parameters in the problem:

```
N.Exp = 10
N.Obs = 5
X.min = 0
X.max = 10
sigma.mu = 0
sigma.sd  = 1
```

Solution 1(a)

Coefficients for 5 observations from 10 experiments

```
Coeff.df <- Experiment(N.Obs, N.Exp, X.min, X.max, sigma.mu
```

```
## Warning in data.frame(beta0, beta1, beta0_lm, beta1_lm):
## from a short variable and have been discarded
```

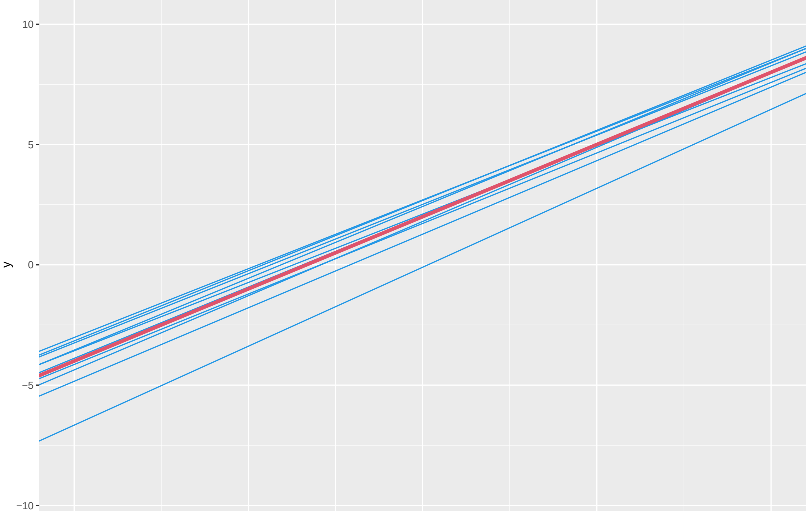
```
Coeff.df
```

```
##         beta0    beta1 beta0_lm beta1_lm
## 1   2.6726237 2.917663 1.718374  2.93079
## 2   2.1061998 2.840345 1.718374  2.93079
## 3   2.5112191 2.880201 1.718374  2.93079
## 4   2.6970558 2.858884 1.718374  2.93079
## 5  -0.1010987 3.281283 1.718374  2.93079
## 6   1.8003284 3.084352 1.718374  2.93079
## 7   2.0384504 2.962657 1.718374  2.93079
## 8   2.4240832 2.984535 1.718374  2.93079
```

Solution 1(a)

Visualize for 5 observations from 10 experiments

```
Visualize(Coeff.df)
```



Lab 1(b)

Now repeat 10 experiments, but with 20 observations each.

Solution 1(b)

Let's assign values to the parameters for problem 1(b):

```
N.Exp = 10
N.Obs = 20
X.min = 0
X.max = 10
sigma.mu = 0
sigma.sd  = 1
```

Solution 1(b)

Coefficients for 20 observations from 10 experiments

```r
Coeff.df <- Experiment(N.Obs, N.Exp, X.min, X.max, sigma.mu
```

```
## Warning in data.frame(beta0, beta1, beta0_lm, beta1_lm):
## from a short variable and have been discarded
```

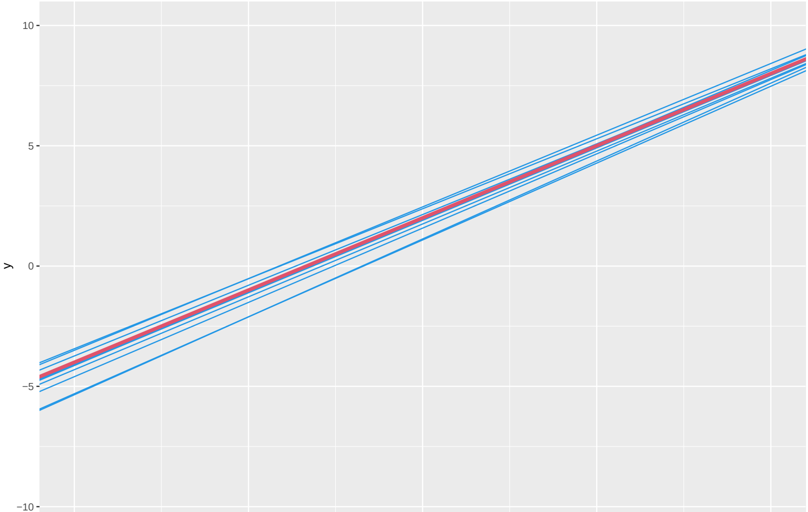
```r
Coeff.df
```

```
##        beta0    beta1 beta0_lm beta1_lm
## 1  1.576291 3.086832  1.99722 3.065514
## 2  2.140904 2.938455  1.99722 3.065514
## 3  2.458869 2.980010  1.99722 3.065514
## 4  1.126846 3.237989  1.99722 3.065514
## 5  1.749408 3.028091  1.99722 3.065514
## 6  1.899592 3.010784  1.99722 3.065514
## 7  1.949224 3.014224  1.99722 3.065514
## 8  1.083933 3.194615  1.99722 3.065514
```

Solution 1(b)

Visualize for 5 observations from 10 experiments

```
Visualize(Coeff.df)
```



Lab 1(c)

For each 5 and 20 observations, use the formula of equations 3.8 to compute the SE for the slope. Why is the SE smaller for 20 observations?

Equation 3.8:

$$SE(\hat{\beta}_0)^2 = \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})}],$$

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