OLS – Lecture 3

Exercise: $Y = \{5, 2, 2, 3\}$, $X = \{5, 3, 5, 3\}$

$$b_1 = \frac{\sum_{i=1}^{n} \left[\left(Y_i - \bar{Y} \right) \left(X_i - \bar{X} \right) \right]}{\sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2} \qquad (4.10)$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

a) Write down the population model.

b) Calculate the OLS estimated slope and intercept.

c) Interpret these estimates.

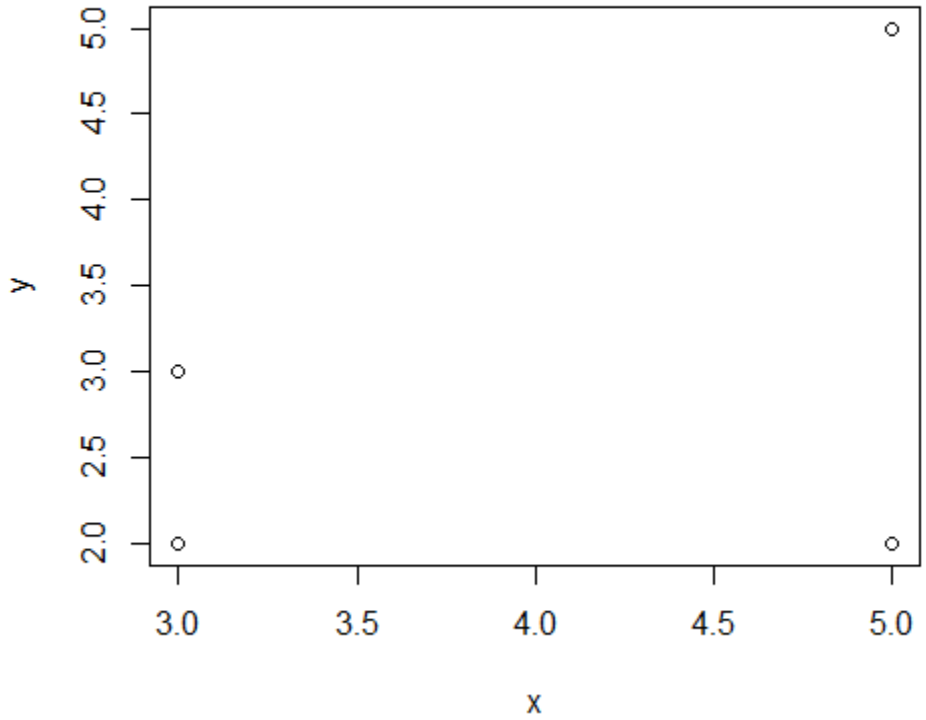
d) How are the formulas for b_1 and b_0 derived?

e) Calculate the OLS predicted values and residuals.

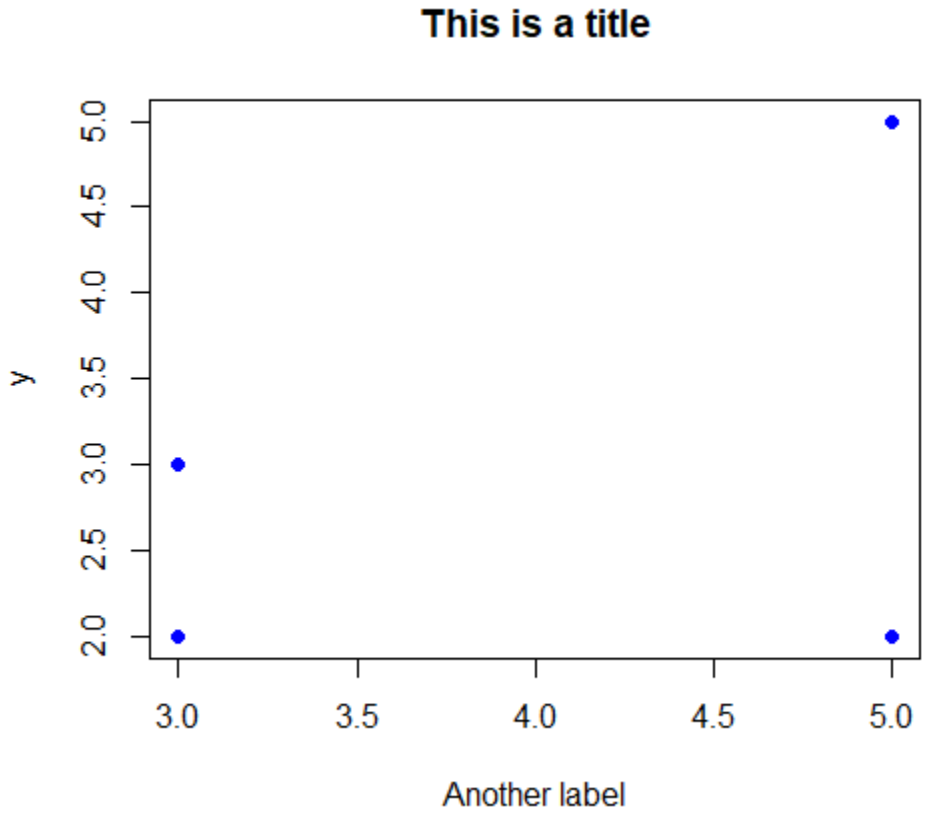
OLS in R

```
y <- c(5,2,2,3)
x <- c(5,3,5,3)
exdata <- data.frame(y, x)
lm(y ~ x, data = exdata)

Call:
lm(formula = y ~ x, data = exdata)

Coefficients:
(Intercept)            x
        1.0          0.5
```

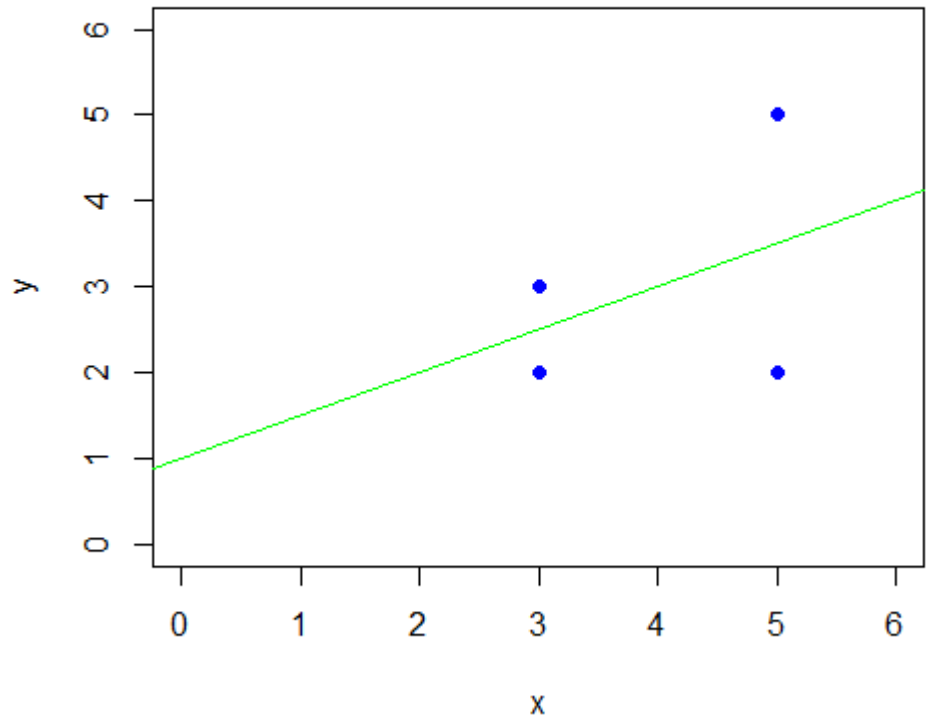
```
plot(exdata$x, exdata$y)
```



```
plot(exdata$x, exdata$y, pch=16, col="blue",
    main="This is a title", xlab="Another
    label")
```

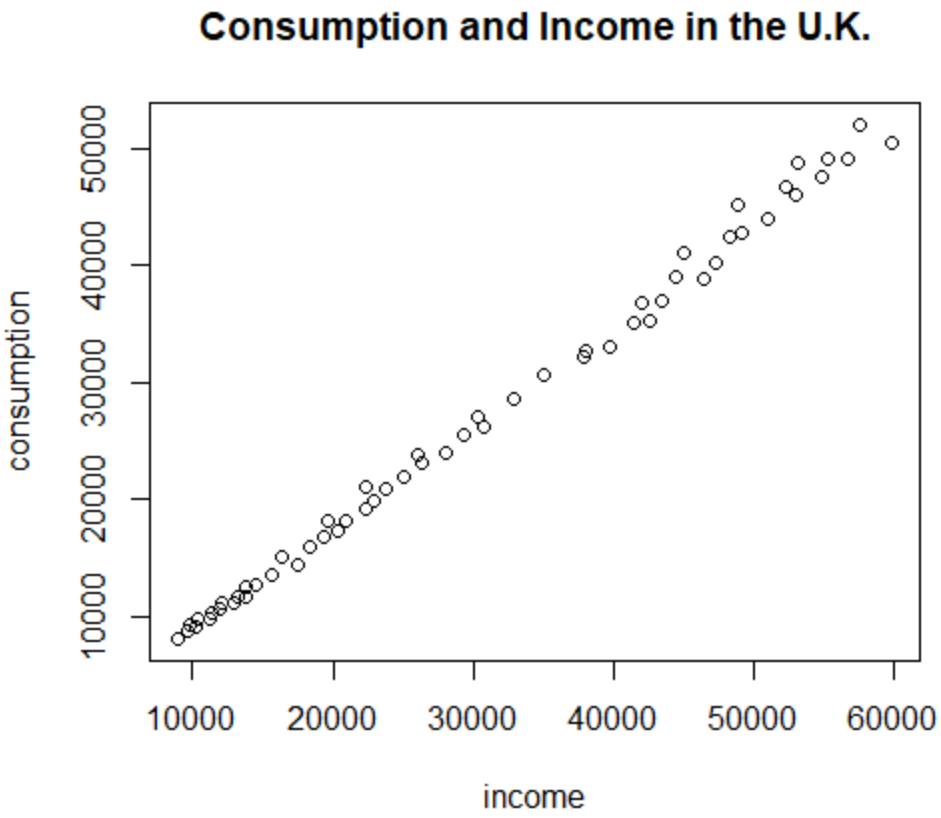


```
plot(exdata$x, exdata$y, pch=16, col="blue",
    xlim=c(0,6), ylim=c(0,6))

abline(lm(y ~ x), col="green")
```

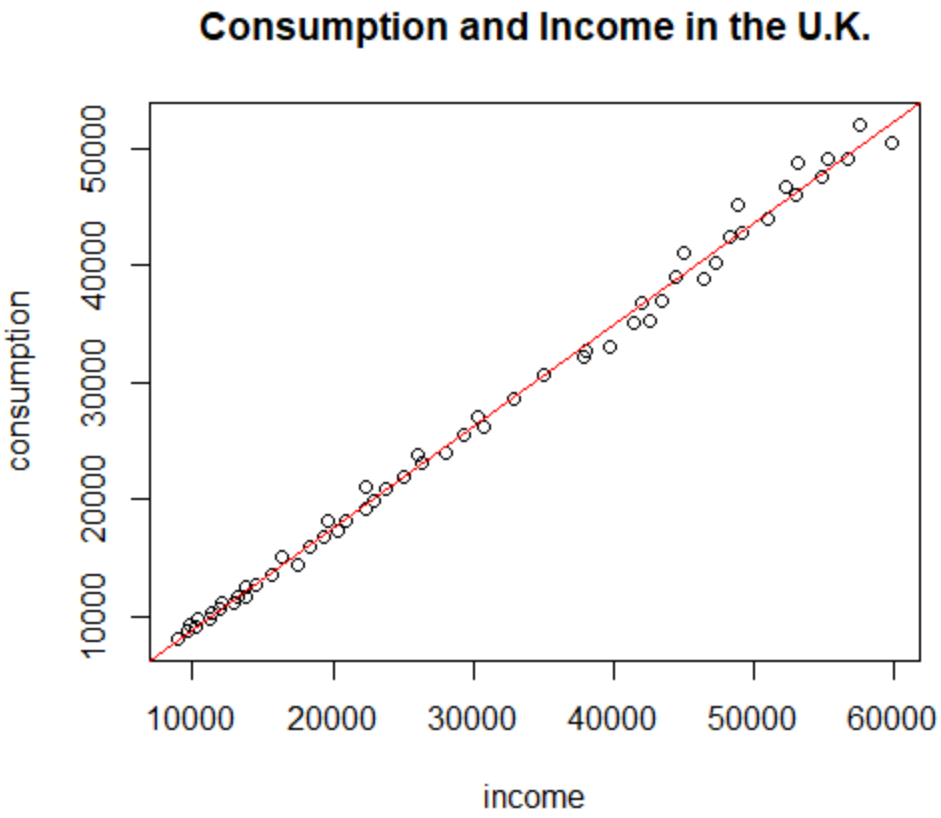


MPC Example

```
mpcdata <-
    read.csv("https://rtgodwin.com/data/mpc.csv")


plot(mpcdata$income, mpcdata$consumption,
    main="Consumption and Income in the U.K.")
```


Consumption and Income in the U.K.
consumption
10000
20000
30000
40000
50000
income
10000
20000
30000
40000
50000
60000

```
lm(consumption ~ income)

Call:
lm(formula = consumption ~ income)

Coefficients:
(Intercept)       income
    176.848        0.869

abline(lm(consumption ~ income), col = "red")
```


Consumption and Income in the U.K.
consumption
10000
20000
30000
40000
50000
income
10000
20000
30000
40000
50000
60000

Some Questions

- What is the estimated value for MPC?
- What is the interpretation of b_1 here?
- What is the interpretation of b_0?
- What is the value of β_1?

4.6 – The Assumptions of OLS

So why should we use OLS?

There are many other options for picking the regression line through the data:

- min. sum of horizontal or orthogonal distances instead of vertical distances
- take the sum of absolute distances instead of squared distances
- instead of squaring the residuals, raise them to the power of 4, 6, etc.
- divide the sample into two parts, take averages, connect the points
- connect any two data points

The reason why we minimize the sum of squared residuals, is because the resulting estimator has good statistical properties (under certain assumptions).

Remember that estimators are random variables!

The OLS slope and intercept estimators have sampling distributions.

Classical Assumptions:

A1: The population model is linear in the βs

A2: No linear combinations among variables (we only have one so far)

A3: The random error term, ϵ, has mean 0: E[ϵ] = 0

A4: ϵ is i.i.d.

A5: ϵ and X are independent: corr(ϵ, X) = 0

A6: ϵ is Normally distributed: $\epsilon \sim N(0, \sigma_\epsilon^2)$

If these assumptions are satisfied, it can be shown that:

$E[b_1] = \beta_1$

OLS is unbiased

OLS is efficient (G-M theorem)

OLS is consistent

However, these assumptions are often unrealistic.

Testing for the validity of these assumptions, re-evaluating the properties of the OLS estimator in the absence of each assumption, and figuring out how to recover unbiasedness\efficiency\consistency, would lead to some different estimators, and would form the basis for future econometrics courses.