Linear Regression

Part 4

The aim of this course is to give you most of the information you need to understand Linear regressions, as well as give you all tools necessary to put it in place. We will focus on the intuition of how it works without ignoring the maths beyond. Each part ends with several exercices to do. Some of them are very easy and other harder. I highly recommend to verify all demonstration presents during the course, for 2 main reasons: verify there is no mistake and help you to understand the objects manipulated. I truly believe that equations manipulation is very helpful in order to have a good intuition of why they are the way they are.

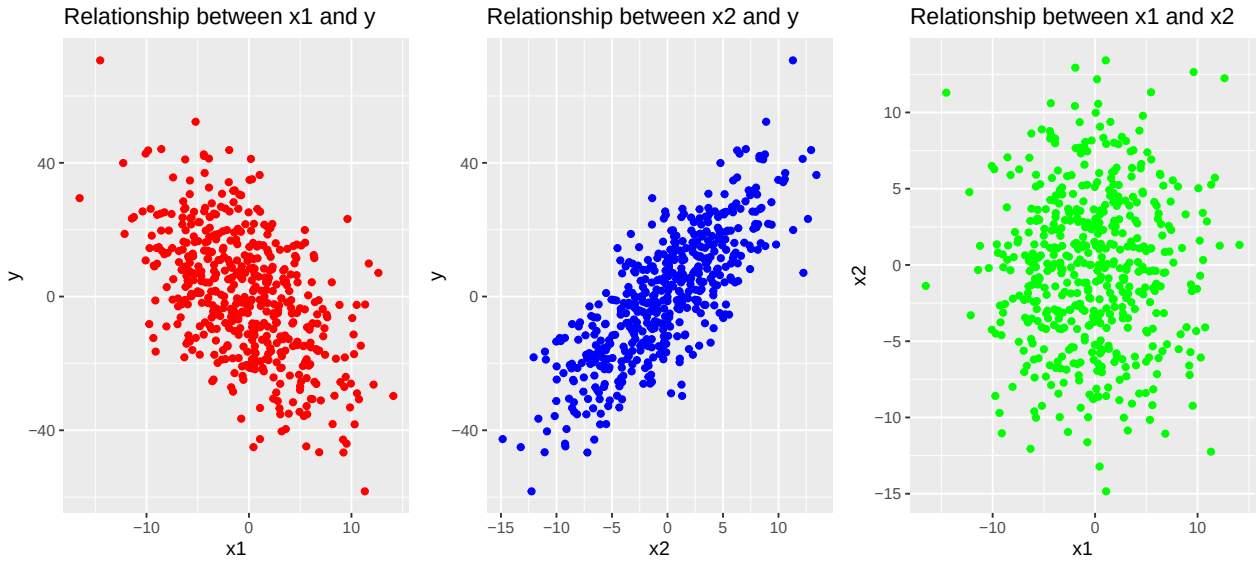
> *"Statistics is about reducing the amount of data."* **R. Fisher**

Part 4

More theory

Frisch–Waugh theorem The Frisch–Waugh(–Lovell) theorem allows us to reduce a multivariate regression analysis to an univariate one. The main idea behind is the fact that there are mutliple ways to estimate a β_1 coefficient in the following model: $y_i = \beta_1 x_i^1 + \beta_2 x_i^2 + \varepsilon_i$. We'll see it with concrete examples where we want to estimate the -2 below.

```
#create the data
n = 500
x1 = rnorm(n=n, sd=5)
x2 = rnorm(n=n, sd=5)
y = -2*x1 + 3*x2 + rnorm(n=n, sd=5)
data = data.frame(x1,x2,y)
```



The theorem says that all of these are equivalent for estimate β_1:
- Estimation by regressing y on $x1$ and $x2$ (1)
- Estimation by regressing y on the residual from the regression of $x1$ on $x2$, generally called *orthogonalization* or *residualization* (2)
- Estimation by regressing the residual from the regression of y on $x2$ on the residual from the regression of $x1$ on $x2$ (3)

From now we will use the `lm()` function since it makes the code a bit easier to read.

Let's compute for the (1) method:

```
#compute beta from regressing y on x1 and x2
reg = lm(data$y ~ data$x1 + data$x2)
beta1_1 = reg$coefficients[2]
beta1_1
```

```
##  data$x1
## -2.03507
```

Now let's compute with the (2) method (*orthogonalization*):

```
#compute the residuals of regressing x1 on x2
reg = lm(data$x1 ~ data$x2)
residuals = reg$residuals

#compute beta from regressing y on the residuals from above
reg = lm(data$y ~ residuals)
beta1_2 = reg$coefficients[2]
beta1_2
```

```
## residuals
##  -2.03507
```

And finally, the (3) method:

```
#compute residuals of regressing y on x2
reg = lm(data$y ~ data$x2)
residuals_1 = reg$residuals

#compute the residuals of regressing x1 on x2
reg = lm(data$x1 ~ data$x2)
residuals_2 = reg$residuals

#compute beta from regressing the residuals_1 on residuals_2
reg = lm(residuals_1 ~ residuals_2)
beta1_3 = reg$coefficients[2]
beta1_3
```

```
## residuals_2
##    -2.03507
```

Wow there all the same (this result is obviously not due to chance)! **I highly encourage** you to try it yourself with another example. You can see that we have reduced a multivariate regression $y_i = \beta_1 x_i^1 + \beta_2 x_i^2 + \varepsilon_i$ to an univariate one.

We talk about *orthogonalization* because we isole the variation of $x1$ that is orthogonal from $x2$. Put another way, when computing the residuals from regressing $x1$ on $x2$, we **keep only the variation of $x1$ that is unexplained by $x2$.**

Proof of the theorem
Before going any further, you might want to check this excellent article about linear regression and projections. We assume the following regression:

$$y = X_1\beta_1 + X_2\beta_2 + r$$

Propertie 1:
If there is no collinarity between our explanatory variables, the best fit to the least squares problem is unique.

Propertie 2:
Any matrix of variables can be split in its projection. For our regression above, it means that we can re-write $X_2 = P_1X_2 + M_1X_2$, where P_1 is the projection to X_1 and M_1 the *residual makers* for X_1. Also, we have: $M_1 = I - P_1$.

Propertie 3:
A regression on orthogonal sets of regressors can be done on each set at a time while still getting the coefficients from the joint regression.

In order to go from $\hat{\beta} = (X^TX)^{-1}X^Ty$ into the projection $\hat{y}$, we only have to multiple it by X, which gives us:

$$P_xy = X(X^TX)^{-1}X^Ty = \hat{y}$$
$$P_x = X(X^TX)^{-1}X^T$$

Now, let's say we have the two following estimations:

$$y = X_1\hat{\beta_1} + X_2\hat{\beta_2} + r_1$$

$$M_{X_2}y = M_{X_2}X_1\hat{\beta_3} + r_2$$

where:

$$M_{X_2} = I - P_{X_2} = I - X_2(X_2^TX_2)^{-1}X_2^T = residual\ makers$$

We want to prove that $\hat{\beta_1} = \hat{\beta_3}$ and $r_1 = r_2$.

1 - Step one, we re-write our first estimation by multiplying it by M_{X_2}:

$$M_{X_2}y = M_{X_2}X_1\hat{\beta_1} + M_{X_2}X_2\hat{\beta_2} + M_{X_2}r_1$$

But we know that $M_{X_2}X_2\hat{\beta_2} = 0$ and $M_{X_2}r_1 = r_1$. **Verify them if needed**. So we have:

$$M_{X_2}y = M_{X_2}X_1\hat{\beta_1} + r_1$$

Also we have:

$$X_1^TM_{X_2}y = X_1^TM_{X_2}X_1\hat{\beta_1} + X_1^Tr_1$$

And because $X^Tr_1 = 0$ (see Exercises section):

$$X_1^TM_{X_2}y = X_1^TM_{X_2}X_1\hat{\beta_1}$$

2 - Step two, we multiple the estimation with $\hat{\beta}_3$ by $(M_{X_2}X)^T$:

$$(M_{X_2}X_1)^T M_{X_2} y = X_1^T M_{X_2}^T M_{X_2} y$$

$$= X^T M_{X_2}^T M_{X_2} X\hat{\beta}_3 + X^T M_{X_2}^T r_2$$

knowing that $X^T M_{X_2}^T r_2 = 0$, we have:

$$X_1^T M_{X_2} y = X_1^T M_{X_2} X_1 \hat{\beta}_3$$

3 - Step three, finishing the proof:

We proved that $X_1^T M_{X_2} y = X_1^T M_{X_2} X_1 \hat{\beta}_1$.
We also proved that $X_1^T M_{X_2} y = X_1^T M_{X_2} X_1 \hat{\beta}_3$.
So we can conclude that $y = X_1\hat{\beta}_1 = X_1\hat{\beta}_3$ and $\hat{\beta}_1 = \hat{\beta}_3$.

As you probably see, this theorem is **not really easy to prove**. However, I highly recommend to do it by yourself with pen and paper one time. It's probably not the most important theorem to know, but it's useful in order to have a **better intuition** of what are the objects we are manipulating.

Some equivalences With all the things we said before (especially with Part 1), we might feel like linear regression and correlation analysis describe a similar thing. And that is true. In fact, there is an equivalence between Pearson correlation coefficient and the estimator of the OLS method. First, let's do the proof of the definition of the estimator that minimizes the squared error with the sum notation:

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad and \quad r_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x$$

$$Argmin_\beta \quad Loss = L(y, x, \beta) = \sum_1^n r_i^2 = \sum_1^n (y - \beta_0 - \beta_1 x)^2$$

$$\frac{\partial L}{\partial \beta_0} = -2\sum_1^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum y_i = \sum_1^n \hat{\beta}_0 + \sum_1^n \hat{\beta}_1 x_i = n\hat{\beta}_0 + \sum_1^n \hat{\beta}_1 x_i$$

$$\frac{1}{n}\sum_1^n y_i = \frac{1}{n} n\hat{\beta}_0 + \frac{1}{n}\sum \hat{\beta}_1 x_i$$

$$\overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

We use this last result for the following:

$$\frac{\partial L}{\partial \beta_1} = \frac{\partial \sum_1^n (y_i - (\overline{y} - \beta_1\overline{x}) - \beta_1 x_i)^2}{\beta_1} = 0$$

$$\frac{\partial \sum_1^n [(y_i - \overline{y}) - \beta_1 (x_i - \overline{x})]^2}{\beta_1} = 0$$

$$-2\sum_1^n [(y_i - \overline{y}) - \hat{\beta}_1 (x_i - \overline{x})]^2 = 0$$

$$-2\sum_1^n [(x_i - \overline{x})((y_i - \overline{y}) - \hat{\beta}_1 (x_i - \overline{x}))] = 0$$

$$\sum_{1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \hat{\beta_1}\sum_{1}^{n}(x_i - \overline{x})^2$$

$$\frac{1}{n}\sum_{1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \hat{\beta_1}\frac{1}{n}\sum_{1}^{n}(x_i - \overline{x})^2$$

$$cov(x, y) = \hat{\beta_1}Var(x) <=> \hat{\beta_1} = \frac{cov(x, y)}{Var(x)}$$

From this and with the fact that $Var(x) = \sigma_x^2$ and $cor(x, y) = \frac{cov(x,y)}{\sigma_x \sigma_y}$, we can re-write the latter:

$$\hat{\beta_1} = \frac{cov(x, y)}{Var(x)} = \frac{cov(x, y)}{\sigma_x \sigma_x} = \frac{cov(x, y)}{\sigma_x \sigma_y}\frac{\sigma_y}{\sigma_x} = cor(x, y)\frac{\sigma_y}{\sigma_x}$$

Exercises

1. With R, show empirically (use the data you want) that we find the (more or less) same $\hat{\beta}$ estimation whether we use the `lm()` function, the formula of the LSO estimator or its equivalence with the Pearson correlation coefficient.
2. Using the Frisch-Waugh-Lowell theorem, show empirically (use the data you want) that when doing the *orthogonalization* method, $x1$ and $x2$ do not need to be vectors but can perfectly be a matrix (i.e: contains multiple variables).
3. What would happen on the difference between method (1) and (2) if $x1$ and $x2$ are orthogonal? Prove it with an example.
4. Prove that $X^T r = 0_k$ from the following regression: $y = X\hat{\beta} + r$ knowing that $\hat{\beta} = (X^T X)^{-1} X^T y$.

References

- *"Understanding the Frisch-Waugh-Lovell Theorem"*, by Matteo Courthoud (2022)
- *"Partial Time Regressions as Compared with Individual Trends"*, by Ragnar Frisch & Frederick Waugh (1933)
- *"Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis"*, by Michael Lovell (1963)