Reminder TD_1

Marion Brouard, Pauline Leveneur

September 26, 2025

In this reminder, we will take the example of predicting the logarithm of the salary based on age and education level *(TD1, question 2)*

- **Objective of the reminder:**
  1. Knowing how to interpret a regression table in R
  2. Knowing how to interpret coefficients depending on the model:
     - Level - Level
     - Log - Level
     - Level - Log
     - Log - Log
  3. Omitted variable bias

R Command used: The `lm` command is used to estimate a regression using ordinary least squares. The code line to obtain the regression table of our example (see below) is as follows: `summary(lm(lnw ~ eduy + age, data = data))`

1 Interpretation of regression outputs in R

1.1 Interpretation of parameters

Figure 1: Regression Table - Estimated Parameters

```
Call:
lm(formula = lnw ~ eduy + age, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1422 -0.1906 -0.0021  0.1924  3.8563

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.9515069  0.0142606   66.72   <2e-16 ***
eduy        0.0629493  0.0007651   82.27   <2e-16 ***
age         0.0145086  0.0002423   59.87   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3553 on 31234 degrees of freedom
Multiple R-squared:  0.2159,Adjusted R-squared:  0.2158
F-statistic:  4300 on 2 and 31234 DF,  p-value: < 2.2e-16
```

- `Call`: The formula used with the dependent variable on the left of $\sim$ and the explanatory variables on the right
- `eduy, age`: Explanatory variables (X_1, X_2)
- `(Intercept)`: The constant (included by default by R)
- `Estimate`: Column of estimated coefficients
  Interpretation: An increase of one year (in age) results, on average, in a 1.45% increase in salary, all else being equal.
- `Std.Error` ("Standard errors"): column of the estimated standard errors of the coefficients.
- `t value` ("t-stat"): Student's t-test statistic (significance test of the coefficient)
  (Reminder: Here we test the hypothesis that the coefficient equals 0. A coefficient is said to be "significant" if it is significantly different from 0.)
- `Pr(>|t|)` ("p-value"): p-value of the t-test
  (*Reminder: The p-value is the minimal risk level such that we reject the null hypothesis. Often used in economics to deduce the significance of the estimated coefficient: If the p-value is less than 5%, the coefficient is significant at the 5% level. The lower the p-value, the more significant the coefficient.*)
- Stars (`Signif. codes - ***`): level of significance of the coefficient. If the coefficient is significant at 10%, we have a `.`, at 5% we have `*`, at 1% we have `**`, and at 0.1% we have `***`
  (*Reminder: As explained above, we can read the significance level directly by looking at the p-value. If the p-value is less than α%, the coefficient is significant at the α% level.*)

1.2 Interpretation of other information

Figure 2: Regression Table - "Other Information"

```
Call:
lm(formula = lnw ~ eduy + age, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1422 -0.1906 -0.0021  0.1924  3.8563

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.9515069  0.0142606   66.72   <2e-16 ***
eduy        0.0629493  0.0007651   82.27   <2e-16 ***
age         0.0145086  0.0002423   59.87   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3553 on 31234 degrees of freedom
Multiple R-squared:  0.2159,Adjusted R-squared:  0.2158
F-statistic:  4300 on 2 and 31234 DF,  p-value: < 2.2e-16
```

- `Residuals (Min 1Q...)`: Statistics from the distribution of the estimated residuals $\hat{\varepsilon}$. Generally less used information. It can be noted that by construction, these residuals are centered around 0 (the median is ≈ 0).
- `Residuals standard error`: unbiased estimator of the variance of the error term. The formula for this estimator is $\hat{\sigma} = \sqrt{\sum \frac{(y_i - \hat{y}_i)^2}{n-p}}$ where n is the number of observations and p is the number of coefficients to estimate. The quantity $n - p$ is called the number of degrees of freedom (`degrees of freedom`).
- `F-statistic`: Fisher's test statistic for the joint nullity of the coefficients. (*Reminder: the statistic $F(q, n-q-1)$ has degrees of freedom q, the number of tested coefficients (here 2, age and education), and $n - q - 1$, with n being the number of observations*)
- `p-value`: p-value associated with Fisher's test. (*Example*: If the p-value is less than 5%, we can reject the null hypothesis of joint nullity of the coefficients at a 5% risk level.)
- `Multiple R-squared` (R^2): The coefficient of determination, i.e., $R^2 \times 100$ is the percentage of the variance in the log-salary sample explained by age and education.
  Reminder 1: $R^2 = \frac{\text{SSE (explained sum of squares)}}{\text{SST (total sum of squares)}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$
  Reminder 2: R^2 increases with the number of explanatory variables: since SSR never increases and often decreases when adding variables.
- `Adjusted R-squared`: R^2 adjusted to account for the number of explanatory variables: it adds a penalty for each additional variable.
  Reminder: $R^2_{adj} = 1 - \frac{(1-R^2)(N-1)}{N-q-1}$

2 Interpretation of coefficients by models

Note: Here we assume X_1 is continuous. The interpretation of β_1 is different if X_1 is discrete.

- **Level(Y)-Level(X) Model:** $Y = \beta_0 + \beta_1 X_1 + \epsilon$
  - Marginal effect of X_1 on $Y = \beta_1$
  - Interpretation: an increase of 1 unit of X_1 results in a change of β_1 units in Y
- **Level(Y)-Level(X) Model:** with squared variables: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$
  - Marginal effect of X_1 on $Y = \beta_1 + 2\beta_2 X_1$
  - To calculate the marginal effect, you need to choose a starting value x_1 of X_1 because the marginal effect is not constant.
  - Alternatively, you can retrieve the average marginal effect (i.e., take the average of the marginal effect applied to each individual) or the marginal effect at the mean (apply the marginal effect to the mean value of X_1).
- **Level(Y)-Log(X) Model:** $Y = \beta_0 + \beta_1 log(X_1) + \epsilon$
  - Interpretation: a 1% increase in X_1 results in an increase of $\beta_1/100$ units of Y
- **Log(Y)-Level(X) Model:** $log(Y) = \beta_0 + \beta_1 X_1 + \epsilon$
  - Interpretation: a 1 unit increase in X_1 results in a $(100 * \beta_1)$
  - Note: β_1 is called the semi-elasticity of Y with respect to X_1
- **Log(Y)-Log(X) Model:** $log(Y) = \beta_0 + \beta_1 log(X_1) + \epsilon$
  - Interpretation: a 1% increase in X_1 results in a β_1% increase in Y
  - Note: β_1 is considered the elasticity of Y with respect to X_1

See the following document for the detailed calculations on interpreting coefficients:
`https://www.parisschoolofeconomics.eu/docs/yin-remi/interpretation-des-coefficients.pdf`

Table 1: Summary of interpretation of coefficients in a linear regression

Model Type	**Dependent Variable**	**Explanatory Variable**	**Interpretation of β_1**
Level - Level	Y	X_1	$\Delta Y = \beta_1 \, \Delta X$ Interpretation: All else being equal, an increase of 1 unit of X_1 is associated with an increase/decrease of β_1 units of Y
Level - Log	Y	$log(X_1)$	$\Delta Y = \frac{\beta_1}{100} \% \, \Delta X$ Interpretation: All else being equal, an increase of 1% of X_1 is associated with an increase/decrease of $\frac{\beta_1}{100}$ units of Y
Log - Level	$log(Y)$	X_1	$\%\Delta Y = (100 \times \beta_1) \, \Delta X$ Interpretation: All else being equal, an increase of 1 unit of X_1 is associated with an increase/decrease of $(100 \times \beta_1)$ % of Y (semi-elasticity)
Log - Log	$log(Y)$	$log(X_1)$	$\%\Delta Y = \beta_1 \% \, \Delta X$ Interpretation: All else being equal, an increase of 1% of X_1 is associated with an increase/decrease of β_1 % of Y (elasticity)

3 Omitted Variable Bias

Using the notations from the lecture (Chapter 1, Prop 4, Slide 22) and the exercise (G is the omitted variable in the long regression), we have:

$$\underbrace{\hat{\beta}_D^S}_{\substack{\text{Estimator of D} \\ \text{Short Regression (Y on D)}}} = \underbrace{\hat{\beta}_D}_{\substack{\text{Estimator of D} \\ \text{Long Regression (Y on D and G)}}} + \overbrace{\underbrace{\hat{\lambda}}_{\substack{\text{Estimator of D} \\ \text{Regression G on D}}} \underbrace{\hat{\beta}_G}_{\substack{\text{Estimator of G} \\ \text{Regression Y on D and G}}}}^{\textbf{Omitted Variable Bias}} \quad \text{with } \hat{\lambda} = \frac{\hat{cov}(D,G)}{\hat{var}(D)}$$

Thus, the omission of the variable G causes bias if:

- Variables G and D are correlated ($\hat{\lambda} \neq 0$)
- and Variables Y and G are correlated ($\hat{\beta}_G \neq 0$)

$\Rightarrow$ **There is an omitted variable bias if the omitted variable is correlated with both the explained**

variable and one of the explanatory variables in the model.

- **Sign of the bias:**

Table 2: Sign of the bias (omitted variable G)

	Corr(D,G)>0	**Corr(D,G)<0**
$\beta_G > 0$	Positive Bias: $\hat{\beta}_D^S$ overestimates β_D	Negative Bias: $\hat{\beta}_D^S$ underestimates β_D
$\beta_G < 0$	Negative Bias: $\hat{\beta}_D^S$ underestimates β_D	Positive Bias: $\hat{\beta}_D^S$ overestimates β_D