Review: Basics of the Linear Regression Model

2025/2026, Semester 1

Ralf A. Wilke
Copenhagen Business School

- A simple model for explaining y given x is:

$$y = \beta_0 + \beta_1 x + u$$

- Some terminology:

  y: Dependent Variable, Explained Variable, Response Variable, Predicted Variable, Regressand

  x: Independent Variable, Explanatory Variable, Control Variable, Predictor Variable, Regressor

  u: Error Term or Disturbance, Key assumption: E(u|x)=0

  It represents factors other than *x* that affect *y*.

  β_1 slope parameter

  β_0 intercept

- The linearity of the model implies that a one-unit change in *x* has the same effect on *y* regardless of the initial value of *x*.
  - This is unrealistic in many economic applications.
- Does β_1 really measure the effect of *x* on *y* if we ignore all other factors? -> causality!
- How to obtain estimates for the coefficients of the population regression function (PRF)?
  - Use OLS!

- Example: CEO Salary and Return on Equity
- The dataset CEOSAL1.dta contains information on 209 CEOs for the year 1990.
- Let us analyse the association between the equity and the wage of a chief executive officer in the population of CEO's:
- *y*: annual *salary* in thousands
  - If *salary*=856.3 then the annual salary is $856,300.
- *x*: average return on equity (*roe*) for the CEO's firm for the previous three years.
  - If *roe*=10, then the average return is 10 percent

- We postulate the simple model:

$$salary = \beta_0 + \beta_1 roe + u$$

- β_1 measures the change in annual salary, in thousands of dollars, when return on equity increases by one percentage point.

- Using an econometric package, the OLS regression line relating *salary* to *roe* is

$$\widehat{salary} = 963.191 + 18.501 roe$$

- How to read these results?
  - If *roe*=0, the predicted salary is $963,191.
  - If roe=30, then

$$\widehat{salary} = 963.191 + 18.501 \times 30 = 1518.221$$

  - Next, $\widehat{\Delta salary} = 18.501 \times \Delta roe$. This means that if the return on equity increases by one percentage point, $\Delta roe = 1$, then *salary* is predicted to change by about $18,500.

- The model with 2 variables can be written as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

where

β_0: intercept

β_1: measures the change in y with respect to x_1 holding other factors fixed

β_2: measures the change in y with respect to x_2 holding other factors fixed

- The key assumption of this model is: $E(u|x_1, x_2) = 0$
  - It means that for any value of x_1 and x_2 in the population, the average unobservable is zero.
  - How to read this? Suppose *u* is ability in a wage equation.

    Then for any combination of *exper* and *educ*, the average ability level is assumed to be the same.

The Model with *k* Independent Variables

- The general multiple regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

- This equation contains *(k+1)* unknown population parameters.
- The key assumption for the general multiple regression model is: $E(u|x_1, x_2, \ldots, x_k) = 0$.
- We still require that *u* is uncorrelated with all independent variables $x_1, \ldots, x_k$.

Generalising functional relationships between variables

- A special case is for $x_2 = x_1^2$:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + u$$

  which is a quadratic function in x_1.
- The multiple regression model therefore incorporates a variety of nonlinear functional forms.
- Please note that the interpretation of the parameter β_1 changes in this case because the other regressor cannot be held constant.
- The key assumption is in the case $x_2 = x_1^2$:

$$E(u|x_1, x_2) = E(u|x_1) = 0$$

- Models involving logarithms have interesting properties:

Model	*Dependent Variable*	*Independent Variable*	*Interpretation of* β_1
Level-level	*y*	*x*	$\Delta y = \beta_1 \Delta x$
Level-log	*y*	log(*x*)	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	log(*y*)	*x*	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	log(*y*)	log(*x*)	$\%\Delta y = \beta_1 \%\Delta x$

- These approximations rely on the properties of the natural logarithm. Less precise the larger the coefficient is.
- Exact formula for the effect of a change of a dummy variable in the log-level model is: $100[exp(\beta_1) - 1]$

- Interpreting the OLS Regression

  - Similar to the simple regression model
  - Now we can explicitly fix all observed regressors, which makes multiple regression analysis so useful.
  - $\hat{\beta}_0$ is the predicted *y* if all regressors *x* are 0.
  - The slope coefficients have partial effect, or ceteris paribus, interpretation:

$$\begin{aligned} \Delta\hat{y} &= \hat{\beta}_1 \Delta x_1 \\ &\vdots \\ \Delta\hat{y} &= \hat{\beta}_k \Delta x_k \end{aligned}$$

    when all other regressors are held constant, i.e. $\Delta x_l = 0$. This means, we control for the other variables.

- Example: Wage Equation (wage1.dta)

$$\widehat{log(wage)} = 0.284 + 0.092educ + 0.0041exper + 0.022tenure$$

with *tenure*: years with the current employer.

- Holding exper and tenure fixed, the coefficient 0.092 means that another year of education is predicted to increase *log(wage)* by 0.092, which translates into a 9.2% increase in *wage*.

Why is this?
Because:
$\%\Delta y \approx (100\beta_1)\Delta x_1$

- This estimate of the return to education at least keeps two important productivity factors fixed.
- We need to discuss in light of the statistical properties of OLS in order to see whether it is a good estimate or not.

- Changing more than one independent variable simultaneously:
  - What is the estimated effect on wage if the individual stays at the same firm for one more year? In this case *exper* and *tenure* increase:

$$\Delta log(\widehat{wage}) = 0.0041\Delta exper + 0.022\Delta tenure = 0.0263$$

This means that we should expect an increase in *wage* by 2.6%.

- The strength of the multiple regression model is that we can compare expected outcomes if we set the independent variables to different values.
- This does not require that the data is experimental, i.e. it can be a random sample with “arbitrary” values of the regressors.

- What is then the meaning of “linear” regression?
  - As we have seen, the general linear regression model also allows for certain nonlinear relationships.
  - What does “linear” mean?
  - The key is that the model is linear in the parameters β_0 and β_1…
  - Warning: the interpretation of the model coefficient does depend on the their definitions.

Inference

- You should be aware of the following:

- Hypothesis testing:
  - one sided alternatives $H_0 : \beta_j \leq 0 \text{ vs. } H_1 : \beta_j > 0$
  - two sided alternatives $H_0 : \beta_j = 0 \text{ vs. } H_1 : \beta_j \neq 0$
    $H_0 : \beta_j = a_j \text{ vs. } H_1 : \beta_j \neq a_j$

  - testing for linear restrictions $H_0 : \beta_1 = \beta_2$

$$H_0 : \beta_i = \beta_j = \beta_k = 0 \text{ vs. } H_1 : H_0 \text{ is not true}$$

- T-values, F-values

- P-values

- Example: test score data: GPA1.dta (sample code inference)

- In the regression output, what is a p-value and a F-value?

- construct t-value, p-value, F-value. (tvalue_example.R, pvalue_code.R, ftest_code.R)

- test *educ*
- test educ=exper
- test_linear_restrictions_example.R

Revision

- Relevant textbook chapters in Wooldridge, Introductory Econometrics:
  - Chapter 2: Simple Regression Model
  - Chapter 3: Multiple Regression: Estimation
  - Chapter 4: Multiple Regression: Inference
- Parts of the content of chapters 3 and 4 will be touched during the next couple of weeks but we will adopt a higher level presentation using matrix notation (as in Appendix E). We will see several proofs of important properties.
- Problem Set 0: Revision of relevant mathematical & statistical tools. (Appendices C-D)