Econometrics Part 2

The linear model in practice

Jose Angel Garcia Sanchez

jagarsanc@gmail.com

Sorbonne Data Analytics

November 2025

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Outline

Université Paris 1 Panthéon Sorbonne

Multicollinearity

- This problem arises if some variable is equal to an exact linear combination of some other variables (e.g. if we have variables like *income*, *income after tax* and *tax*)
- This variable is unnecessary because it contains repetitive information
- Worse, it prevents the computation of the OLS estimator, since X is no longer full rank and we cannot compute $\hat{b} = (X'X)^{-1}X'y$
- If the software detects multicollinearity, it arbitrarily removes one repetitive variable

Near multicollinearity

- This problem arises if some variable is *almost* equal to an exact linear combination of some other variables
- In other words, it means that they are very strongly correlated
- We *can* compute $\hat{b} = (X'X)^{-1}X'y$
- But : $\hat{b}$ is very unstable and the estimator is not very reliable (great variance)

How to detect near multicollinearity

- The VIF (Variance Inflation Factor) indicates by how much the variance of a parameter b_i is inflated due to the correlation of variable X_i with all the other explanatory variables
- Python command : `from statsmodels.stats.outliers_influencers import variance_inflation_factor`
- Say our model has k variables : constant, X_1, X_2, ... X_{k-1}
- Python will run the regression of each X_i on all the other X_j's, save its R^2 called R_i^2
- Example : $X_1 = a_1 + a_2 X_2 + ... + a_{k-1} X_{k-1} + \varepsilon$
- Next, VIF for each variable i is computed as : $VIF_i = \frac{1}{1-R_i^2}$
- One can prove that in the original regression $y = b_0 + b_1 X_1 + ... + b_{k-1} X_{k-1}$:
- $V(\hat{b}_j) = \frac{\sigma^2}{(N-1)Var(X_j)} VIF_j$
- $VIF_i > 5$ indicates a high risk of multicollinearity between variable i and the other variables
- If you don't see any big issue, do not remove variables : risk of omitted variable bias (see below)

Dummy variables : 2 categories

- Let's say we have N individual observations providing income, gender and the number of years of education.
- We want to explain individual income y by education x and gender z : $y_i = a + b.x_i + c.z_i + u_i$.
- z is a *categorical variable*, ideally coded with 0/1
- Say $z = 1$ codes for males
- c represents the extra money that provides the fact of being a male with respect to being a female
- It would be irrelevant to have a male *and* a female variable because it would cause *exact multicollinearity* (besides the fact that it is unnecessary)
- Remark : setting a categorical variable as the dependent variable needs other type of modeling, such as Probit and Logit (see later chapters)

Dummy variables : more than 2 categories

- Let's say we have N individual observations providing an index of life satisfaction, income and type of environment (*big city, small city, rural*), respectively coded with values 1, 2 and 3.
- We want to explain life satisfaction y by income x and environment z
- z is still a *categorical variable*, but we have 3 categories : we have to choose a category with respect to which parameters will be computed, say *rural*
- The model should be written :
  $y_i = a + b.x_i + c_1.z_{1i} + c_2.z_{2i} + u_i$, with z_1 and z_2 are dummy variables that code for *big city* and *small city* respectively
- These dummy variables should be created by hand or through pandas
- Python command : `pd.get_dummies(df['environment'], drop_first=True)` or using `statsmodels` formula with `C(environment)`
- Do not keep original coding (1, 2, 3), which would be meaningless in a regression

Dummy variables : outliers

- Let's say we have N observations for a country that experienced a shock.
- We want to explain life expectancy y by number of doctors per capita x : $y_t = a + b.x_t + u_t$.
- Let's say that the shock took place on year i : this unusual observation is likely to bias the whole regression, but we don't want to delete it
- We will take this event into account with a dummy variable :
- $D_t = 1$ if $t = i$, $D_t = 0$ otherwise.
- The model is now written : $y_t = a + b.x_t + c.D_t + u_t$.
- c represents the impact of the shock

Omitted variable bias

- It can be easily understood that estimating the previous model with and without dummy D can lead to changes even in parameters a and b
- Omitting a relevant variable (here D) leads to *omitted variable bias* that affects all estimated parameters
- Proof : in the finite sample case with the Frish-Waugh theorem (see Dormont), and proof involving consistency in later chapters (on endogeneity)
- This is a very important issue in applied work, so try not to forget important variables in models

The Frish-Waugh theorem

- Let Z be a vector of variables.
- M_Z is the orthogonal projection matrix on $L^{\perp}(Z)$.
- Let's define 2 models :
  1. $y = Xb_{(i)} + Zc + u$
  2. $M_Z y = M_Z X b_{(ii)} + M_Z u$.
- The Frish-Waugh theorem states that $\hat{b}_{(i)} = \hat{b}_{(ii)}$
- Important consequence : estimating the model with both sets of variables X and Z **or** with X alone will not provide the same results.
- This is called the "omitted variable bias"

On which condition will results be the same ?

Answer

- Estimating the model with both sets of variables X and Z **or** with X alone will yield the same results only if vectors X and Z are perpendicular, i.e. totally uncorrelated
- i.e. $X'Z = 0$
- Which is quite unlikely

A side effect of the FW theorem : centered variables

- Estimating the simple model $y_t = a + b.x_t + u_t$ with or without centering observations (i.e., replacing each value y_i by $y_i - \bar{y}$, same for x) will give the same result for $\hat{b}$.
- Centering observations will make the constant unnecessary

Advantages of using logs

1. Taking the *log* of the dependent variable can make the residuals look more "normal"
2. It can as well reduce heteroskedasticity
3. Parameters can be interpreted as percentages or elasticities that can be of interest

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

About the log transformation

- The average predicted y should be equal to the average observed y
- Here, we can compute a prediction of $log(y)$
- Can we get back to raw predicted wage simply by taking the exponential of the predicted y in log ?

Log and predictions (1)

- Let's use model $\forall i,\ log(y_i) = a + bx_i + u_i$ (what matters is that we take the *log* of y)
- Taking the exponential :
  $y_i = exp(a + bx_i + u_i) = exp(a + bx_i).exp(u_i)$
- So if the u's are independent from the x's, we get :
  $E[y_i] = E[exp(a + bx_i).exp(u_i)] = E[exp(a + bx_i)].E[exp(u_i)]$
- We usually consider the x's as fixed, so :
  $E[y_i] = exp(a + bx_i).E[exp(u_i)]$
- A prediction for y is thus : $\hat{y}_i = exp(\hat{a} + \hat{b}x_i).\hat{E}[exp(u_i)]$,
- But how to estimate $\hat{E}[exp(u_i)]$ (the retransformation parameter) ?

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Log and predictions (2)

It can be proven that if $u \hookrightarrow N(0, \sigma^2)$, then $E[exp(u_i)] = exp(\frac{\sigma^2}{2})$
This can be estimated using $exp(\frac{\hat{\sigma}^2}{2}) = \hat{\Psi}$
We thus have : $\hat{y}_i = exp(\hat{a} + \hat{b}x_i).\hat{\Psi}$
To get back to levels when using a prediction in *logs*, we cannot simply take the exponential of the prediction : there is an additional retransformation parameter to take into account.
The preceding formula holds only if the u's are normal : if their distribution is unknown but homoscedastic (same variance), we can use the non-parametric[1] *smearing* estimator :
$\hat{\Psi} = \frac{\Sigma exp(\hat{u}_i)}{N}$

[1] Because no assumptions on the type of distribution

Are all observations given the same weight ?

We have :

$$\hat{b}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

So :

$$\hat{b}_1 = \frac{\sum(x_i - \bar{x})^2 \frac{(y_i - \bar{y})}{(x_i - \bar{x})}}{\sum(x_i - \bar{x})^2} = \sum p_i \frac{(y_i - \bar{y})}{(x_i - \bar{x})}$$

Calling $p_i = \frac{(x_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$.
$\frac{(y_i - \bar{y})}{(x_i - \bar{x})}$ is the slope of the line drawn from the point corresponding to individual i to the sample average, and p_i is an increasing function of $(x_i - \bar{x})$.
Estimate $\hat{b}_1$ is thus highly influenced by extreme points (see Anscombe's quartet of identical regressions).

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Outliers (1)

- Outlier : observation with a large residual compared to other observations
- They indicate poor goodness of fit
- How to detect them : using a residuals plot
- Goal : estimate how far the observed value is from the predicted value
- Or more rigorously : using the *studentized* residuals, that adjust for the *leverage* of that particular point

Outliers (2)

- Outliers are observations that have a large residual
- However, if the observation is highly influential, it has pulled the predicted value towards itself
- In other words, the variances of the residuals differ, even though the variances of the true errors are all equal to each other
- Consequence : need for studentization
- The *studentized* residuals adjust for the leverage of that particular point (see what h is below)

$$rstud_i = \frac{\hat{u}_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Leverage (1)

- High leverage observation : has an explanatory variable that takes a value very different from the sample mean
- In a simple regression model, could be (for example) at the far right and would influence the slope of the regression line
- How to detect them : the "hat statistic"

Leverage (2)

- We know that $\hat{y} = P_X y$
- P_X can be called the "hat" matrix (puts a hat on y)
- Let's call h_i the i^{th} element of the diagonal of the hat matrix
- h_i is the *leverage* of observation i : if it is large, then observation i has a high leverage
- $0 < h_i < 1$

Leverage (3)

- We know that $\hat{u} = (I - P_X)u$, so $V(\hat{u}) = \sigma^2(I - P_X)$
- So : $V(\hat{u}_i) = \sigma^2(1 - h_i)$
- The greater the leverage h_i, the smaller the variance of the residual for observation i
- Meaning : high leverage observations tend to bring the regression line close to them
- High leverage observations are not necessarily influential, but low leverage observations won't be influential
- For the simple linear model : $h_i = \frac{1}{N} + \frac{(x_i - \bar{x})}{\Sigma(x_i - \bar{x})^2}$

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Influence measures (1)

- Influential observation : if removed from the regression, estimated parameters vary a lot
- How to detect them : the *Dfbeta* statistic
- For every variable k, its *Dfbeta* measures how much its coefficient would change if we removed one observation
- Every *Dfbeta* possible is computed, and the software provides its maximum and minimum
- The change in $\hat{b}_k$ if observation i is dropped is :

$$Dfbeta_{i,k} = \frac{\hat{b}_k - \hat{b}_{k,(-i)}}{\hat{\sigma}_{k,(-i)}}$$

with $\hat{b}_{k,(-i)}$ and $\hat{\sigma}_{k,(-i)}$ computed without observation i. It represents the number of se's by which the $\hat{b}_k$ changes when observation i is removed.

UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Influence measures (2)

- *Dffit* : measures how much the prediction for observation i changes when the estimation is made *without* observation i
- $Dffit = \frac{\hat{y}_i - \hat{y}_{i,(-i)}}{\hat{\sigma}_{(-i)}\sqrt{h_i}}$
- *Covratio* statistic : measures the change in the determinant of the covariance matrix of the estimates by deleting the ith observation (indicates loss in precision)
- *Cook's distance* : measures the influence of observation i on the model as a whole, not on a single coefficient (it combines all the *Dfbeta*'s)
- $D_i = \frac{(\hat{b}-\hat{b}_{(-i)})'(X'X)(\hat{b}-\hat{b}_{(-i)})}{k\hat{\sigma}^2}$

Cutoffs

- Leverage h_i of observation i : $\bar{h} = k/N$, $0 \leq h_i \leq 1$. If for observation i, $h_i > \bar{h}$ then it might be influential.
- Change in predicted value *Dffit* : observation i is influential if $|Dffit_i| > 2$ (general cutoff) or $> 2\sqrt{k/N}$
- Change in a parameter *Dfbeta* : observation i is influential if $|Dfbeta_{k,i}| > 2$ (general cutoff) or $> 2/\sqrt{N}$ (adj. for large sample)
- Impact on the whole model (Cook's distance D) : observation i is influential if $D_i > 1$
- "Covratio" statistic : observation i is influential if $|covratio - 1| > 3k/N$