Statistical Literacy — MINT

Lecture 11: Linear Multivariate Regression (and some special cases), OLS assumptions

Rémi Viné

The Graduate Institute | Geneva

December 4th, 2023

Outline

Housekeeping

- Problem set 11 (last one!) is now available
- The mock exam will be available today at 4pm

Multivariate regression

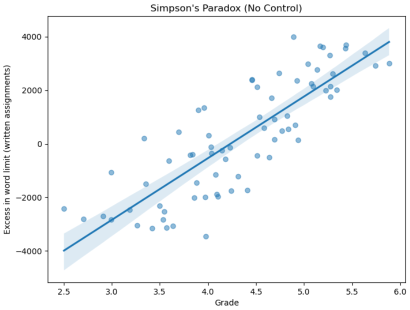
What for?

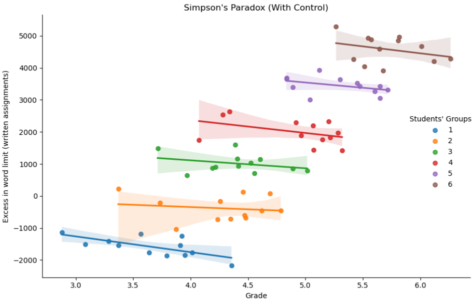
- A univariate analysis tends to be poor, *e.g.*, is education the *only* factor playing a role on wage?
  - Most likely not!
  - Experience, sector, country, size of company, share of blue collars, *etc.*

Multivariate regression

What for? The Simpson's paradox

- Missing some explanatory variables can lead to opposite relationships because of *confounding factors*





Multivariate regression

New interpretation of estimators

- Multivariate regressions add explanatory variables (there are k explanatory variables)
  - $y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k + e$
  - Now, the interpretation of an estimated coefficient requires to fix all the other explanatory variables - **ceteris paribus**
  - For a change of x_2 by one unit, the corresponding change of y is by b_2, provided that x_1, x_k and all other explanatory variables but x_2 are unchanged

Multivariate regression

New model fit: the adjusted R^2

- Incorporating variables can, randomly, capture some variance
- The R^2 can be artificially inflated just because of this
- Hence, correct the $R^2 = \frac{ExplainedVariance}{UnexplainedVariance}$ by the number of estimated coefficients (no need to remember the formula by heart):

$$Adjusted\ R^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

  - $R^2 \uparrow \implies Adjusted\ R^2 \uparrow$
  - $n \uparrow \implies Adjusted\ R^2 \uparrow$
  - $k \uparrow \implies Adjusted\ R^2 \downarrow$

Selection of the best model

- What is best?
  - Include all possible variables?
  - Only include the variable(s) of interest?
- In practice, one can use **backward selection** or **forward selection** to keep only the most relevant variables (statistically-wise)
- **Beware**: a coefficient strongly significant statistically speaking may not be strong in magnitude
  - *E.g.*, imagine a coefficient between monthly wage in CHF (Y) and education in years (X) of 0.003 and a $p - value < 0.00$
  - Then the statistical significance can be deemed as strong, but the link is weak in magnitude (to more education does correspond higher wages but only marginally higher)

Univariate & multivariate regressions: a few changes

	Univariate regression	Multivariate regression
Regression equation	$y = b_0 + b_1 x_1 + e$	$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k + e$
Interpretation b_0	y when $x_1 = 0$ *(only meaningful if 0 is in the range of x_1)*	y when $x_1 = ... = x_k = 0$ *(virtually meaningless)*
Interpretation b_1	Δy when $\Delta x_1 = 1$	Δy when $\Delta x_1 = 1$ with all other x_j constant *(change of y with one x,* **ceteris paribus***)*
Inference on b coefficients *(Slope = 0?)*	$CV = t_{\alpha/2}^{n-2}$ $test - statistic = \frac{b_1 - 0}{\frac{s_1}{\sqrt{n}}}$	$CV = t_{\alpha/2}^{n-k-1}$ $test - statistic = \frac{b_k - 0}{\frac{s_k}{\sqrt{n}}}$
R-squared	Baseline R-squared	Adjusted R-squared *(needed to account for addition of new variables)*
F-test	Not very useful as close to inference on b_1	Very useful to select the best model

Example of model selection

Backward elimination: drop one by one non-significant variables (for a given α, here 5%)

	Explained variable: Share of deaths by homicide, percent			
OLS models	Model 1	Model 2	Model 3	Model 4
	(1)	(2)	(3)	(4)
Per capita annual CO2 emissions	0.0066	0.0128	0.0170	
	(0.0402)	(0.0389)	(0.0183)	
Corruption index	-0.0159**	-0.0133**	-0.0135**	-0.0112**
	(0.0072)	(0.0060)	(0.0058)	(0.0052)
Per capita annual GHG emissions	0.0050	0.0032		
	(0.0264)	(0.0262)		
Gini coefficient	8.3288***	8.1905***	8.2393***	8.1489***
	(1.3406)	(1.3202)	(1.2537)	(1.2493)
Mean daily income ($)	0.0064			
	(0.0101)			
const	-1.6128***	-1.6023***	-1.6039***	-1.5837***
	(0.5859)	(0.5845)	(0.5825)	(0.5818)
Observations	156	156	156	156
R^2	0.2922	0.2903	0.2902	0.2862
Adjusted R^2	0.2686	0.2715	0.2762	0.2768
F Statistic	12.3835*** (df = 5.0; 150.0)	15.4389*** (df = 4.0; 151.0)	20.7145*** (df = 3.0; 152.0)	30.6669*** (df = 2.0; 153.0)

Note: *p<0.1; **p<0.05; ***p<0.01

Note 2: Standard errors in parentheses

- Do not forget the units (also to get a sense of magnitudes)!
  - Emissions in tons, income in USD, Gini is an indicator between zero (perfect equality) and 1 (perfect inequality), corruption index is an indicator between 0 (high perceived corruption) to 100 (no corruption perceived)

Multivariate analysis: Interpret and infer

	Explained variable: Share of deaths by homicide, percent			
OLS models	Model 1	Model 2	Model 3	Model 4
	(1)	(2)	(3)	(4)
Per capita annual CO2 emissions	0.0066	0.0128	0.0170	
	(0.0402)	(0.0389)	(0.0183)	
Corruption index	-0.0159**	-0.0133**	-0.0135**	-0.0112**
	(0.0072)	(0.0060)	(0.0058)	(0.0052)
Per capita annual GHG emissions	0.0050	0.0032		
	(0.0264)	(0.0262)		
Gini coefficient	8.3288***	8.1905***	8.2393***	8.1489***
	(1.3406)	(1.3202)	(1.2537)	(1.2493)
Mean daily income ($)	0.0064			
	(0.0101)			
const	-1.6128***	-1.6023***	-1.6039***	-1.5837***
	(0.5859)	(0.5845)	(0.5825)	(0.5818)
Observations	156	156	156	156
R^2	0.2922	0.2903	0.2902	0.2862
Adjusted R^2	0.2686	0.2715	0.2762	0.2768
F Statistic	12.3835*** (df = 5.0; 150.0)	15.4389*** (df = 4.0; 151.0)	20.7145*** (df = 3.0; 152.0)	30.6669*** (df = 2.0; 153.0)

Note: *p<0.1; **p<0.05; ***p<0.01
Note 2: Standard errors in parentheses

- In model 1, when the corruption index increases by 1 unit, **all other explanatory variable equal**, the corresponding change is a decrease in the share of death due to homicide by 0.0159 percentage points
- This predictor is statistically significant for $\alpha = 5\%$ but not for $\alpha = 1\%$

Multivariate analysis: Predict

	Explained variable: Share of deaths by homicide, percent			
OLS models	Model 1	Model 2	Model 3	Model 4
	(1)	(2)	(3)	(4)
Per capita annual CO2 emissions	0.0066	0.0128	0.0170	
	(0.0402)	(0.0389)	(0.0183)	
Corruption index	-0.0159**	-0.0133**	-0.0135**	-0.0112**
	(0.0072)	(0.0060)	(0.0058)	(0.0052)
Per capita annual GHG emissions	0.0050	0.0032		
	(0.0264)	(0.0262)		
Gini coefficient	8.3288***	8.1905***	8.2393***	8.1489***
	(1.3406)	(1.3202)	(1.2537)	(1.2493)
Mean daily income ($)	0.0064			
	(0.0101)			
const	-1.6128***	-1.6023***	-1.6039***	-1.5837***
	(0.5859)	(0.5845)	(0.5825)	(0.5818)
Observations	156	156	156	156
R^2	0.2922	0.2903	0.2902	0.2862
Adjusted R^2	0.2686	0.2715	0.2762	0.2768
F Statistic	12.3835*** (df = 5.0; 150.0)	15.4389*** (df = 4.0; 151.0)	20.7145*** (df = 3.0; 152.0)	30.6669*** (df = 2.0; 153.0)

Note: *p<0.1; **p<0.05; ***p<0.01

Note 2: Standard errors in parentheses

- In model 4, for a Gini coefficient $= 0.49$ and a corruption index $= 35$ (the case of Brazil), the predicted share of deaths due to homicides is
  - $\hat{y} = -1.5827 - 0.0112 \times 35 + 8.1489 \times 0.49 \approx 2.02\%$
  - The observed share of deaths by homicide is 4.67% so the model *underestimates* the deaths by homicide in Brazil

Binary (or dichotomous) variables

- Assume the following model is run:
  $Wage = b_0 + b_1 YearsEducation + b_2 WhiteCollar + e$
  - $Wage$ and $YearsEducation$ are numerical variables
  - $WhiteCollar$, is a binary variable: it is 0 if the individual is a blue collar and 1 if the individual is a white collar
    - b_2 captures the predicting role of being a white collar as compared to blue collar in the wage of the individual
  - Here, the interpretation of the intercept becomes more intuitive: b_0 is the predicted wage when the individual has no education and is a blue collar ($y = b_0 + b_1 \times 0 + b_2 \times 0 + e$)
  - For uneducated white collars ($= 1$), the predicted wage is $b_0 + b_2$

$\Rightarrow$ Binary variables are very frequently used and allow to study the heterogeneity of a relationship (*e.g.*, $\rightarrow$ *treated versus non-treated*)

Nominal variables

- The heterogeneity of the relationship might relate to more than two categories (sector worked, type of job, *etc.*)
- Trick: create as many binary variables as categories and include all but one in the OLS
  - Do you see why all categories **but one**?
- Say, there are 4 sectors (A, B, C, D) in the economy, then the regression equation could be:

$$Wage = b_0 + b_1 YearsEducation + b_2 B + b_3 C + b_4 D + e$$

Example of a multivariate regression with a categorical variable

	Explained variable: Share of deaths by homicide, percent
	(1)
Corruption index	-0.0161***
	(0.0049)
Gini coefficient	4.6777***
	(1.3346)
Asia	0.0895
	(0.2271)
Europe	0.2061
	(0.2758)
North America	2.3088***
	(0.2891)
Oceania	0.1253
	(0.4996)
South America	1.7068***
	(0.3315)
const	-0.4691
	(0.5993)
Observations	156
R^2	0.5375
Adjusted R^2	0.5156
Residual Std. Error	0.9421(df = 148)
F Statistic	24.5695*** (df = 7.0; 148.0)

Note: *p<0.1; **p<0.05; ***p<0.01

Note 2: Standard errors in parentheses

- "Africa" is the reference region
- Compared to Africa, the share of death by homicide is higher (give the specific value in percentage points) in the Americas
- Compare this model, and model (4)
  - $Adjusted\ R^2$ and $F - statistic$

OLS Assumptions

- So far, we have avoided a crucial point: what are the underlying assumptions of OLS, and why do they matter?

OLS is BLUE

Best Linear Unbiased Estimator (BLUE)

- If the 4 assumptions are met, then OLS is the **BLUE**

  $\rightarrow$ *This is called the Gauss-Markov theorem*

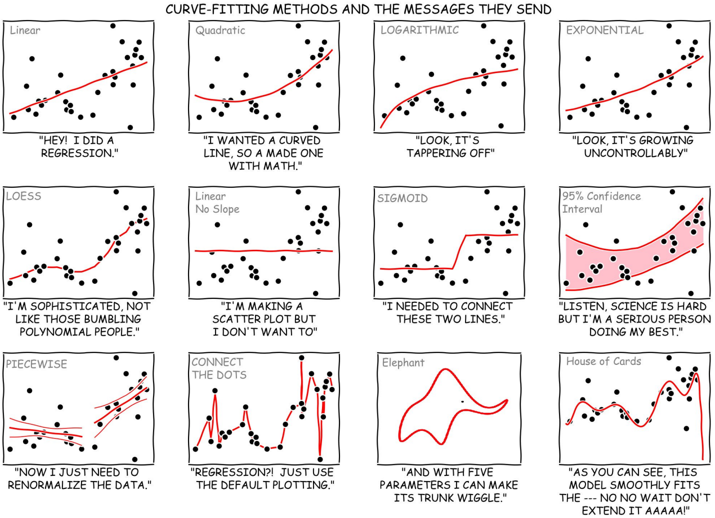
  - *Unbiased* means that the predicted estimator is, in expected value, to be the population parameter, here: $E[\hat{\beta_1}] = \beta_1$

The assumptions are:

1. There is a linear relationship between the explanatory variable(s) and the explained variable
2. The errors and the explanatory variables are not correlated
3. The explanatory variables are not *too* linearly linked (correlated) with each other
4. The errors look random

1. OLS Assumption and diagnostics

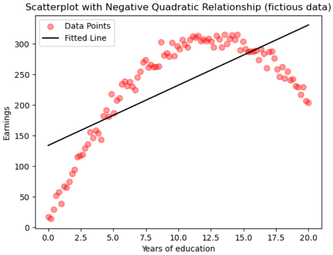
Linearity: only a **visual** verification (not a statistical test!)



by Douglas Higinbotham in Python inspired by https://xkcd.com/2048

1. OLS Assumption and diagnostics

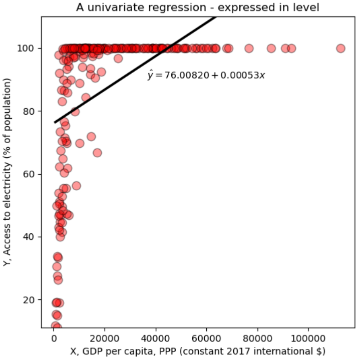
Linearity

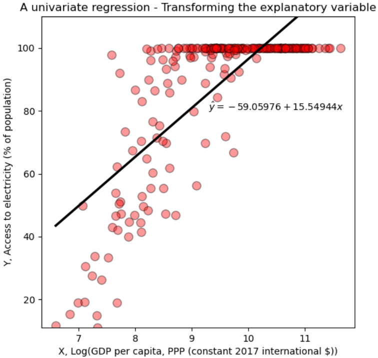


- There is a relationship between the two variables, but mostly a non-linear relationship
- In such case, $y = b_0 + b_1 x + e$ is not a suitable model
- BUT, a **transformed model** $y = b_0 + b_1 x^2 + b_2 x + e$ would be linear

- In practice, transforming a model is very common to approach linearity
- The most common transformation is the **logarithm transformation**

1. A log transformation (real data example from WDI)





- Beware of the change in the axis of the transformed variable(s)
  - X, Y can be transformed separately, or together, depending on where the linearity is best
- Note that, here, linearity is not fully satisfactory
  - Do you see why? (Hence the need for initial descriptive statistics)

2. Errors and explanatory variables are not correlated

- This assumption is not *testable*, but remains extremely important - as estimates are **biased** if not verified
- If verified, there is **exogeneity**; if not, **endogeneity**
- Formally, it writes *(no need to remember)*: $E[\varepsilon|x_1, ..., x_k] = 0$

  $\rightarrow$ Conditional on the explanatory variables, the expected error is zero

  - Using the example: $Wage = b_0 + b_1 \times education + e$, exogeneity is (totally) unrealistic:
    - **Omitted variable bias:** Many other factors might play a role: type of diploma, country, experience, *etc.*
      $\rightarrow E[\varepsilon|x_1, ..., x_k] \neq 0$
    - **Measurement errors** of "education" might lead part of education variance in the error $\rightarrow E[\varepsilon|x_1, ..., x_k] \neq 0$
    - **Reverse causality:** With vocational training, the direction of the relationship is not obvious

2. Errors and explanatory variable not correlated

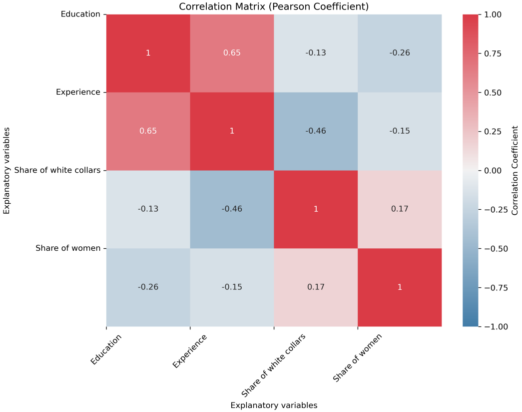
- In practice, ensuring the absence of endogeneity is challenging
- Takeaways (for this class):
  - Never conclude on a causal effect based on regressions (and correlations)
  - Remain aware of the risk of inaccurate estimations (biases)
    - In fact, in some extreme cases even the sign of the relationship can be wrong(!)
  - Stick to the simple interpretation (non-causal) provided in this class

3. Explanatory variables not too linked: no multicollinearity

- A regression can have many explanatory variables: $y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k + e$
- **Linear correlation** across explanatory variables (r_{x_i, x_j}) should not be too high (it is ok is not linearly linked).
  - If more than 90% of the variance of one explanatory variable can be captured by the other explanatory variables, then we consider that there is multicollinearity
- The risk of multicollinearity is serious, because the OLS model cannot be computed at all
- Examples of perfect collinearity:
  - Include both: x_1 "distance to work in km", and x_2 "distance to work in m"
  - Include both: x_1 "person aged more than 20 yo", and x_2 "person aged less than or equal to 20 yo"

3. Explanatory variables not too linked: no multicollinearity

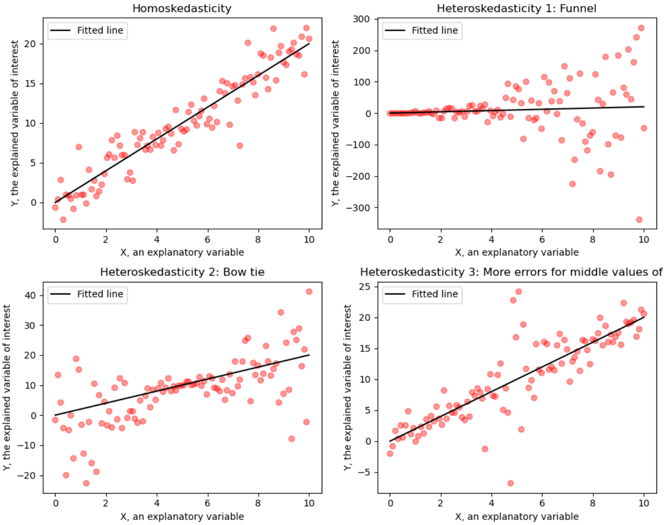
A good practice: prepare a correlation table of the explanatory variables



- Go beyond (not included in exam): you might want to compute the *VIF (Variance Inflation Factor)* and verify what variable(s) should be dropped from the model (if > 10)

4. Randomness of errors: Homoskedasticity

No autocorrelation is another criterion, but not covered in this class - needed for time series



Important takeaways on diagnostics

- If these assumptions are not met, OLS is not the **BLUE**, and may even provide estimates very far from the true parameters
- **Beware:** A visual check is by no means a statistical test
- It is just a help to suspect or not some potential issues

Next session

- Next session is a wrap-up session, we can discuss lectures 10 and 11 contents if needed (feel free to ask), discussing the exam, giving general feedback on assignments.