Applied Statistical Analysis I/ Quantitative Methods I POP77003/77051 Fall 2024

Week 11

Yao (Sara) HAN

hany3@tcd.ie

Department of Political Science

Trinity College Dublin

20 November, 2024

Today's Agenda

(1) Lecture recap

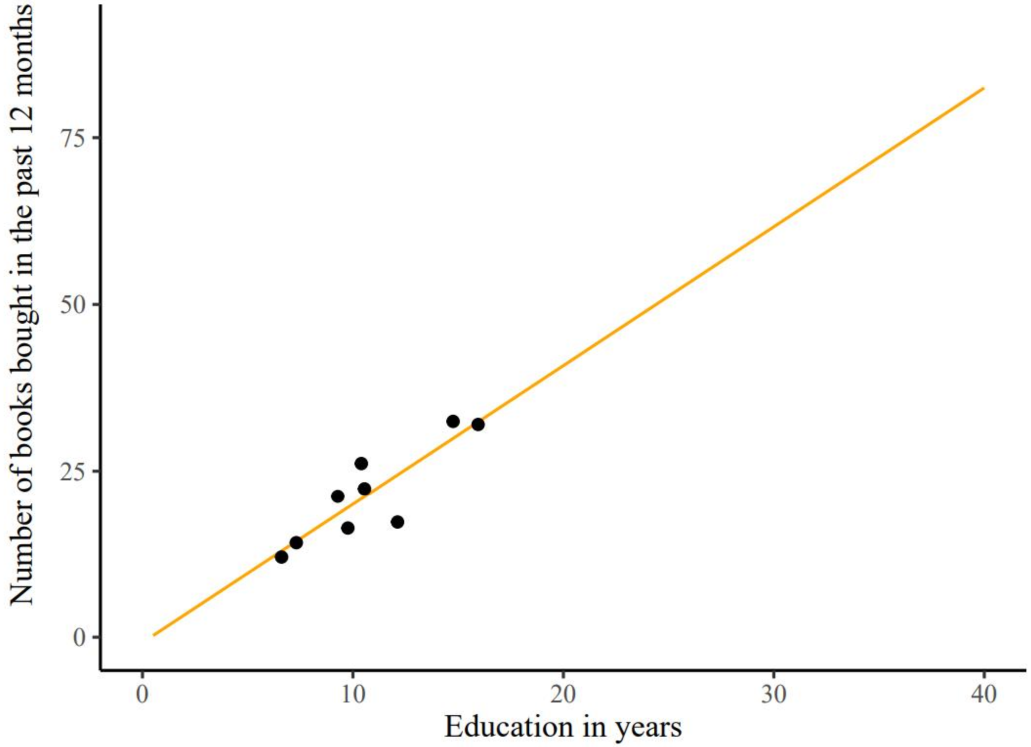
(2) Tutorial exercises

Discrepancy, Leverage and Influence

What are influential cases/outliers?

Discrepancy, Leverage and Influence

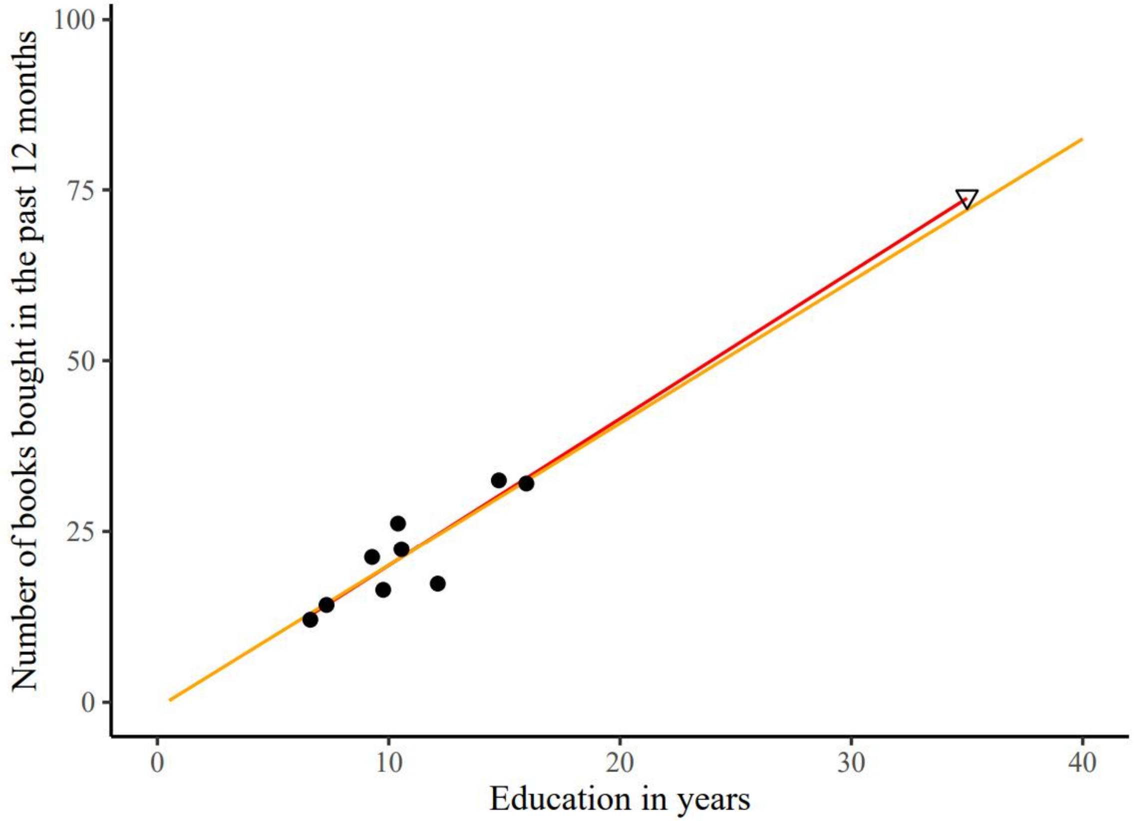
Not all outliers are concerning, because leverage $\neq$ influence, and discrepancy $\neq$ influence. $\longrightarrow$ Influence = leverage x discrepancy



*These are fictional data.

Leverage

Observation is unusual in its value on X, has high leverage, but low discrepancy. $\longrightarrow$ Low influence



$\rightarrow$ Hat values (h_i), distance of each observation from the data center

- **Hat value:**

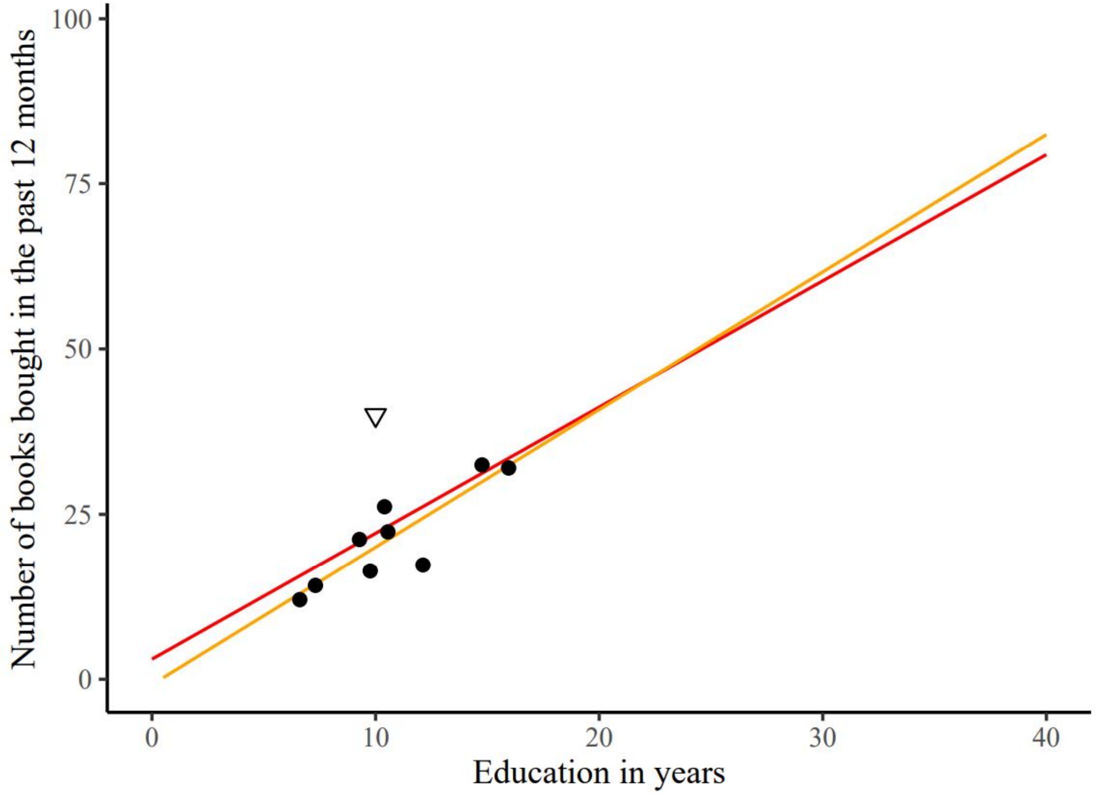
$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2}$$

h_i measures distance from center of cloud of points in X space

  - Outcome, Y, values are not involved in determining leverage
  - Leverage is a statement about the X space
  - A high hat value h_i equates to high leverage

Discrepancy

Observation is unusual in its value on Y, given its value on X, has high discrepancy, but low leverage. $\longrightarrow$ Low influence



$\rightarrow$ Standardized ($\hat{\epsilon}_i{}'$) and studentized residuals ($\hat{\epsilon}_i{}^*$), because ϵ_i is scale-dependent and high leverage leads to low ϵ_i

- We can form a **standardized residual** $\hat{\varepsilon}_i'$ which all have equal variance as

$$\hat{\varepsilon}_i' = \frac{\hat{\varepsilon}_i}{se(\hat{\varepsilon}_i)}$$

where $se(\hat{\varepsilon}_i) = \hat{\sigma}\sqrt{1 - h_i} = \sqrt{\frac{RSS}{n-k-1}}\sqrt{1 - h_i}$

- However, the distribution of $\hat{\varepsilon}_i'$ is not a t-distribution because the numerator and denominator are not independent

What to do? Studentized residuals

- Estimate the standard deviation of $se(\hat{\varepsilon}_i)$ with a sum of squares that does not include the *i*th residual
- Delete the *i*th observation, and refit the model based on $n-1$ observations, and get

$$\hat{\sigma}_{(-i)} = \sqrt{\frac{RSS}{n-1-k-1}}$$

- This gives us the **studentized residual**

$$\hat{\varepsilon}_i^* = \frac{\hat{\varepsilon}_i}{se(\hat{\varepsilon}_i)_{(-i)}}$$

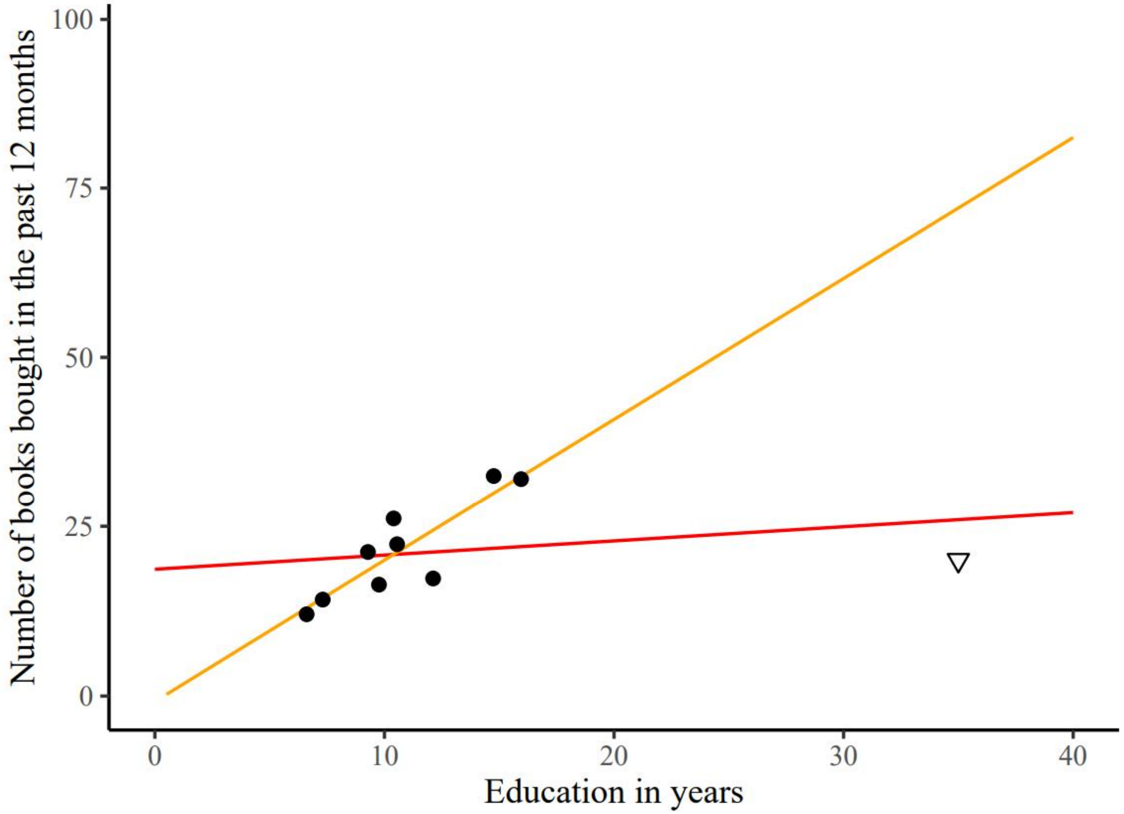
- Now, the $\hat{\sigma}$ in the denominator is not correlated with the numerator

$$\hat{\varepsilon}_i^* \sim t_{n-1-k-1}$$

- We also can look at adjusted p-value
  - Bonferroni correction is multiplying the p-values by the number of residuals

Influence

Observation has high leverage and discrepancy, an unusual value on X and Y. $\longrightarrow$ High influence



Influence

Validate through

1. Cook's Distance, difference in predicted values when observation i is included and not included
2. Difference in betas (DFBeta), difference in coefficients when observation i is included and not included
3. Leverage versus residual plot

Remedies

1. Check for coding errors
2. Think carefully about omitted variables

- Cook's D $= \frac{\sum_j (\hat{Y}_j - \hat{Y}_{j(-i)})^2}{(k+1)se^2} = \frac{(\hat{\varepsilon}_i^*)^2}{k+1} \frac{h_i}{1-h_i}$
- Cook's distance is the effect of ith observation on all fitted values
- Cook's distance can be high if h_i is very large (close to 1) and $(\hat{\varepsilon}_i^*)^2$ is moderate
  - Or if $(\hat{\varepsilon}_i^*)^2$ is very large and h_i is moderate, or if they are both extreme
- COOKSD $> \frac{4}{n-k-1}$ is unusually influential case

We can use leave-one-out deletion diagnostics, delete an observation, and see how much the fitted regression coefficients change

- Difference $= \hat{\beta}_j - \hat{\beta}_{j(-i)}$
- A large change suggests high influence

- Check influence: Difference in betas $= \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{se(\hat{\beta}_{j(-i)})}$
- Difference in betats is the effect of *i*th observation on a single estimated coefficient
- $|\text{DFBETAS}| > 1$ is considered large in a small or medium sized sample
- $|\text{DFBETAS}| > 2n^{-1/2}$ is considered large in a big sample

OLS assumptions

上

What are the assumptions of linear regression?

Assumptions of linear regression

Assumptions about the error (ϵ_i):

$$\epsilon_i \sim N(0, \sigma^2)$$

* ϵ_i is normally distributed
* $E(\epsilon_i) = 0$, no bias
* ϵ_i has constant variance σ^2 (Homoscedasticity)
* No autocorrelation
* X values are measured without error

(Kellstedt and Whitten 2018, 190–194)

Assumptions of linear regression

Assumptions about the model specification:

* No causal variables left out and no noncausal variables included
* Parametric linearity

(Kellstedt and Whitten 2018, 190–194)

Assumptions of linear regression

Minimal mathematical requirements:

* X must vary
* Number of observations must be larger than the number of predictors
* In multiple regression: No perfect multicollinearity

(Kellstedt and Whitten 2018, 190–194)

ϵ_i is normally distributed

Validate through

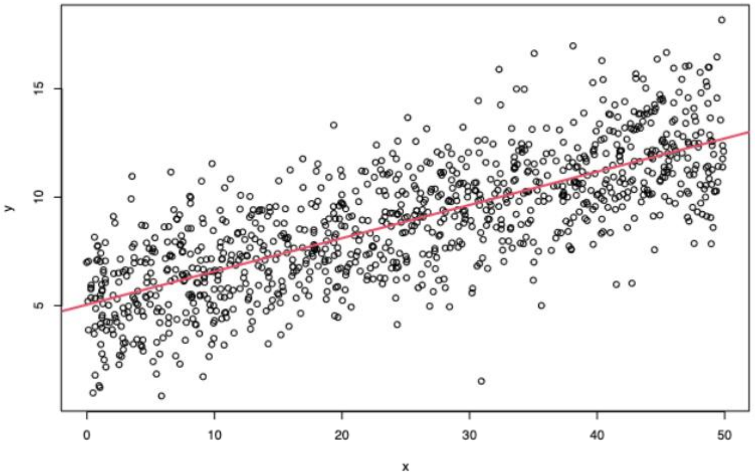
1. Histogram for ϵ_i
2. QQ (Quantile-quantile) plot

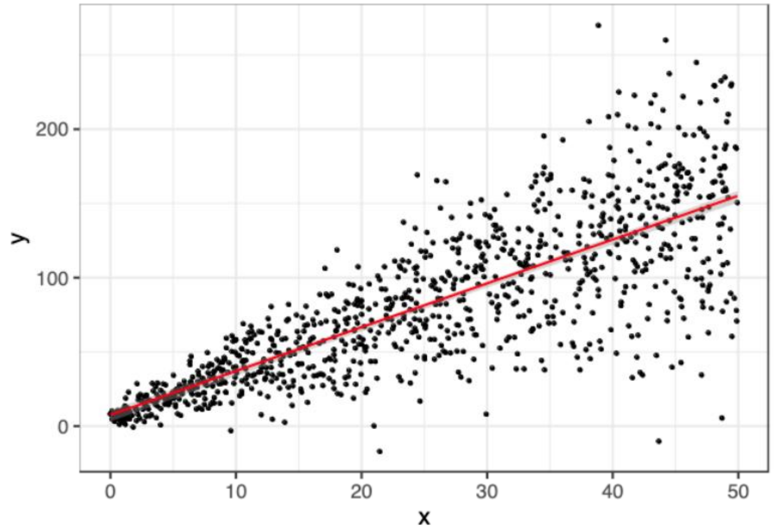
$\rightarrow$ If violated, standard errors are unreliable

Remedies

1. Gather more data

ϵ_i has constant variance σ^2





ϵ_i has constant variance σ^2

Validate through

1. Residual versus fitted plot

$\rightarrow$ If violated, standard errors are unreliable

Remedies

1. Log-transform Y
2. Robust standard errors

Parametric linearity

Validate through

1. Scatter plot
2. Residual plot

$\rightarrow$ If violated, slope coefficients are unreliable

Remedies

1. Transform X

No perfect multicollinearity

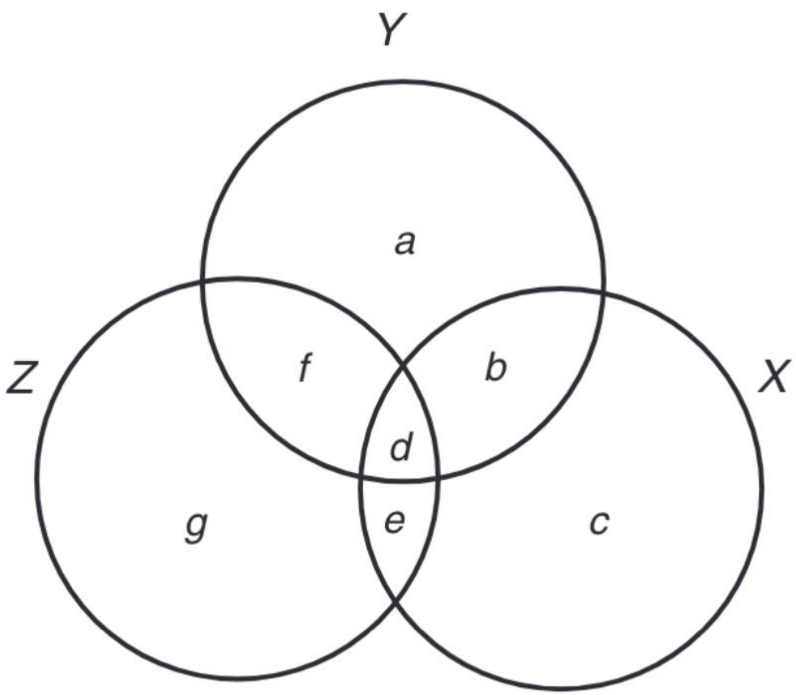


Figure 9.1. Venn diagram in which *X*, *Y*, and *Z* are correlated.

(Kellstedt and Whitten 2018, 212).

Multicollinearity

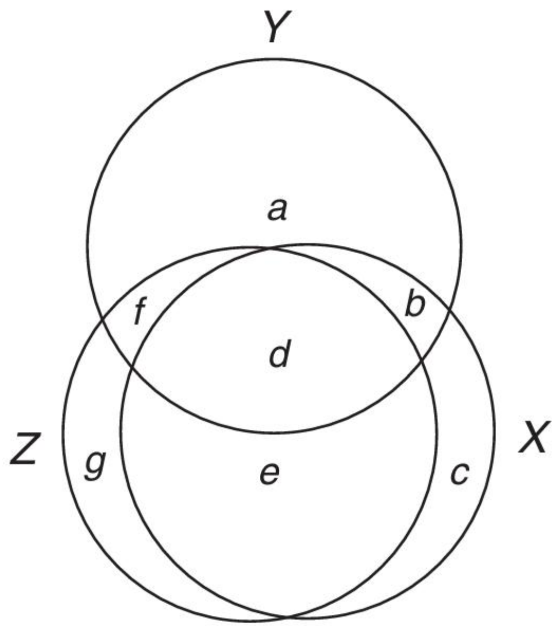


Figure 11.6 Venn diagram with multicollinearity

(Kellstedt and Whitten 2018, 212).

No perfect multicollinearity

Validate through

1. Correlation matrix
2. Variance Inflation Factor (VIF), indicates how much variation in X is explained by other independent variables

$\rightarrow$ Mathematical requirement, slope cannot be estimated

Remedies

1. Gather more data
2. Combine variables in index

References I

Kellstedt, Paul M., and Guy D. Whitten. 2018. *The fundamentals of political science research.* Cambridge: Cambridge University Press.