Applied Statistical Analysis I

Regression diagnostics

Hannah Frank
frankh@tcd.ie

Department of Political Science
Trinity College Dublin

November 29, 2023

Today's Agenda

(1) Validating quadratic effects (from last week)

(2) Lecture recap

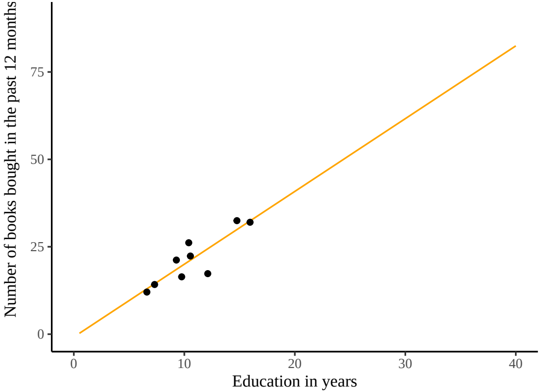
(3) Tutorial exercises: What is the relationship between education and Euroscepticism?

Discrepancy, Leverage and Influence

What are influential cases/outliers?

Discrepancy, Leverage and Influence

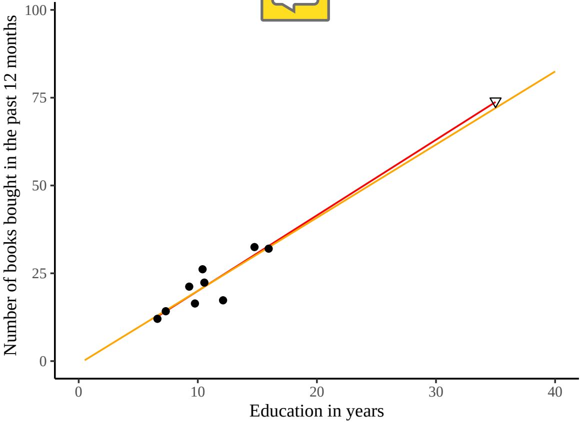
Not all outliers are concerning, because leverage $\neq$ influence, and discrepancy $\neq$ influence. $\longrightarrow$ Influence = leverage x discrepancy



*These are fictional data.

Leverage

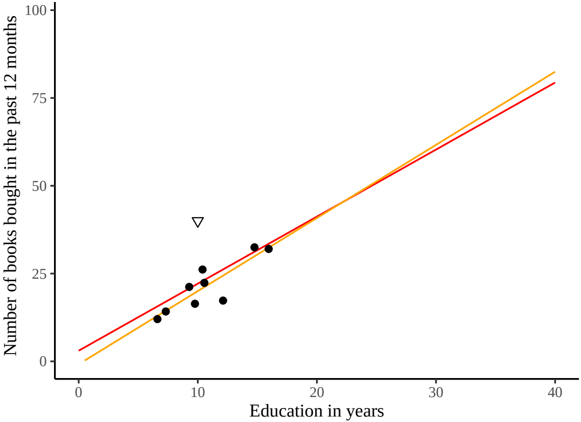
Observation is unusual in its value on X, has high leverage, but low discrepancy. $\longrightarrow$ Low influence



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

Discrepancy

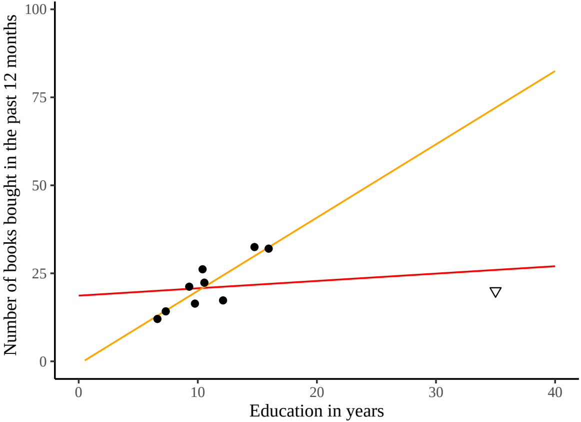
Observation is unusual in its value on Y, given its value on X, has high discrepancy, but low leverage. $\longrightarrow$ Low influence



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

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

OLS assumptions

What are the assumptions of linear regression?

Assumptions of linear regression

Assumptions about the error (ϵ_i), $Y_i = \alpha + \beta X_i + \epsilon_i$

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

* ϵ_i is normally distributed $\rightarrow$ needed for inference
* $E(\epsilon_i) = 0$, no bias $\rightarrow$ violated if error is not random, but correlated with omitted variable
* ϵ_i has constant variance σ^2 (Homoscedasticity $\leftrightarrow$ Heteroscedasticity)
* No autocorrelation, ocorrelation occurs when the stochastic terms for any two or more cases are systematically related to each other".
* X values are ured without error

(Kellstedt and Whitten 2018, 190–194)

Assumptions of linear regression

Assumptions about the model specification, $Y_i = \alpha + \beta X_i + \epsilon_i$

* 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, $Y_i = \alpha + \beta X_i + \epsilon_i$

* 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

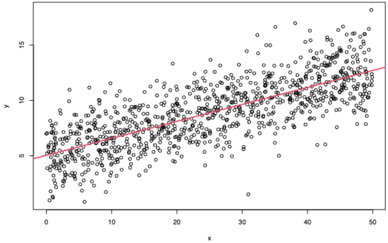
1. Histogram for ϵ_i
2. QQ (Quantile-quantile) plot

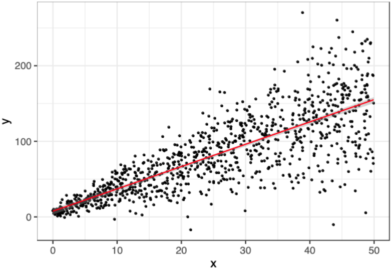
$\rightarrow$ If violated, standard errors are unreliable

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. Roust standard er

Parametric linearity

Validate through

1. Scatter plot
2. Residual plot

$\rightarrow$ If violated, slope coefficients are unreliable

Remedies

1. Transform X

No perfect multicollinearity

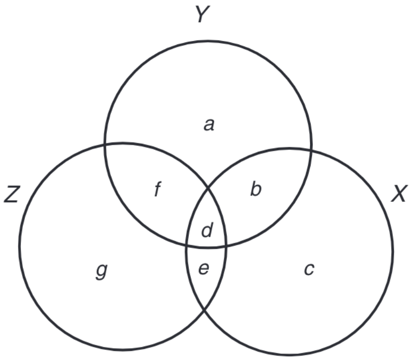


Figure 9.1. Venn diagram in which X, Y, and Z are correlated.

(Kellstedt and Whitten 2018, 212).

No perfect multicollinearity

Validate through

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

→ 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.