Secondary Data in Education Research

National Pupil Database and R practicals

Dr Yiyang Gao

Durham University Evidence Centre for Education

Epiphany Term Jan 25 2024

Outline

- National Pupil Database
- R session: Data preparation, OLS regression, and Diagnostic plots
- Summary

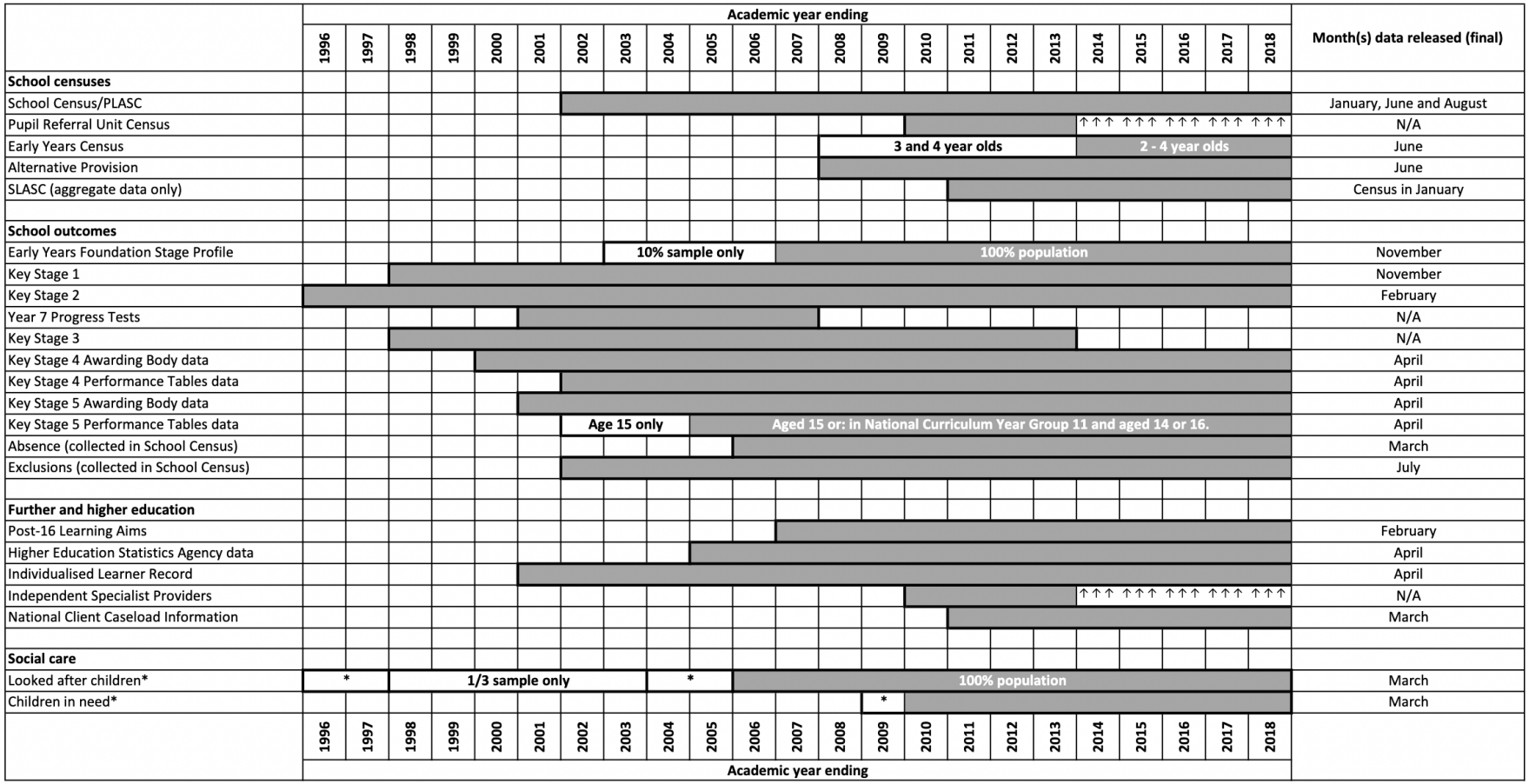
National Pupil Database

- What is the NPD?
- Find and explore data in the NPD

What is the NPD?

- The National Pupil Database (NPD) is a record-level administrative data resource curated by the UK government's Department for Education that is used for funding purposes, school performance tables, policy making, and research (Jay et al., 2019).

Data Structure



Source: (Jay et al., 2019)

Find and explore data in the NPD

- Find the NPD data
- https://www.find-npd-data.education.gov.uk/

Find and explore data in the NPD

- Open access school census: Schools, pupils and their characteristics
- https://explore-education-statistics.service.gov.uk/find-statistics/school-pupils-and-their-characteristics

Find and explore data in the NPD

- Open access Key Stages results
- https://explore-education-statistics.service.gov.uk/find-statistics/a-level-and-other-16-to-18-results
- https://explore-education-statistics.service.gov.uk/find-statistics/key-stage-4-performance-revised

R practicals

- The percentage of students getting an A* in the given subject (Mathematics, Politics, Sociology, Music) by county.
- Part 1: Data preparation, Q1 - Q6
- Part 2: OLS regression, Q7 - Q9
- Part 3: Diagnostic plots, Q10 - Q14

R practicals

Table: A-Level Grade Descriptions

Grade	Description
A*	This is the highest grade, indicating exceptional performance. It was introduced to differentiate the very top students.
A	This is a high performance grade, indicating a thorough understanding and high level of competence in the subject.
B	This grade signifies a good level of competence and a strong understanding of the subject.
C	Represents a competent performance with a good understanding of the subject.
D	This grade indicates a satisfactory level of performance.
E	The minimum passing grade, showing an adequate level of performance.
U (Unclassified)	This indicates that the required standard was not met and no grade is awarded.

Model interpretation

Table: OLS Regression Results

	Estimate	Std. Error	t-value	$Pr(\|>t\|)$
(Intercept)	28.77	5.59	5.15	0.00 ***
Politics	0.10	0.23	0.437	0.66
Observations		46		
Residual Standard Error		11.42		
R-squared		0.004		
Adjusted R-squared		-0.018		
F-statistic		0.19 on 1 and 45 DF		
P-value		0.66		

Model interpretation

Table: OLS Regression Results

	Estimate	Std. Error	t-value	$Pr(\|>t\|)$
(Intercept)	28.77	5.59	5.15	0.00 ***
Politics	0.10	0.23	0.437	0.66
Observations		46		
Residual Standard Error		11.42		
R-squared		0.004		
Adjusted R-squared		-0.018		
F-statistic		0.19 on 1 and 45 DF		
P-value		0.66		

- Predict the percentage of students getting an A* in Music A level based on the percentage of students getting an A* in Politics A level.

Model interpretation

Table: OLS Regression Results

	Estimate	Std. Error	t-value	$Pr(\|>t\|)$
(Intercept)	28.77	5.59	5.15	0.00 ***
Politics	0.10	0.23	0.437	0.66
Observations		46		
Residual Standard Error		11.42		
R-squared		0.004		
Adjusted R-squared		-0.018		
F-statistic		0.19 on 1 and 45 DF		
P-value		0.66		

- This is the expected value of Music when Politics is 0. It means that if none of the students get an A* in Politics A level, the model predicts that, on average, about 28.77 percent of students will get an A* in Music A level.

Model interpretation

Table: OLS Regression Results

	Estimate	Std. Error	t-value	$Pr(\|>t\|)$
(Intercept)	28.77	5.59	5.15	0.00 ***
Politics	0.10	0.23	0.437	0.66
Observations		46		
Residual Standard Error		11.42		
R-squared		0.004		
Adjusted R-squared		-0.018		
F-statistic		0.19 on 1 and 45 DF		
P-value		0.66		

- This is the slope of the line and represents the change in Music for a one-unit change in Politics.
- For every additional percentage point of students getting an A* in Politics A level, the percentage of students getting an A* in Music A level is expected to increase by 0.10 points, according to the model.

Model interpretation

Table: OLS Regression Results

	Estimate	Std. Error	t-value	$Pr(\|>t\|)$
(Intercept)	28.77	5.59	5.15	0.00 ***
Politics	0.10	0.23	0.437	0.66
Observations		46		
Residual Standard Error		11.42		
R-squared		0.004		
Adjusted R-squared		-0.018		
F-statistic		0.19 on 1 and 45 DF		
P-value		0.66		

- The standard errors of the coefficients suggest the level of uncertainty around these coefficient estimates. The larger the standard error, the less certain we are about the coefficient.

Model interpretation

Table: OLS Regression Results

	Estimate	Std. Error	t-value	$Pr(\|>t\|)$
(Intercept)	28.77	5.59	5.15	0.00 ***
Politics	0.10	0.23	0.437	0.66
Observations		46		
Residual Standard Error		11.42		
R-squared		0.004		
Adjusted R-squared		-0.018		
F-statistic		0.19 on 1 and 45 DF		
P-value		0.66		

- The t-values are not explicitly listed. However, we can calculate the t-values from the information given in the output.
- The t-value for each coefficient in a linear regression model is calculated by dividing the estimate of the coefficient by its standard error.

Model interpretation

Table: OLS Regression Results

	Estimate	Std. Error	t-value	$Pr(\|>t\|)$
(Intercept)	28.77	5.59	5.15	0.00 ***
Politics	0.10	0.23	0.437	0.66
Observations		46		
Residual Standard Error		11.42		
R-squared		0.004		
Adjusted R-squared		-0.018		
F-statistic		0.19 on 1 and 45 DF		
P-value		0.66		

$$t_{\text{intercept}} = \frac{\text{Estimate}}{\text{Standard Error}} = \frac{28.77}{5.59} = 5.15 \quad (1)$$

$$t_{\text{politics}} = \frac{\text{Estimate}}{\text{Standard Error}} = \frac{0.10}{0.23} = 0.43 \quad (2)$$

Model interpretation

Table: OLS Regression Results

	Estimate	Std. Error	t-value	$Pr(\|>t\|)$
(Intercept)	28.77	5.59	5.15	0.00 ***
Politics	0.10	0.23	0.437	0.66
Observations		46		
Residual Standard Error		11.42		
R-squared		0.004		
Adjusted R-squared		-0.018		
F-statistic		0.19 on 1 and 45 DF		
P-value		0.66		

- t-value: The t-statistic tests whether the coefficient is significantly different from 0. For Politics, the t-value is 0.43.

Model interpretation

Table: OLS Regression Results

	Estimate	Std. Error	t-value	$Pr(\|>t\|)$
(Intercept)	28.77	5.59	5.15	0.00 ***
Politics	0.10	0.23	0.437	0.66
Observations		46		
Residual Standard Error		11.42		
R-squared		0.004		
Adjusted R-squared		-0.018		
F-statistic		0.19 on 1 and 45 DF		
P-value		0.66		

- $Pr(|>t|)$: This p-value tells us about the probability of observing any value equal to or more extreme than the one actually observed if the null hypothesis (no effect) is true. For Politics, the p-value is 0.66, which is much higher than the typical alpha level of 0.05. This suggests that the coefficient for Politics is not statistically significant.

How $Pr(>|t|)$ is calculated

- The p-values (denoted as $Pr(>|t|)$ in regression outputs) associated with t-statistics in a regression analysis are calculated using the t-distribution.

How $Pr(>|t|)$ is calculated

- **Step 1: Null Hypothesis (H_0):**
  The null hypothesis for each coefficient is that there is no relationship between the independent variable (associated with that coefficient) and the dependent variable. In other words, the coefficient is equal to zero. Mathematically,
  $$H_0 : \beta_i = 0$$
  where β_i is the coefficient of the i-th independent variable.
- **Step 2: Alternative Hypothesis (H_a):**
  The alternative hypothesis is that there is a relationship between the independent variable and the dependent variable. Mathematically,
  $$H_a : \beta_i \neq 0$$
  (for a two-tailed test, which is the most common case).

How $Pr(> |t|)$ is calculated

- **Step 3: T-Statistic Calculation:**
  The t-statistic for each coefficient is calculated using the following formula:

$$t_i = \frac{\hat{\beta}_i}{\text{SE}(\hat{\beta}_i)}$$

- $\hat{\beta}_i$ is the estimated coefficient for the i-th independent variable.
- $\text{SE}(\hat{\beta}_i)$ is the standard error of the estimated coefficient $\hat{\beta}_i$.

How $Pr(> |t|)$ is calculated

- **Step 4: Degree of Freedom:**
  Degrees of Freedom (df): To use the t-distribution, you need the degrees of freedom, which in the context of regression is typically the number of observations minus the number of estimated parameters. If you have n observations and are estimating two parameters (intercept and slope), the degrees of freedom would be n - 2.

How $Pr(> |t|)$ is calculated

- **Step 5: P-value Calculation:**
  The p-value is calculated based on the t-statistic and the degrees of freedom (df). The p-value represents the probability of observing a t-statistic as extreme as the one calculated under the null hypothesis.
- For a two-tailed test, the p-value is the probability of observing a t-statistic as extreme as the one calculated on both tails of the t-distribution.
- For a one-tailed test (where you are only interested in one direction, either positive or negative), the p-value is the probability of observing a t-statistic as extreme as the one calculated in that specific tail.

How $Pr(> |t|)$ is calculated

- **Step 6: Interpretation:**
  If the p-value is smaller than a predetermined significance level (commonly 0.05), then the null hypothesis is rejected, and you conclude that there is a statistically significant relationship between the independent variable and the dependent variable. If the p-value is greater than the significance level, you fail to reject the null hypothesis, indicating that there is no statistically significant relationship.

How $Pr(> |t|)$ is calculated

- For example, if we have a t-statistic of 0.43 (for Politics), the DF is 46 - 2 = 44, and we decide we are conducting a two-tailed test...
- Assume a chosen significance level (α) of 0.05 for a 95 percent confidence level. Find the critical t-value (t_{crit}) for a two-tailed test at $\alpha/2$.
- Divide the significance level by 2 to account for two tails: $\alpha/2 = 0.05/2 = 0.025$.
- Calculate the $Pr(> |t|)$ in R

Model interpretation

Table: OLS Regression Results

	Estimate	Std. Error	t-value	$Pr(>\|t\|$
(Intercept)	28.77	5.59	5.15	0.00 ***
Politics	0.10	0.23	0.437	0.66
Observations		46		
Residual Standard Error		11.42		
R-squared		0.004		
Adjusted R-squared		-0.018		
F-statistic		0.19 on 1 and 45 DF		
P-value		0.66		

- This is the estimate of the standard deviation of the residuals. It shows the average amount that the response will deviate from the true regression line.

Model interpretation

Table: OLS Regression Results

	Estimate	Std. Error	t-value	$Pr(\|>t\|)$
(Intercept)	28.77	5.59	5.15	0.00 ***
Politics	0.10	0.23	0.437	0.66
Observations		46		
Residual Standard Error		11.42		
R-squared		0.004		
Adjusted R-squared		-0.018		
F-statistic		0.19 on 1 and 45 DF		
P-value		0.66		

- This value indicates how much of the variability in Music is explained by Politics. In this case, only about 0.42 percent of the variance in Music is explained, which is very low.

Model interpretation

Table: OLS Regression Results

	Estimate	Std. Error	t-value	$Pr(\|>t\|)$
(Intercept)	28.77	5.59	5.15	0.00 ***
Politics	0.10	0.23	0.437	0.66
Observations		46		
Residual Standard Error		11.42		
R-squared		0.004		
Adjusted R-squared		-0.018		
F-statistic		0.19 on 1 and 45 DF		
P-value		0.66		

- This adjusts the R-squared value for the number of predictors in the model. It's slightly negative, indicating that the model is not useful for explaining the variance in the response.

Model interpretation

Table: OLS Regression Results

	Estimate	Std. Error	t-value	$Pr(\|>t\|)$
(Intercept)	28.77	5.59	5.15	0.00 ***
Politics	0.10	0.23	0.437	0.66
Observations		46		
Residual Standard Error		11.42		
R-squared		0.004		
Adjusted R-squared		-0.018		
F-statistic		0.19 on 1 and 45 DF		
P-value		0.66		

- These are used to determine whether the model is statistically significant. The F-statistic is 0.1912 and the p-value is 0.664.
- Since the p-value is much higher than 0.05, we fail to reject the null hypothesis that the model with predictors is no better than a model with just the intercept.

Model interpretation

- The model suggests that there is no significant linear relationship between the percentage of students getting an A* in Politics A level and the percentage of students getting an A* in Music A level.
- The very low R-squared value indicates that the model does not explain much of the variability in the response, and the non-significant p-value for the Politics coefficient further suggests that changes in Politics scores do not have a significant linear effect on Music scores.

Diagnostic plots

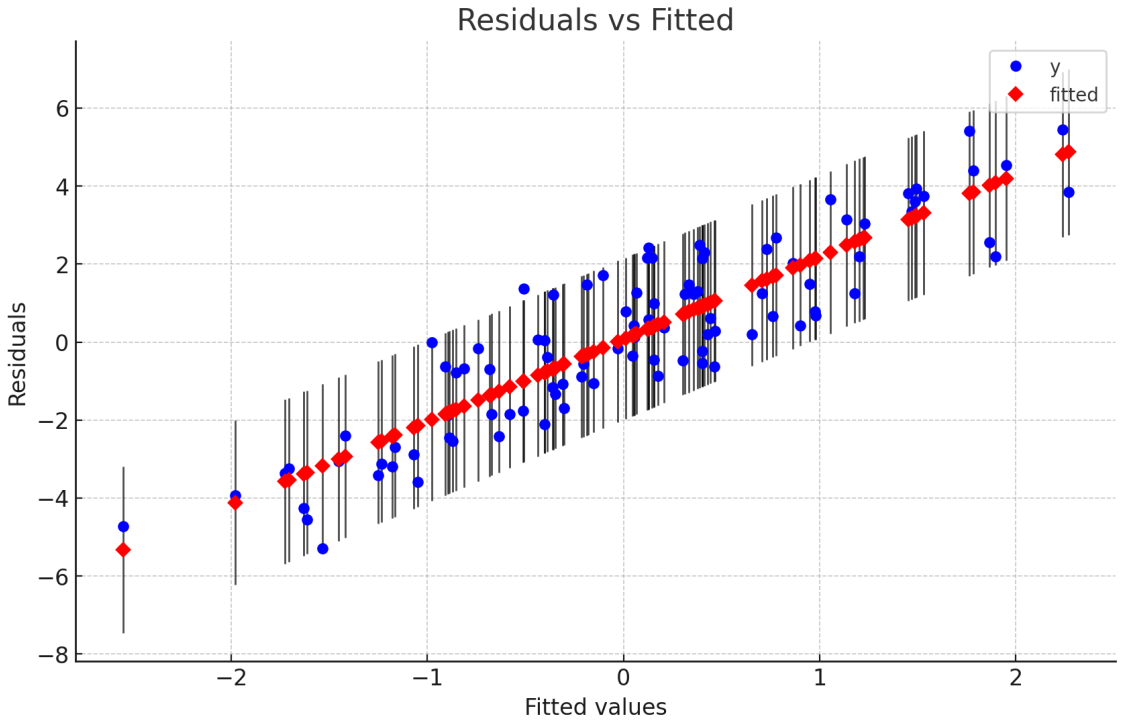
- In Ordinary Least Squares (OLS) regression, diagnostic plots are used to check various assumptions and potential issues with the regression model.

Diagnostic plots

- Residuals vs Fitted Plot (plot(model, which = 1)
- Assumption Tested: Linearity and Homoscedasticity.
- Purpose: This plot shows if residuals have non-linear patterns. The residuals should be randomly dispersed around the horizontal axis; if there's a pattern, it suggests a non-linear relationship.
- Homoscedasticity implies that the residuals have constant variance across all levels of the independent variables. If the variance changes (e.g., a funnel shape), it suggests heteroscedasticity.

Residuals vs Fitted Plot

- The residuals are randomly scattered around the horizontal axis (zero line). There should be no discernible pattern or curve, indicating a good fit with no obvious violations of linearity or homoscedasticity.

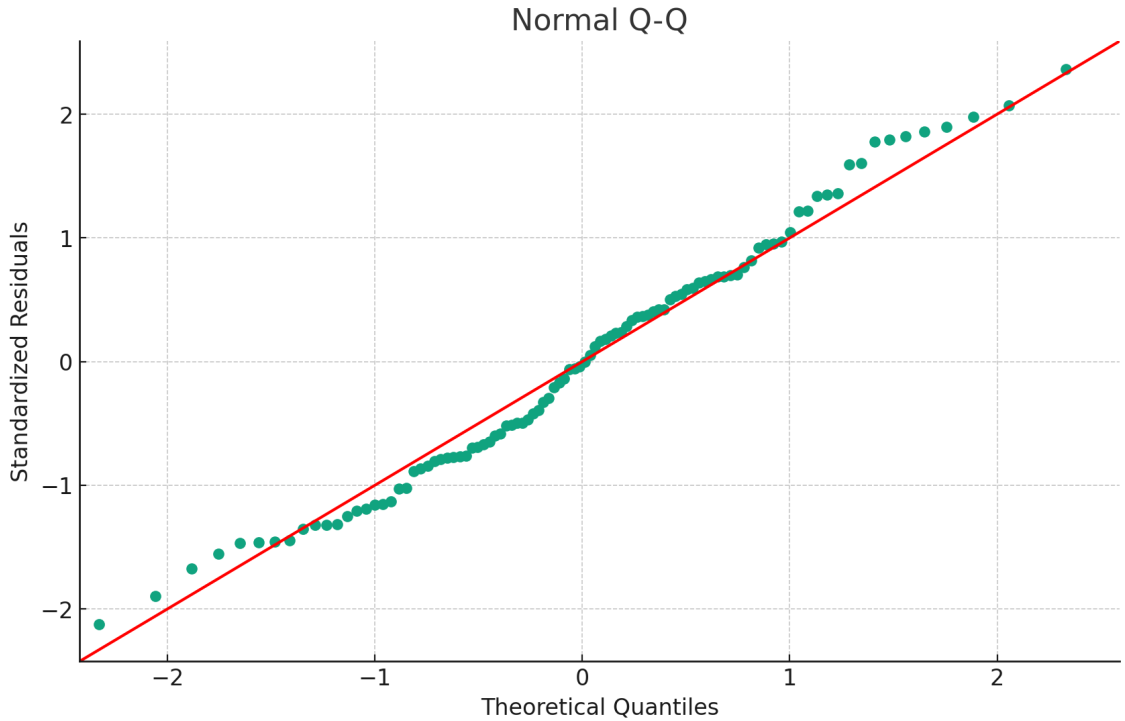


Diagnostic plots

- Normal Q-Q Plot (plot(model, which = 2)
- Assumption Tested: Normality of Residuals.
- Purpose: This plot tests if the residuals are normally distributed. In a well-fitting model, the points should lie approximately along a straight line. Significant deviations from this line suggest non-normality.

Normal Q-Q Plot

- The points should fall approximately along the straight diagonal line. This indicates that the residuals are normally distributed. Small deviations might occur, especially at the tails, but overall, the points should follow the line closely.

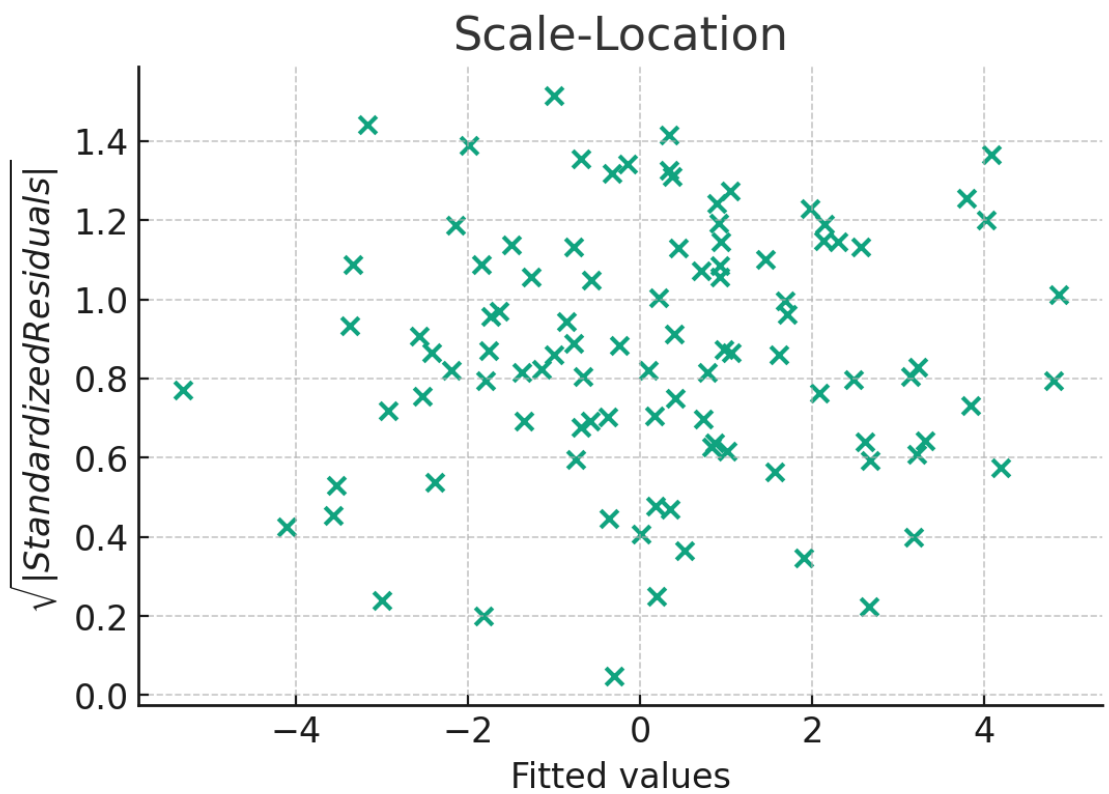


Diagnostic plots

- Scale-Location Plot (plot(model, which = 3)
- Assumption Tested: Homoscedasticity.
- Purpose: Similar to the Residuals vs Fitted plot, this plot checks for constant variance of residuals (homoscedasticity). It's another way to visualize if the spread of the residuals is consistent across the range of predictors.

Scale-Location Plot

- The spread of the residuals should be roughly the same across the entire range of fitted values (i.e., a horizontal band). This suggests homoscedasticity, meaning the variance of the errors is constant across levels of the predictor variable.

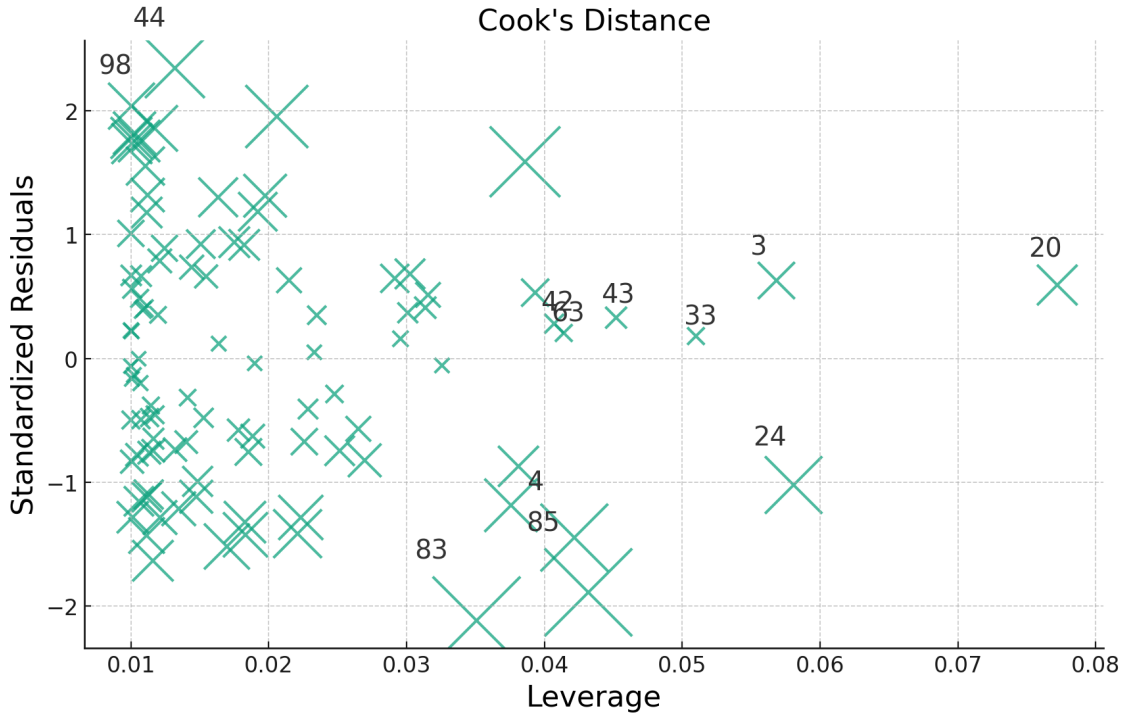


Diagnostic plots

- Cook's Distance Plot (plot(model, which = 4)
- Assumption Tested: Influence of Individual Data Points.
- Purpose: This plot identifies influential observations that have a significant impact on the regression coefficients. Large values of Cook's distance suggest that the corresponding observations have a large impact on the estimated regression coefficients.

Cook's Distance Plot

- Most points should have low Cook's distance values, typically below the threshold (commonly 1 or 0.5). There should be no points with a substantially higher Cook's distance compared to the rest, indicating no single observation unduly influences the model.

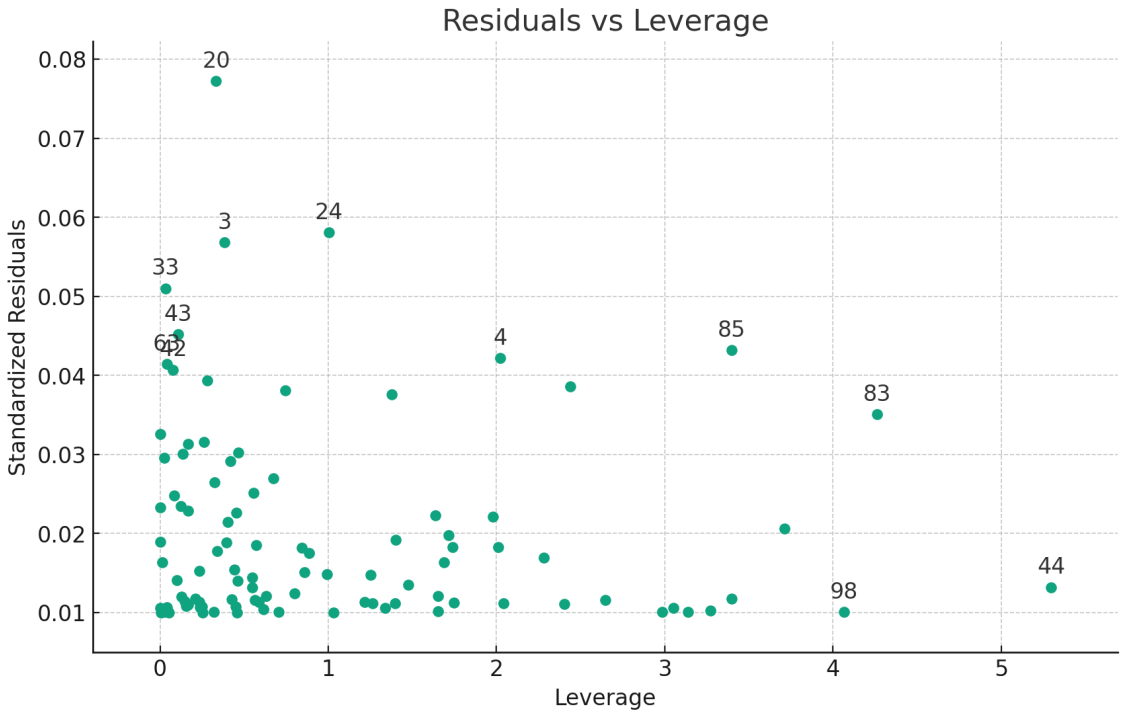


Diagnostic plots

- Residuals vs Leverage Plot (plot(model, which = 5)
- Assumption Tested: Influence of Individual Data Points.
- Purpose: This plot helps to find influential cases in the data set, i.e., observations that affect the regression line's slope significantly. Points with high leverage can significantly alter the position of the regression line if removed.

Residuals vs Leverage Plot

- The leverage values (on the x-axis) should not be extreme, and there should be no points that stand out from the rest. The plot might also include Cook's distance contours to help identify influential points. Ideally, no points should be outside the contours or in the top-right or bottom-right corners.



Wrap-up

- Administrative data
- Secondary data analysis with R

Thank you