Checking conditions for MLR

Prof. Maria Tackett

[Click here for PDF of slides]

Example: SAT Averages by State

- This data set contains the average SAT score (out of 1600) and other variables that may be associated with SAT performance for each of the 50 U.S. states. The data is based on test takers for the 1982 exam.
- Response variable:
  - **SAT**: average total SAT score

Data comes from **`case1201`** data set in the **`Sleuth3`** package

SAT Averages: Predictors

- **`Takers`**: percentage of high school seniors who took exam
- **`Income`**: median income of families of test-takers ($ hundreds)
- **`Years`**: average number of years test-takers had formal education in social sciences, natural sciences, and humanities
- **`Public`**: percentage of test-takers who attended public high schools
- **`Expend`**: total state expenditure on high schools ($ hundreds per student)
- **`Rank`**: median percentile rank of test-takers within their high school classes

Model

term	estimate	std.error	statistic	p.value
(Intercept)	-94.659	211.510	-0.448	0.657
Takers	-0.480	0.694	-0.692	0.493
Income	-0.008	0.152	-0.054	0.957
Years	22.610	6.315	3.581	0.001
Public	-0.464	0.579	-0.802	0.427
Expend	2.212	0.846	2.615	0.012
Rank	8.476	2.108	4.021	0.000

Model conditions

1. **Linearity:** There is a linear relationship between the response and **each** predictor variable

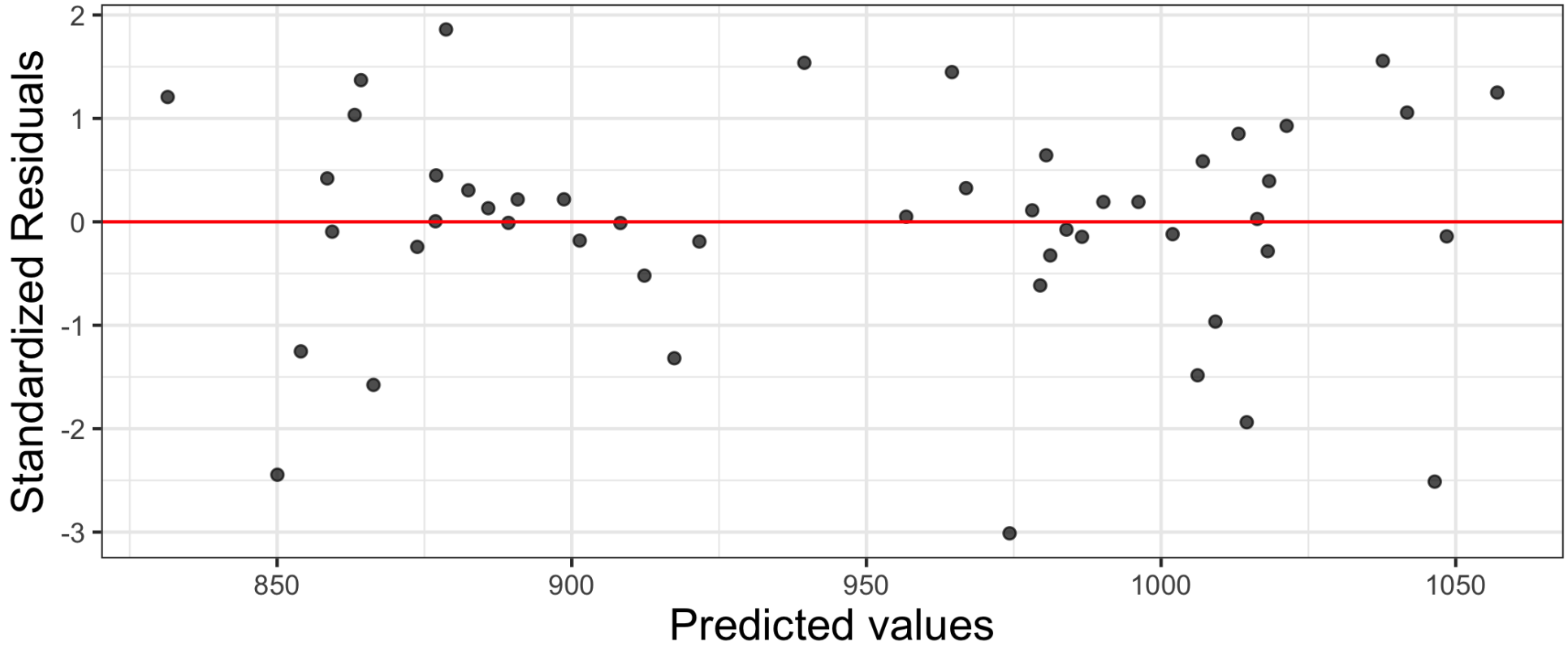
2. **Constant Variance:** The variability of the errors is equal for all values of the predictor variable.

3. **Normality:** The errors follow a normal distribution.

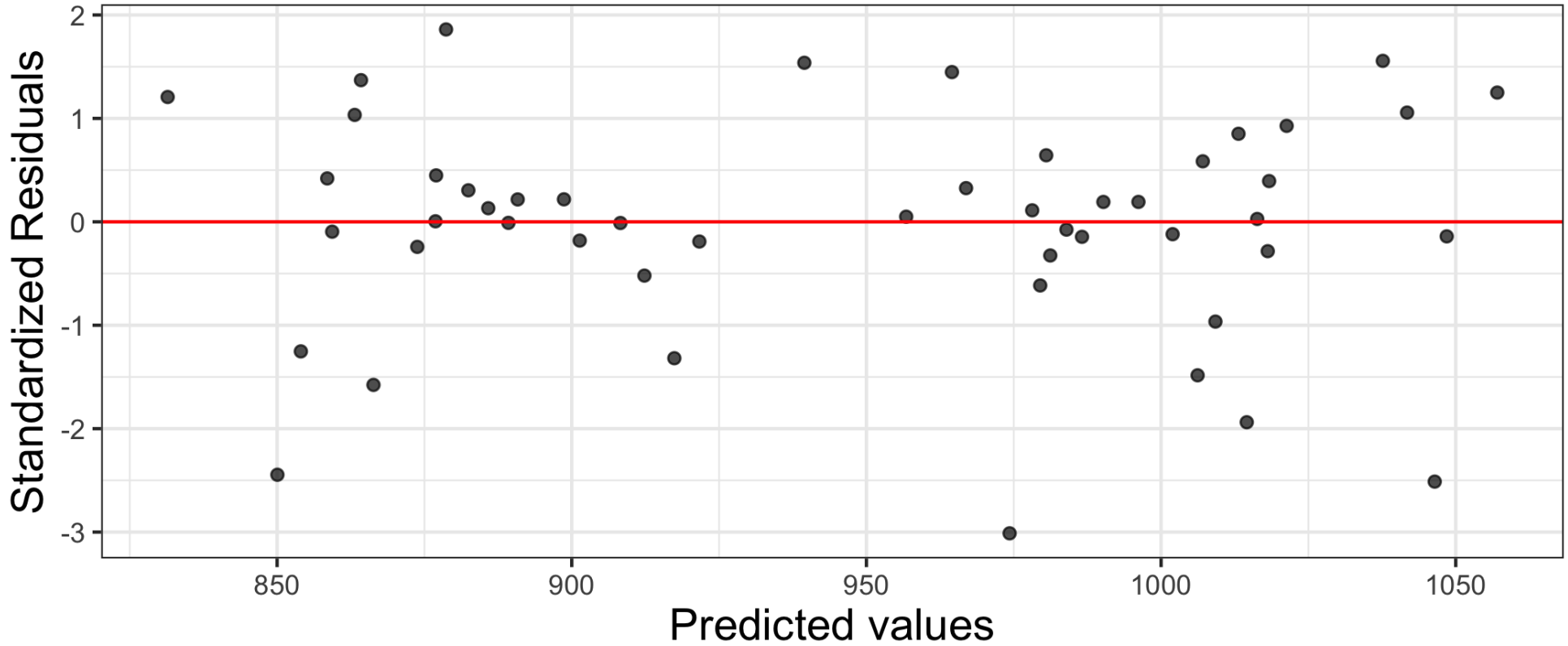
4. **Independence:** The errors are independent from each other.

Use plots of the standardized residuals to check conditions.

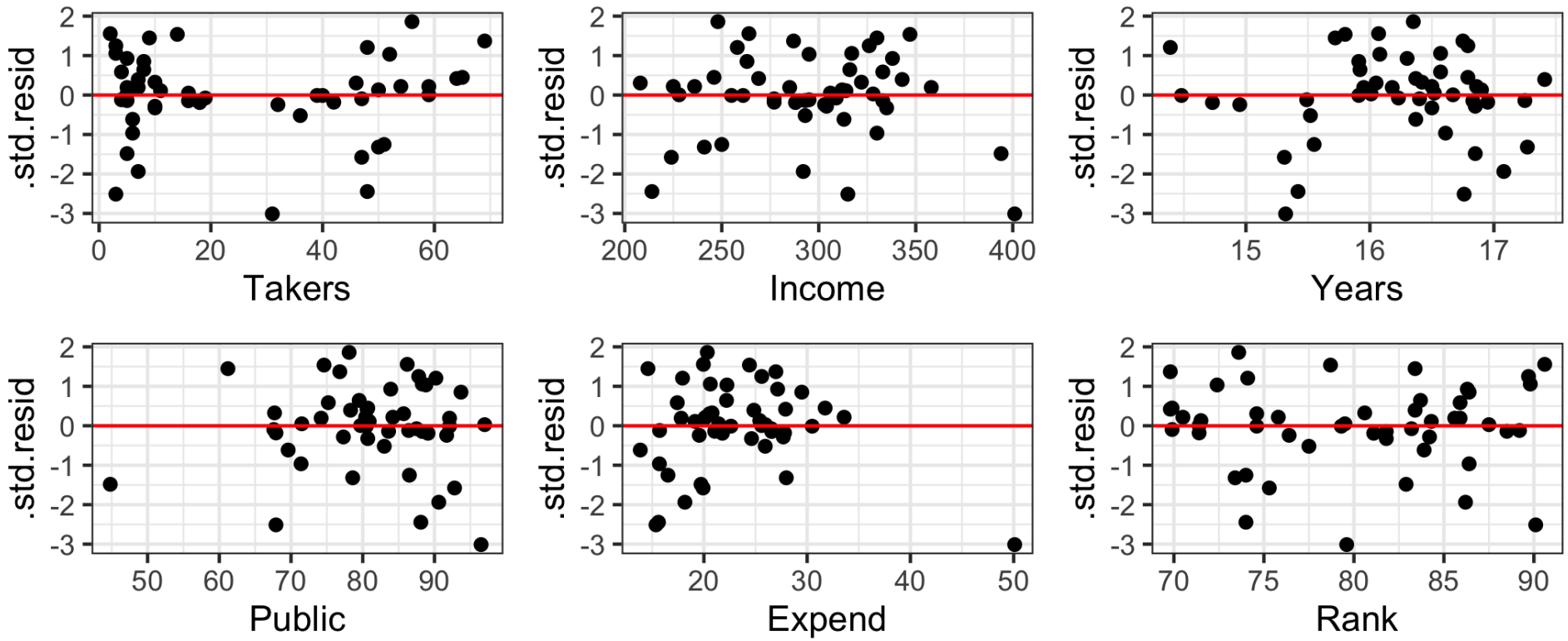
Standardized residuals vs. predicted values



Checking linearity: Std. residuals vs. predicted



Checking linearity: Std. residuals vs. each predictor



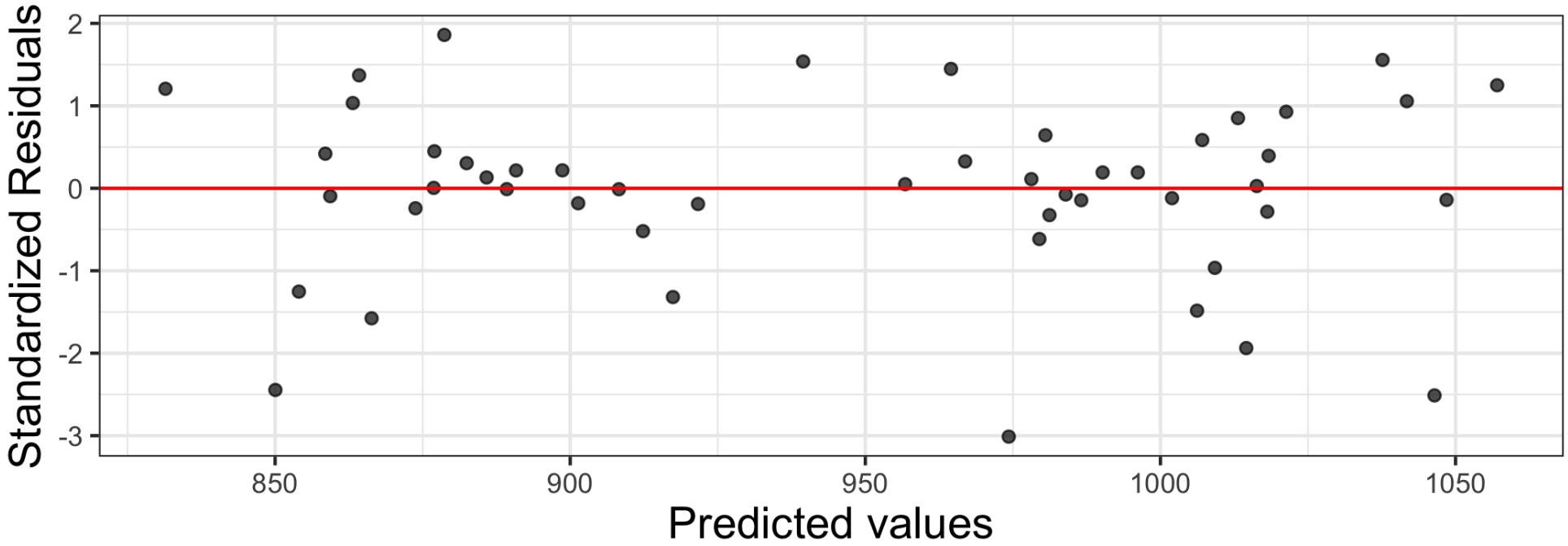
Checking linearity

✅ The plot of standardized residuals vs. predicted shows no distinguishable pattern

✅ The plots of standardized residuals vs. each predictor variable show no distinguishable pattern

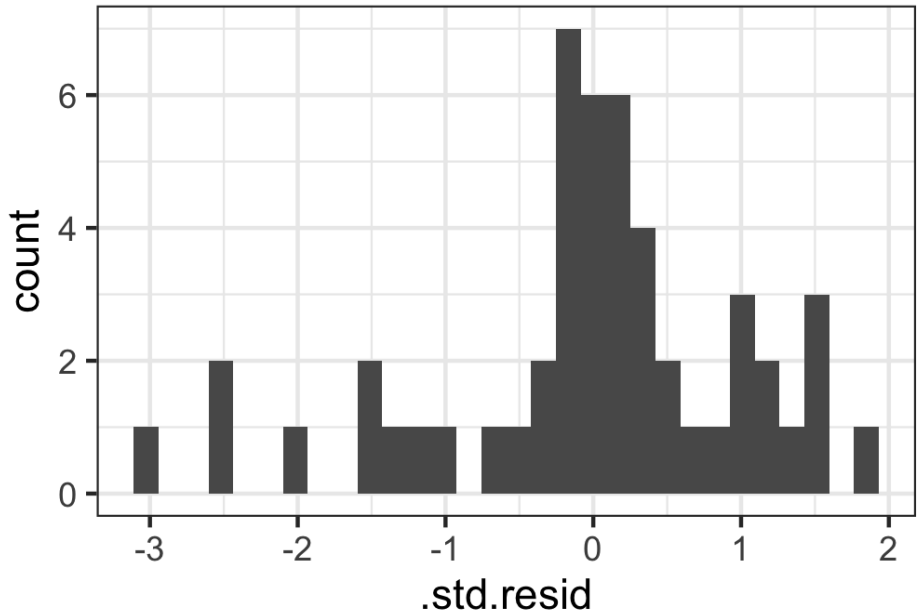
The linearity condition is satisfied.

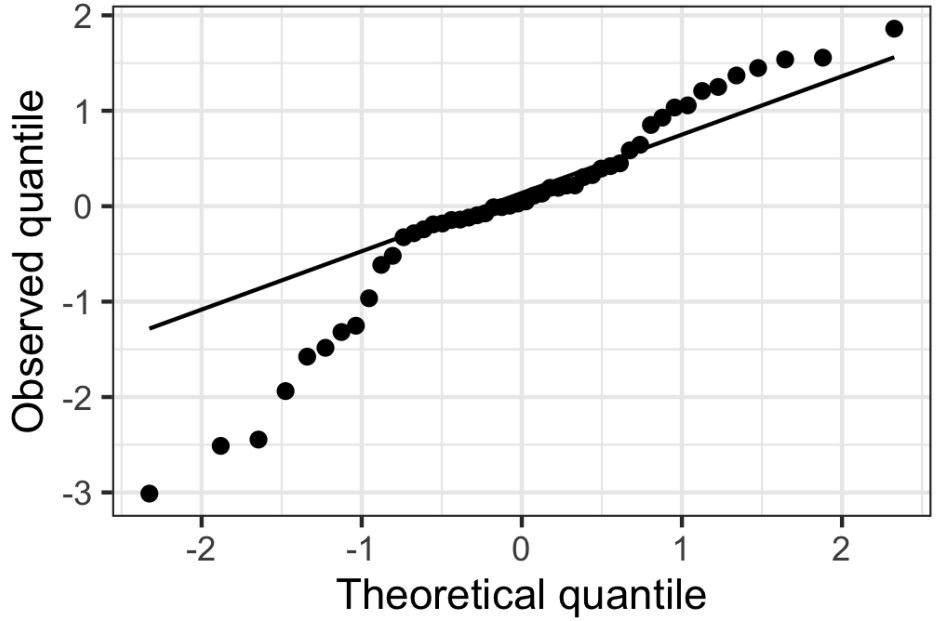
Checking constant variance



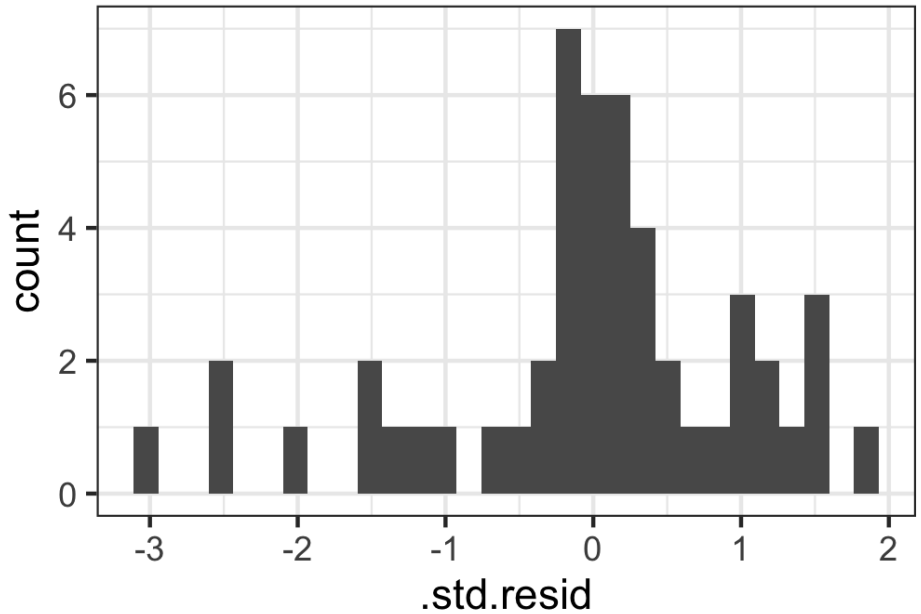
✅ The vertical spread of the residuals is relatively constant across the plot. **The constant variance condition is satisfied.**

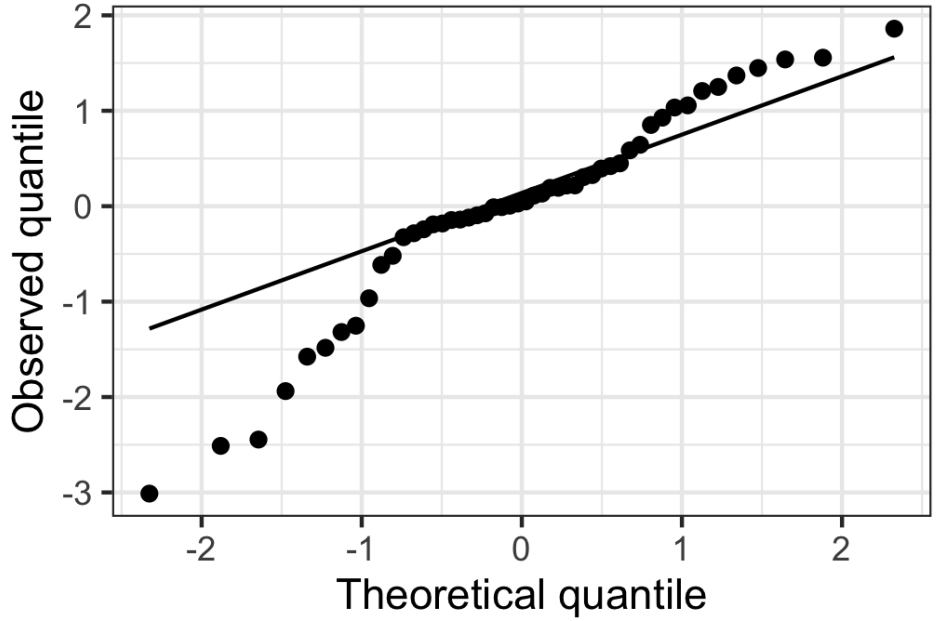
Checking normality





Checking normality





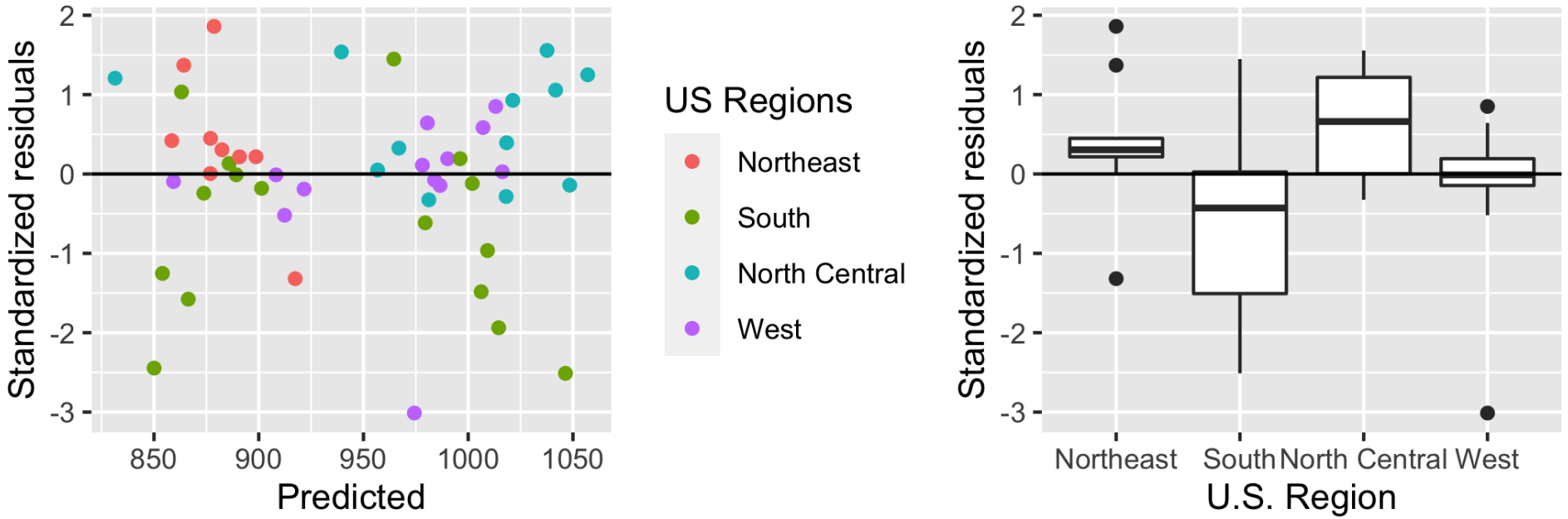
⚠️ **Normality is not satisfied**; however, $n > 30$, so our sample is large enough that we can relax the Normality condition and proceed.

Checking independence

- We can often check the independence condition based on the context of the data and how the observations were collected.
- If the data were collected in a particular order, examine a scatterplot of the standardized residuals versus order in which the data were collected.

Checking independence

Since the observations are US states, let's take a look at the standardized residuals by region.



Checking independence

❌ The model tends to overpredict for states in the South and underpredict for states in the North Central, so the **independence condition is not satisfied**.

Multiple linear regression is **not** robust to violations of independence, so we need to fit a new model that includes region as a predictor to account for the systematic differences by region.

Next, check the model diagnostics

Once you've assessed the conditions for multiple linear regression, then you can use the model diagnostics to detect influential points or multicollinearity.