MLR: Checking conditions & mulitcollinearity

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

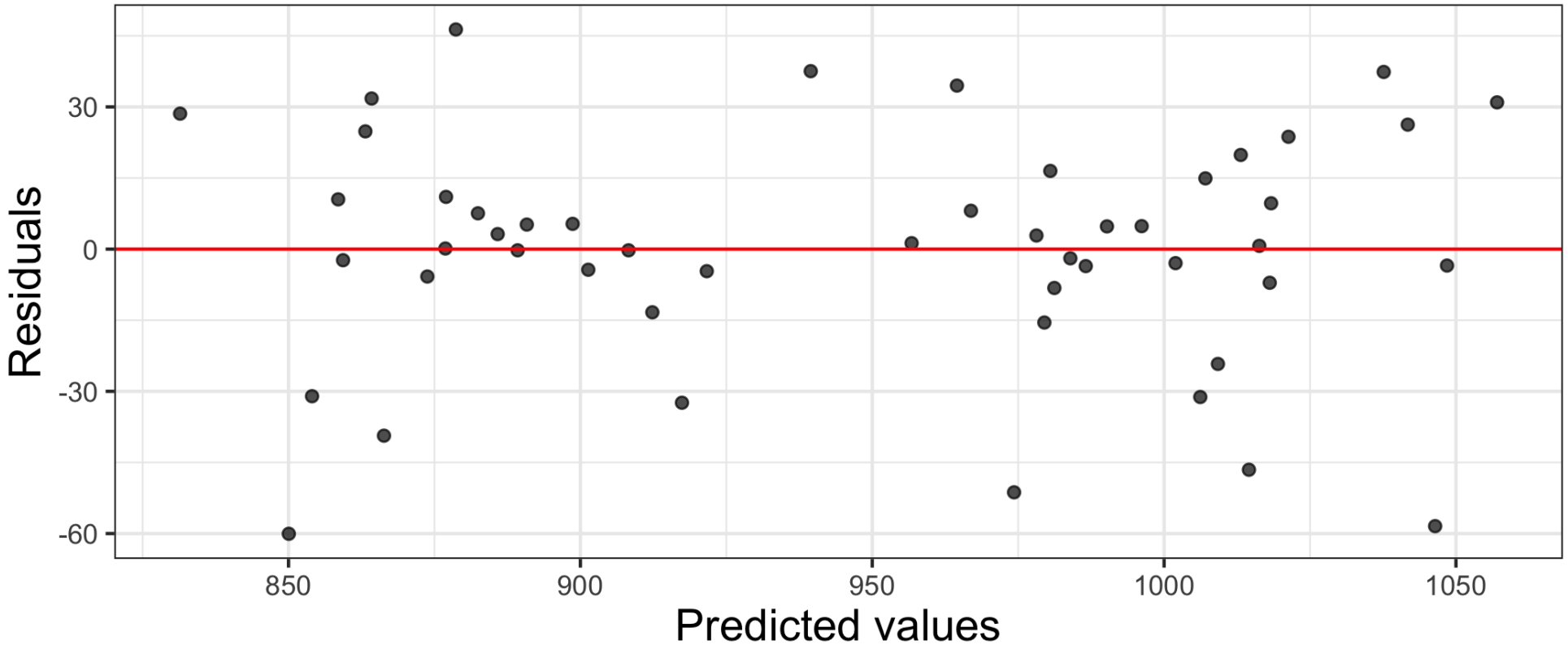
1. **Linearity:** There is a linear relationship between the response and predictor variables.

2. **Constant Variance:** The variability about the least squares line is generally constant.

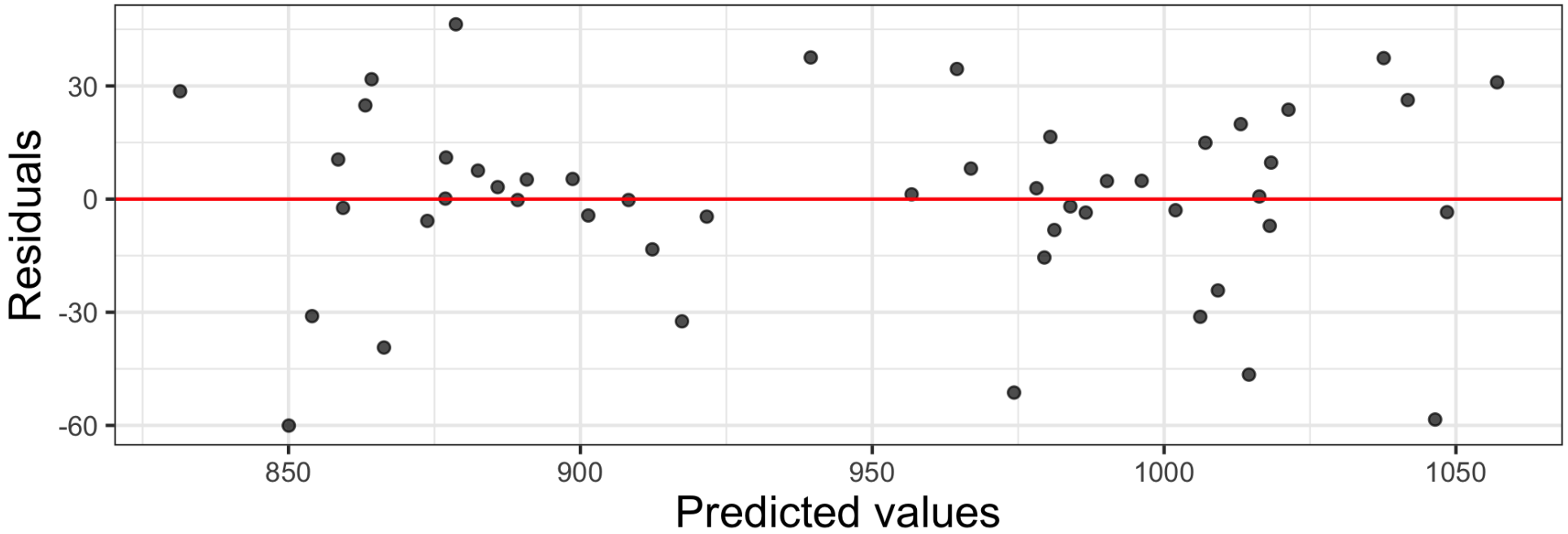
3. **Normality:** The distribution of the residuals is approximately normal.

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

Residuals vs. predicted values



Linearity: Residuals vs. predicted



Linearity: Residuals vs. each predictor

If there is some pattern in the plot of residuals vs. predicted values, you can look at individual plots of residuals vs. each predictor to try to identify the issue.

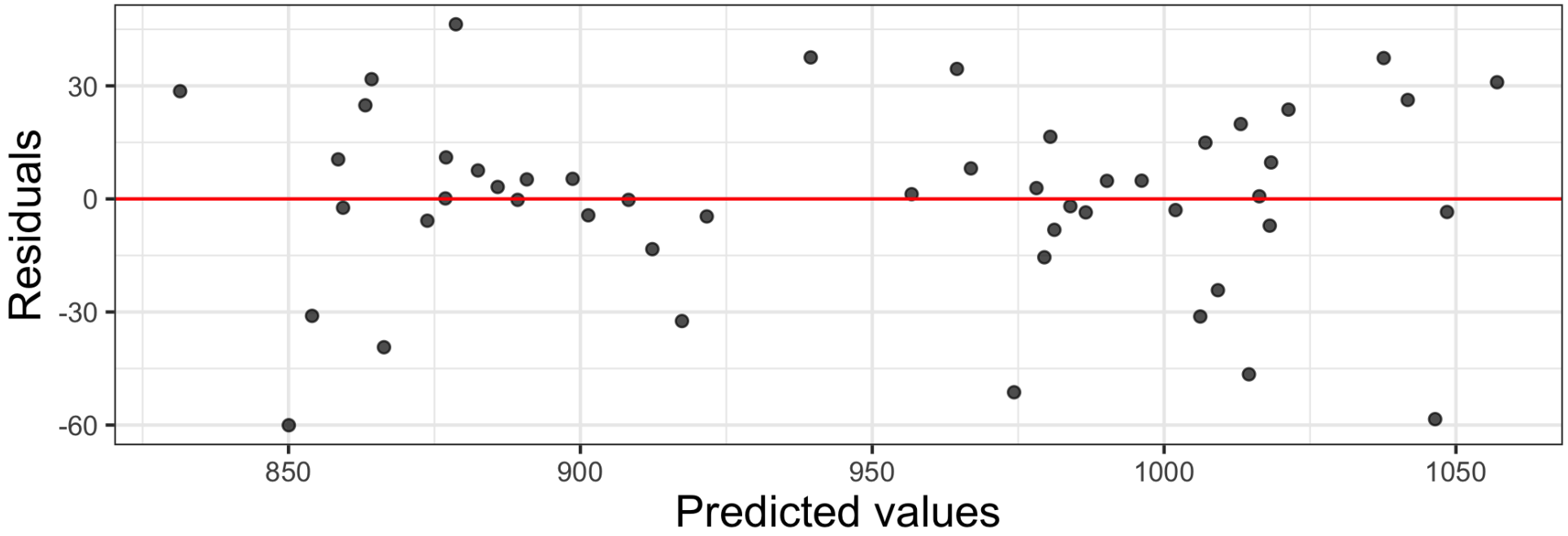
Checking linearity

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

✅ The plots of residuals vs. each predictor variable are generally fine; perhaps look into **`Years`** more closely.

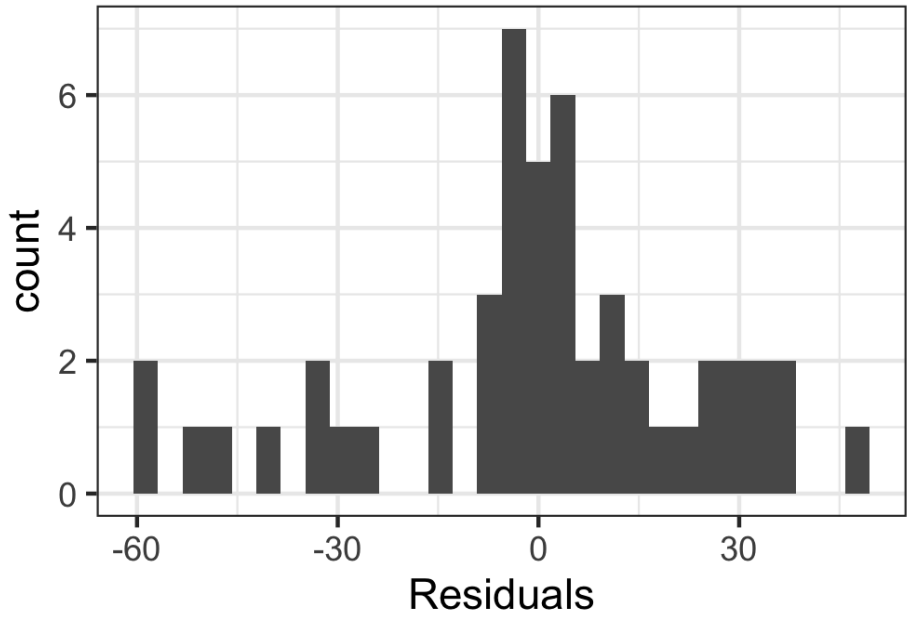
The linearity condition is generally satisfied.

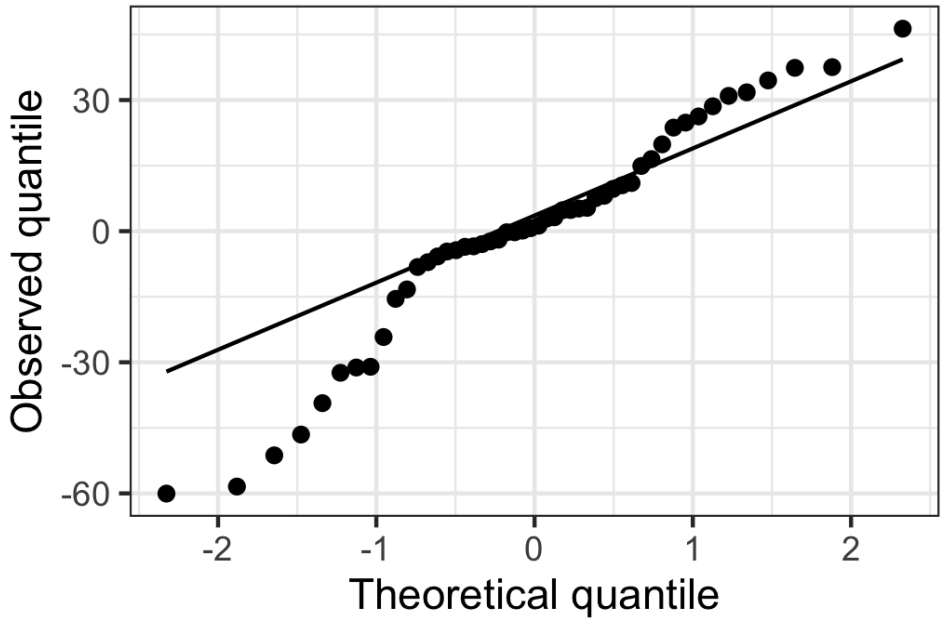
Checking constant variance



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

Checking normality





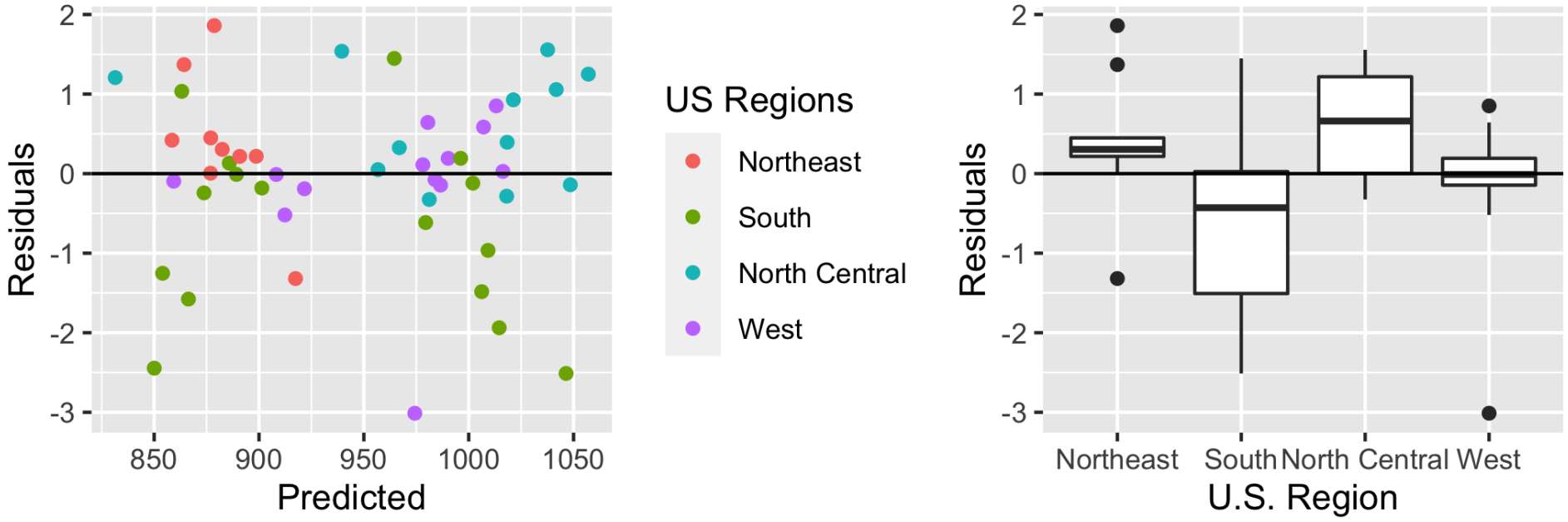
⚠️ **Normality is not satisfied**. However, $n > 30$, so by the Central Limit Theorem, we can still do inference about the model parameters.

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 residuals versus order in which the data were collected.
- If there is a grouping variable lurking in the background, check the residuals based on that grouping variable.

Checking independence

Since the observations are US states, let's take a look at the 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 before moving forward, we should try fitting a model that includes **region** to account for these differences by region.

Multicollinearity

Why multicollinearity is a problem

- We can't include two variables that have a perfect linear association with each other
- If we did so, we could not find unique estimates for the model coefficients

Example

Suppose the true population regression equation is $y = 3 + 4x$

- Suppose we try estimating that equation using a model with variables x and $z = x/10$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 z$$

$$= \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 \frac{x}{10}$$

$$= \hat{\beta}_0 + \left(\hat{\beta}_1 + \frac{\hat{\beta}_2}{10}\right) x$$

Example

$$\hat{y} = \hat{\beta}_0 + \left(\hat{\beta}_1 + \frac{\hat{\beta}_2}{10}\right)x$$

- We can set $\hat{\beta}_1$ and $\hat{\beta}_2$ to any two numbers such that $\hat{\beta}_1 + \frac{\hat{\beta}_2}{10} = 4$
- Therefore, we are unable to choose the "best" combination of $\hat{\beta}_1$ and $\hat{\beta}_2$

Why multicollinearity is a problem

- When we have almost perfect collinearities (i.e. highly correlated predictor variables), the standard errors for our regression coefficients inflate
- In other words, we lose precision in our estimates of the regression coefficients
- This impedes our ability to use the model for inference or prediction

Detecting Multicollinearity

Multicollinearity may occur when...

- There are very high correlations ($r > 0.9$) among two or more predictor variables, especially when the sample size is small
- One (or more) predictor variables is an almost perfect linear combination of the others
- Include a quadratic in the model mean-centering the variable first
- Including interactions between two or more continuous variables

Detecting multicollinearity in the EDA

✅ Look at a correlation matrix of the predictor variables, including all indicator variables

- Look out for values close to 1 or -1

✅ Look at a scatterplot matrix of the predictor variables

- Look out for plots that show a relatively linear relationship

Detecting Multicollinearity (VIF)

Variance Inflation Factor (VIF): Measure of multicollinearity in the regression model

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

where $R^2_{X_j|X_{-j}}$ is the proportion of variation X that is explained by the linear combination of the other explanatory variables in the model.

Detecting Multicollinearity (VIF)

Typically $VIF > 10$ indicates concerning multicollinearity

- Variables with similar values of VIF are typically the ones correlated with each other

Use the `vif()` function in the **rms** R package to calculate VIF

VIF For SAT Model

```
vif(sat_model) %>% tidy() %>% kable()
```

names	x
Takers	16.478636
Income	3.128848
Years	1.379408
Public	2.288398
Expend	1.907995
Rank	13.347395

Takers and **Rank** are correlated. We need to remove one of these variables and refit the model.

Model without `Takers`

term	estimate	std.error	statistic	p.value
(Intercept)	-213.754	122.238	-1.749	0.087
Income	0.043	0.133	0.322	0.749
Years	22.354	6.266	3.567	0.001
Public	-0.559	0.559	-0.999	0.323
Expend	2.094	0.824	2.542	0.015
Rank	9.803	0.872	11.245	0.000

```
## # A tibble: 1 × 3
##   adj.r.squared   AIC   BIC
##           <dbl> <dbl> <dbl>
## 1         0.863  476.  489.
```

Model without **Rank**

term	estimate	std.error	statistic	p.value
(Intercept)	535.091	164.868	3.246	0.002
Income	-0.117	0.174	-0.675	0.503
Years	26.927	7.216	3.731	0.001
Public	0.536	0.607	0.883	0.382
Expend	2.024	0.980	2.066	0.045
Takers	-3.017	0.335	-9.014	0.000

```
## # A tibble: 1 × 3
##   adj.r.squared   AIC   BIC
##           <dbl> <dbl> <dbl>
## 1         0.814  491.  505.
```

Choosing a model

Model with **Takers** removed:

adj.r.squared	AIC	BIC
0.863	476.031	489.415

Model with **Rank** removed:

adj.r.squared	AIC	BIC
0.8141061	491.4388	504.8229

Based on Adjusted R^2, AIC, and BIC, the model with **Takers** removed is a better fit. Therefore, we choose to remove **Takers** from the model and leave **Rank** in the model to deal with the multicollinearity.