Multiple Linear Regression

Model Selection & Diagnostics

Prof. Maria Tackett

03.04.20

STA 210

Click for PDF of slides

Announcements

- Project proposal **due Thursday, March 5 at 11:59p**
- Reading 07 today
- Sign Up for DataFest!

🗓️ April 3 - 5

📍 Penn Pavilion

🔗 stat.duke.edu/datafest

> Have a great spring break! ☀️

R packages

```r
library(tidyverse)
library(knitr)
library(broom)
library(patchwork)
```

Today's Agenda

- Model Selection
    - R^2 vs. Adj. R^2
    - AIC & BIC
    - Strategies
- Model diagnostics

Model Selection

Which variables should be in the model?

- This is a very hard question that is the subject of a lot of statistical research
- There are many different opinions about how to answer this question
- This lecture will mostly focus on how to approach variable selection
    - We will introduce some specific methods, but there are many others out there

Which variables should you include?

- It depends on the goal of your analysis
- Though a variable selection procedure will select one set of variables for the model, that set is usually one of several equally good sets
- It is best to start with a well-defined purpose and question to help guide the variable selection

Prediction

- Goal: to calculate the most precise prediction of the response variable
- Interpreting coefficients is **not** important
- Choose only the variables that are strong predictors of the response variable
  - Excluding irrelevant variables can help reduce widths of the prediction intervals

One variable's effect

- Goal: Understand one variable's effect on the response after adjusting for other factors

- Only interpret the coefficient of the variable that is the focus of the study

  - Interpreting the coefficients of the other variables is **not** important

- Any variables not selected for the final model have still been adjusted for, since they had a chance to be in the model

Explanation

- **Goal:** Identify variables that are important in explaining variation in the response
- Interpret any variables of interest
- Include all variables you think are related to the response, even if they are not statistically significant
    - This improves the interpretation of the coefficients of interest
- Interpret the coefficients with caution, especially if there are problems with multicollinearity in the model

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.

- Data: `case1201` data set in the `Sleuth3` package

- Response variable:

  - **SAT**: average total SAT score

SAT Averages: Explanatory Variables

- `State`: U.S. State
- `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

In-Class Exercise:

Select the primary modeling objective for each scenario

http://bit.ly/sta210-sp20-selection

Use **NetId@duke.edu** for your email address.

If you finish early, discuss a modeling strategy for each scenario.

04:00

Model selection criterion

R^2

- **Recall**: R^2 is the proportion of the variation in the response variable explained by the regression model

- R^2 will always increase as we add more variables to the model
  - If we add enough variables, we can always achieve $R^2 = 100\%$

- If we only use R^2 to choose a best fit model, we will be prone to choose the model with the most predictor variables

Adjusted R^2

- **Adjusted R^2**: a version of R^2 that penalizes for unnecessary predictor variables

- Similar to R^2, it is a measure of the amount of variation in the response that is explained by the regression model

- Differs from R^2 by using the mean squares rather than sums of squares and therefore adjusting for the number of predictor variables

R^2 and Adjusted R^2

$$R^2 = \frac{\text{Total Sum of Squares} - \text{Residual Sum of Squares}}{\text{Total Sum of Squares}}$$

$$Adj.R^2 = \frac{\text{Total Mean Square} - \text{Residual Mean Square}}{\text{Total Mean Square}}$$

- $Adj.R^2$ can be used as a quick assessment to compare the fit of multiple models; however, it should not be the only assessment!
- Use R^2 when describing the relationship between the response and predictor variables

SAT:

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

```
glance(sat_model)
```

```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>  <dbl> <dbl> <d
## 1     0.879         0.862  26.3      51.9 4.16e-18     7  -231.  477.  4
## # … with 2 more variables: deviance <dbl>, df.residual <int>
```

Selection Criteria: AIC & BIC

- Akaike's Information Criterion (AIC):

$$AIC = n\log(RSS) - n\log(n) + 2(p+1)$$

- Schwarz's Bayesian Information Criterion (BIC):

$$BIC = n\log(RSS) - n\log(n) + log(n) \times (p+1)$$

See the supplemental note on AIC & BIC for derivations.

Selection Criteria: AIC & BIC

$$AIC = n\log(RSS) - n\log(n) + 2(p+1)$$
$$BIC = n\log(RSS) - n\log(n) + \log(n) \times (p+1)$$

- First Term: Decreases as p increases
- Second Term: Fixed for a given sample size n
- Third Term: Increases as p increases

Selection Criteria: Using AIC & BIC

$$AIC = n\log(RSS) - n\log(n) + 2(p+1)$$
$$BIC = n\log(RSS) - n\log(n) + \log(n)\times(p+1)$$

- Choose model with smallest AIC or BIC
- If $n \geq 8$, the penalty for BIC is larger than that of AIC, so BIC tends to favor *more parsimonious* models (i.e. models with fewer terms)

Selection Process: Backward Selection

- Start with model that includes all variables of interest
- Drop variables one at a time that are deemed irrelevant based on some criterion. Common criterion include
  - Drop variable that results in the model with the highest Adj. $\boxed{\text{R\^{}}}$ ***or***
  - Drop variable that results in the model with the lowest value of AIC or BIC
- Stop when no more variables can be removed from the model based on the criterion

Selection Process: Forward Selection

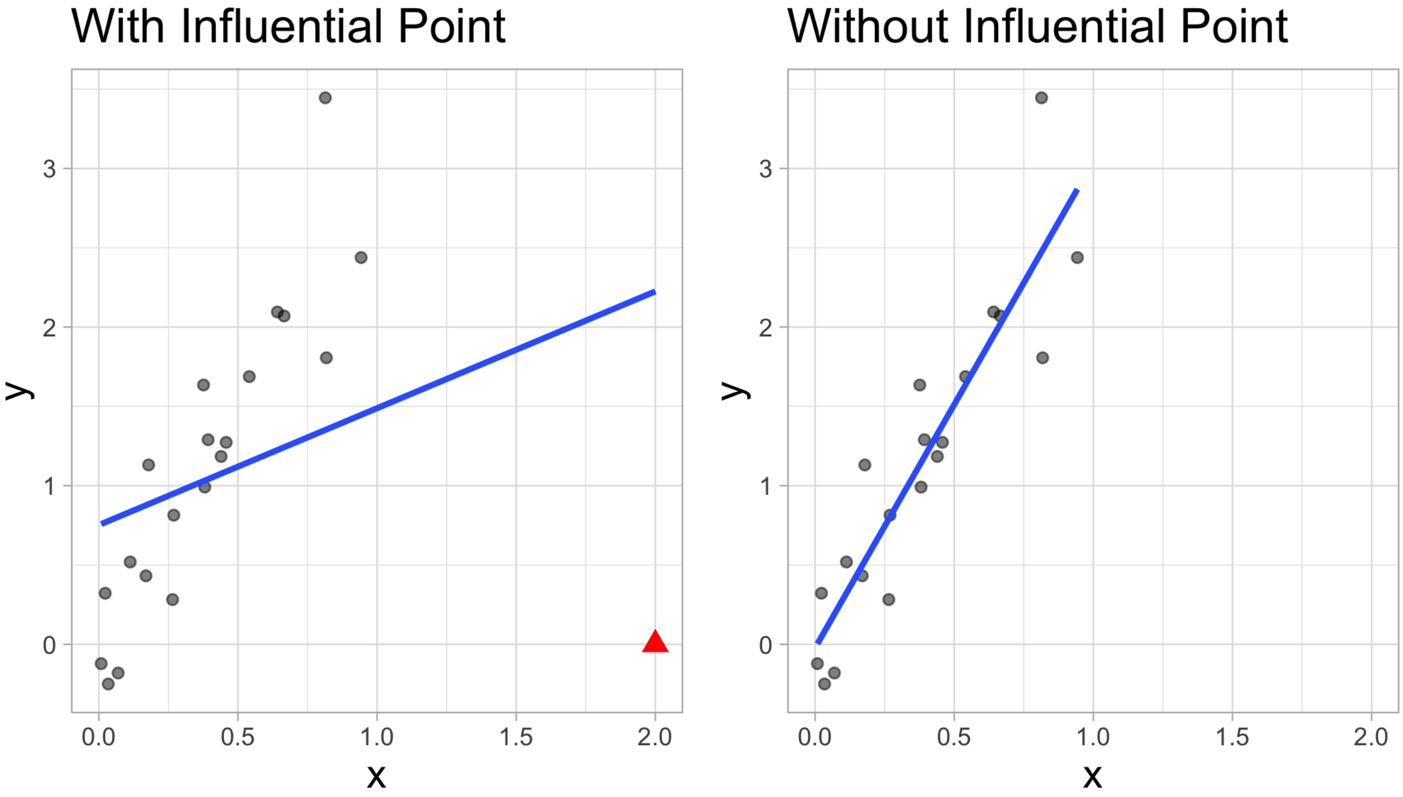
- Start with the intercept-only model
- Include variables one at a time based on some criterion. Common criterion include
  - Add variable that results in the model with highest Adj. R^2 ***or***
  - Add variable that results in the model with the lowest value of AIC or BIC
- Stop when no more variables can be added to the model based on the criterion

Model Diagnostics

Influential and Leverage Points

Influential Observations

An observation is **influential** if removing it substantially changes the coefficients of the regression model



Influential Observations

- In addition to the coefficients, influential observations can have a large impact on the standard errors
- Occasionally these observations can be identified in the scatterplot
  - This is often not the case - especially when dealing with multivariate data
- We will use measures to quantify an individual observation's influence on the regression model
  - **leverage**, **standardized residuals**, and **Cook's distance**

Model diagnostics in R

- Use the **`augment`** function in the broom package to output the model diagnostics (along with the predicted values and residuals)

- Output from `augment` :

  - response and predictor variables in the model
  - `.fitted`: predicted values
  - `.se.fit`: standard errors of predicted values
  - `.resid`: residuals
  - .vocab[`.hat`]: leverage
  - `.sigma`: estimate of residual standard deviation when corresponding observation is dropped from model
  - .vocab[`.cooksd`]: Cook's distance
  - .vocab[`.std.resid`]: standardized residuals

SAT: Augmented Data

```r
sat_aug <- augment(sat_model) %>%
  mutate(obs_num = row_number()) #add observation number for plots
```

```r
glimpse(sat_aug)
```

```
## Observations: 50
## Variables: 15
## $ SAT        <int> 1088, 1075, 1068, 1045, 1045, 1033, 1028, 1022, 1017,
## $ Takers     <int> 3, 2, 3, 5, 5, 8, 7, 4, 5, 10, 5, 4, 9, 8, 7, 3, 6, 1
## $ Income     <int> 326, 264, 317, 338, 293, 263, 343, 333, 328, 304, 358
## $ Years      <dbl> 16.79, 16.07, 16.57, 16.30, 17.25, 15.91, 17.41, 16.5
## $ Public     <dbl> 87.8, 86.2, 88.3, 83.9, 83.6, 93.7, 78.3, 75.2, 97.0,
## $ Expend     <dbl> 25.60, 19.95, 20.62, 27.14, 21.05, 29.48, 24.84, 17.4
## $ Rank       <dbl> 89.7, 90.6, 89.8, 86.3, 88.5, 86.4, 83.4, 85.9, 87.5,
## $ .fitted    <dbl> 1057.0438, 1037.6261, 1041.7431, 1021.3039, 1048.4680
## $ .se.fit    <dbl> 8.976321, 10.838317, 8.737717, 6.472356, 9.224889, 12
## $ .resid     <dbl> 30.9562319, 37.3739084, 26.2569334, 23.6961288, -3.46
## $ .hat       <dbl> 0.11609974, 0.16926150, 0.11000956, 0.06036139, 0.122
## $ .sigma     <dbl> 26.16716, 25.89402, 26.30760, 26.38760, 26.64972, 26.
## $ .cooksd    <dbl> 2.931280e-02, 7.051849e-02, 1.970989e-02, 7.901850e-0
## $ .std.resid <dbl> 1.24986670, 1.55651598, 1.05649773, 0.92792786, -0.14
```

Leverage

- **Leverage:** measure of the distance between an observation's values of the predictor variables and the average values of the predictor variables for the whole data set

- An observation has high leverage if its combination of values for the predictor variables is very far from the typical combinations in the data

    - It is potentially an influential point, i.e. may have a large impact on the coefficient estimates and standard errors

- **Note:** Identifying points with high leverage has nothing to do with the values of the response variables

Calculating Leverage

- **Simple Regression:** leverage of the i^{th} observation is

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

- **Multiple Regression:** leverage of the i^{th} observation is the i^{th} diagonal of

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

- **Note**: Leverage only depends on values of the **predictor** variables

STA 210

High Leverage

- Values of leverage are between $\frac{1}{n}$ and 1 for each observation
- The average leverage for all observations in the data set is $\frac{(p+1)}{n}$

An observation has **high leverage** if

$$h_i > \frac{2(p+1)}{n}$$

- Note: We can't rely on residuals alone to identify these points , since observations with high leverage tend to have small residuals

High Leverage

- Questions to check if you identify points with high leverage:
  - Are they a result of data entry errors?
  - Are they in the scope for the individuals for which you want to make predictions?
  - Are they impacting the estimates of the model coefficients, especially for interactions?
- Just because a point has high leverage does not necessarily mean it will have a substantial impact on the regression. Therefore you should check other measures.

SAT: Leverage

```
(leverage_threshold <- 2*(7+1)/nrow(sat_aug))
```

```
## [1] 0.32
```

```
ggplot(data = sat_aug, aes(x = obs_num, y = .hat)) +
  geom_point(alpha = 0.7) +
  geom_hline(yintercept = leverage_threshold, color = "red")+
  labs(x = "Observation Number",y = "Leverage",title = "Leverage")
  geom_text(aes(label=ifelse(.hat > leverage_threshold, as.charact
```

Points with high leverage

```
sat_aug %>% filter(.hat > leverage_threshold) %>%
  select(obs_num, Takers, Income, Years, Public, Expend, Rank)
```

```
## # A tibble: 2 x 7
##   obs_num Takers Income Years Public Expend  Rank
##     <int>  <int>  <int> <dbl>  <dbl>  <dbl> <dbl>
## 1      22      5    394  16.8   44.8   19.7  82.9
## 2      29     31    401  15.3   96.5   50.1  79.6
```

Why do you think these points have high leverage?

Standardized & Studentized Residuals

- What is the best way to identify outliers (points that don't fit the pattern from the regression line)?
- Look for points that have large residuals
- We want a common scale, so we can more easily identify "large" residuals
- We will look at each residual divided by its standard error

Standardized Residuals

$$std.res_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

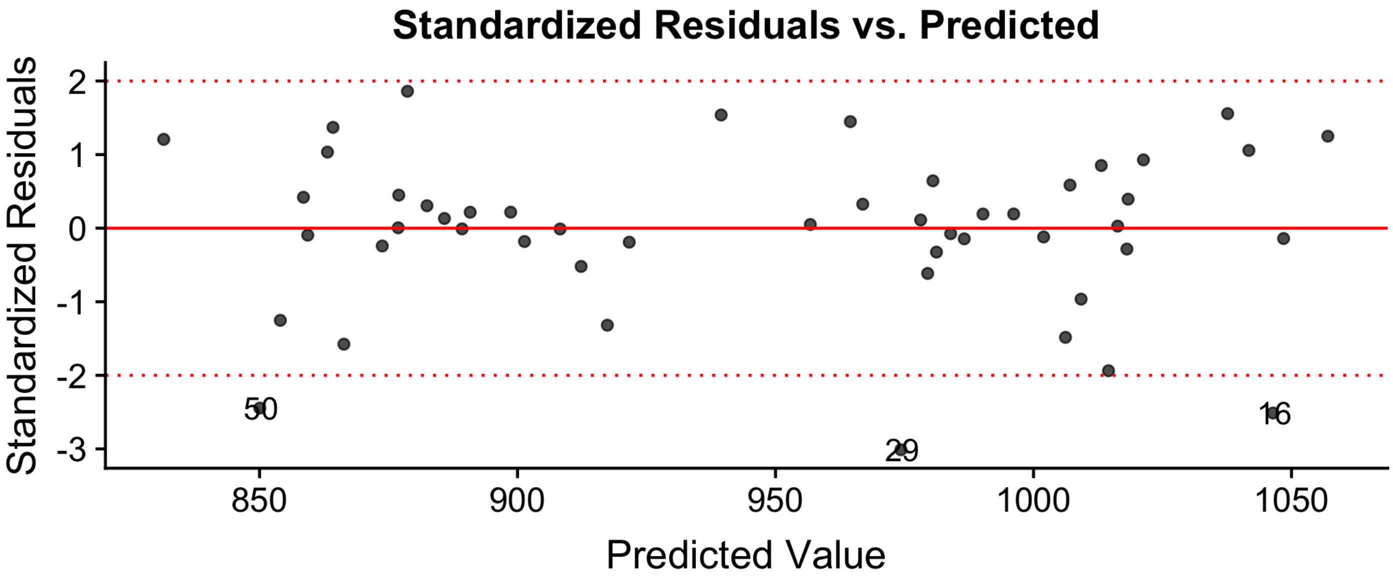
- The standard error of a residual, $\hat{\sigma}\sqrt{1 - h_i}$ depends on the value of the predictor variables
- Residuals for observations that are high leverage have smaller variance than residuals for observations that are low leverage
  - This is because the regression line tries to fit high leverage observations as closely as possible

Standardized Residuals

- Values with very large standardized residuals are outliers, since they don't fit the pattern determined by the regression model
- Observations with standardized residuals of magnitude > 2 should be examined more closely
- Observations with large standardized residuals are outliers but may not have an impact on the regression line
- **Good Practice:** Make residual plots with standardized residuals
    - It is easier to identify outliers and check for constant variance assumption

SAT: Standardized Resdiuals vs. Predicted

```
ggplot(data = sat_aug, aes(x = .fitted,y = .std.resid)) +
  geom_point(alpha = 0.7) +
  geom_hline(yintercept = 0,color = "red") +
  geom_hline(yintercept = -2,color = "red",linetype = "dotted") +
  geom_hline(yintercept = 2,color = "red",linetype = "dotted") +
  labs(x ="Predicted Value",y ="Standardized Residuals",title = "S
  geom_text(aes(label = ifelse(abs(.std.resid) > 2,as.character(ob
```



Motivating Cook's Distance

- If a observation has a large impact on the estimated regression coefficients, when we drop that observation...
    - The estimated coefficients should change
    - The predicted Y value for that observation should change
- One way to determine each observation's impact could be to delete it, rerun the regression, compare the predicted Y values from the new and original models
    - This could be very time consuming
- Instead, we can use Cook's Distance which gives a measure of the change in the predicted Y value when an observation is dropped

Cook's Distance

- **Cook's Distance:** Measure of an observation's overall impact, i.e. the effect removing the observation has on the estimated coefficients

- For the i^{th} observation, we can calculate Cook's Distance as

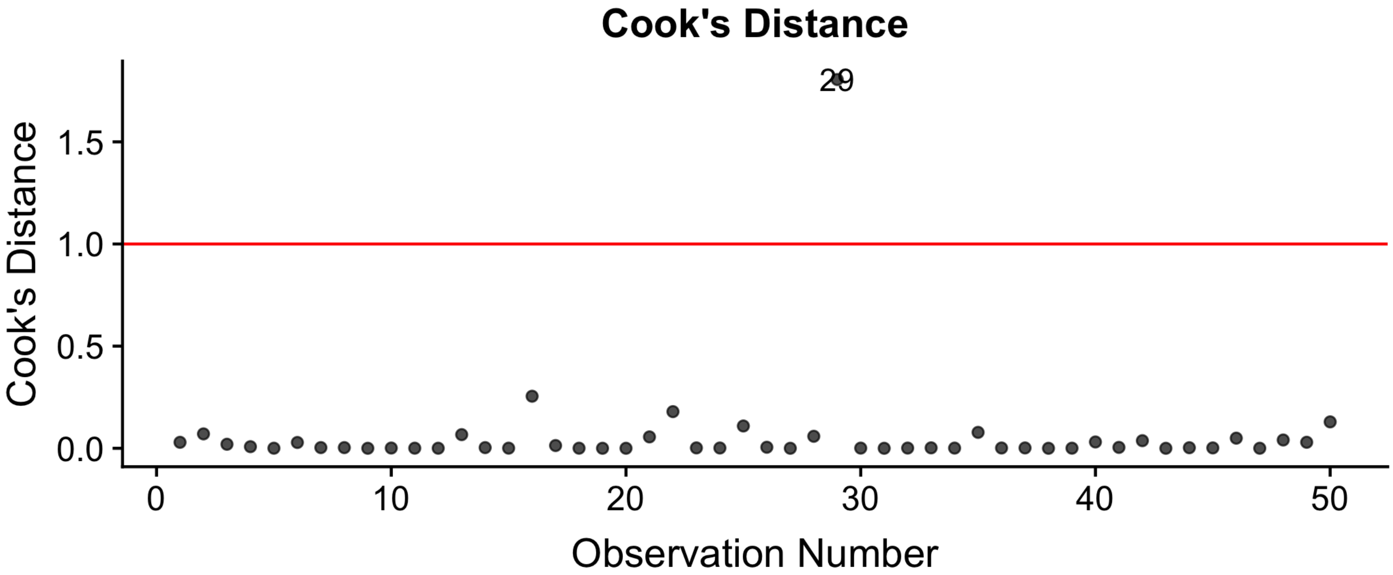
$$D_i = \frac{1}{p}(std.res_i)^2\left(\frac{h_i}{1-h_i}\right)$$

- *Note:* Cook's distance, D_i, incorporates both the residual and the leverage for each observation

- An observation with large D_i is said to have a strong influence on the predicted values

Cook's Distance

```
ggplot(data = sat_aug, aes(x = obs_num, y = .cooksd)) +
  geom_point(alpha = 0.7) +
  geom_hline(yintercept=1,color = "red")+
  labs(x= "Observation Number",y = "Cook's Distance",title = "Cook
  geom_text(aes(label = ifelse(.cooksd > 1,as.character(obs_num),'
```



Influential point: Alaska

```
## # A tibble: 1 x 7
##   obs_num Takers Income Years Public Expend  Rank
##     <int>  <int>  <int> <dbl>  <dbl>  <dbl> <dbl>
## 1      29     31    401  15.3   96.5   50.1  79.6
```

With Alaska

term	estimate
(Intercept)	-94.659
Takers	-0.480
Income	-0.008
Years	22.610
Public	-0.464
Expend	2.212
Rank	8.476

Without Alaska

term	estimate
(Intercept)	-203.926
Takers	0.018
Income	0.181
Years	16.536
Public	-0.443
Expend	3.730
Rank	9.789

Using these measures

- Standardized residuals, leverage, and Cook's Distance should all be examined together
- Examine plots of the measures to identify observations that may have an impact on your regression model
- Some thresholds for flagging potentially influential observations:
  - **Leverage**: $h_i > \frac{2(p+1)}{n}$ (some software uses $2p/n$)
  - **Standardized Residuals**: $|std.res_i| > 2$
  - **Cook's Distance**: $D_i > 1$

What to do with outliers/influential observations?

- It is OK to drop an observation based on the predictor variables if...
  - It is meaningful to drop the observation given the context of the problem
  - You intended to build a model on a smaller range of the predictor variables. Mention this in the write up of the results and be careful to avoid extrapolation when making predictions
- It is not OK to drop an observation based on the response variable
  - These are legitimate observations and should be in the model
- You can try transformations or increasing the sample size by collecting more data
- In either instance, you can try building the model with and without the outliers/influential observations

See the supplemental notes Details on Model Diagnostics for more details about standardized residuals, leverage points, and Cook's distance.

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 pick a unique best fit model

Why multicollinearity is a problem

- Ex. Suppose the true population regression equation is $y = 3 + 4x$
- Suppose we try estimating that regression model using the variables x and $z = x/10$

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

- We can set $\hat{\beta}_1$ and $\hat{\beta}_2$ to any two numbers such that $\hat{\beta}_1 + \frac{\hat{\beta}_2}{10} = 4$
    - We are unable then 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 explanatory variables), the standard errors for our regression coefficients inflate

- In other words, we lose precision in our estimates of the regression coefficients

Detecting Multicollinearity

Multicollinearity may occur when...

- There are very high correlations $(r > 0.9)$ among two or more explanatory variables, especially for smaller sample sizes
- One (or more) explanatory variables is an almost perfect linear combination of the others
- Include quadratic terms without first mean-centering the variables before squaring
- Including interactions with two or more continuous variables

Detecting Multicollinearity

- Look at a correlation matrix of the predictor variables, including all indicator variables
  - Look out for values close to 1 or -1
- If you think one predictor variable is an almost perfect linear combination of other predictor variables, you can run a regression of that predictor variable vs. the others and see if R^2 is close to 1

Detecting Multicollinearity (VIF)

- Variance Inflation Factor (VIF): Measure of multicollinearity

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

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.

- 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` package to calculate VIF

Tips VIF

- Calculate VIF using the **vif** function in the rms package

```
library(rms)
tidy(vif(sat_model))
```

```
## # A tibble: 6 x 2
##   names       x
##   <chr>   <dbl>
## 1 Takers 16.5
## 2 Income  3.13
## 3 Years   1.38
## 4 Public  2.29
## 5 Expend  1.91
## 6 Rank   13.3
```

Takers and Rank are correlated. Should refit the model with one of these variables removed.

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 x 2
##   adj.r.squared   AIC
##           <dbl> <dbl>
## 1         0.863  476.
```

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 x 2
##   adj.r.squared   AIC
##           <dbl> <dbl>
## 1         0.814  491.
```