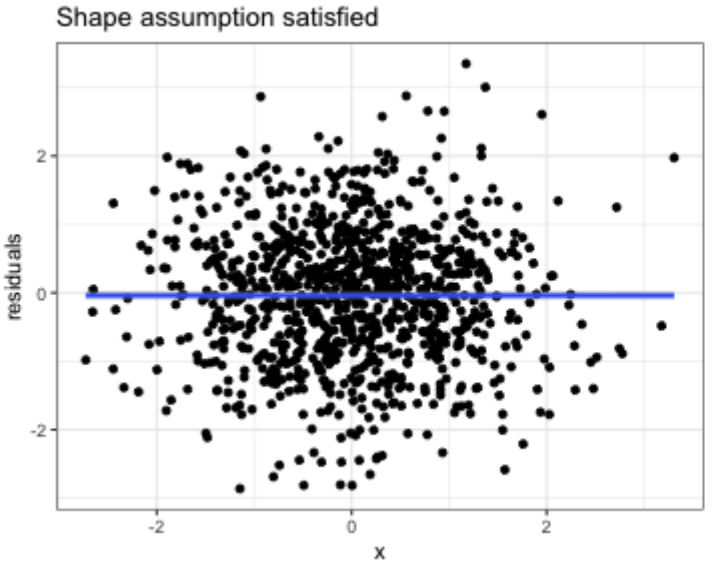
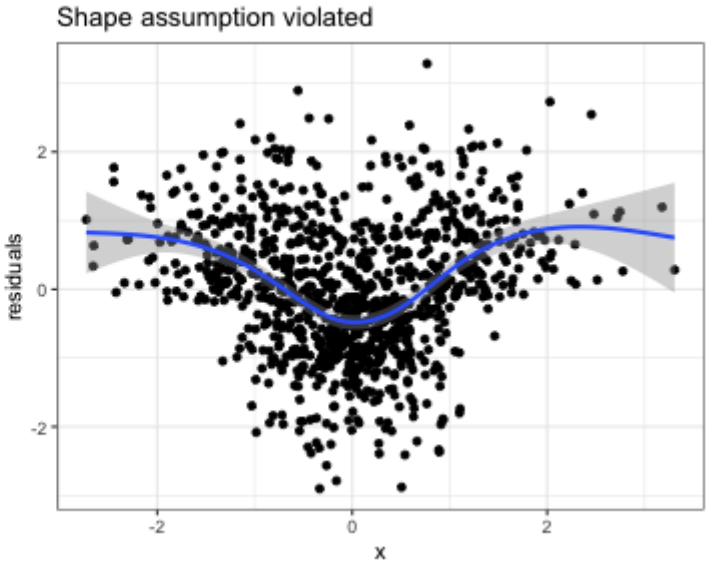
Logistic regression assumptions and diagnostics

- Project 1, Part 2 released on course website
  - Only research question 1
  - Fit a model, perform diagnostics, test hypotheses

Recap: quantile residual plots





- no pattern to residuals
- random scatter around 0

- pattern to the residuals (here it looks like ∪)

Multicollinearity

What is multicollinearity?

Definition: Multicollinearity occurs when one explanatory variable can be approximated by a linear combination of other explanatory variables in the data

Example:

$$Y_i \sim \text{Bernoulli}(\pi_i)$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}$$

worst case: $X_{i1} = \alpha_2 X_{i2} + \alpha_3 X_{i3}$

$$\Rightarrow \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + (\beta_1 \alpha_2 + \beta_2) X_{i2} + (\beta_1 \alpha_3 + \beta_3) X_{i3}$$

Too many unknowns (βs), can't estimate coefficients

higher multicollinearity $\Rightarrow$ more problems with estimation

Class activity, Part I

https://sta214-s23.github.io/class_activities/ca_lecture_11.html

- Simulate correlated data
- Assess the impact on estimated coefficients

The impact of multicollinearity

> How does correlation between the explanatory variables impact the fitted model?

- when we have perfect multicollinearity, can't estimate some coefficient
- increased standard errors for $\hat{\beta}$ as multicollinearity increases
  => less power for hypothesis
- difficult to interpret $\hat{\beta}$

Example: College scorecard data

The Department of Education compiles the College Scorecard, which is used to help prospective college students compare schools. For each school, variables include:

- CONTROL: whether the school is public or private
- SATVRMID: midpoint of SAT critical reading scores of students attending the school
- ACTCMMID: midpoint of the ACT cumulative scores
- UGDS: number of undergraduate students at the school
- NPT4: average cost to attend the school
- PCTFLOAN: fraction of undergraduates receiving a federal student loan
- MD_EARN_WNE_P10: median salary of students 10 years after graduation

test scores are probably highly correlated

Example: College scorecard data

- CONTROL: whether the school is public or private
- SATVRMID: midpoint of SAT critical reading scores of students attending the school
- ACTCMMID: midpoint of the ACT cumulative scores
- UGDS: number of undergraduate students at the school
- NPT4: average cost to attend the school
- PCTFLOAN: fraction of undergraduates receiving a federal student loan
- MD_EARN_WNE_P10: median salary of students 10 years after graduation

> Which of these variables may suffer from multicollinearity?

Diagnosing multicollinearity

> How do I detect multicollinearity in my data?

- correlation matrix (shows correlation for each pair of quantitative explanatory variables)
- Scatterplot matrix (scatterplot of each pair of explanatory variables)
- VIF (variance inflation factor)
  - can look @ more than two variables at a time

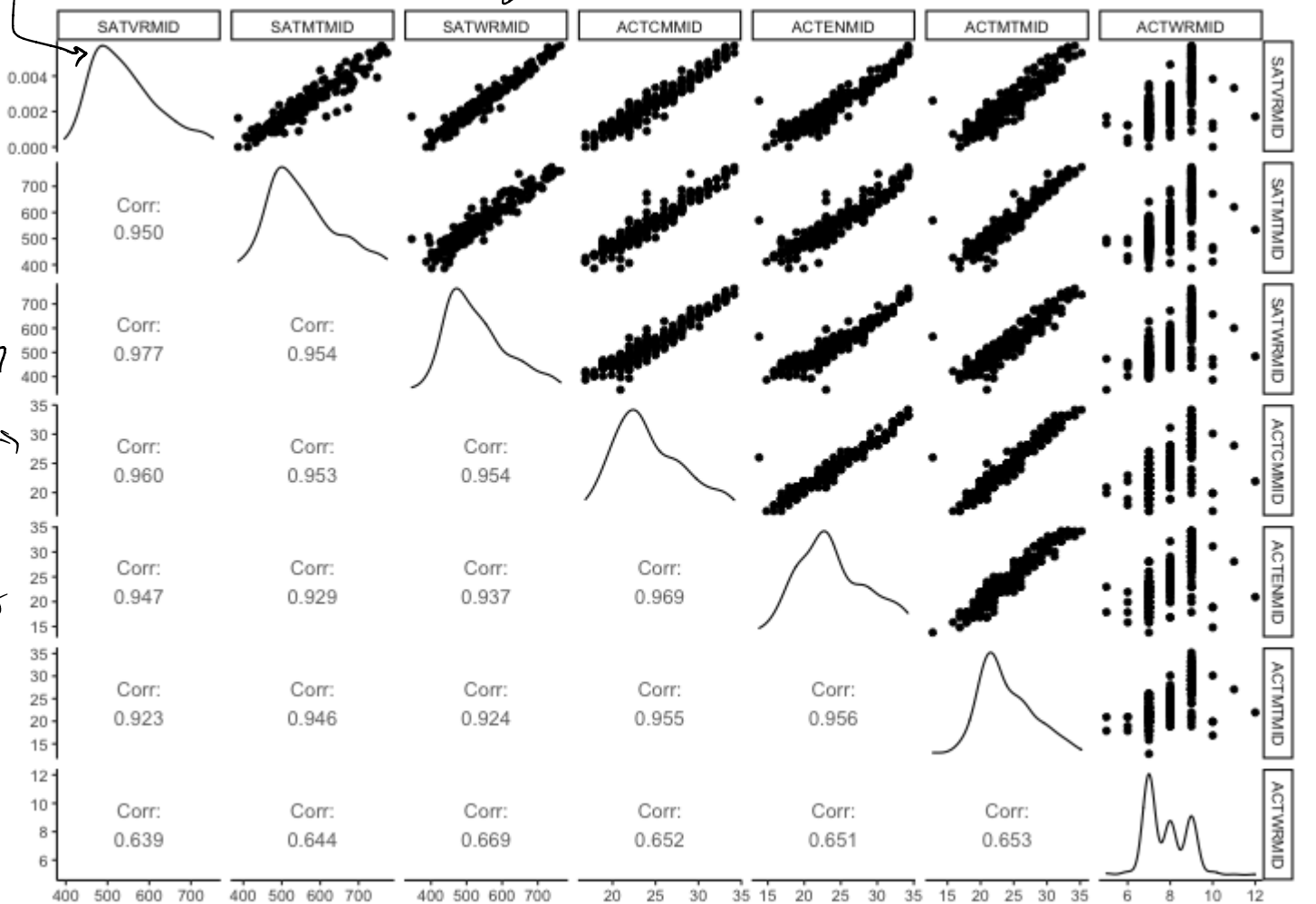
Scatterplot and correlation matrix

(pairs)

distribution of each variable

scatterplots for each pair of variables

correlation between each pair of variables



Variance inflation factors

$$VIF_j = \frac{\text{Var}(\hat{\beta}_j) \text{ using all explanatory variables in model}}{\text{Var}(\hat{\beta}_j) \text{ using only } j^{th} \text{ explanatory variable}}$$

$VIF_j \geq 1$

$VIF_j \approx 1$ if variable j is unrelated to other variables

$$VIF_j = \frac{1}{1 - R_j^2}$$

$R_j^2 = R^2$ for linear regression of j^{th} explanatory variable on other explanatory variables

High VIF $\Rightarrow$ high multicollinearity

Thresholds: usually concerned if VIF > threshold (e.g. 5 or 10)

Variance inflation factors in R

```
library(car)

m1 <- glm(RPY_3YR_70 ~ CONTROL + SATVRMID +
            SATMTMID + SATWRMID + ACTCMMID + ACTENMID +
            ACTMTMID + ACTWRMID + UGDS +
            PCTFLOAN + MD_EARN_WNE_P10 + NPT4,
          data = scorecard, family = binomial)
vif(m1)
```

```
##          CONTROL        SATVRMID        SATMTMID        SATWRMID             ACT
##         3.511039        14.752767        11.246146        13.763868        11.6
##         ACTENMID        ACTMTMID        ACTWRMID             UGDS             PCT
##        12.258720         8.837329         1.671265         2.426494         2.0
## MD_EARN_WNE_P10             NPT4
##         1.333261         2.536284
```

Almost all test score variables have high multicollinearity!

Addressing model issues

> How should we handle multicollinearity in the explanatory variables? Discuss with a neighbor for a few minutes, then we will discuss as a group.

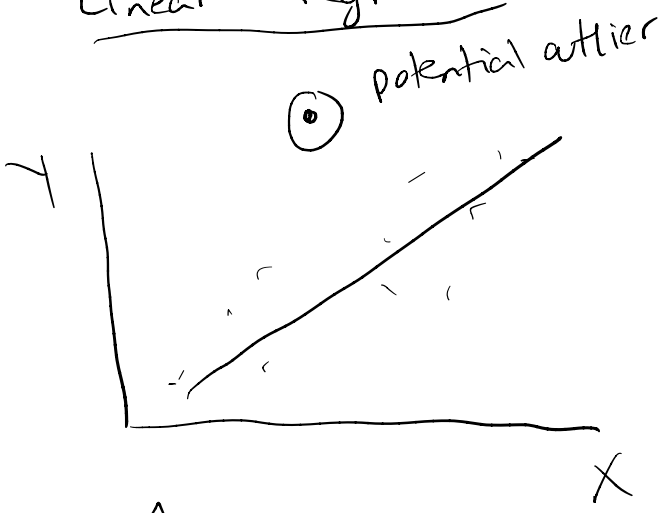
- remove some variables (e.g., choose one test score to represent all test scores)
- combine some variables (e.g. weight & height => BMI)
- ignore! (if we only care about predictions)
- add penalty term

Outliers and influential points

What is an outlier?

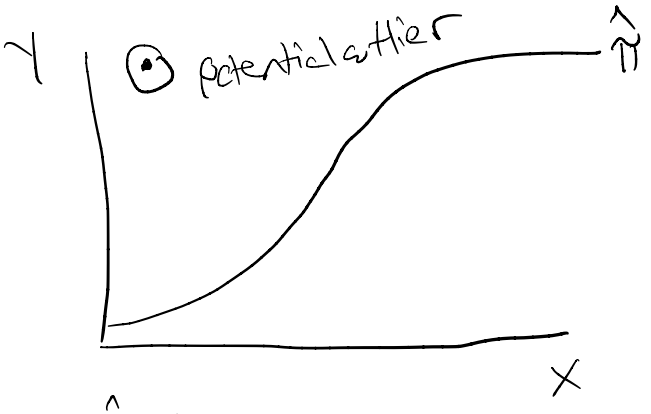
Outlier: A point which doesn't follow the same trends as the rest of the data

Linear regression:



$y_i - \hat{y}_i$ is large

Logistic regression:



$\hat{\pi}_i$ is close to 0 but $y_i = 1$

or $\hat{\pi}_i$ is close to 1 but $y_i = 0$

Class activity, Part II

https://sta214-s23.github.io/class_activities/ca_lecture_11.html

- Simulate data with a potential outlier
- Assess the impact on estimated coefficients

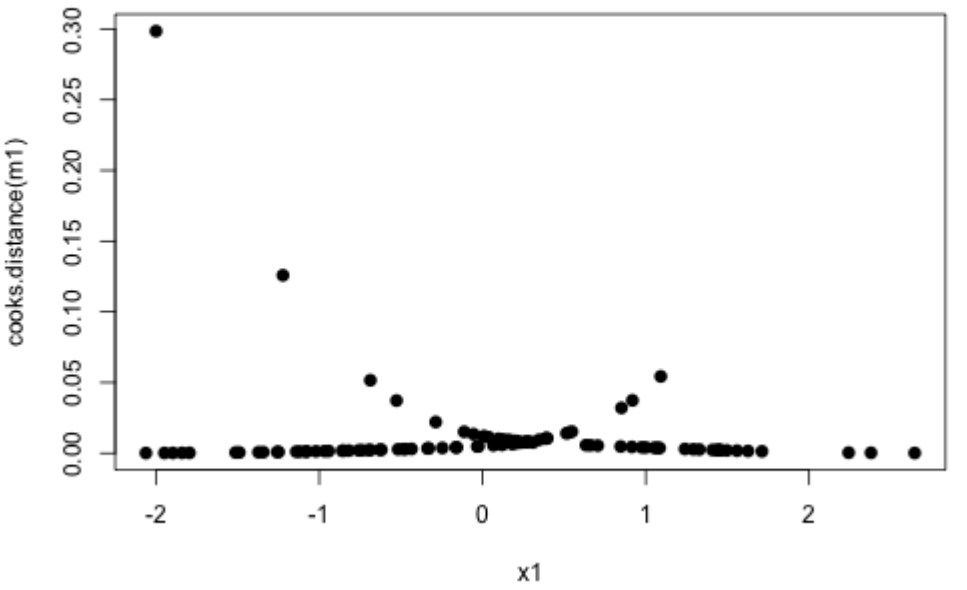
Class activity

> How does an outlier influence the fitted regression model?

Cook's distance

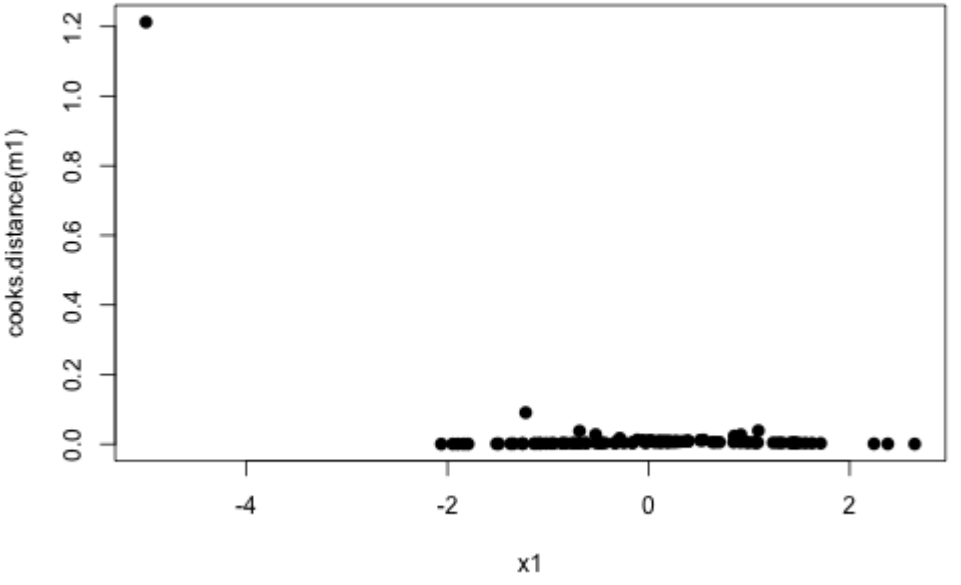
Cook's distance in R

```
x1 <- c(x, -2)
y1 <- c(y, 1)
m1 <- glm(y1 ~ x1, family = binomial)

plot(x1, cooks.distance(m1))
```



Cook's distance in R

```
x1 <- c(x, -5)
y1 <- c(y, 1)
m1 <- glm(y1 ~ x1, family = binomial)

plot(x1, cooks.distance(m1))
```



Addressing model issues

> How should we handle outliers and influential points? Discuss with a neighbor for a few minutes, then we will discuss as a group.