Logistic regression assumptions and diagnostics

Motivating example: Dengue data

Data: Data on 5720 Vietnamese children, admitted to the hospital with possible dengue fever. Variables include:

- *Sex*: patient's sex (female or male)
- *Age*: patient's age (in years)
- *WBC*: white blood cell count
- *PLT*: platelet count
- other diagnostic variables...
- *Dengue*: whether the patient has dengue (0 = no, 1 = yes)

Previously: Logistic regression model

$$Y_i = \text{dengue status } (0 = \text{negative}, 1 = \text{positive})$$

$$Y_i \sim Bernoulli(p_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 WBC_i$$

> What assumptions does this logistic regression model make? How should we assess these assumptions? Discuss with your neighbor for 2--3 minutes, then we will discuss as a group.

Assumptions

- **Independence**: All Y_i's are independent (needed for MLE)
- **Shape**: the specified shape of the systematic component is correct (e.g., log odds really are a linear function of WBC)
- **No perfect collinearity** (check multicollinearity in our explanatory variables)
- **Lack of outliers**: all responses generated from the same process

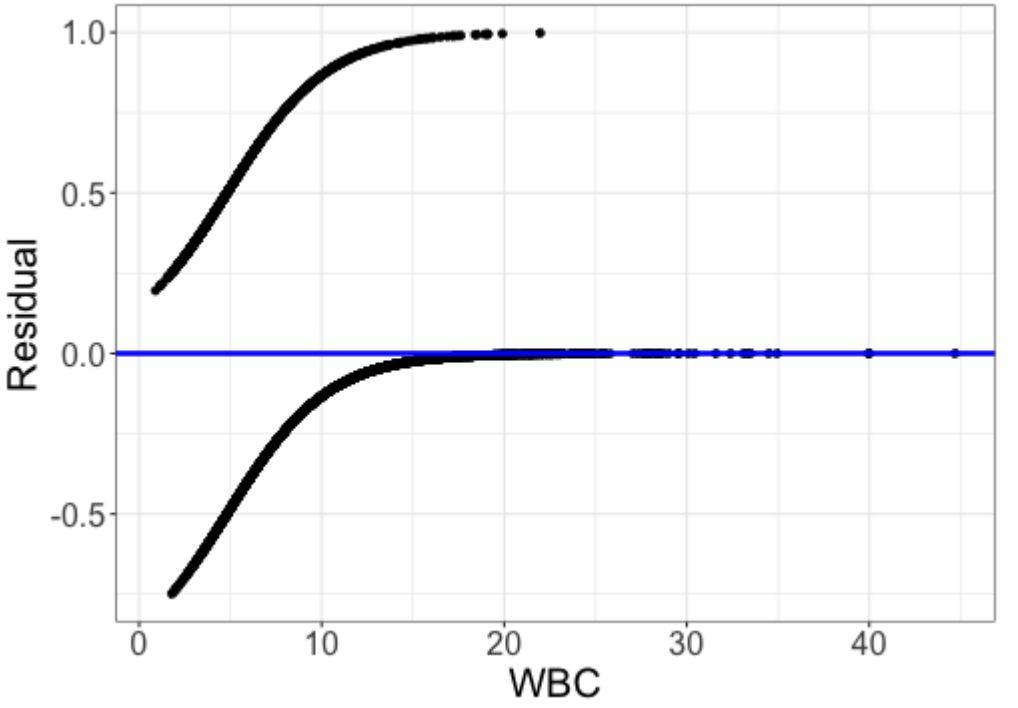
Assessing

- **Independence**: think about data generating process
- **Shape**: some kind of plot (like a residual plot) ← today
- **outliers**: leverage & cook's distance ← next time
- **collinearity**: correlation plots, variance inflation factor ← next time

Don't use usual residuals for logistic regression

Fitted model: $\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = 1.737 - 0.361\, WBC_i$

Residuals $Y_i - \hat{p}_i$:



Assessing shape with empirical logit plots

Example: Putting data. Interested in the relationship between the length of a putt, and whether it was made:

$$Y_i \sim Bernoulli(p_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1\, Length_i$$

Length	3	4	5	6	7
Number of successes	84	88	61	61	44
Number of failures	17	31	47	64	90
Total	101	119	108	125	134

Idea: Estimate $\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$, plot against Length!

Empirical logits

Step 1: estimate the probability of success for each length of putt

Length	3	4	5	6	7
Number of successes	84	88	61	61	44
Number of failures	17	31	47	64	90
Total	101	119	108	125	134
Probability of success $\hat{p}$	0.832	0.739	0.565	0.488	0.328

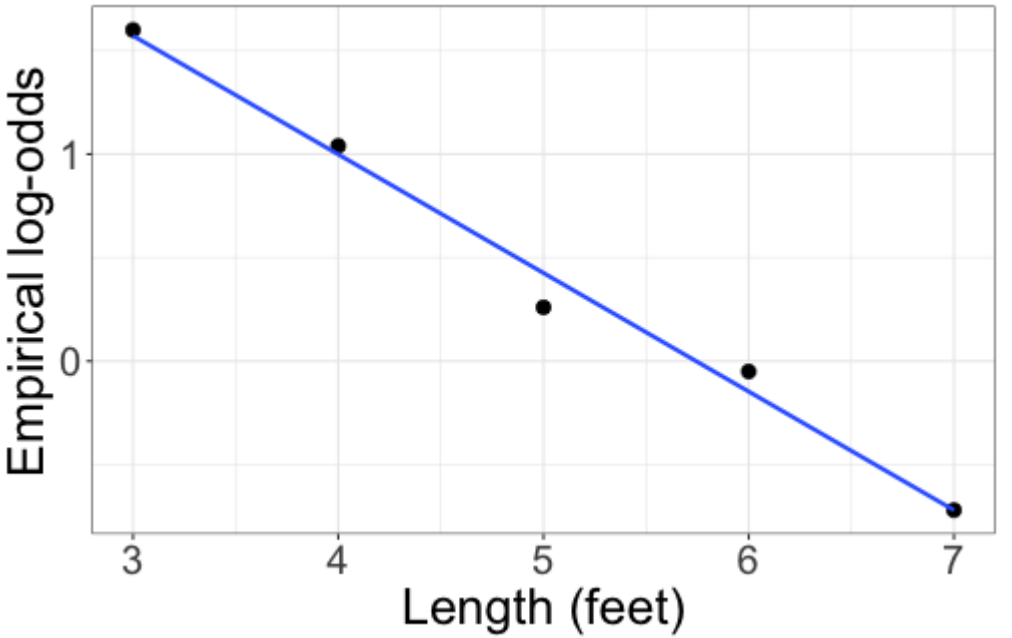
Empirical logits

Step 2: convert empirical probabilities to empirical log odds

Length	3	4	5	6	7
Number of successes	84	88	61	61	44
Number of failures	17	31	47	64	90
Total	101	119	108	125	134
Probability of success $\hat{p}$	0.832	0.739	0.565	0.488	0.328
Odds $\frac{\hat{p}}{1-\hat{p}}$	4.941	2.839	1.298	0.953	0.489
Log-odds $\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$	1.60	1.04	0.26	-0.05	-0.72

Empirical logits

Step 3: plot empirical log-odds against predictor, and add a least-squares line



Looks pretty good!

Does it seem reasonable that the log-odds are a linear function of length?

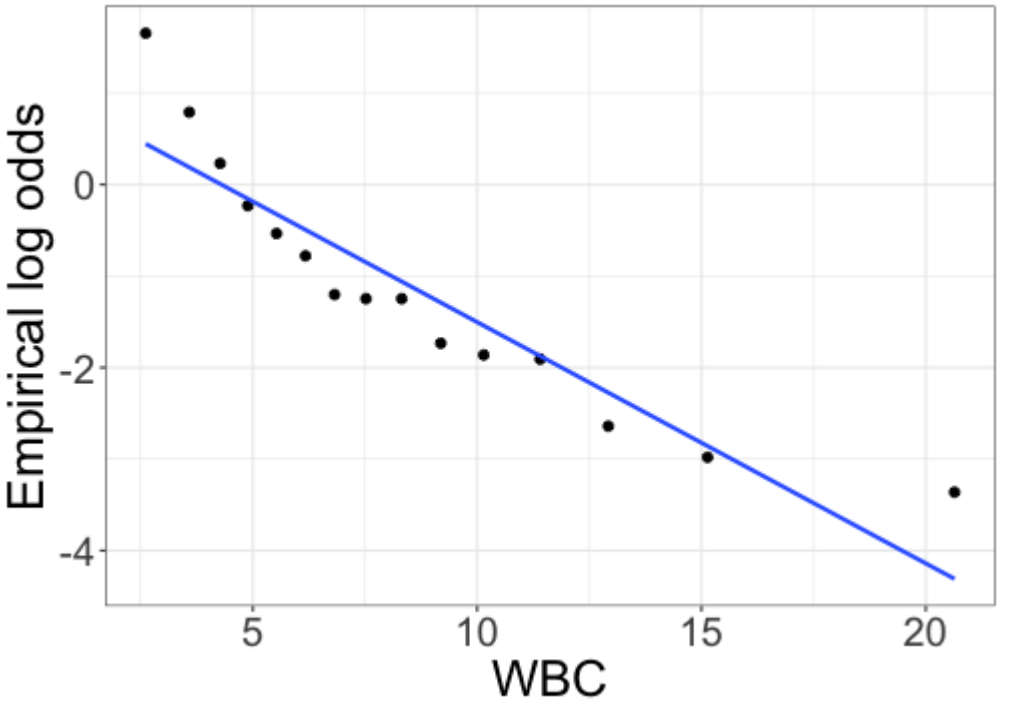
Back to the dengue data...

WBC	0.90	1.15	1.23	1.25	1.54	1.58	...
Dengue = 0	0	0	0	0	0	0	...
Dengue = 1	1	2	1	1	3	1	...

> What problem do I run into?

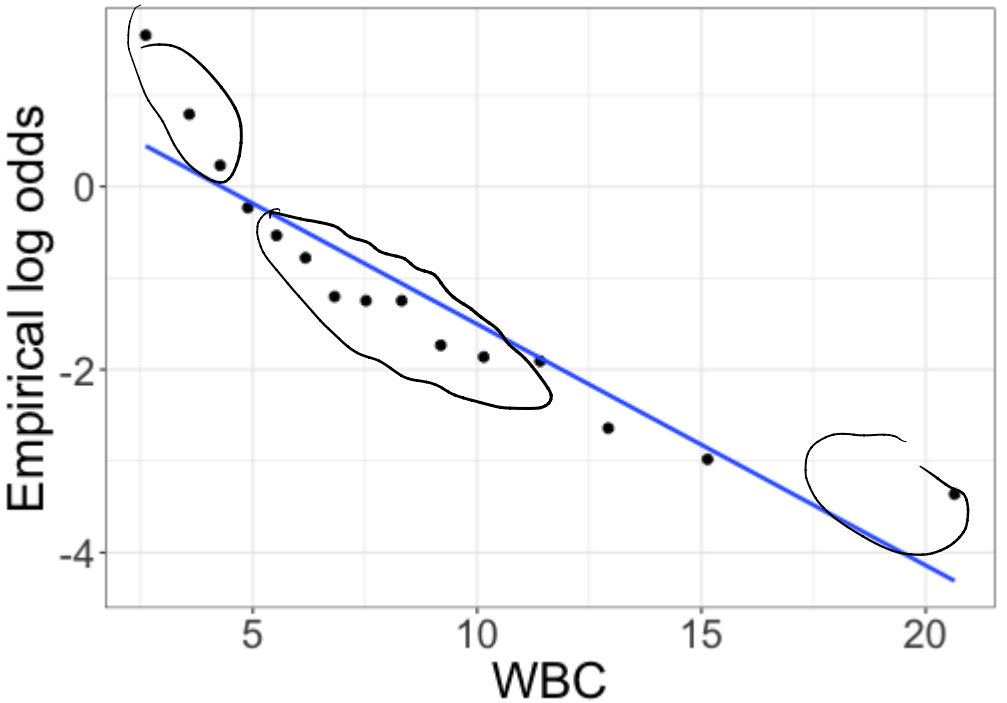
Too few observations @ each WBC to estimate log odds!

Binned empirical logit plots



1) Specify # bins (usually want at least 8-10 bins)
2) Divide data into bins based on WBC
3) Calculate empirical log odds in each bin, & average WBC in each bin
4) Plot!

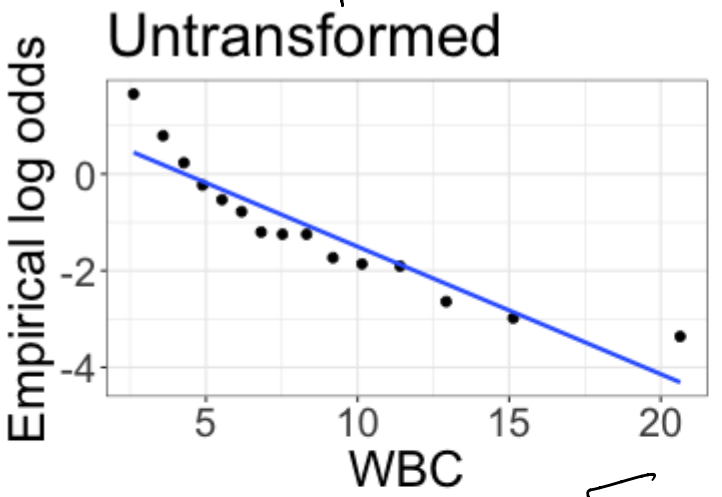
Binned empirical logit plots



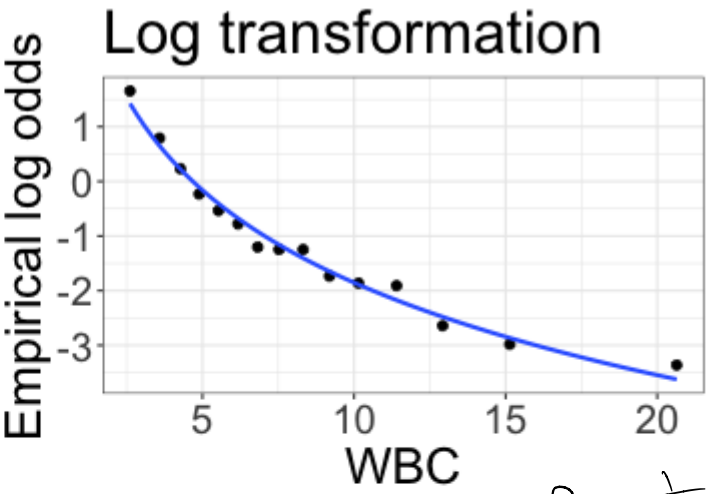
Does it seem reasonable that the log-odds are a linear function of WBC?

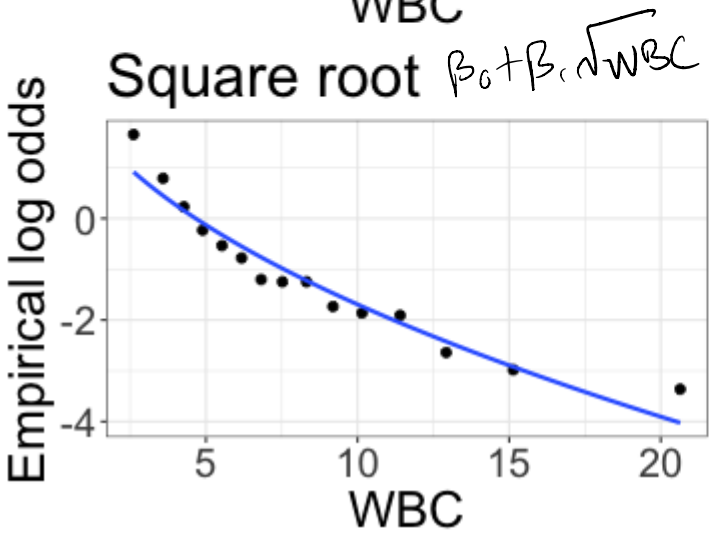
Trying some transformations

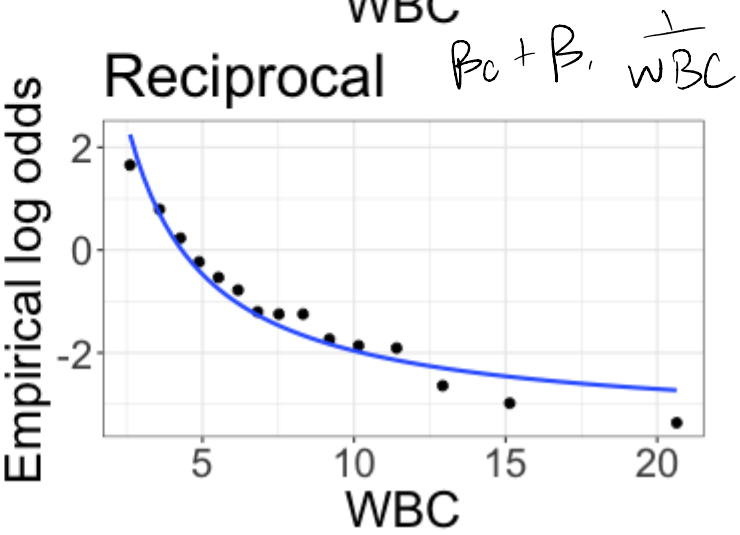
$\beta_0 + \beta_1 WBC$



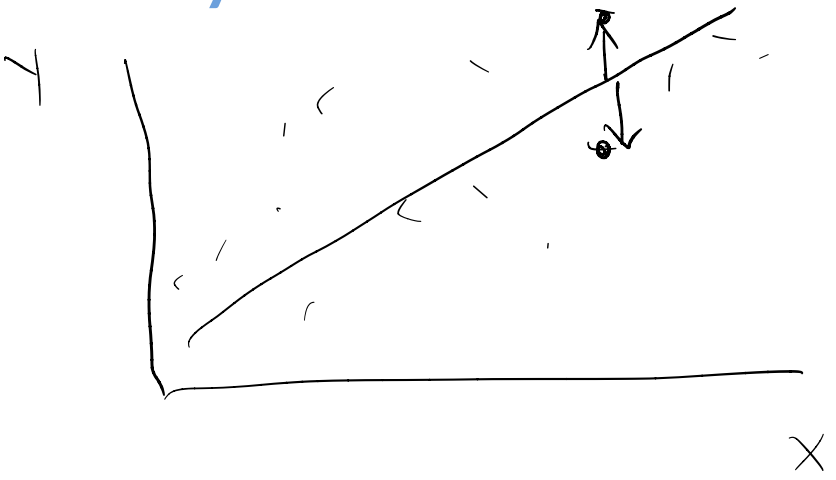
$\beta_0 + \beta_1 \log WBC$







Why residuals in linear regression are nice



$r_i = Y_i - \hat{Y}_i$

$r_i > 0 \Rightarrow$ underestimate

$r_i < 0 \Rightarrow$ overestimate

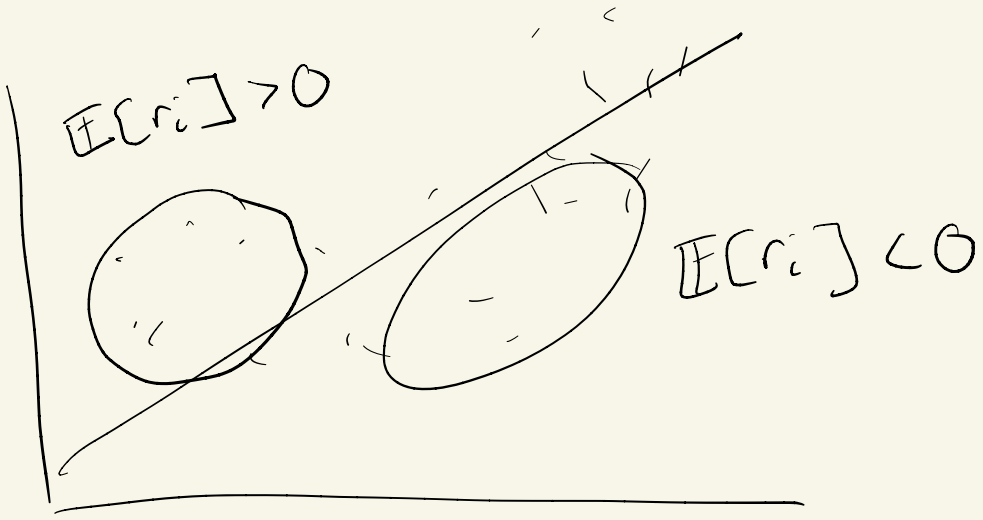
want: for each x_i,

$\mathbb{E}[r_i | x_i] \approx 0$

<u>If</u> line is a good fit, $\mathbb{E}[r_i | x_i] \approx 0 \ \forall i$

residual plot

random scatter



logistic regression:

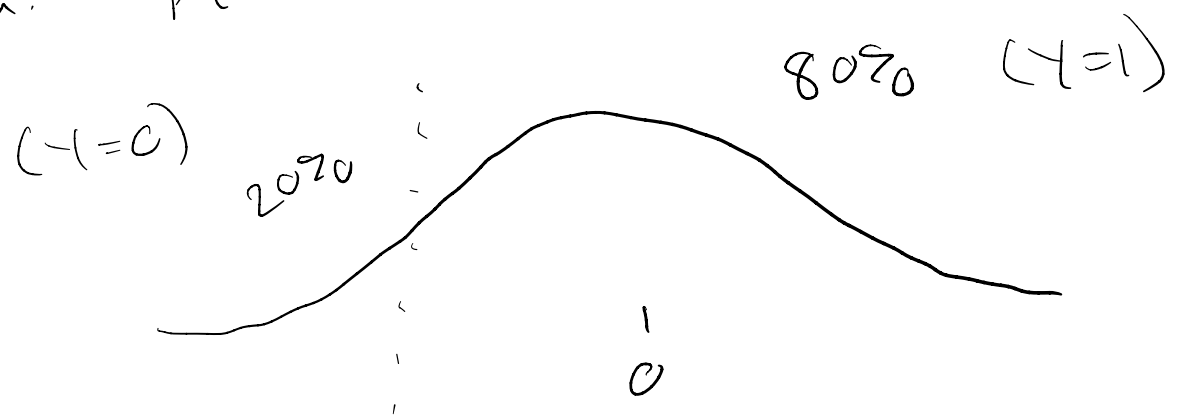
Quantile residuals for logistic regression

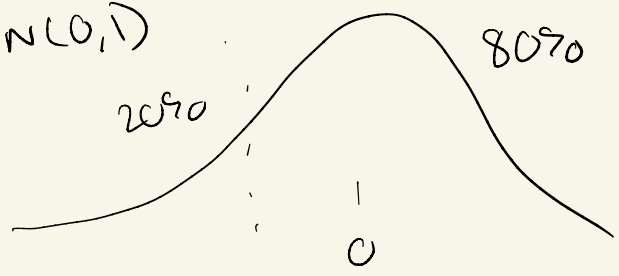
Motivation: Suppose $\hat{p}_i = 0.8$

Want: define residual r_Q st

- If $\hat{p}_i \approx p_i \Rightarrow \mathbb{E}[r_Q | x_i] \approx 0$
- If $\hat{p}_i > p_i \Rightarrow \mathbb{E}[r_Q | x_i] < 0$
- If $\hat{p}_i < p_i \Rightarrow \mathbb{E}[r_Q | x_i] > 0$

Idea: $\hat{p}_i = 0.8$. Divide standard Normal into 2 regions:





1) Sample $Y_i \sim \text{Bernoulli}(p_i)$

2) If $Y_i = 0$, sample r_Q from left side

If $Y_i = 1$, sample r_Q from right side

If $\hat{p}_i \approx p_i$, then $r_Q \sim N(0,1)$

$\hat{p}_i < p_i$, then we sample from right side more often than we expect

$\Rightarrow \mathbb{E}[r_Q] > 0$

$\hat{p}_i > p_i$, " left side

more often

$\Rightarrow \mathbb{E}[r_Q] < 0$

observe $(X_1, Y_1), \dots, (X_n, Y_n)$

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

$\log\left(\frac{p_i}{1-p_i}\right) = \dots$

Quantile residuals example (in R)

```
x <- rnorm(1000)

p <- exp(-1+2*x)/(1+exp(-1+2*x))

plot(x,p)

y <- rbinom(1000, 1, p)

m1 <- glm(y ~ x, family = binomial)

data.frame(x = x, residuals = qresid(m1)) %>%
    ggplot(aes(x = x, y = residuals)) +
    geom_point() +
    geom_smooth() +
    theme_bw()
```

Class activity, Part I

https://sta712-f22.github.io/class_activities/ca_lecture_5.html

Class activity, Part II

https://sta712-f22.github.io/class_activities/ca_lecture_5.html