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

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

$$Y_i \sim Bernoulli(\pi_i)$$

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

> 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.

Logistic regression assumptions

- response variable is binary (could also generalize to binomial)
- need the relevant explanatory variables in model to estimate relationship
- random sample or data come from random process
- no outliers / weird observations — all observations come from the same process
- Shape: linear relationship between log odds & explanatory variables
- Independence (observations are independent) (needed for MLE)

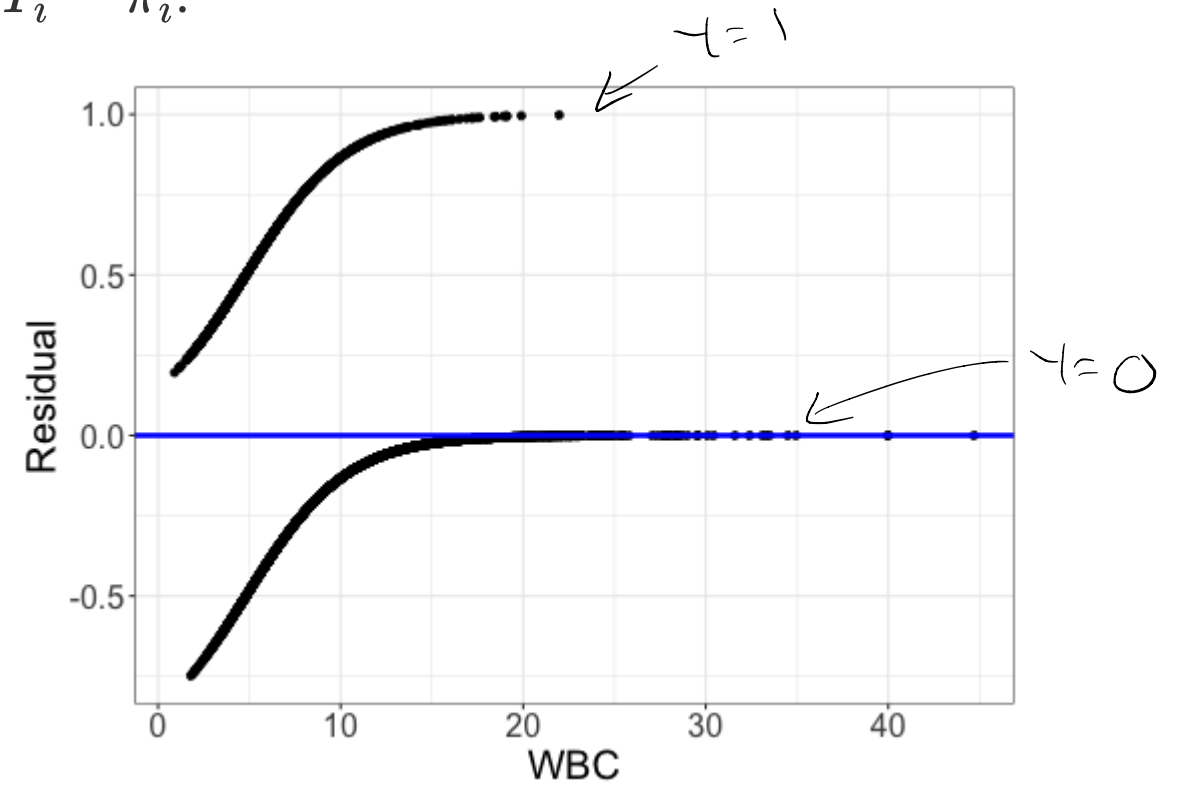
Assess

- response: check whether binary
- EDA to explore explanatory
- randomness & independence: think about data generating process
- shape: empirical logit plots, quantile residuals (today)
- check multicollinearity (next week)

Don't use raw residuals for logistic regression

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

Residuals $Y_i - \widehat{\pi}_i$:



Assessing shape with empirical logit plots

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

$$Y_i \sim Bernoulli(\pi_i)$$

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

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 $\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)$ for each length and plot against length, check if linear

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 $\widehat{\pi}$	0.832	0.739	0.565	0.488	0.328

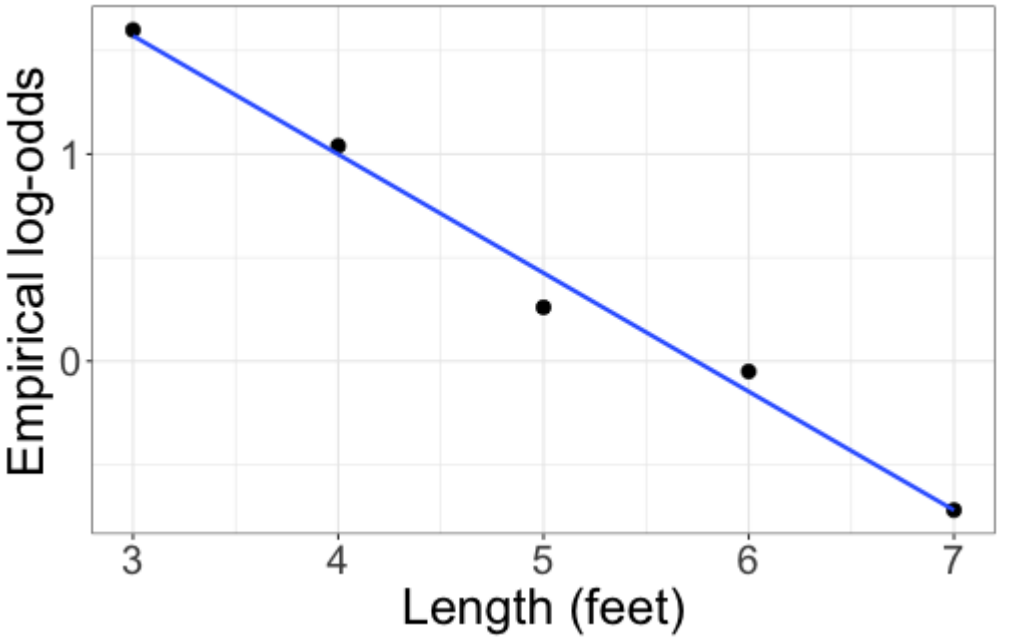
Empirical logits

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 $\widehat{\pi}$	0.832	0.739	0.565	0.488	0.328
Odds $\frac{\widehat{\pi}}{1-\widehat{\pi}}$	4.941	2.839	1.298	0.953	0.489
Log-odds $\log\left(\frac{\widehat{\pi}}{1-\widehat{\pi}}\right)$	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



linearity looks pretty good!

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

Back to the dengue data...

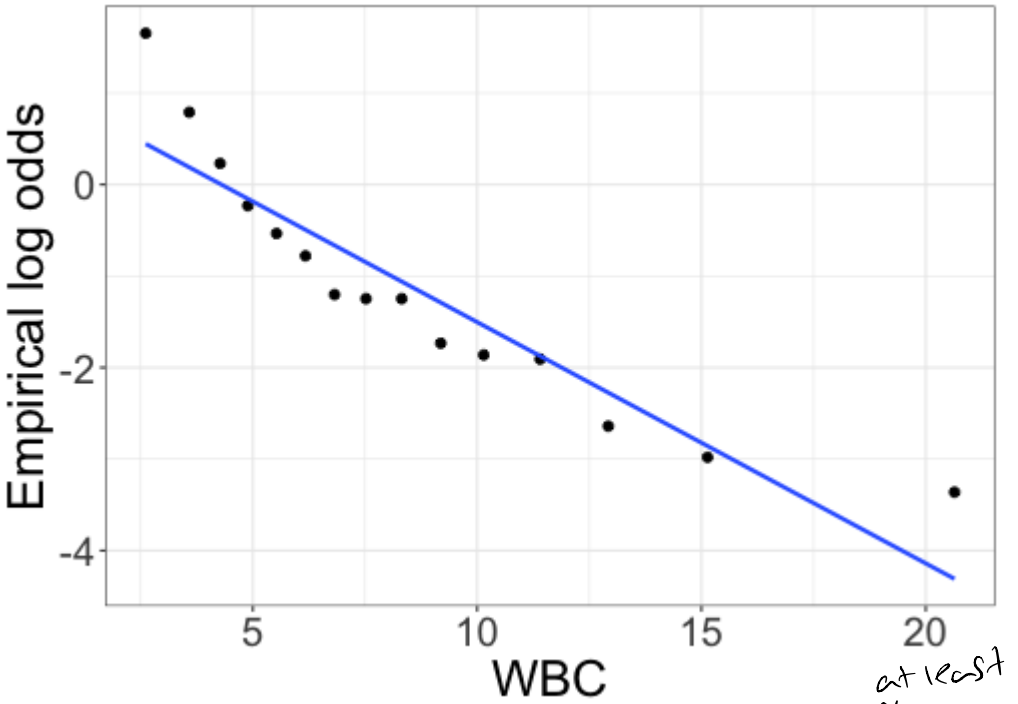
WBC	0.90	1.15	1.23	1.25	1.54	1.58	...
Dengue = 0	0	0	0	0	0	0	...
Dengue = 1	1	2	1	1	3	1	...

> What problem do I run into?

- Many different values of WBC, few observations at each value

=> $\log(1)$ $\log(0)$ $\log(\frac{1}{2})$ $\log(1)$

Binned empirical logit plots

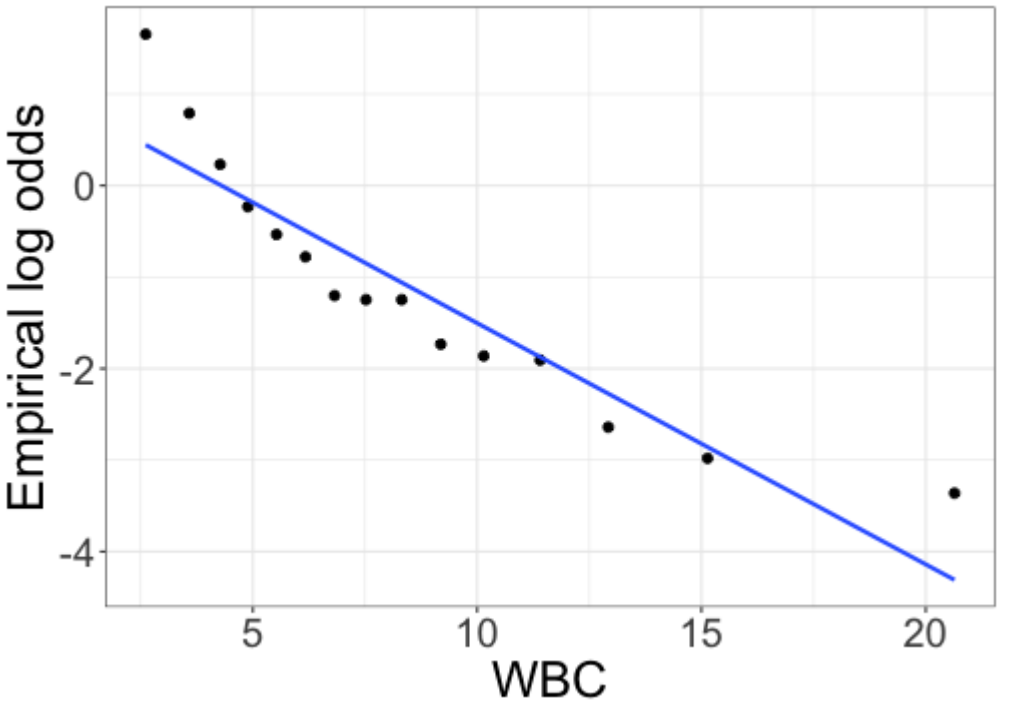


1) Specify number of bins (usually want at least 8-10, but depends on data size)

2) Divide data into the different bins based on WBC (e.g. WBC $\in [0,1]$, WBC $\in [1,2]$, WBC $\in [2,5]$)

3) Calculate empirical log odds in each bin

Binned empirical logit plots



pretty close to linear

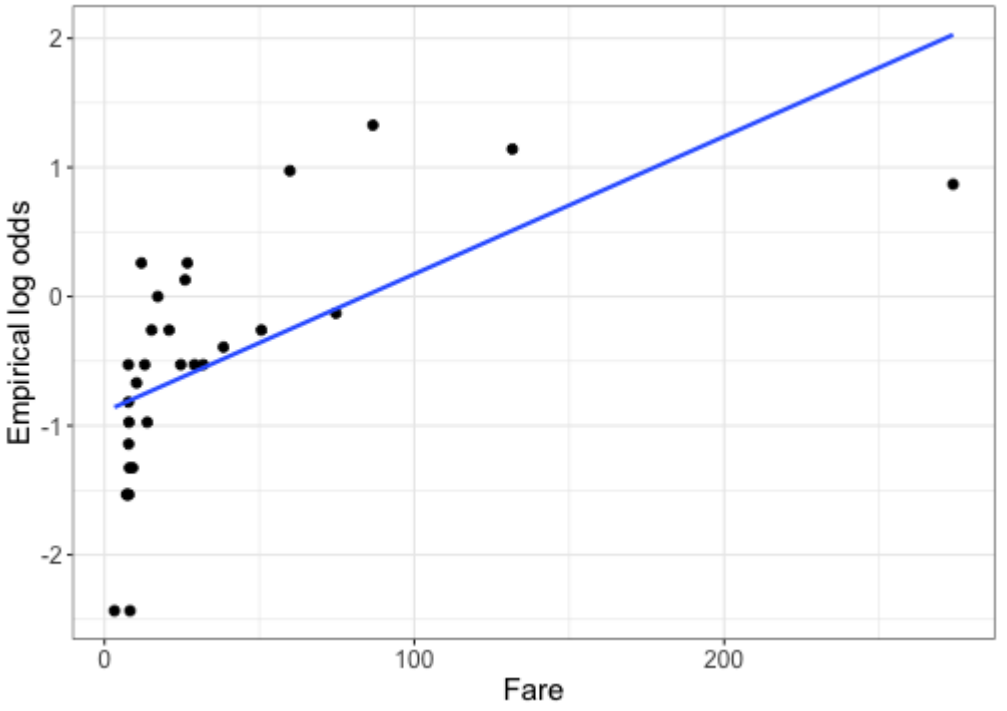
Maybe a little nonlinear

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

Class activity, Part I

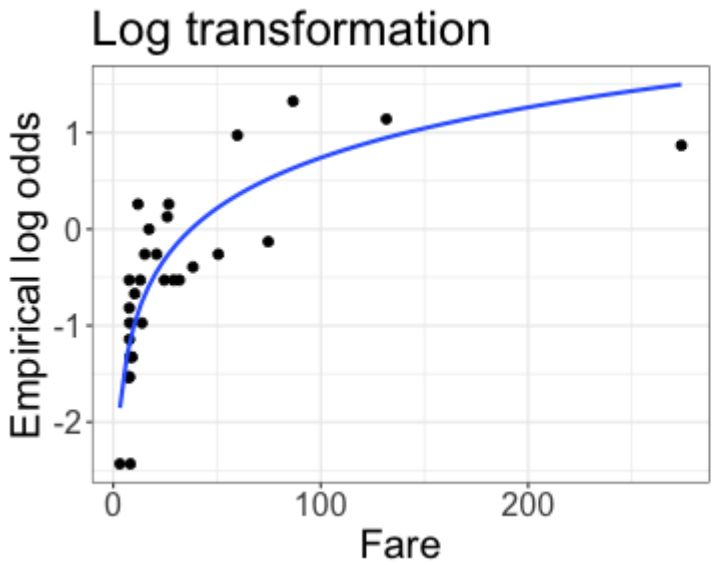
https://sta214-s23.github.io/class_activities/ca_lecture_10.html

Class activity

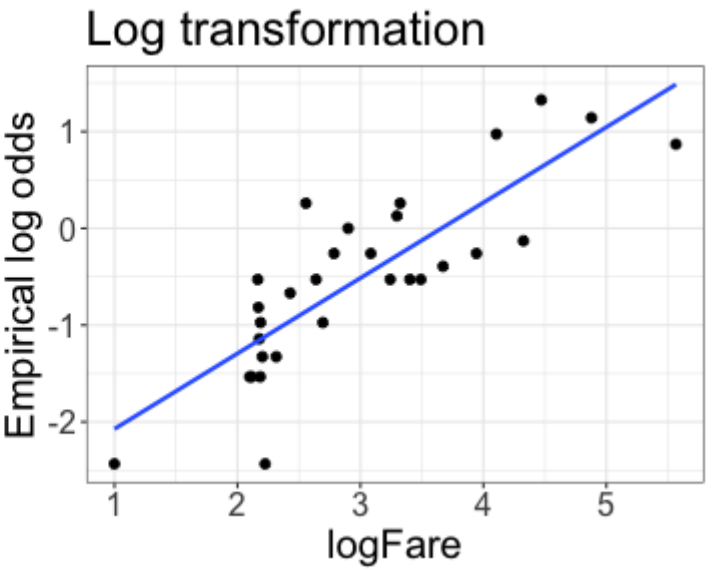


Class activity

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \log(\text{Fare}_i)$$

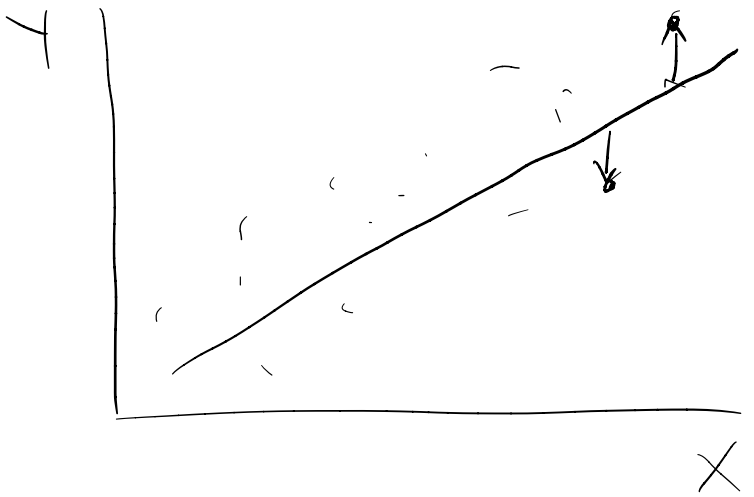


plot against Fare



plot against log(Fare)

Why residuals in linear regression are nice



$r_i = y_i - \hat{y}_i$

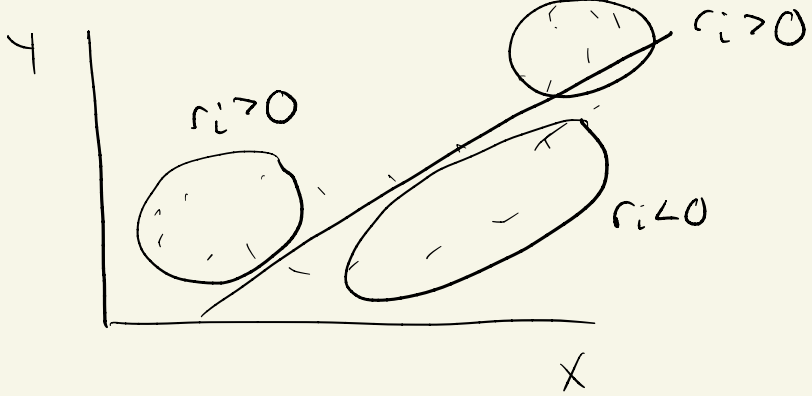
$r_i > 0 \Rightarrow$ underestimate

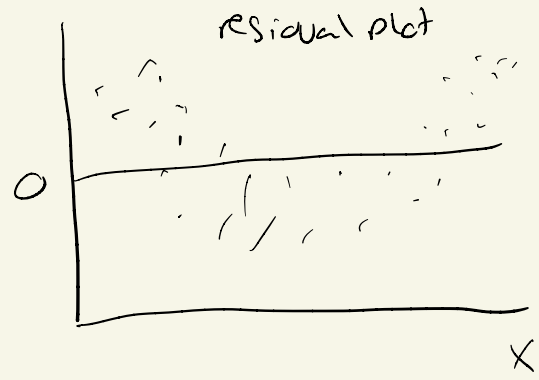
$r_i < 0 \Rightarrow$ overestimate

want $r_i \approx 0$ on average

If line is a good fit, r_i is scattered around 0 for all X_i

residual plot:





pattern!

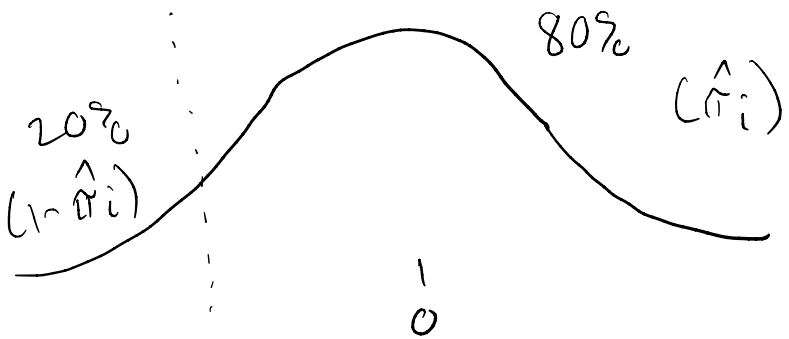
(randomized)

Quantile residuals for logistic regression

Motivation: Suppose $\hat{\pi}_i = 0.8$. we want: a residual r_Q such that

- if $\hat{\pi}_i \approx \pi_i$ (good estimate), then $r_Q \approx 0$ (on average)
- if $\hat{\pi}_i > \pi_i$ (overestimate) $r_Q < 0$ (on average)
- if $\hat{\pi}_i < \pi_i$ (underestimate) $r_Q > 0$ (on average)

Idea: $\hat{\pi}_i = 0.8$. Divide $N(0,1)$ into 2 regions



if $Y_i = 1$, simulate r_Q from right side

$Y_i = 0$, simulate r_Q from left side

if $\hat{\pi}_i \approx \pi_i$, then $r_Q \approx N(0,1)$

Class activity, Part II

https://sta214-s23.github.io/class_activities/ca_lecture_10.html