Logistic regression assumptions and diagnostics

Last time: IRLS for logistic

weights, vector of working responses

$$\beta^{(r+1)} = (X^T W^{(r)} X)^{-1} X^T W^{(r)} Z^{(r)}$$

$$Z^{(r)} = X\beta^{(r)} + (W^{(r)})^{-1}(Y - p^{(r)})$$

Initialization:

- $$p_i^{(0)} = \begin{cases} 0.25 & Y_i = 0 \\ 0.75 & Y_i = 1 \end{cases}$$

$\log\left(\frac{p_i}{1-p_i}\right) = \beta^T X_i$

- $$[X\beta^{(0)}]_i = \log\left(\frac{p_i^{(0)}}{1-p_i^{(0)}}\right)$$

- $$[W^{(0)}]_{ii} = p_i^{(0)}(1-p_i^{(0)})$$

Plan going forward

So far: Parameter estimation w/ MLE, fitting logistic regression models (HW 1-3)

Exam 1: tentatively released Friday, Feb. 10
- take home, probably closed notes

Next up:
- Diagnostics for logistic regression
- Properties of MLEs

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

- **Shape**:
  - log odds really are a linear function of the explanatory variables
  - $p_i \in (0,1)$
- **Independence**: Y_i are independent
- **Lack of outliers**: All responses are generated from the same process (same βs work for all observations)
- Binary response

$$Y_i \sim N(\mu_i, \sigma_\varepsilon^2)$$

$$Var[Y_i | X_i] = \sigma_\varepsilon^2$$

Diagnostics

- Some kind of plot? some kind of residuals? (today)
- Think about data generating process
- Leverage & Cook's distance (next time)

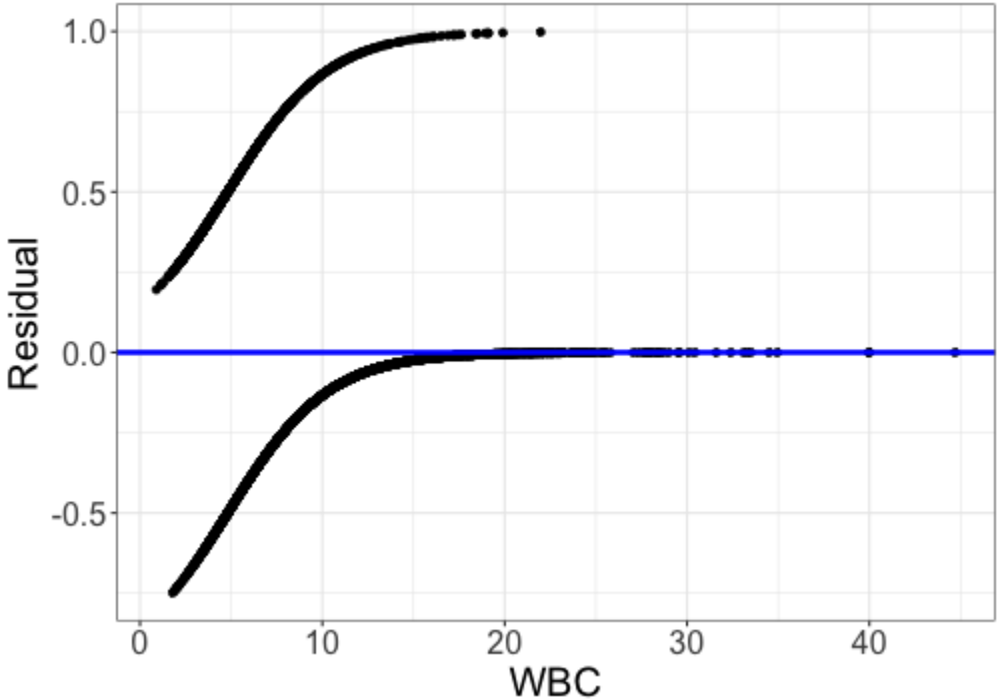
$$Y_i \sim Bernoulli(p_i)$$

$$Var[Y_i | X_i] = p_i(1-p_i)$$

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 $\underbrace{\log\left(\frac{\hat{p}}{1-\hat{p}}\right)}_{\text{empirical logits}}$ and 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

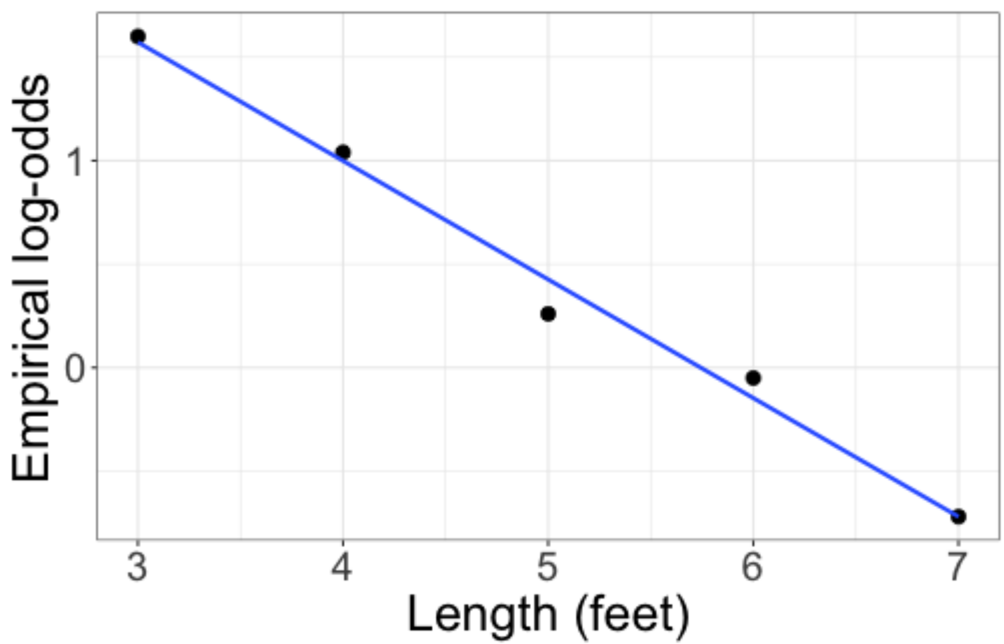
Empirical logits

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



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?

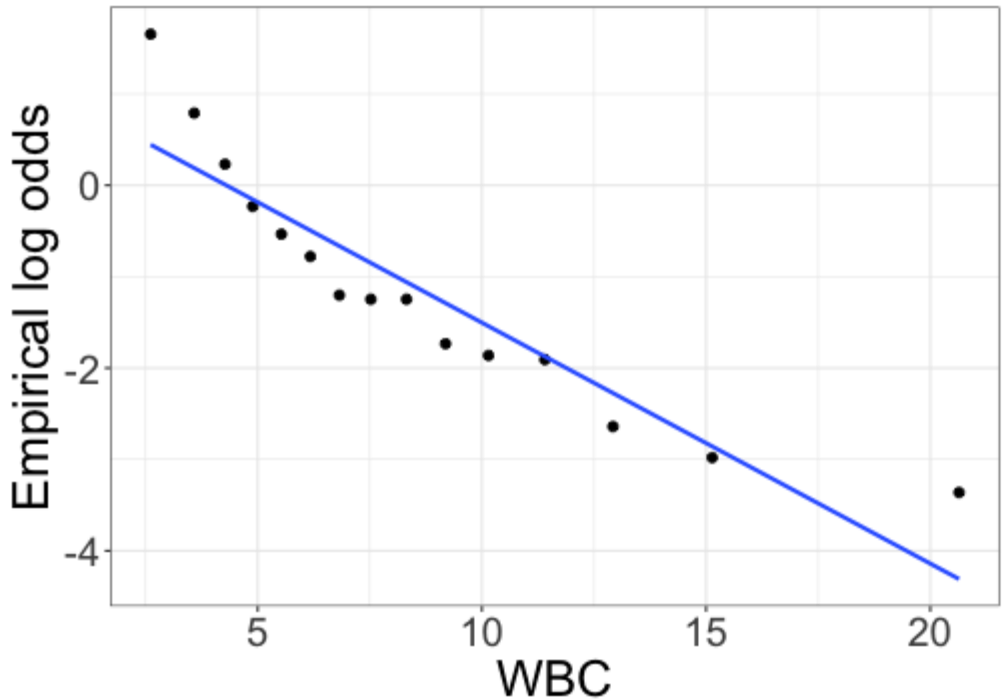
Too few observations @ each WBC to estimate log odds

Categorical variable (hair color, e.g.) $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 Red_i + \beta_2 Blonde_i + \beta_3 Black_i$
(no shape assumption for linearity)

Binned empirical logit plots

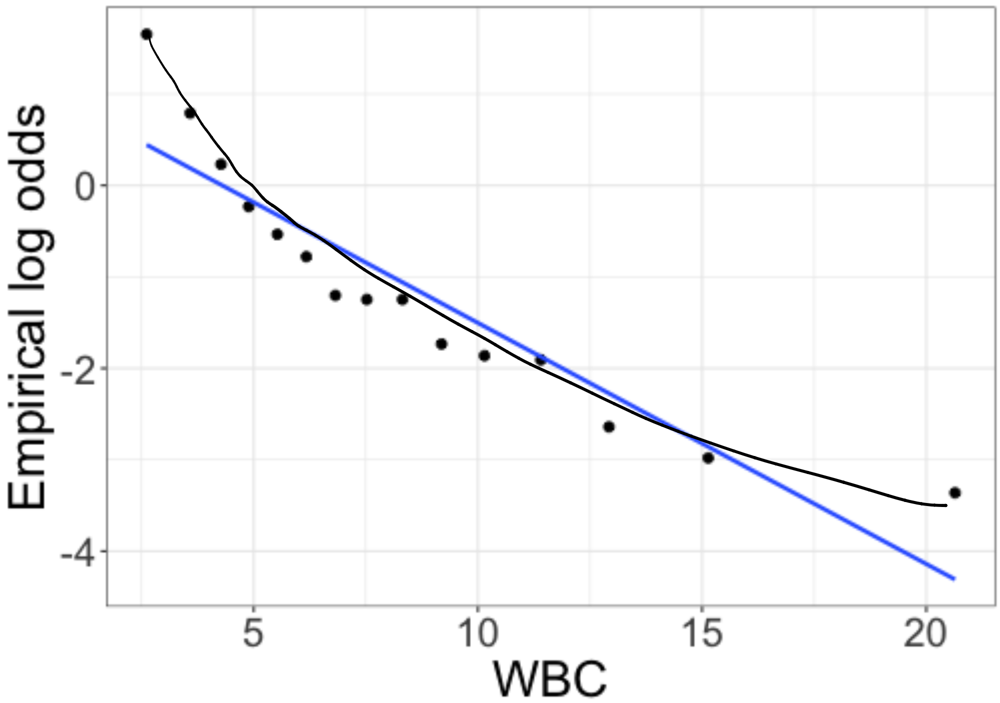
$= 1$ if hair = red

$= 0$ if hair $\neq$ red



1) Specify n_{bins} (usually want at least 8-10, but depends on data size)
2) Divide data into n_{bins} groups based on WBC
3) In each bin, calculate empirical log odds
4) Plot!

Binned empirical logit plots

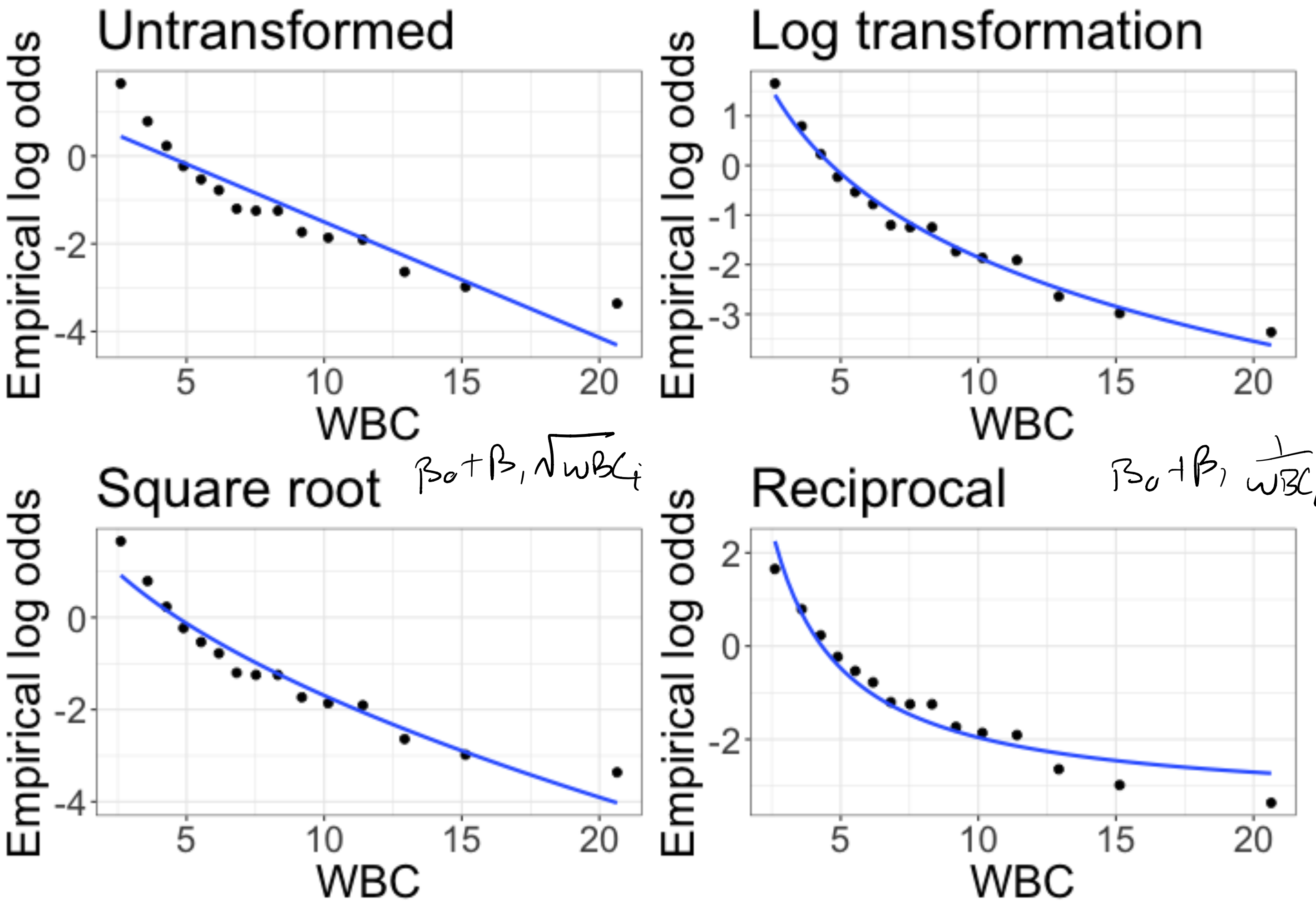


· maybe slightly nonlinear
But a line does a good job

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

Trying some transformations

$\beta_0 + \beta_1 \log WBC_i$



$\beta_0 + \beta_1 \sqrt{WBC_i}$

$\beta_0 + \beta_1 \frac{1}{WBC_i}$

Why residuals in linear regression are nice

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

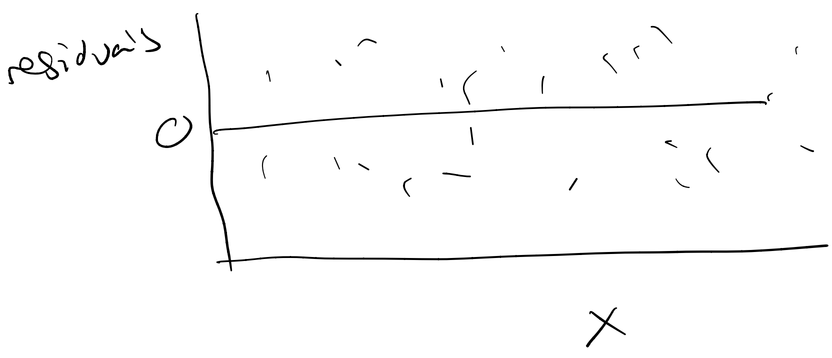
$r_i > 0 \Rightarrow$ underestimate

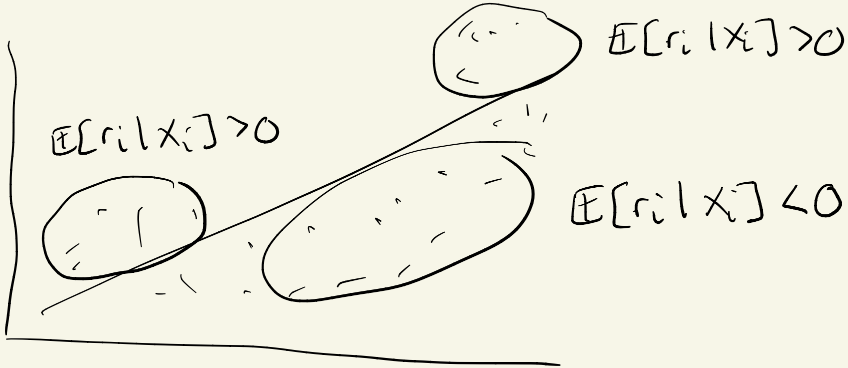
$r_i < 0 \Rightarrow$ overestimate

want $r_i \approx 0$ on average, for each value of X

<u>If</u> the line is a good fit, $\mathbb{E}[r_i \mid X_i] = 0 \quad \forall X_i$

random scatter





pattern!

- patterns in residual plot indicate issues with our model

- residuals are continuous

Quantile residuals for logistic regression

Class activity

https://sta711-s23.github.io/class_activities/ca_lecture_9.html