Empirical log odds plots

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

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

$$Y_i \sim Bernoulli(\overset{\pi_i}{\cancel{p_i}})$$

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

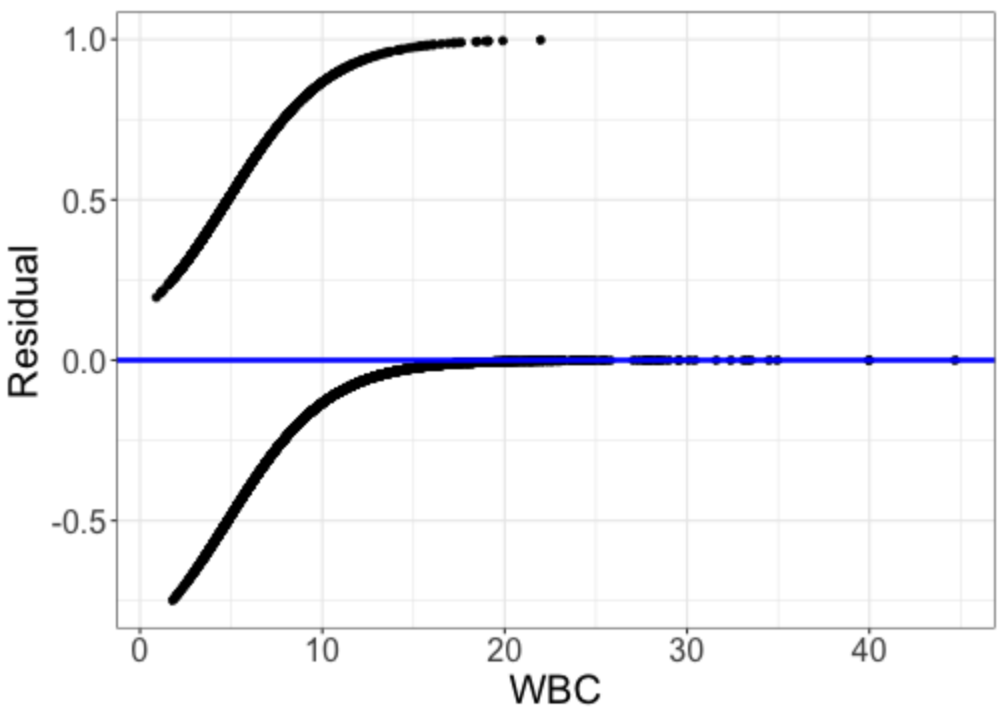
> What assumptions does this logistic regression model make? How should we assess these assumptions? Discuss with your neighbor for 2--3 minutes, then we will discuss as a group.

1) randomness (data come from some random process)
2) log odds are a <u>linear</u> function of WBC (<u>shape assumption</u>)
3) independence (observations Y_i are independent)

Don't use usual residuals for logistic regression

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

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



Assessing shape with empirical log odds plots

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

$$Y_i \sim Bernoulli(\overset{\tilde{\pi}_i}{\cancel{p_i}})$$

$$\log\left(\frac{\overset{\hat{\pi}_i}{\cancel{p_i}}}{1 - \underset{\hat{\pi}_i}{\cancel{p_i}}}\right) = \beta_0 + \beta_1\, Length_i$$

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

Empirical log odds

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

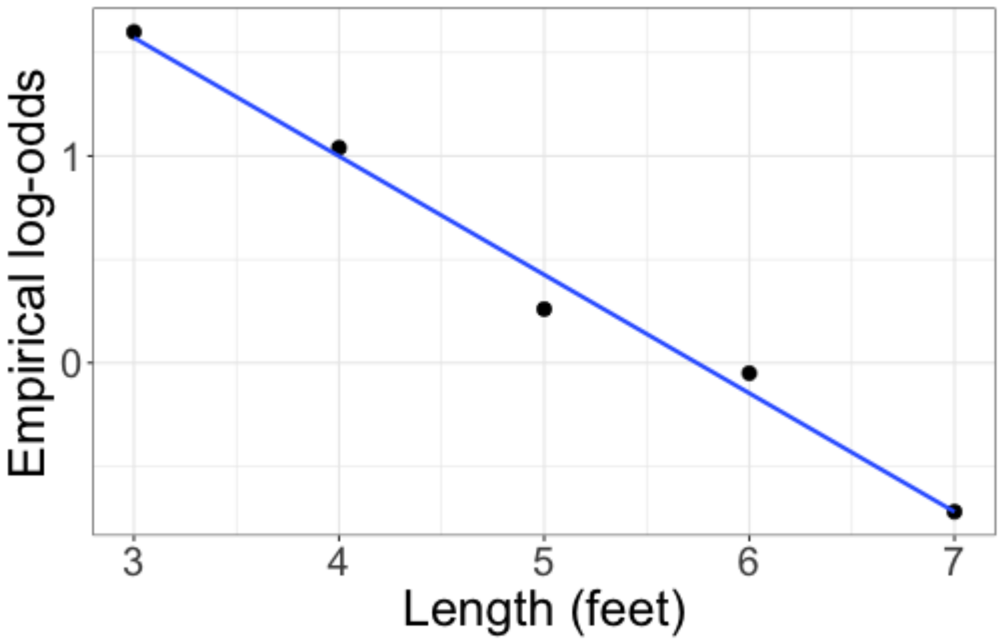
Empirical log odds

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 $\dfrac{\hat{p}}{1-\hat{p}}$	4.941	2.839	1.298	0.953	0.489
Log-odds $\log\left(\dfrac{\hat{p}}{1-\hat{p}}\right)$	1.60	1.04	0.26	-0.05	-0.72

Empirical log odds

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



seems reasonable!

> 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	...

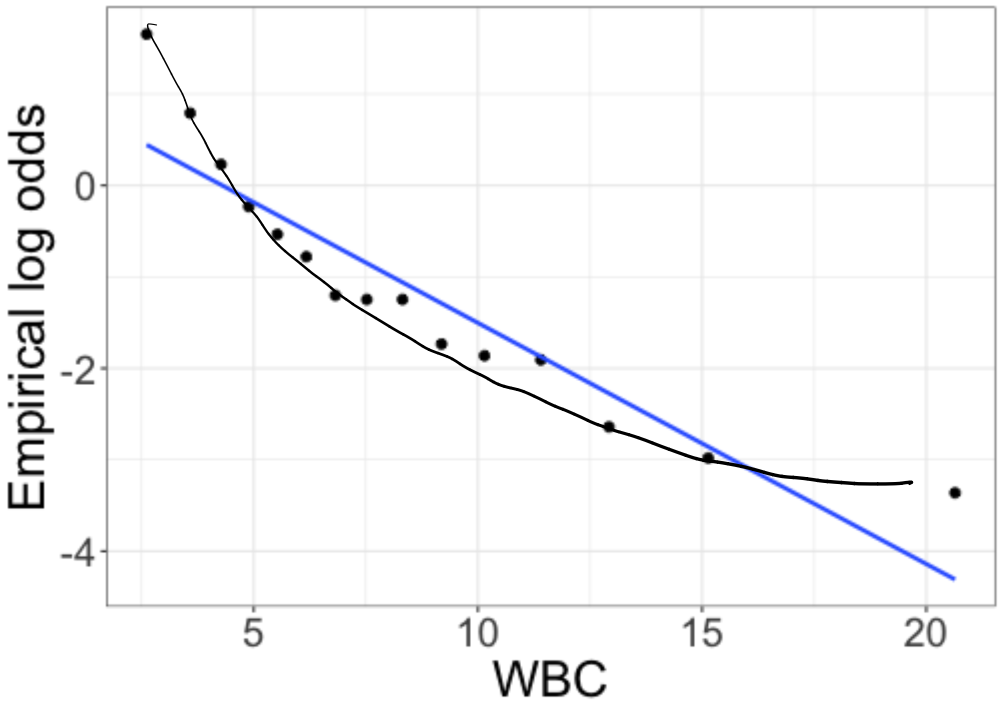
> What problem do I run into?

$0.90 \Rightarrow \hat{\pi} = 1 \qquad \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = \log\left(\frac{1}{0}\right) = \infty$

we have too few observations @ each WBC!

Solution:
1) Bin observations
2) calculate log odds in each bin

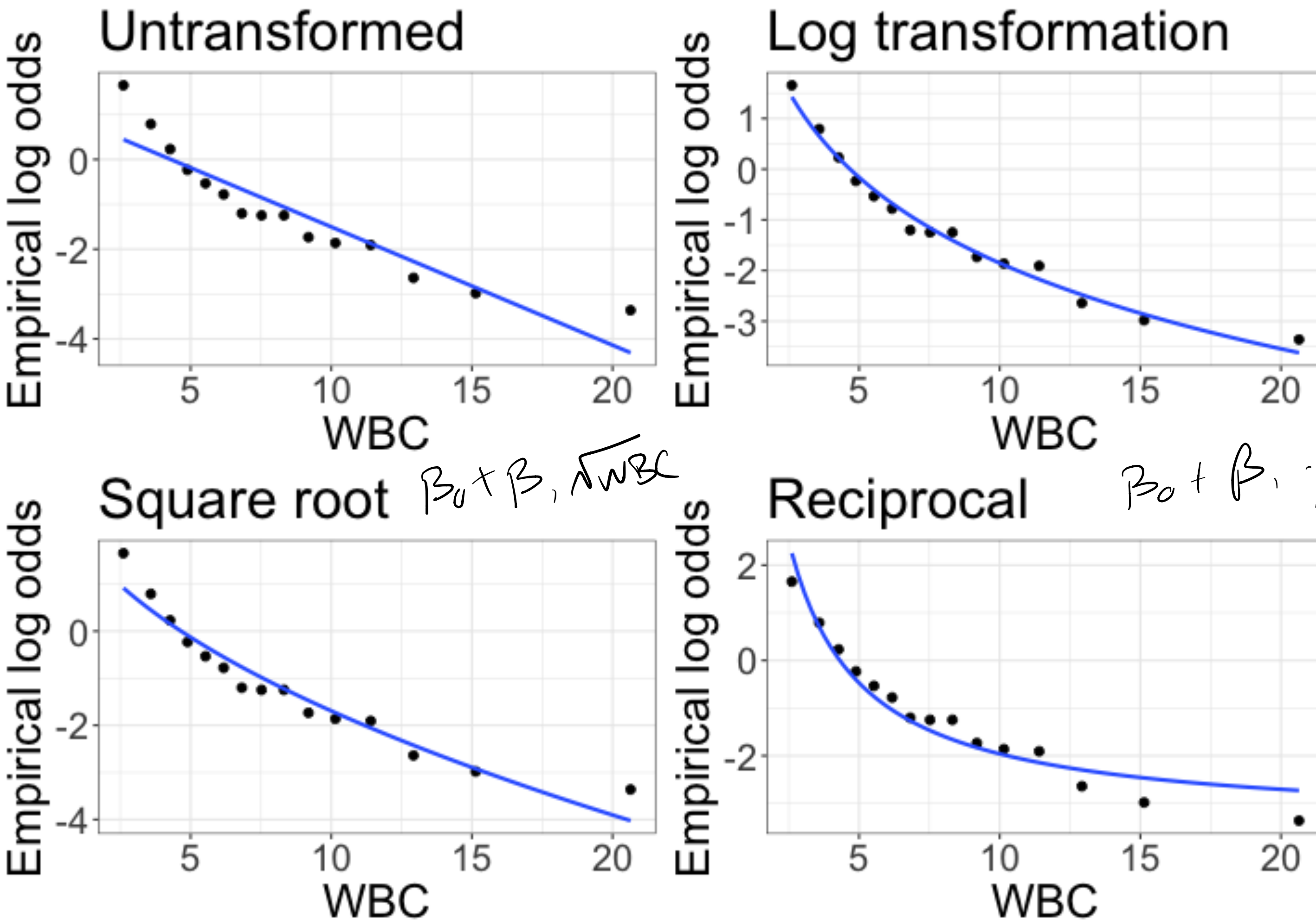
Binned empirical log odds plots



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

Trying some transformations

$\beta_0 + \beta_1 \log(WBC)$



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

$\beta_0 + \beta_1 \cdot \frac{1}{WBC}$

Lab 2

https://sta214-f22.github.io/labs/lab_2.html