

# Logistic regression assumptions and diagnostics

# Plan going forward

# 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

$Y_i$  = dengue status (0 = negative, 1 = positive)

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

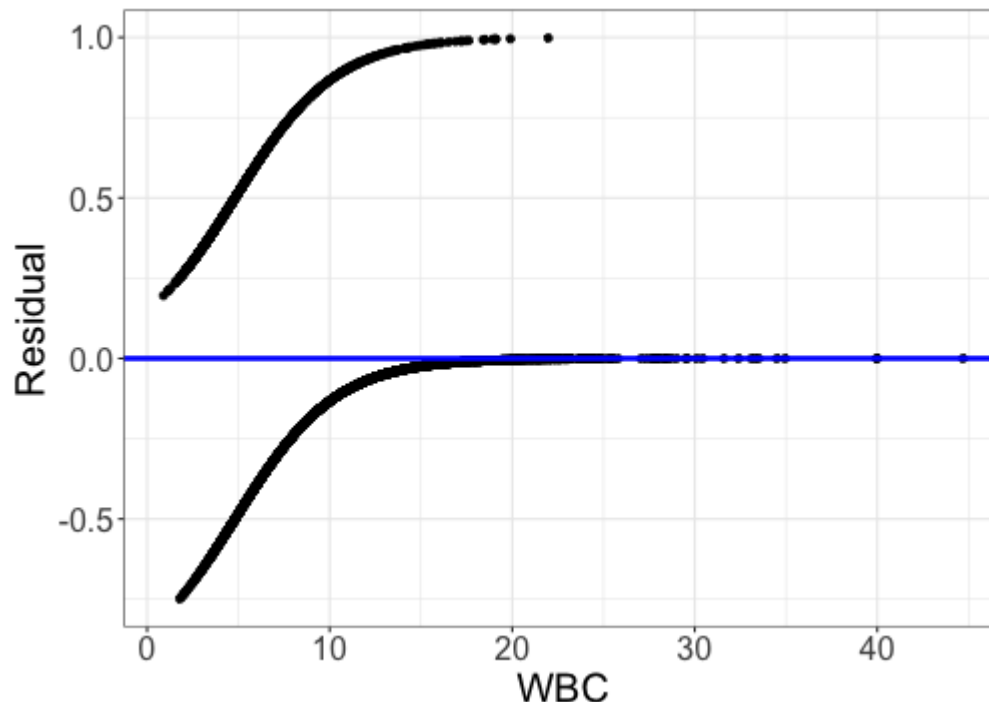
$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 WBC_i$$

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.

# 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:

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

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

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 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

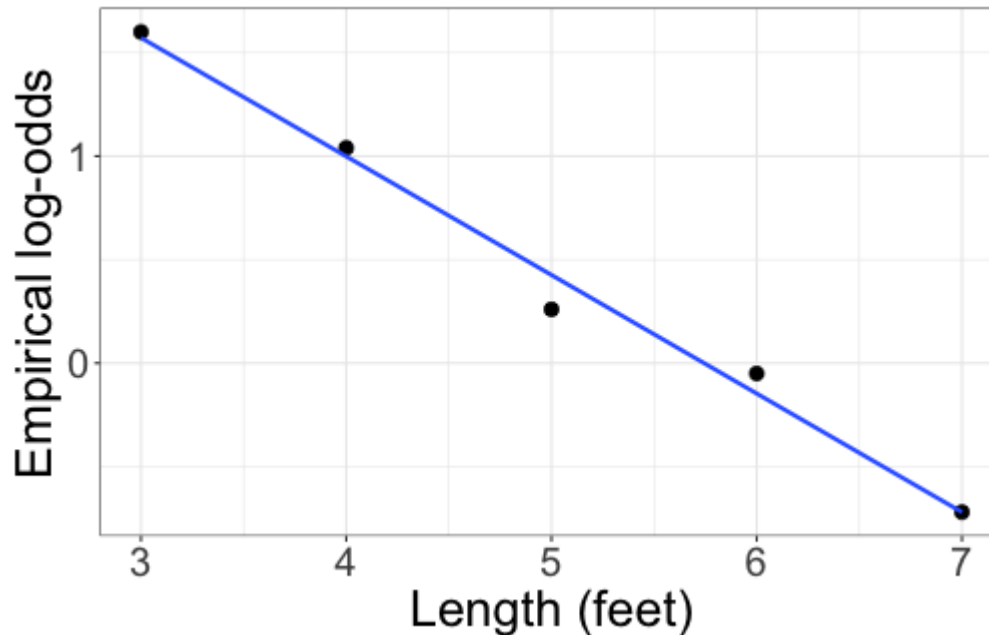
**Step 2:** convert empirical probabilities to empirical log odds

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



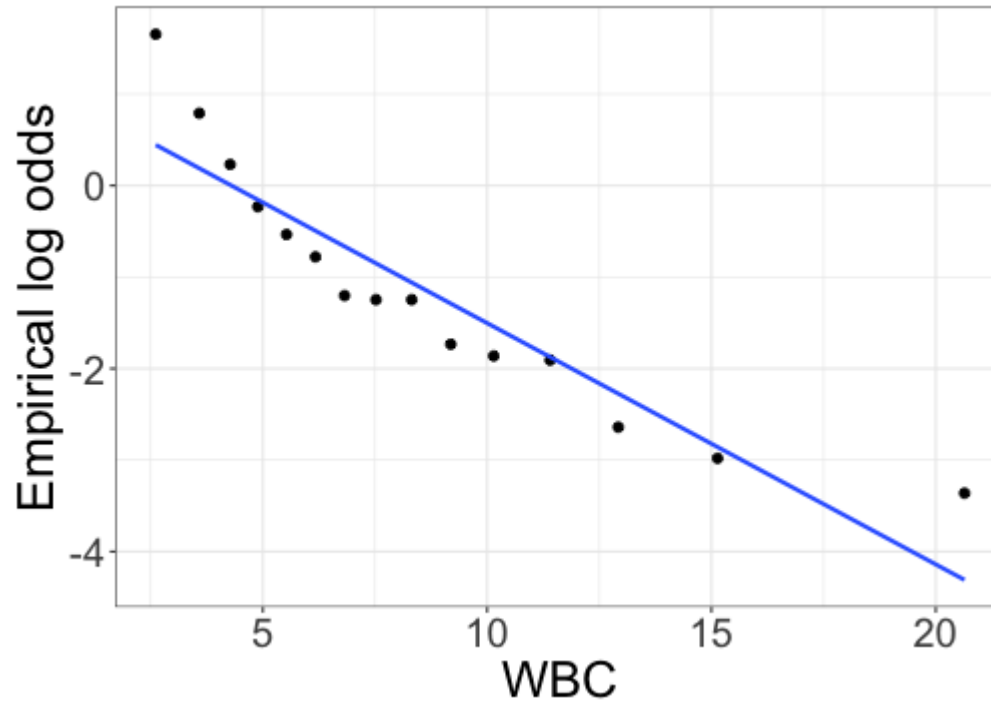
Does it seem reasonable that the log-odds are a linear function of length?

## Back to the dengue data...

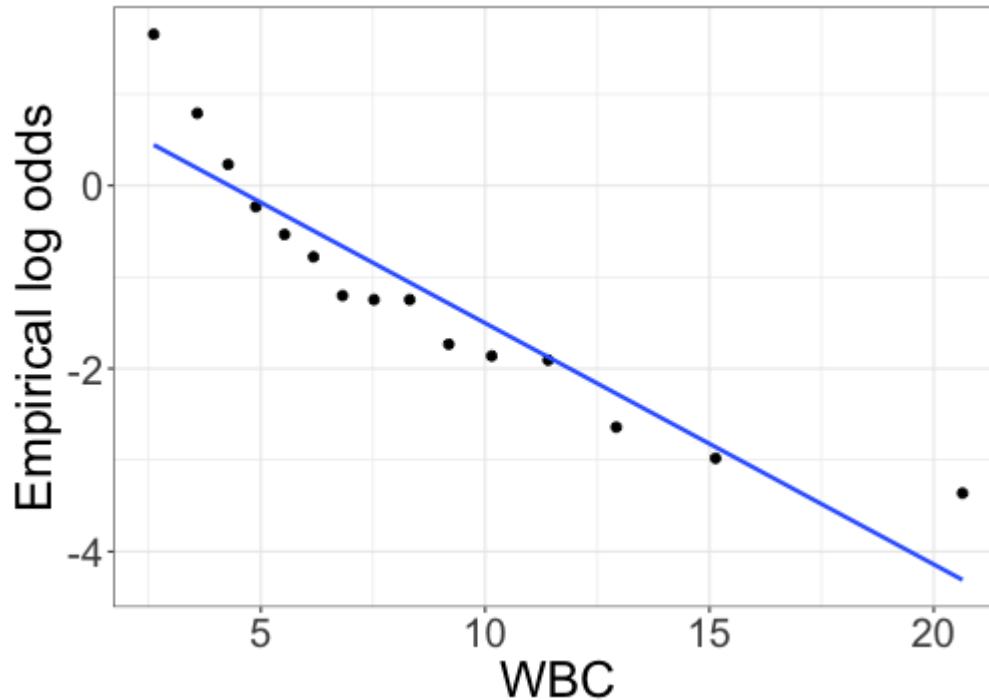
WBC	0.90	1.15	1.23	1.25	1.54	1.58	...
Dengue = 0	0	0	0	0	0	0	...
Dengue = 1	1	2	1	1	3	1	...

What problem do I run into?

## Binned empirical logit plots

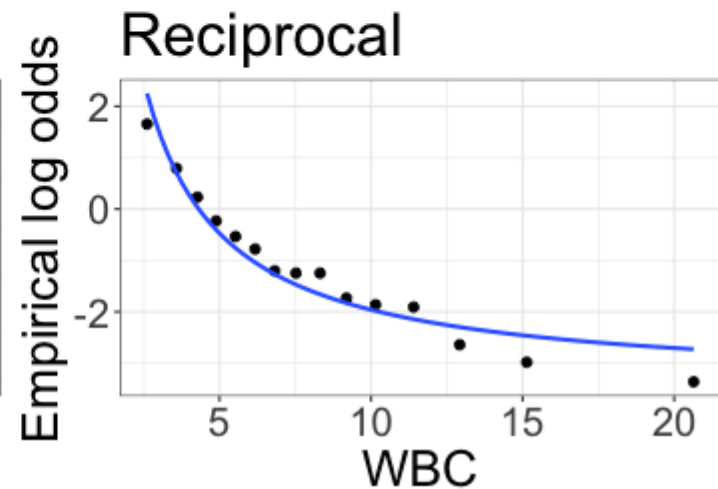
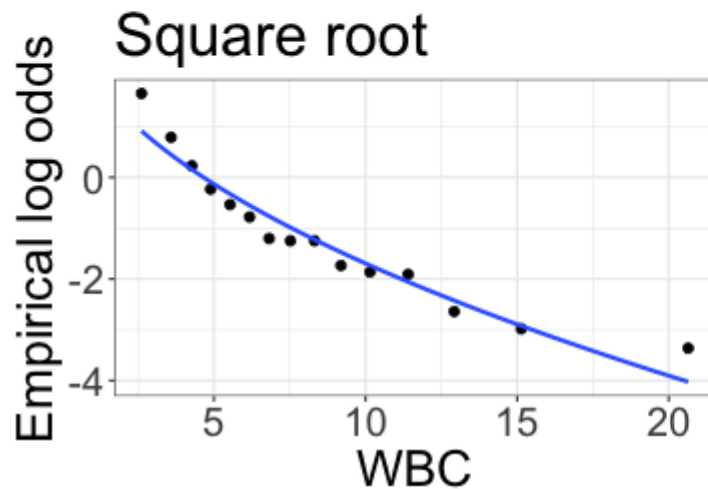
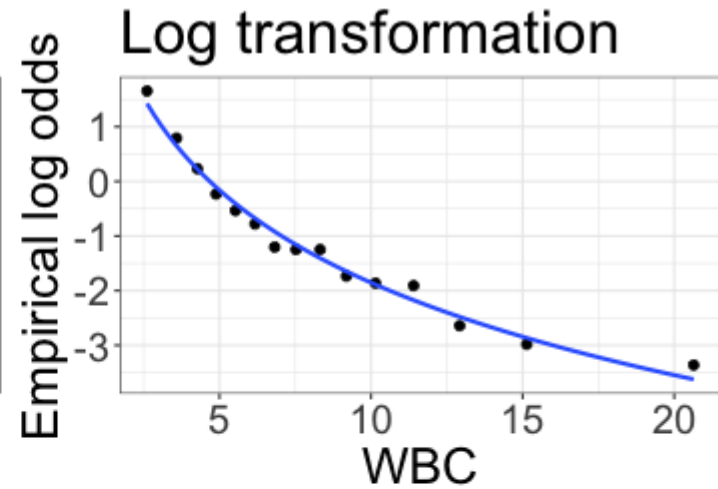
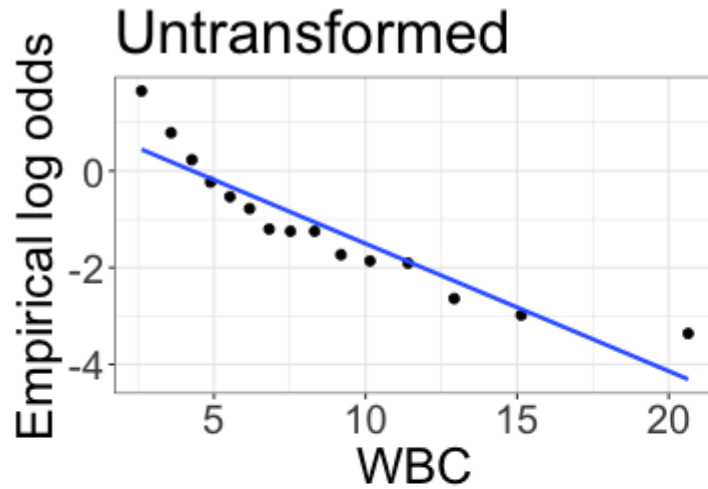


## Binned empirical logit plots



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

# Trying some transformations



# Why residuals in linear regression are nice

# Quantile residuals for logistic regression

# Class activity

[https://sta711-s23.github.io/class\\_activities/ca\\_lecture\\_9.html](https://sta711-s23.github.io/class_activities/ca_lecture_9.html)