Logistic regression assumptions and diagnostics

Plan going forward

So far:

- Estimation (MLE)
- Asymptotics
- Hypothesis testing

Next up:

- assumptions & diagnostics
- confidence intervals
- comparing different estimators

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**: (X_i, Y_i) are independent (usual for MLE)
- **shape**: log odds really are a linear function of the explanatory variable(s)

$$g(p_i) = \beta_0 + \beta_1 X_i$$

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

- **Lack of outliers**: all responses are generated from the same process
- Binary response

$Y_i \sim N(\mu_i, \sigma^2)$

$\mu_i = \beta_0 + \beta_1 X_i$

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

Diagnostics

- think about data generating process
- some kind of plot?
- Leverage & Cooks distance

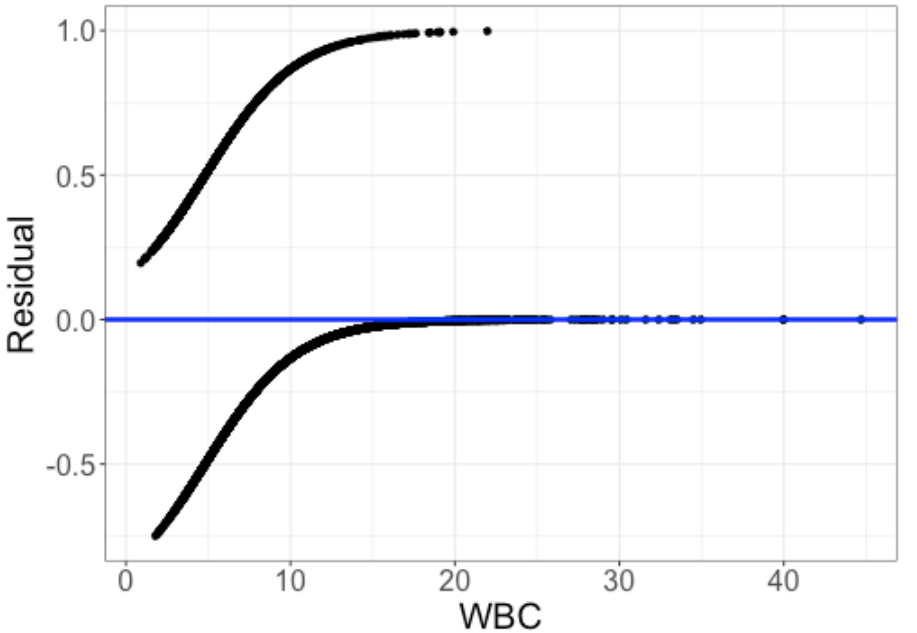
$$\log\left(\frac{\hat{Y}_i}{Y_i}\right)$$

Don't use usual residuals for logistic regression

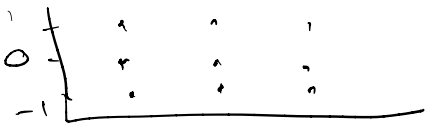
residuals in linear: $Y_i - \hat{\mu}_i$

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

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



$Y_i - \hat{Y}_i \in \{-1, 0, 1\}$



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 $\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right)$ for each length

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

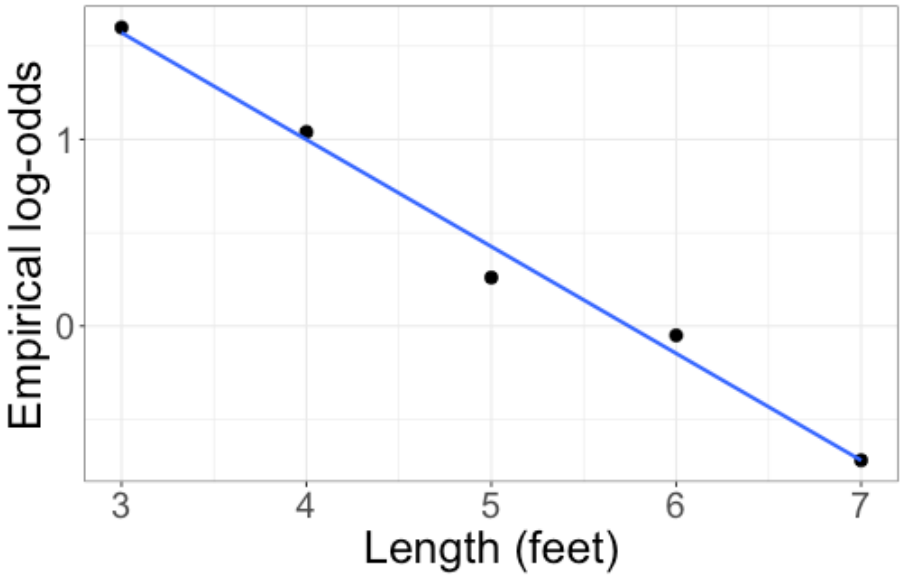
Empirical logits

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 for the log odds looks pretty good!

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

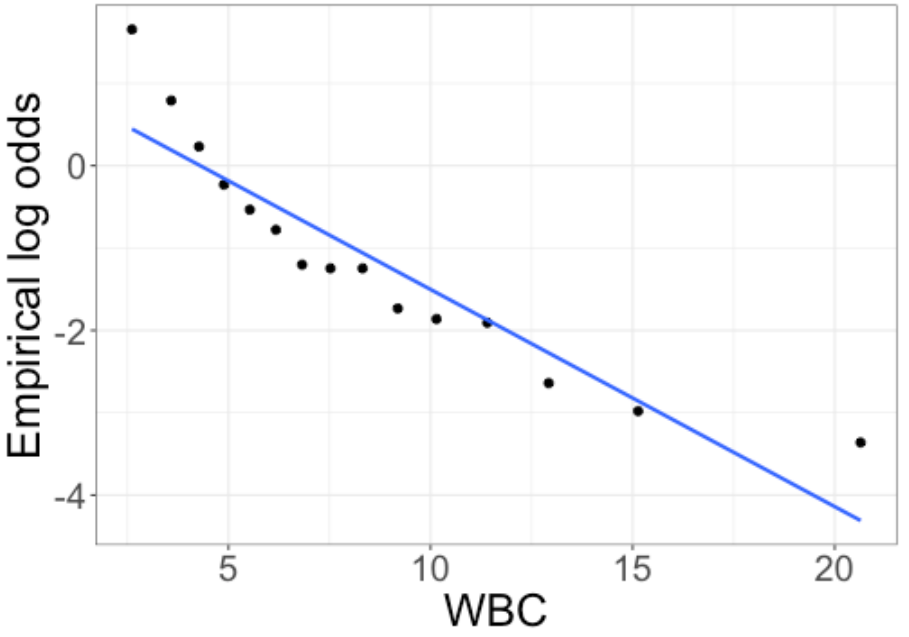
Back to the dengue data...

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 at each WBC to estimate log odds

Binned empirical logit plots



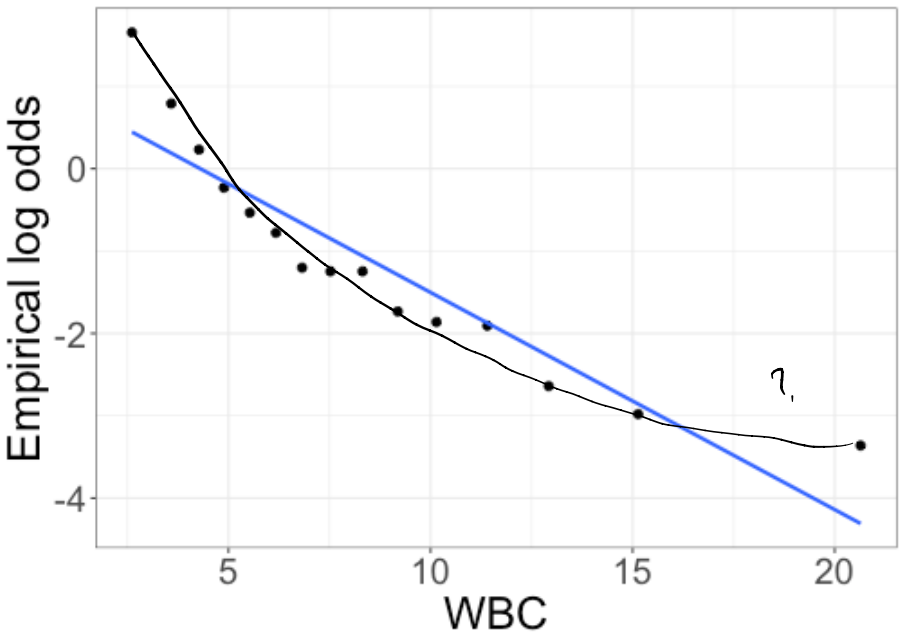
1) Specify n_{bins} (usually want at least 8-10, but depends on data size)

2) Divide data into n_{bins} groups based on WBC

3) In each bin, calculate empirical log odds

4) Plot empirical log odds vs. center of each bin

Binned empirical logit plots

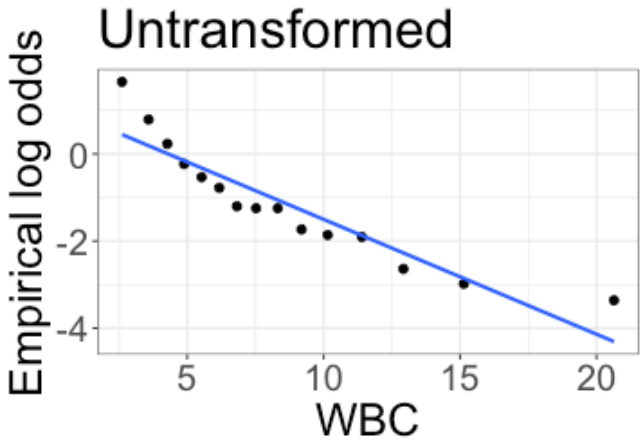


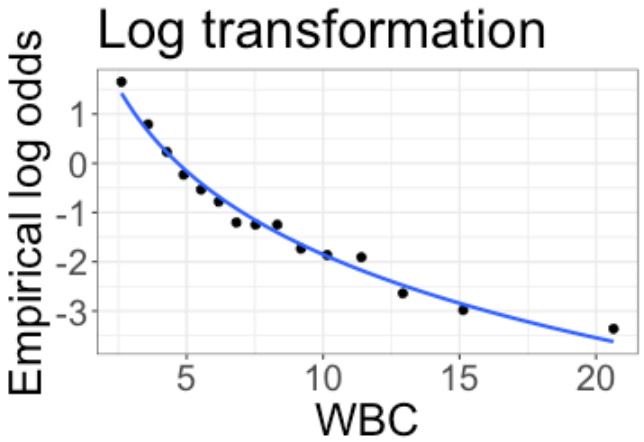
maybe slightly nonlinear

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

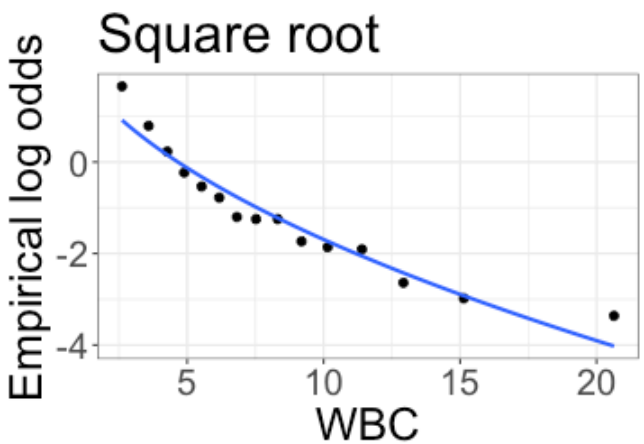
Trying some transformations

$\beta_0 + \beta_1 WBC_i$

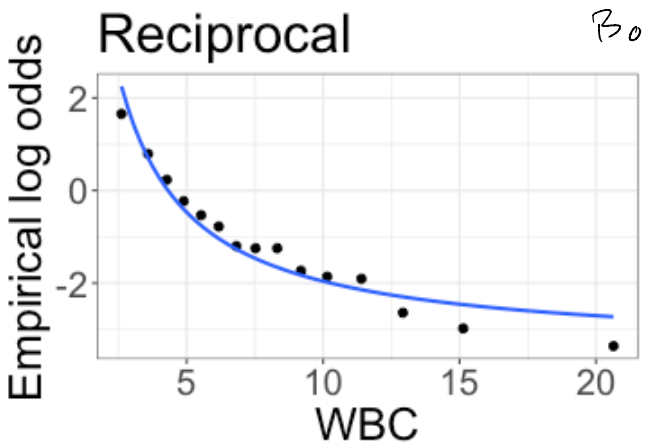




$\beta_0 + \beta_1 \log(WBC_i)$



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



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

$Y_i \sim N(\mu_i, \sigma^2)$ Y_i is continuous

$Y_i - \hat{\mu}_i$ is continuous

Why residuals in linear regression are nice

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

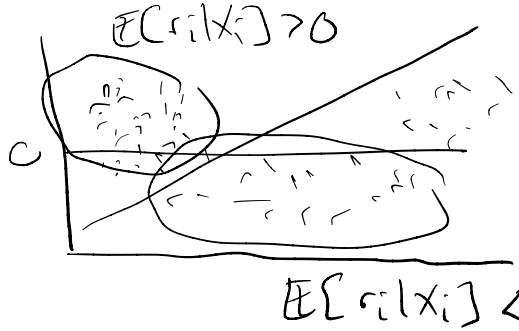
$r_i > 0 \Rightarrow$ underestimate

$r_i < 0 \Rightarrow$ overestimate

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

If the line is a good fit, $E[r_i | X_i] = 0 \quad \forall X_i$

residual plot



- patterns in residual plots indicate issues with model
- residuals are continuous

Quantile residuals for logistic regression

We want: residuals which are

- continuous
- defined for each point
- mean 0 when assumptions are met
  ($\mathbb{E}[r_i \mid x_i] = 0 \quad \forall x_i$)
- approximately normal?