Lecture 1

Logistic regression recap

Recall the dengue data from last semester:

- Data on Vietnamese children admitted to hospital with possible dengue fever
- Variables include:
  - Age
  - White blood cell count (WBC)
  - Platelet count (PLT)
  - Dengue status (0 = no dengue, 1 = dengue)

Logistic regression recap

I want to model dengue status, with Age, WBC, and PLT as explanatory variables.

What does my model look like?

Dengue status $\hookrightarrow Y_i \sim \text{Bernoulli}(p_i)$ $\leftarrow p(Y_i = 1)$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{WBC}_i + \beta_3 \text{PLT}_i$$

Logistic regression recap

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

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{WBC}_i + \beta_3 \text{PLT}_i$$

Question: How do I interpret a regression coefficient (e.g. β_1)?

β_1 = (additive) change in log odds of dengue associated with a one-year increase in Age, holding WBC & PLT fixed

e^{β_1} = (multiplicative) change in odds

Logistic regression recap

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

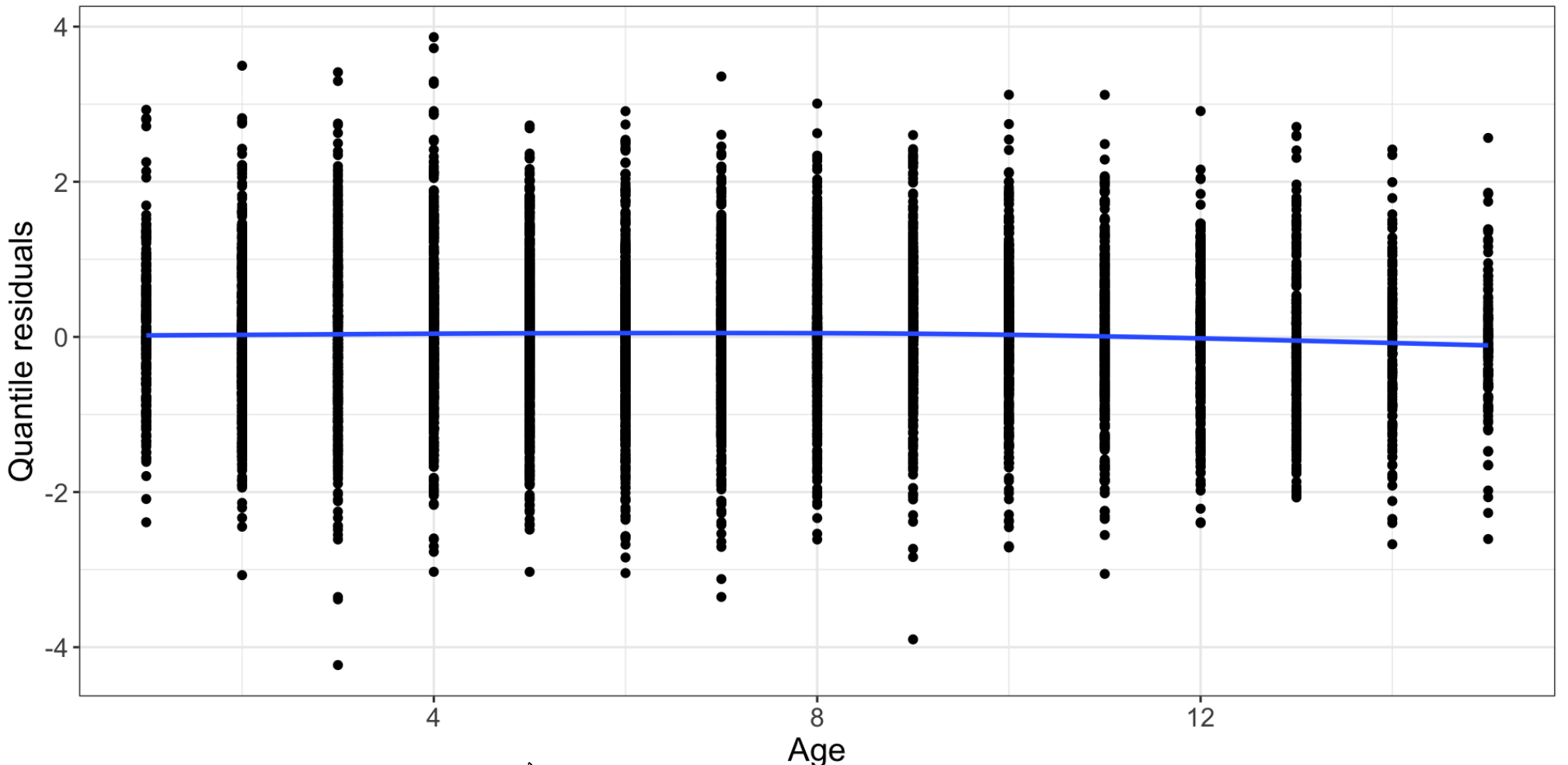
$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{WBC}_i + \beta_3 \text{PLT}_i$$

Question: What assumptions does this model make?

- Binary outcome
- independence (observations are independent)
  - → think about data generating process
- Shape (linearity in log odds) ← quantile residual plots, empirical logit plots
- lack of outliers (all observations come from the same process)
  - Cook's distance

Logistic regression recap

Quantile residual plot:



Looks good!

Logistic regression recap

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

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{WBC}_i + \beta_3 \text{PLT}_i$$

Question: I want to know whether there is relationship between Age and Dengue status, after accounting for WBC and PLT. How can I address this question?

$H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$

- Wald test
- LRT

Logistic regression recap

```r
m1 <- glm(Dengue ~ Age + WBC + PLT, data = dengue,
          family = binomial)
summary(m1)
```

```
Call:
glm(formula = Dengue ~ Age + WBC + PLT, family = binomial, data =
dengue)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.2525593  0.1548038   8.091  5.9e-16 ***
Age          0.1383186  0.0099763  13.865  < 2e-16 ***
WBC         -0.2523294  0.0135371 -18.640  < 2e-16 ***
PLT         -0.0060276  0.0006113  -9.860  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Logistic regression recap

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

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{WBC}_i + \beta_3 \text{PLT}_i$$

Question: Suppose now I want to test $H_0 : \beta_2 = \beta_3 = 0$. How do I carry out the likelihood ratio test?

LRT test stat $G = 2\log\left(\frac{L_{full}}{L_{reduced}}\right)$

under H_0, $G \approx \chi^2_q$ ← # parameters tested

For binary logistic regression: $-2\log L$ = deviance

$G = \text{Deviance}_{reduced} - \text{Deviance}_{full}$

Logistic regression recap

```r
m1 <- glm(Dengue ~ Age + WBC + PLT, data = dengue, family = binomial)
m1$deviance
```

```
[1] 5200.823
```

```r
m2 <- glm(Dengue ~ Age, data = dengue, family = binomial)
m2$deviance
```

```
[1] 6272.458
```

Test statistic = 6272.458 - 5200.823 = 1071.6

Logistic regression recap

```r
m1 <- glm(Dengue ~ Age + WBC + PLT, data = dengue, family = binomial)
m1$deviance
```

```
[1] 5200.823
```

```r
m2 <- glm(Dengue ~ Age, data = dengue, family = binomial)
m2$deviance
```

```
[1] 6272.458
```

Test statistic = $\text{deviance}_{\text{reduced}} - \text{deviance}_{\text{full}} = 1071.6$

How do I calculate a p-value? use χ^2_2 distribution

Logistic regression recap

```
1 m1 <- glm(Dengue ~ Age + WBC + PLT, data = dengue, family = binomial)
2 m1$deviance
```

```
[1] 5200.823
```

```
1 m2 <- glm(Dengue ~ Age, data = dengue, family = binomial)
2 m2$deviance
```

```
[1] 6272.458
```

Test statistic = $\text{deviance}_{\text{reduced}} - \text{deviance}_{\text{full}} = 1071.6$

```
1 pchisq(1071.6, df=2, lower.tail=F)
```

```
[1] 2.018443e-233
```

p-value ≈ 0

Logistic regression recap

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

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{WBC}_i + \beta_3 \text{PLT}_i$$

Question: The researchers are interested in whether their model does a good job identifying patients with dengue. Do our hypothesis tests address that question?

No. Need to be able to assess predictive ability

Rough course plan

- Logistic regression recap
- Prediction and model selection
- Supplementary skills (research papers, SAPs, simulation)
- Poisson regression and EDMs
- Mis-specified models (overdispersion, zero-inflation, etc.)
- Correlated data

Course components

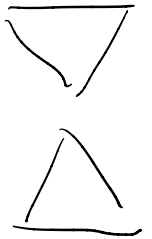
- Homework assignments (graded on completion)
- Challenge assignments (graded on mastery)
- Data analysis projects (graded on mastery)
- Semester research project (graded on mastery)
  - Group project
  - Involves written report, final presentation, and intermediate check-points
  - Due next Monday: group members and tentative topic

Reading a research paper

Research papers in the sciences and social sciences typically contain:

- Abstract
- Introduction
- Methods
- Results
- Discussion
- Conclusion

Reading a research paper



- **Abstract:** overview and key points
- **Introduction:** motivation, background, overview of work
- **Methods:** details on study design, data, statistical analysis
- **Results:** summary of results, including figures, tables, p-values, etc.
- **Discussion:** discussion of results in context of research question
- **Conclusion:** short summary of paper and key results; connection to broader research

Class activity

Reading the original dengue paper:

https://sta712-f23.github.io/class_activities/ca_1.pdf

For next class:

- finish reading the paper and working through the class activity
- we will discuss the paper on Wednesday