Linear and logistic regression

Ciaran Evans

Last time: parameter estimation for linear regression

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

$$SSE(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$

$$\frac{\partial}{\partial \beta} SSE(\beta) =$$

Parameter estimation for linear regression

$$\frac{\partial}{\partial \beta} SSE(\beta) \quad = \quad -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) \quad \stackrel{set}{=} \quad \mathbf{0}$$

Warmup

Work on the warmup activity:

https://sta711-s26.github.io/class_activities/ca_02.html

Submit your work on Canvas.

Warmup

```
library(Stat2Data)
data("FirstYearGPA")

lm(GPA ~ HSGPA + SATM, data = FirstYearGPA) |> coef()
```

```
##  (Intercept)         HSGPA          SATM
## 0.7579761649 0.5305150632 0.0007984613
```

```
y <- FirstYearGPA$GPA
X <- cbind(1, FirstYearGPA$HSGPA, FirstYearGPA$SATM)
solve(t(X) %*% X) %*% t(X) %*% y
```

```
##                [,1]
## [1,] 0.7579761649
## [2,] 0.5305150632
## [3,] 0.0007984613
```

Regression assumptions

$$\text{GPA}_i = \beta_0 + \beta_1 \text{HSGPA}_i + \beta_2 \text{SATM}_i + \varepsilon_i$$

Question: What assumptions do we often make when fitting a linear regression model?

Regression assumptions

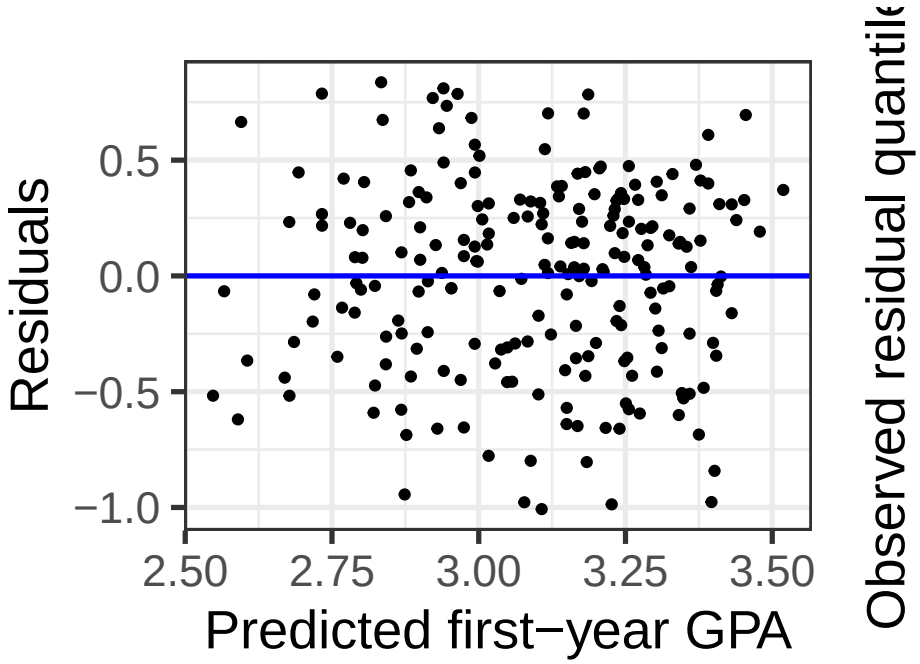
$$\text{GPA}_i = \beta_0 + \beta_1 \text{HSGPA}_i + \beta_2 \text{SATM}_i + \varepsilon_i$$

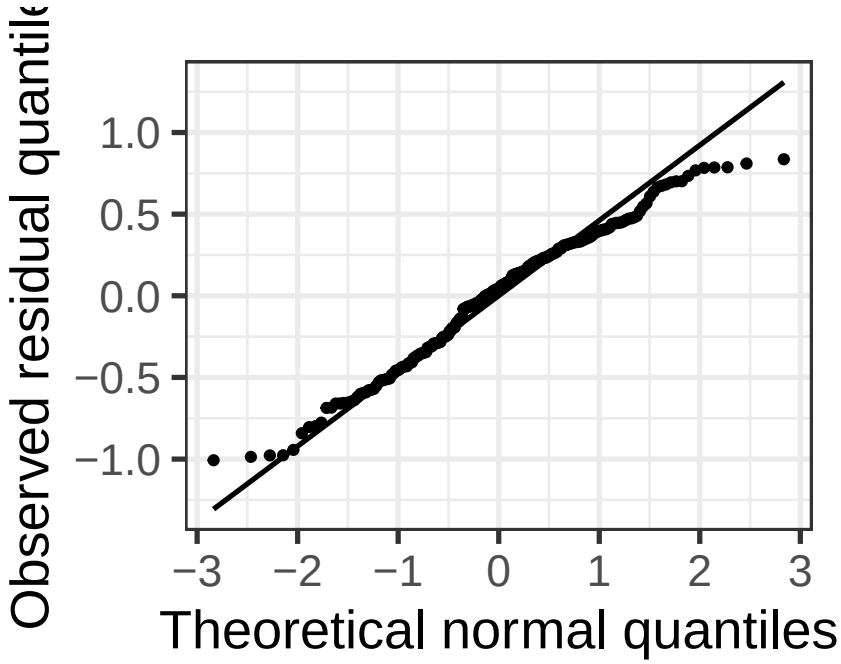
Some common assumptions:

- Shape
- Constant variance (variance of ε_i is the same for all observations)
- Normality (ε_i comes from a normal distribution)
- Independence

Question: How do we assess these assumptions?

Diagnostic plots

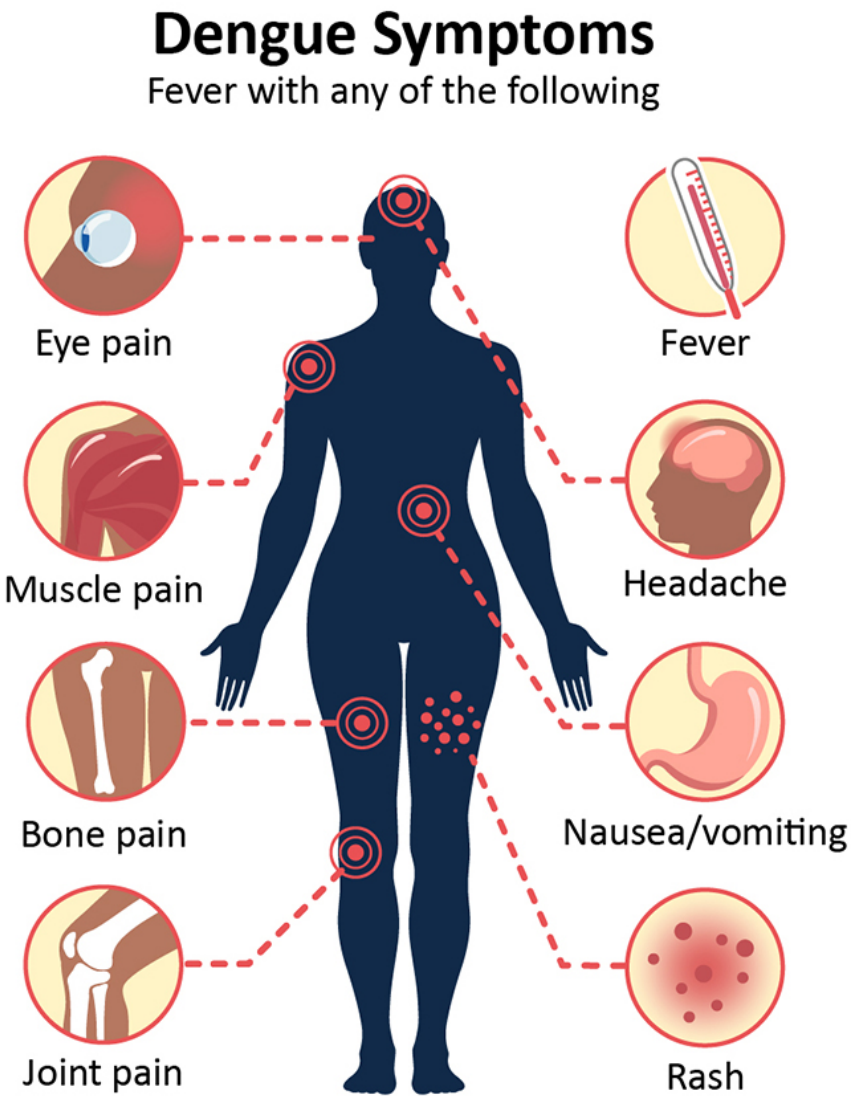




Question: Do the regression assumptions seem reasonable here?

Another motivating example: Dengue fever

Dengue fever: a mosquito-borne viral disease affecting 400 million people a year



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)

Research goal: Predict dengue status using diagnostic measurements

Fitting a model: initial attempt

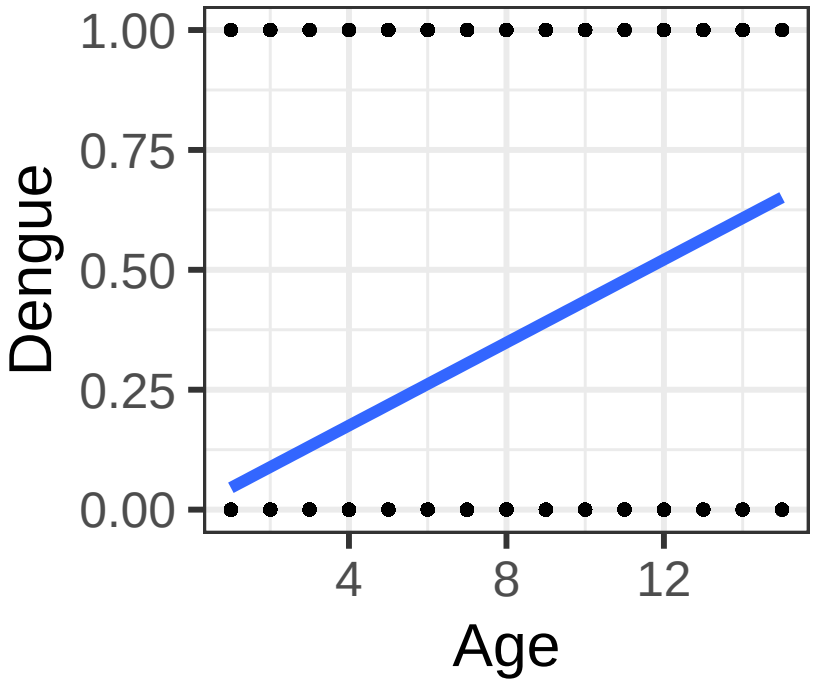
What if we try a linear regression model?

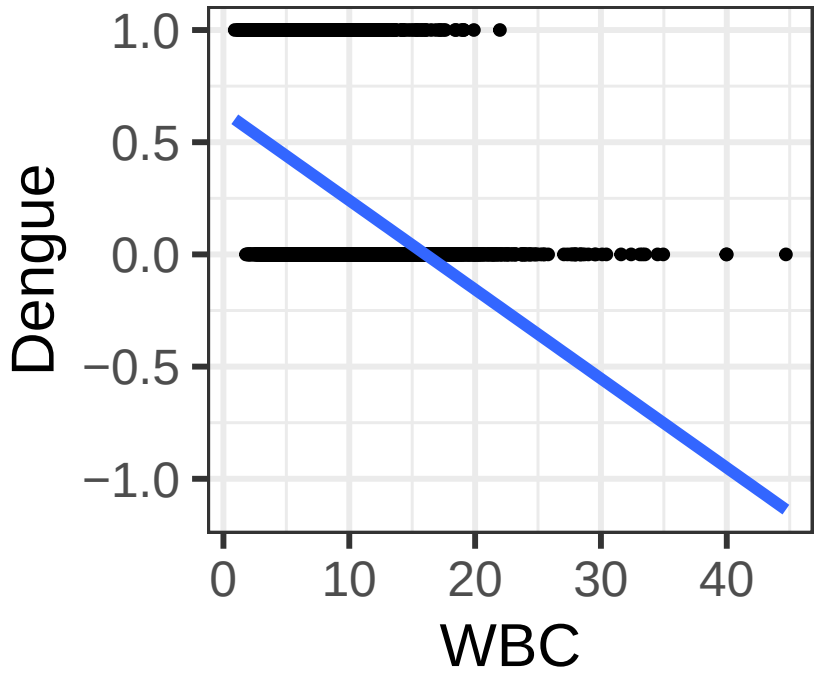
$$Y_i = \text{dengue status of } i\text{th patient}$$

$$Y_i = \beta_0 + \beta_1 WBC_i + \varepsilon_i \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

What are some potential issues with this linear regression model?

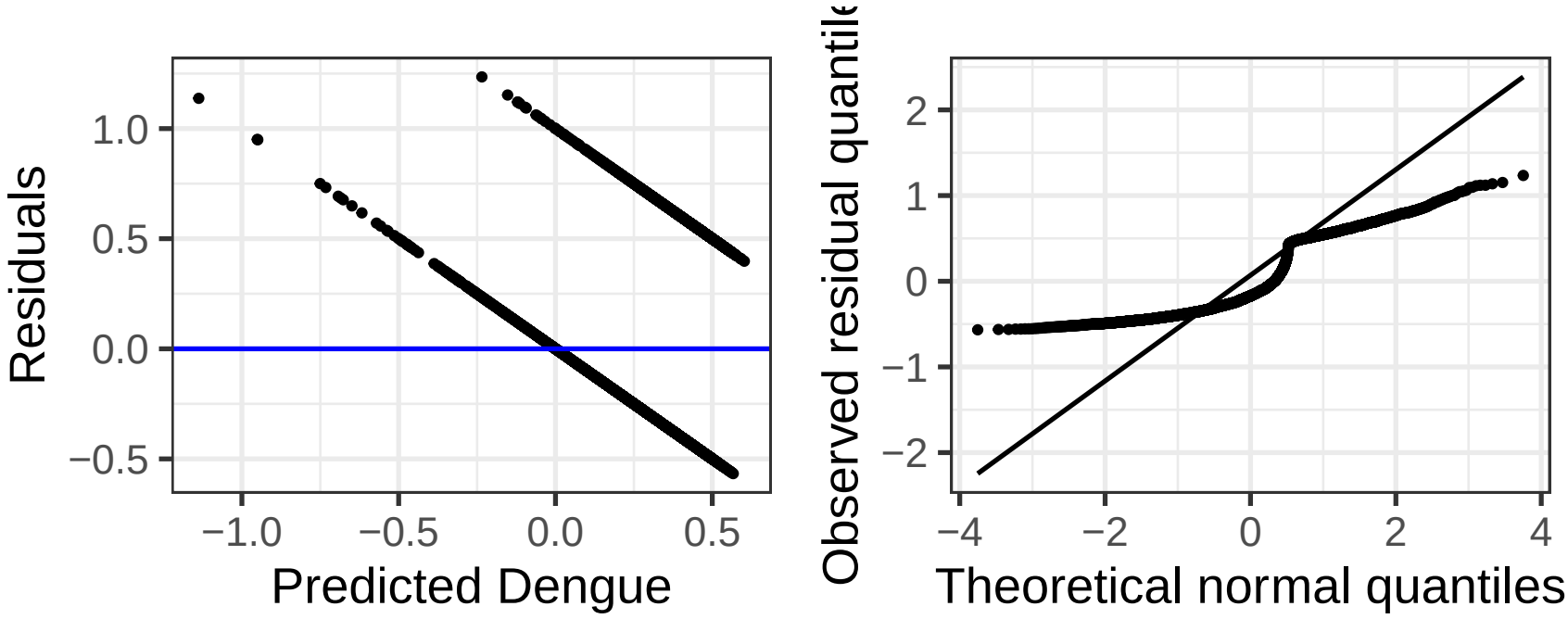
Don't fit linear regression with a binary response





Don't fit linear regression with a binary response

Diagnostic plots:



Second attempt

Let's rewrite the linear regression model:

Second attempt

$$Y_i|WBC_i \sim Bernoulli(p_i) \qquad p_i = \mathbb{P}(Y_i = 1|WBC_i)$$

$$p_i = \beta_0 + \beta_1 WBC_i$$

Are there still any potential issues with this approach?

Fixing the issue: logistic regression

$$Y_i | WBC_i \sim Bernoulli(p_i)$$

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

where $g : (0, 1) \to \mathbb{R}$ is unbounded.

Usual choice: $g(p_i) = \log\left(\frac{p_i}{1 - p_i}\right)$

Logistic regression model

$$Y_i | WBC_i \sim Bernoulli(p_i)$$

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

Why is there no noise term ε_i in the logistic regression model? Discuss for 1–2 minutes with your neighbor, then we will discuss as a class.