Linear and logistic regression

Ciaran Evans

Last time: parameter estimation for linear regression

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

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

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

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

Parameter estimation for linear regression

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

$$\Rightarrow \quad X^T(y - X\beta) = 0$$

$$\Rightarrow \quad X^Ty - X^TX\beta = 0$$

$$\Rightarrow \quad X^TX\beta = X^Ty$$

$$\Rightarrow \quad \hat{\beta} = (X^TX)^{-1}X^Ty$$

least squares estimator of β

Warmup

Work on the warmup activity:

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

Submit your work on Canvas.

Warmup

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

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

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

```r
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

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

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

Regression assumptions

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

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

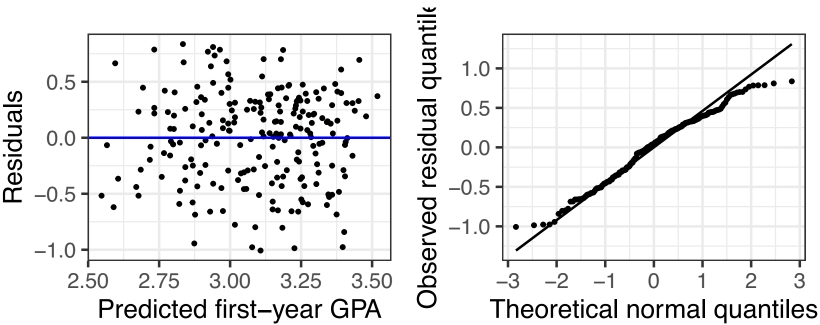
Some common assumptions:

- 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?

- normality: Q Q plot
- independence: think about how data were generated
- shape & constant variance: residual plot

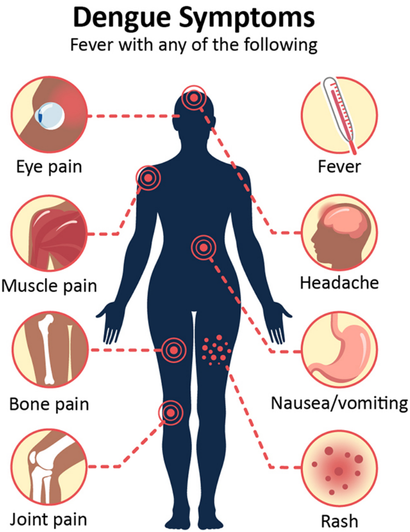
Diagnostic plots



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?

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

(1 = dengue, 0 = no dengue)

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

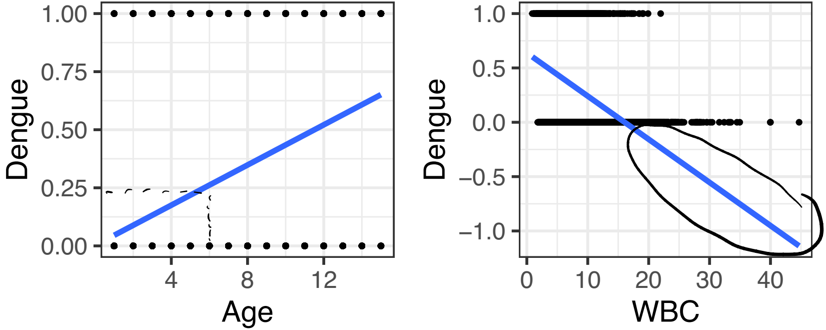
What are some potential issues with this linear regression model?

$Y_i \in \{0, 1\}$

$\beta_0 + \beta_1 WBC_i + \varepsilon_i \quad \in (-\infty, \infty)$ (potentially)

(continuous & possibly unbounded)

Don't fit linear regression with a binary response



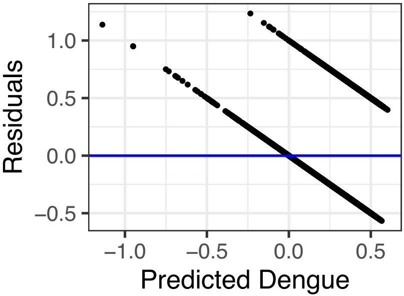
Age=6 → $\hat{\text{Dengue}}$ = 0.25

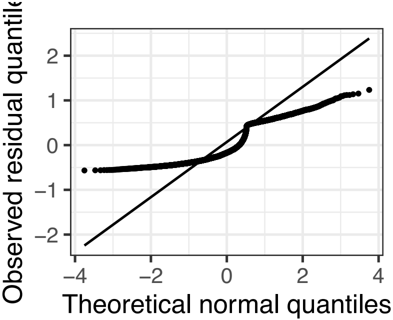
maybe this is a 25% chance of having Dengue?

if WBC > 15, $\hat{\text{Dengue}}$ < 0 ??

Don't fit linear regression with a binary response

Diagnostic plots:





Second attempt

if $V \sim N(\mu, \sigma^2)$ $\quad a, b \in \mathbb{R}$

$a + bV \sim N(a + b\mu, b^2\sigma^2)$

Let's rewrite the linear regression model:

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

$$\Rightarrow \quad Y_i | WBC_i \sim N(\beta_0 + \beta_1 WBC_i, \sigma^2)$$

or written in two pieces:

describes distribution of $Y_i | WBC_i$

$$Y_i | WBC_i \sim N(\mu_i, \sigma^2) \quad \text{(random component)}$$

$$\mu_i = \beta_0 + \beta_1 WBC_i \quad \text{(systematic component)}$$

specifies how distribution depends on WBC_i

Problem: $Y_i \in \{0, 1\}$ $\Rightarrow$ $Y_i | WBC_i$ is not normal!

Use Bernoulli instead!