Parametric models and logistic regression

Data

Information on 911 Indonesian husband-wife couples, with the wife aged between 20 and 35, and variables including:

- Contraceptive method used (0 = none, 1 = some use)
- Wife's age (in years)
- Husband's age (in years)
- Wife's education (1 = low, 2, 3, 4 = high)
- Husband's education (1 = low, 2, 3, 4 = high)
- Number of children ever born

Notation: Let Y = contraceptive use (0 or 1), and Age = wife's age. Let (Age_i, Y_i) be the observations for couple i ($i = 1, \ldots, n$)

Regression Modeling

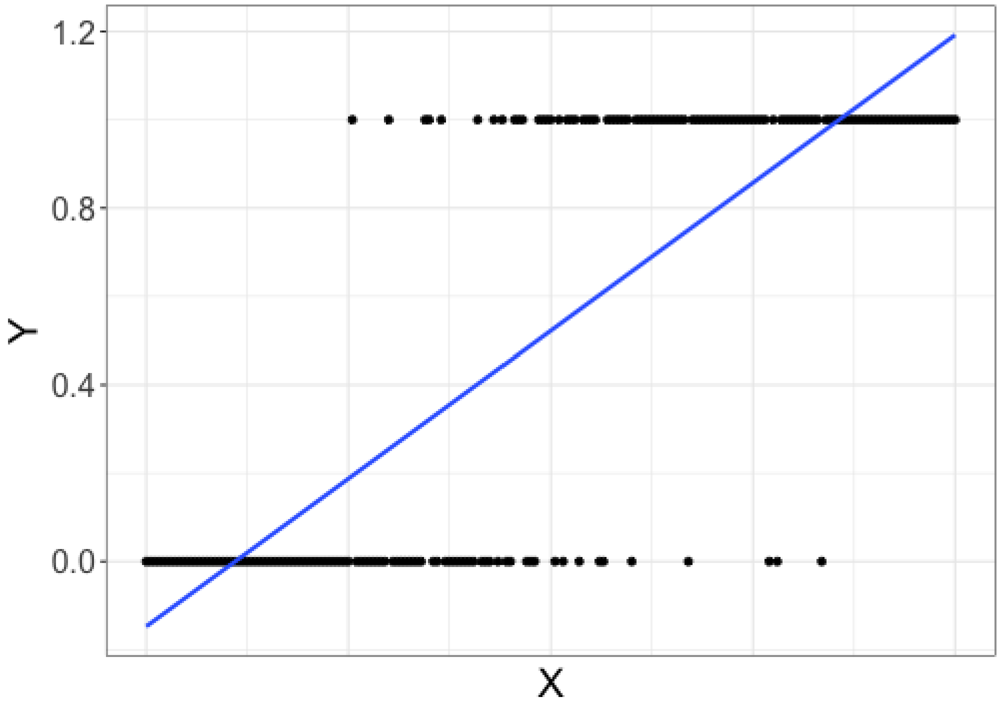
Goal: The goal of a regression model is to describe the relationship between the predictor and the response.

Example: linear regression

$$Y_i = \beta_0 + \beta_1 Age_i + \varepsilon_i \quad \text{where} \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$$

> Here $Y_i = 0$ or 1. Is a linear regression appropriate?

Linear regression is not appropriate for binary data



Revisiting the linear regression model

Parametric modeling

A regression model is an example of a more general process called **parametric modeling**

- **Step 1:** Choose a reasonable distribution for Y_i
- **Step 2:** Build a model for the parameters of interest
- **Step 3:** Fit the model

Step 1: Choose a reasonable distribution for Y_i

> What do I mean by a *distribution*?

Step 1: Choose a reasonable distribution for Y_i

> What do I mean by a *distribution*?

- A **distribution** tells us what outcomes are possible for Y_i, and how often these outcomes occur.

Here the possible values of Y_i are 0 (no contraceptive use) and 1 (some use).

> How often do these values occur in the population?

Step 1: Choose a reasonable distribution for Y_i

> What do I mean by a *distribution*?

- A **distribution** tells us what outcomes are possible for Y_i, and how often these outcomes occur.

Here the possible values of Y_i are 0 (no contraceptive use) and 1 (some use).

> How often do these values occur in the population?

- We don't know, so we will estimate from the sample
- We assume the probability $Y_i = 1$ depends on Age_i

Step 1: Choose a reasonable distribution for Y_i

Bernoulli distribution

Definition: Let Y_i be a binary random variable, and $\pi_i = P(Y_i = 1)$. Then $Y_i \sim Bernoulli(\pi_i)$.

> What do I mean by a *random variable*?

Bernoulli distribution

Definition: Let Y_i be a binary random variable, and $\pi_i = P(Y_i = 1)$. Then $Y_i \sim Bernoulli(\pi_i)$.

> What do I mean by a *random variable*?

A **random variable** is an event that has a set of possible outcomes, but we don't know which one will occur

- Here $Y_i = 0$ or 1
- Our goal is to use the observed data to estimate $\pi_i = P(Y_i = 1)$

Step 2: Build a model

- $Y_i =$ contraceptive use (0 = none, 1 = some)
- $Y_i \sim Bernoulli(\pi_i)$
- Our parameter is π_i, which we assume depends on Age_i. For a binary response, we will use a **logistic regression** model

Logistic regression model

$Y_i =$ contraceptive use (0 = none, 1 = some)

$Age_i =$ wife's age (in years)

Step 1: $Y_i \sim Bernoulli(\pi_i)$

Step 2: $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \ Age_i$

> 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.

A note on parameters

$$Y_i \sim Bernoulli(\pi_i) \qquad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1\ Age_i$$

- π_i: parameter for the distribution of Y_i. Depends on Age_i
- β_0, β_1: parameters for the (unknown) relationship between Age_i and π_i

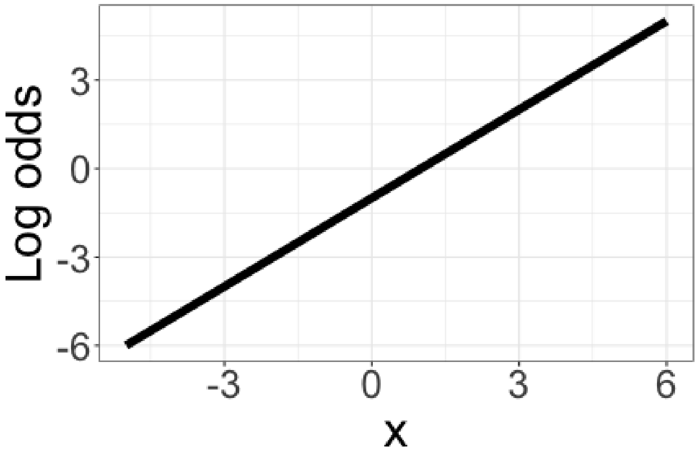
Modeling π_i

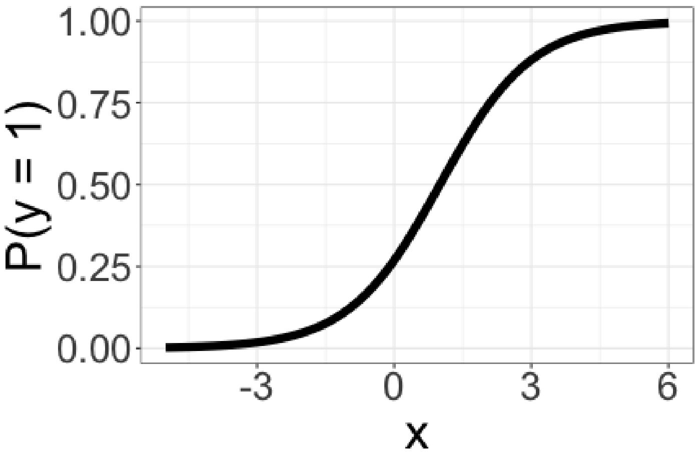
$$Y_i \sim Bernoulli(\pi_i) \qquad \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1\ Age_i$$

> What if I want the model in terms of π_i, instead of the log odds?

Shape of the regression curve

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1\ X_i \qquad \pi_i = \frac{e^{\beta_0+\beta_1\ X_i}}{1+e^{\beta_0+\beta_1\ X_i}}$$



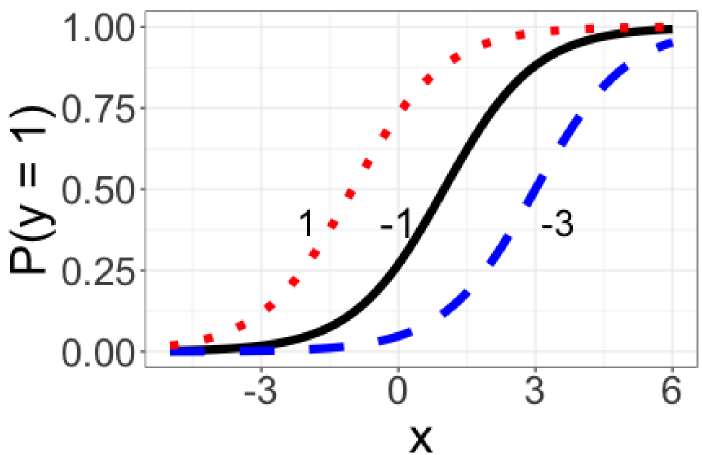


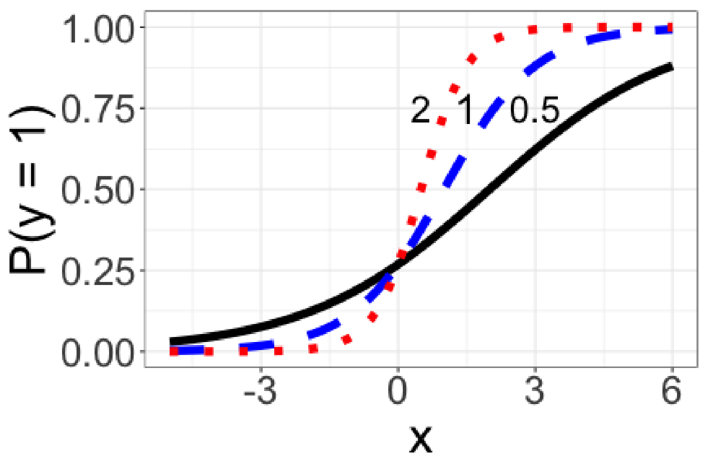
Shape of the regression curve

How does the shape of the fitted logistic regression depend on β_0 and β_1?

$$\pi_i = \frac{\exp\{\beta_0 + x_i\}}{1+\exp\{\beta_0 + x_i\}} \quad \text{for} \quad \beta_0 = -3, -1, 1$$

$$\pi_i = \frac{\exp\{-1 + \beta_1\, x_i\}}{1+\exp\{-1 + \beta_1\, x_i\}} \quad \text{for } \beta_1 = 0.5, 1, 2$$





Parametric modeling

$Y_i =$ contraceptive use (0 = none, 1 = some)

$Age_i =$ wife's age (in years)

Step 1: $Y_i \sim Bernoulli(\pi_i)$

Step 2: $\log\left(\dfrac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \, Age_i$

Step 3: Fitting the model

$$\log\left(\frac{\widehat{\pi}_i}{1-\widehat{\pi}_i}\right) = -0.976 + 0.052 \, Age_i$$

Class Activity, Part I

https://sta214-f22.github.io/class_activities/ca_lecture2.html

- Spend 5--7 minutes working in pairs on questions 1 -- 5
- Solutions are provided for 1 -- 3
- We will discuss 4 and 5 as a class

Class Activity

What is the predicted probability of contraception use if the wife is 30 years old?

Class Activity

> Suppose that researchers want to follow up with couples for whom the probability of contraception use is less than 60%. Which age range should they target?

Class Activity, Part II

https://sta214-f22.github.io/class_activities/ca_lecture2.html

- Spend 3--5 minutes working in pairs on questions 6 -- 8
- Solutions are provided for 6 and 7
- We will discuss 8 as a class

Interpretation

Fitted model: log odds form

$$\log\left(\frac{\widehat{\pi}_i}{1-\widehat{\pi}_i}\right) = -0.976 + 0.052\ Age_i$$

- *Interpretation:* For every one-year increase in age, we predict that the log odds of contraception use increase by 0.052

Interpretation

Fitted model: log odds form

$$\log\left(\frac{\widehat{\pi}_i}{1-\widehat{\pi}_i}\right) = -0.976 + 0.052\ Age_i$$

- *Interpretation:* For every one-year increase in age, we predict that the log odds of contraception use increase by 0.052

Fitted model: odds form

$$\frac{\widehat{\pi}_i}{1-\widehat{\pi}_i} = e^{-0.976+0.052\ Age_i} = e^{-0.976}e^{0.052\ Age_i}$$

- *Interpretation:* For every one-year increase in age, we predict that the odds of contraception use get multiplied by $e^{0.052} = 1.053$

Comparing linear and logistic regression

- We built the logistic regression model using steps for building a parametric model
- We can use the same steps for linear regression:
  - **Step 1:** $Y_i \sim N(\mu_i, \sigma^2)$
  - **Step 2:** $\mu_i = \beta_0 + \beta_1 X_i$
- Choosing logistic vs. linear regression depends on the distribution of Y_i
  - As we move through the course, we will see other distributions for Y_i too