# 12.1: The Simple Linear and Logistic Regression Models

Taylor

University of Virginia

# Introduction

So far we've been assuming that we have a random sample (independent and *identically distributed*)

$$Y_1, \ldots, Y_n \overset{iid}{\sim} \text{something}(\theta)$$

# Introduction

So far we've been assuming that we have a random sample (independent and *identically distributed*)

$$Y_1, \dots, Y_n \overset{iid}{\sim} \text{something}(\theta)$$

Now we'll talk about when our data are still mutually independent, but not *identically distributed*

$$Y_1 \sim \text{something}_1(\theta), \dots, Y_n \sim \text{something}_n(\theta)$$

# Introduction

So far we've been assuming that we have a random sample (independent and *identically distributed*)

$$Y_1, \ldots, Y_n \overset{iid}{\sim} \text{something}(\theta)$$

Now we'll talk about when our data are still mutually independent, but not *identically distributed*

$$Y_1 \sim \text{something}_1(\theta), \ldots, Y_n \sim \text{something}_n(\theta)$$

We still need to fit/learn parameters, and we'll also observe predictors/inputs/covariates $X_1, \ldots, X_n$.

# Introduction

We want to develop a framework for modelling a *probabilistic* dependence between a **dependent response** ($Y$) and an **independent, explanatory predictor** ($X$). The general form of our model is

$$Y = f(x) + \epsilon$$

where $f(\cdot)$ is a *deterministic* function and $\epsilon$ is a random error (random variable).

# Introduction

How do we get an idea for what $f(\cdot)$ should be? Typically we plot $x_1, \ldots, x_n$ against $y_1, \ldots, y_n$. This is called a **scatterplot**.

Also, theoretical/scientific justification is often necessary/advised.

# Simple Linear Regression

When there looks to be a linear relationship, we can use this model, the **simple linear regression model**.

$$Y = \beta_0 + \beta_1 x + \epsilon$$

and

1. $\epsilon \sim \mathcal{N}(0, \sigma^2)$ (does not depend on what $x$ is)

# Simple Linear Regression

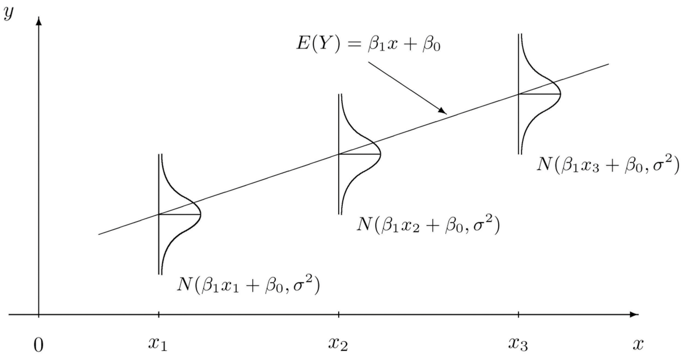
Remember we think of all the $X$ data and the coefficients as constants...

$$\begin{aligned} E[Y] &= E[\beta_0 + \beta_1 x + \epsilon] \\ &= \beta_0 + \beta_1 x + E[\epsilon] \\ &= \beta_0 + \beta_1 x \end{aligned}$$

$$\begin{aligned} V[Y] &= V[\beta_0 + \beta_1 x + \epsilon] \\ &= V[\epsilon] \\ &= \sigma^2 \end{aligned}$$

So now $Y$ is still normal, but has a mean that changes with an input

# Simple Linear Regression

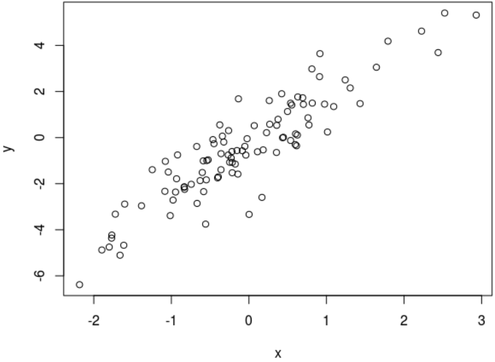



So if I tell you $X = 4$, you can answer questions like what's the probability that $Y > 5$...

Also, keep in mind $\beta_0$, $\beta_1$ and $\sigma^2$ are population parameters...we don't know them yet.

# Simple Linear Regression

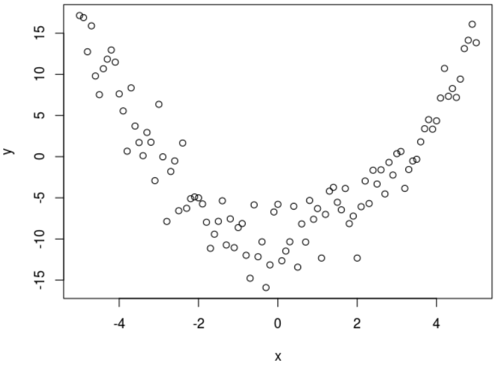
If our scatterplot looks like this:




then we suspect a good model fit. The output of the regression software will be the estimates $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$.

# Simple Linear Regression

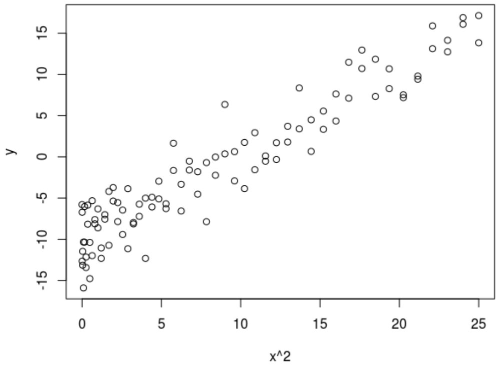
If our scatterplot looks like this:




then we shouldn't just plug in the x and y data into the regression software. What can we do?

# Simple Linear Regression

Instead of plugging in the data $\{x_i, y_i\}$, we should plug in the *transformed* data $\{x_i^2, y_i\}$! Here is the scatterplot of the transformed data:



# Logistic Regression

So far we've assumed $Y$ is normal. In particular it is continuous and takes values on $\mathbb{R}$.

What if $Y \in \{0, 1\}$? If $Y$ is a Bernoulli rv, it's average $p \in [0, 1]$, but $\beta_0 + \beta_1 x$ might not be...

# Logistic Regression

So far we've assumed $Y$ is normal. In particular it is continuous and takes values on $\mathbb{R}$.

What if $Y \in \{0, 1\}$? If $Y$ is a Bernoulli rv, it's average $p \in [0, 1]$, but $\beta_0 + \beta_1 x$ might not be...

Well we have to transform $p$ a bit. We still assume it depends on $x$. So let's call it $p(x)$ now.

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

# Logistic Regression

$\log(p/(1-p))$ is called the **logit function**. You can interpret it as the log of the odds ratio.

Its inverse is called the **logistic (or sigmoid or sigmoidal logistic, etc.)** function. You can write it like this

$$\frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}.$$

# Logistic Regression

So we can write our logistic model two different ways:

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

or

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

Note that we don't have any additive noise or $\epsilon$s around