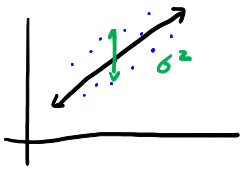
Linear Regression:
$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
$\varepsilon_i \sim N(0, \sigma^2)$

12 Logistic Regression

↳ Dichotomous Response

Recall the simple linear regression model:



$$Y = \beta_0 + \beta_1 x + \epsilon = E[Y] + \epsilon \tag{12.1}$$

with,

$$\epsilon \overset{\text{iid}}{\sim} N(0, \sigma^2).$$

Note: In model (12.1), the response variable (Y) is assumed to be a quantitative continuous variable. The covariate (x) can be quantitative or qualitative (dummy variable).

The distributional assumptions placed on the errors (ϵ) form the basis of all inferential procedures related to the simple linear regression model. More specifically, normality of the errors allows us to construct confidence intervals for $E[Y_p]$ and run testing procedures on the slope parameter β_1.

An equivalent form of model (12.1) is given by:

$$Y \overset{\text{iid}}{\sim} N(\beta_0 + \beta_1 x, \sigma^2). \tag{12.2}$$

Y is continuous.

- From model statements (12.1) and (12.2), notice that the mean of the response variable is equal to a deterministic function. That is,

$$E[Y] = \beta_0 + \beta_1 x.$$

- A statistical model can be specified by the response variable's mean and some distributional assumption.

Motivation of the logistic regression model

Question: How do we define a simple (one covariate) regression model that allows for a qualitative response variable?

To answer this question, first recall the Bernoulli random variable:

DEFINITION 12.1 Any rv whose possible values are 0 and 1 is called a **Bernoulli random variable**.

Also recall that the **expected value** (or true mean) of a Bernoulli random variable is its success probability. That is, if $Y \sim Bern(p)$, then

expected value

$$E[Y] = p$$

only yes or no

success probability

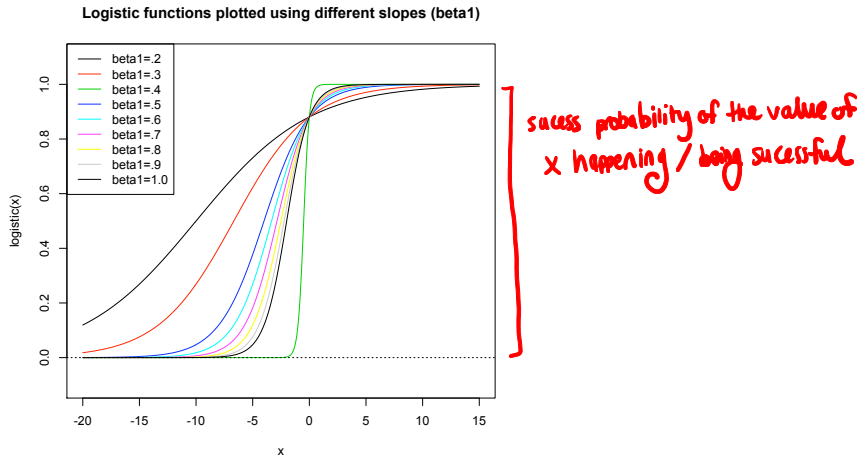
Logistic Regression
$Y_1, \ldots, Y_n$ Bernoulli with $E[Y_i] = p_i$

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \qquad \exp(a) = e^a$$

Answer: Regress a logistic function $p = f(x)$ on covariate x.

Note: The logistic function is bounded between 0 and 1 ($0 < f(x) < 1$).
The success probability of a Bernoulli random variable is bounded between 0 and 1 ($0 < p < 1$).

The logistic function



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

The logistic statistical model:

Let $Y_1, Y_2, \ldots, Y_n$ be independently distributed Bernoulli random variables with respective success probabilities $p_1, p_2, \ldots, p_n$. Then the **logistic regression model** is:

$$E[Y_i] = p_i = \frac{e^{(\beta_0+\beta_1 x_i)}}{1+e^{(\beta_0+\beta_1 x_i)}}, \quad i = 1, 2, \ldots, n. \tag{12.3}$$

The estimated logistic model:

$$\hat{p}_i = \frac{e^{(\hat{\beta}_0+\hat{\beta}_1 x_i)}}{1+e^{(\hat{\beta}_0+\hat{\beta}_1 x_i)}}, \quad i = 1, 2, \ldots, n. \tag{12.4}$$

Note: The quantities $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimated intercept and slope.

Note: Estimating model parameters β_1 and β_2 requires statistical software. (SPSS, SAS, R, Statcrunch, etc..)

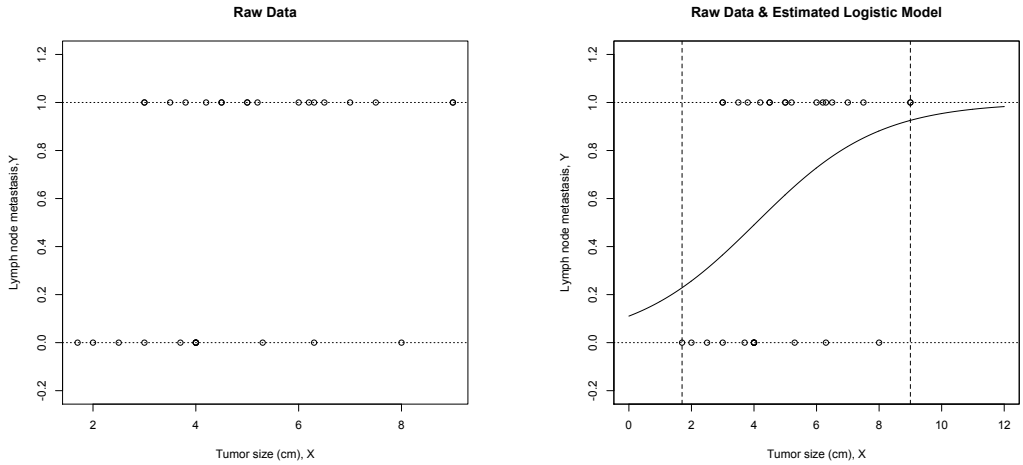
Note: Since the response variable Y is Bernoulli, the $Y_i's$ consist of zeros and ones.

Example 71

Esophageal cancer is a serious and very aggressive disease. Scientists conducted a study of 31 patients with esophageal cancer in which they studied the relationship between the size of the tumor that a patient had and whether or not the cancer had spread (metastasized) to the lymph nodes of the patient. In this study the response variable is dichotomous: $Y = 1$ if the cancer had spread to the lymph nodes and $Y = 0$ if not. The predictor variable is the size (recorded as the maximum dimension, in cm) of the tumor found in the esophagus. The data are given below:

Patient number	Tumor Size (cm), X	Lymph node metastasis, Y	Patient number	Tumor Size (cm), X	Lymph node metastasis, Y
1	6.5	1	17	6.2	1
2	6.3	0	18	2.0	0
3	3.8	1	19	9.0	1
4	7.5	1	20	4.0	0
5	4.5	1	21	3.0	1
6	3.5	1	22	6.0	1
7	4.0	0	23	4.0	0
8	3.7	0	24	4.0	0
9	6.3	1	25	4.0	0
10	4.2	1	26	5.0	1
11	8.0	0	27	9.0	1
12	5.2	1	28	4.5	1
13	5.0	1	29	3.0	0
14	2.5	0	30	3.0	1
15	7.0	1	31	1.7	0
16	5.3	0			

① Need to convert logistic variable to 1s & 0s. (Spreading or Not Spreading)
↳ Can Use If Else



Standard logistic regression output

	Estimate	Std. Error	z value	Pr($>|z|$) or Sig
(Intercept)	$\hat{\beta}_0$	$\sigma_{\hat{\beta}_0}$	$z_{calc} = \frac{\hat{\beta}_0}{\sigma_{\hat{\beta}_0}}$	P-value
x	$\hat{\beta}_1$	$\sigma_{\hat{\beta}_1}$	$z_{calc} = \frac{\hat{\beta}_1}{\sigma_{\hat{\beta}_1}}$	P-value

Example 71 continued

	Estimate	Std. Error	z value	Pr($>|z|$) or Sig
(Intercept)	-2.0858	1.2256	-1.702	0.0888
x	0.5117	0.2561	1.998	0.0457*

$$\hat{p} = \frac{\exp(-2.0856 + .517x)}{1 + \exp(-2.0856 + .517x)}$$

Estimate parameters

- Goal is to estimate β_0 & β_1
- The pmf of Y_i is $f(y_i|p_i) = p_i^{y_i}(1-p_i)^{1-y_i} = \begin{cases} p_i & y_i = 1 \\ 1-p_i & y_i = 0 \end{cases}$ piecewise

Note:

Regression

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon$ ← $E\varepsilon = 0$

$EY_i = \beta_0 + \beta_1 X_i$

link function is

$f(u) = u$

expected value of Bernoulli

- Set up the log likelihood

$$\ell(\beta_0, \beta_1) = \log\left[\prod_{i=1}^{n} f(y_i|\beta_0\beta_1)\right]$$
$$= \sum_{i=1}^{n} \log(f(y_i|\beta_0, \beta_1))$$
$$= \sum_{i=1}^{n} [y_i \log p_i + (1-y_i)\log(1-p_i)]$$
$$= \sum_{i=1}^{n} \left[y_i \log\left(\frac{p_i}{1-p_i}\right) + \log[1-p_i]\right]$$
$$= \sum_{i=1}^{n} [y_i(\beta_0+\beta_1 x_i) + \log(1+\exp(\beta_0+\beta_1 x_i))]$$

log odds

logit link

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

Example 71 continued

Find the estimated probability that cancer will spread to the lymph nodes for someone whose tumor is 7 cm.

Note: To be *good statistician*, we should always report the point estimate with its margin of error.

Odds and Log-Odds

Rearranging Equation (12.3) gives

$$\frac{p_i}{1-p_i} = e^{\beta_0+\beta_1 x_i}, \quad i = 1, 2, \ldots, n.$$

The equation above relates the *odds* of event $\{Y = 1\}$ occurring to a deterministic *exponential function*.

Taking the natural log of both sides gives:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i, \quad i = 1, 2, \ldots, n.$$

The equation above relates the *log-odds* of event $\{Y = 1\}$ occurring to a deterministic *linear function*.

Example 71 continued

Interpretation of the slope parameter β_1 (or $\hat{\beta}_1$)

Consider a 1 unit increase in the covariate.

1. The odds of event $\{Y = 1\}$ occurring when the covariate is fixed at x is:

$$odds_1 = \frac{p_1}{1 - p_1} = e^{\beta_0 + \beta_1(x)}$$

2. The odds of event $\{Y = 1\}$ occurring when the covariate is fixed at $x + 1$ is:

$$odds_2 = \frac{p_2}{1 - p_2} = e^{\beta_0 + \beta_1(x+1)}$$

3. Thus:

$$\text{“odds ratio”} = \Theta = \frac{odds_2}{odds_1} = \frac{e^{\beta_0 + \beta_1(x+1)}}{e^{\beta_0 + \beta_1 x}} = \frac{e^{\beta_0 + \beta_1 x + \beta_1}}{e^{\beta_0 + \beta_1 x}} = \frac{e^{\beta_0 + \beta_1 x} e^{\beta_1}}{e^{\beta_0 + \beta_1 x}} = e^{\beta_1}$$

4. Equivalently:

$$odds_2 = e^{\beta_1} \cdot (odds_1)$$

Two equivalent interpretations of $\hat{\beta}_1$

- *“The odds ratio of event $\{Y = 1\}$ occurring for covariate fixed at $x+1$ verses covariate fixed at x is equal to $e^{\hat{\beta}_1}$.”*
- *“The odds of event $\{Y = 1\}$ occurring are multiplied by $e^{\hat{\beta}_1}$ for every 1 unit increase in x.”*

Inference on the regression parameter β_1

Consider testing the null hypothesis:

$$H_0 : \beta_1 = 0$$

The statistic used for this hypothesis test is:

$$z_{calc} = \frac{\hat{\beta}_1}{\sigma_{\hat{\beta}_1}}$$

The rejection rules and p-value computations are the same as any z-test.

Note: If $H_0 : \beta_1 = 0$ is true, then $\Theta = e^{\beta_1} = e^0 = 1$.

Example 71 continued

Logistic regression with a binary covariate

Let $Y_1, Y_2, \ldots, Y_n$ be independently distributed Bernoulli random variables with respective success probabilities $p_1, p_2, \ldots, p_n$. Also suppose that $x_1, x_2, \ldots, x_n$ is a sequence of 0's and 1's (the covariate is a binary variable). For this situation, the simple logistic regression model is:

$$E[Y_i] = p_i = \frac{e^{(\beta_0+\beta_1 x_i)}}{1+e^{(\beta_0+\beta_1 x_i)}}, \quad i = 1, 2, \ldots, n,$$

where

$$x_i = \begin{cases} 1 & \text{if the } i^{th} \text{ case is in the exposed group} \\ 0 & \text{if the } i^{th} \text{ case is in the unexposed group.} \end{cases}$$

Note: In this model, both the x and y values consist of 0's and 1's.

Note: A data set of this nature can be organized with a two-by-two table.

Example 72

Note: This is Example 60 from the categorical data analysis section.

The health histories of 11,900 middle-aged men were tracked over many years. During the study 126 of the men developed lung cancer, including 89 men who were smokers and 37 men who were former smokers. The following contingency table summarizes the data:

		lung cancer	No lung cancer
Smoker	Yes	89	6,063
	No	37	5,711

The response variable and covariate are respectively defined by:

$$y_i = \begin{cases} 1 & \text{if the } i^{th} \text{ case has lung cancer} \\ 0 & \text{if the } i^{th} \text{ case does not have lung cancer} \end{cases} \qquad x_i = \begin{cases} 1 & \text{if the } i^{th} \text{ case is a current smoker} \\ 0 & \text{if the } i^{th} \text{ case is not a current smoker} \end{cases}$$

Logistic output

	Estimate	Std. Error	z value	Pr($>\|z\|$) or Sig
(Intercept)	−5.0392	0.1649	−30.559	2e-16 ***
x	0.8179	0.1965	4.163	3.13e-5 ***