Lecture 8: Simple Logistic Regression
STAT 452, Spring 2021

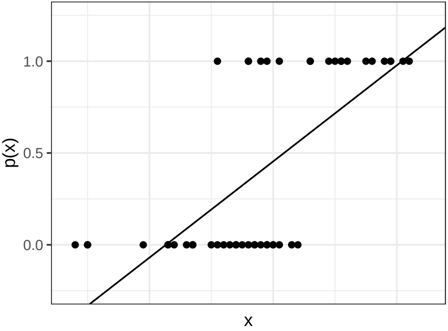
Simple Logistic Regression

- Simple logistic regression is a method to model a binary response variable, $y \in \{0, 1\}$, using a single predictor variable x.
- Specifically, the method models $p(x) = Pr(y = 1|x)$, the probability $y = 1$ given predictor x.

Simple Logistic Regression

Why not use linear regression to represent these probabilities?

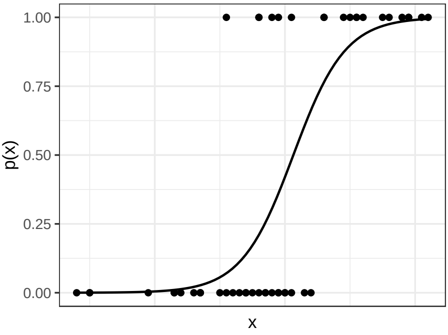
$$p(x) = Pr(y = 1|x) = \beta_0 + \beta_1 x$$



Simple Logistic Regression

The **logistic function** is commonly used to model $p(x)$ since it always gives outputs between 0 and 1.

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



Simple Logistic Regression

Two ways to express the simple logistic regression model:

Probability form:

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

which can be interpreted as the probability $y = 1$ for a given value x of the predictor.

Logit form:

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

The left-hand side is called the *logit* or *log-odds*. Logistic regression expressed in terms of the logit is linear in its parameters.

Simple Logistic Regression

Some algebraic manipulation can be used to show that the two representations are equivalent:

$$
\begin{aligned}
p &= \frac{1}{1 + e^{-\beta_0 - \beta_1 x}} \\
\frac{1}{p} &= 1 + e^{-\beta_0 - \beta_1 x} \\
\frac{1 - p}{p} &= e^{-\beta_0 - \beta_1 x} \\
\frac{p}{1 - p} &= e^{\beta_0 + \beta_1 x} \\
\log\left(\frac{p}{1 - p}\right) &= \beta_0 + \beta_1 x
\end{aligned}
$$

Here we are letting $p = p(x)$ to simplify notation.

Inference

Hypothesis test for β_1:
$H_0 : \beta_1 = 0$
$H_A : \beta_1 \neq 0$

Test statistic:

$$z = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$$

This is sometimes referred to as the Wald z-statistic.

A $1 - \alpha$ confidence interval for β_1:

$$\hat{\beta}_1 \pm z_{\alpha/2} se(\hat{\beta}_1)$$

Example

- Here we consider a data set form the National Health and Nutrition Examination Survey (NHANES).[1]
- The data set can be accessed from the R package `NHANES`, and documentation is available in the help menu (type `help(NHANES)`).
- For this example, we fit a simple logistic regression model to predict the probability that a person has diabetes using BMI (body mass index) as an explanatory variable.[2]

[1] https://www.cdc.gov/nchs/nhanes/about_nhanes.htm

[2] https://en.wikipedia.org/wiki/Body_mass_index

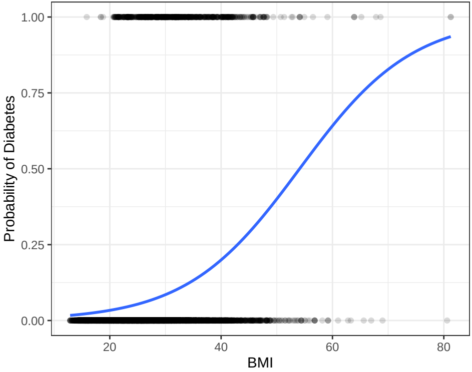
```
> library(tidyverse)
> library(NHANES)

# pre-processing:
# 1) remove missing data
# 2) recode Diabetes (1=Yes, 0=No)
> nhanes2 <- NHANES %>%
    select(Diabetes, BMI) %>%
    na.omit() %>%
    mutate(Diabetes = ifelse(Diabetes == "Yes", 1, 0))

> table(nhanes2$Diabetes)
   0    1
8880  749

> table(nhanes2$Diabetes) / nrow(nhanes2)
         0          1
0.92221414 0.07778586
```

```
ggplot(nhanes2, aes(x = BMI, y = Diabetes)) + geom_point(alpha = 0.15) +
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se=F) +
  labs(x = "BMI", y = "Probability of Diabetes") + theme_bw()
```



```
> glm1 <- glm(Diabetes ~ BMI, family = "binomial", data = nhanes2)


> summary(glm1)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.33055    0.15305  -34.83   <2e-16 ***
BMI          0.09853    0.00475   20.75   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


> confint(glm1, level = 0.95)
                  2.5 %     97.5 %
(Intercept) -5.63383508 -5.0336943
BMI          0.08927634  0.1079021
```

The fitted logistic regression model in terms of the logit:

$$\log\left(\frac{\hat{p}(x)}{1-\hat{p}(x)}\right) = \hat{\beta}_0 + \hat{\beta}_1 x = -5.3305 + 0.0985x$$

For a person with a BMI $= 30$, the prediction for the logit is

$$-5.3305 + 0.0985(30) = -2.3755$$

The fitted logistic regression model in probability from:

$$\hat{p}(x) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}} = \frac{e^{-5.3305 + 0.0985x}}{1 + e^{-5.3305 + 0.0985x}}$$

For a person with BMI $= 30$ the prediction for the probability of having diabetes is

$$\hat{p}(30) = \frac{e^{-5.3305 + 0.0985(30)}}{1 + e^{-5.3305 + 0.0985(30)}} = \frac{e^{-2.3755}}{1 + e^{-2.3755}} = 0.08506$$

In R, the prediction for the logit can be obtained with the command:

```
> new_x <- data.frame(BMI = 30)
> predict(glm1, newdata = new_x)
        1
-2.374496
```

The prediction for the probability can be obtained with the command

```
> predict(glm1, newdata = new_x, type="response")
         1
0.08513827
```

Any difference from the manual calculations are due to rounding.

Interpreting the Coefficients

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

In terms of the *logit* we have the following interpretation:

An one unit increase in x is associated with a change in the log-odds, or logit, by β_1.

Going back to the example, a one-unit increase in `BMI` is associated with a $\hat{\beta}_1 = 0.0985$ increase in the log-odds.

Interpreting Coefficients

More intuitively, the sign of β_1 has meaningful interpretation:

- If $\beta_1 > 0$, then increasing x will be associated with increasing the probability $p(x)$.
- If $\beta_1 < 0$, then increasing x will be associated with decreasing the probability $p(x)$.

Your Turn

(a) Fit a logistic regression model for `Diabetes` using `Age` as a predictor. Use `ggplot2` to plot the fitted logistic regression curve.

(b) What is predicted probability that a 30 year old has diabetes? What is the predicted probability that a 60 year old has diabetes?