Lecture 9: Multiple Logistic Regression
STAT 452, Spring 2021

Multiple Logistic Regression

- Multiple logistic regression is a method to model a binary response variable, $y \in \{0, 1\}$, using predictor variables $x_1, x_2, \cdots, x_p$.
- Specifically, the method models $p(\boldsymbol{x}) = Pr(y = 1|\boldsymbol{x})$, the probability $y = 1$ given predictors $\boldsymbol{x} = \begin{pmatrix} x_1 & x_2 & \cdots & x_p \end{pmatrix}'$.

Multiple Logistic Regression

Two ways to express multiple logistic regression model:

Probability form:

$$p(\boldsymbol{x}) = Pr(y = 1|\boldsymbol{x}) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}} = \frac{1}{1 + e^{-\beta_0-\beta_1 x_1-\cdots-\beta_p x_p}}$$

Logit form:

$$\log\left(\frac{p(\boldsymbol{x})}{1 - p(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

Example

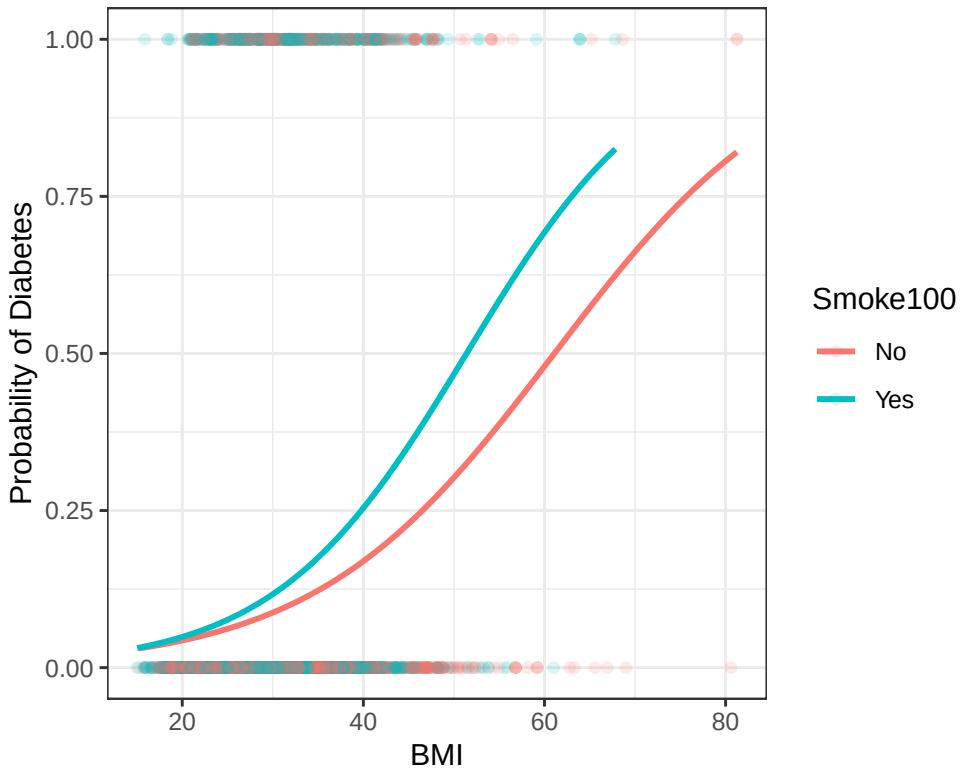
Fit a multiple logistic regression model with `Diabetes` (1=Yes, 0=No) as the response variable, and the following predictors:

- `BMI`: body mass index
- `Smoke100`: participant has smoked at least 100 cigarettes (Yes/No)

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

# pre-processing:
# 1) remove missing data
# 2) recode Diabetes (1=Yes, 0=No)
> nhanes2 <- NHANES %>%
    select(Diabetes, BMI, Smoke100) %>%
    na.omit() %>%
    mutate(Diabetes = ifelse(Diabetes == "Yes", 1, 0))
```

```
ggplot(nhanes2, aes(x = BMI, y = Diabetes, color = Smoke100)) +
  geom_point(alpha = 0.15) +
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se=F) +
  ylab("Probability of Diabetes") + theme_bw()
```



```
> glm2 <- glm(Diabetes ~ BMI + Smoke100,
              family = "binomial", data = nhanes2)
> summary(glm2)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.866222   0.177946 -27.347  < 2e-16 ***
BMI          0.083493   0.005227  15.975  < 2e-16 ***
Smoke100Yes  0.351101   0.080231   4.376 1.21e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> confint(glm2, level = 0.95)
                   2.5 %      97.5 %
(Intercept) -5.21818545 -4.52044380
BMI          0.07328657  0.09378203
Smoke100Yes  0.19392476  0.50853544
```

The equation for the fitted logistic regression model in probability form:

$$\hat{p}(x_1, x_2) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2}} = \frac{e^{-4.866+0.083 x_1+0.351 x_2}}{1+e^{-4.866+0.083 x_1+0.351 x_2}}$$

For a person with `BMI = 30` and `Smoke100 = Yes (1)`, the predicted probability of diabetes is

$$\hat{p}(30, 1) = \frac{e^{-4.866+0.083(30)+0.351(1)}}{1+e^{-4.866+0.083(30)+0.351(1)}} = 0.12$$

For a person with `BMI = 30` and `Smoke100 = No (0)`, the predicted probability of diabetes is

$$\hat{p}(30, 0) = \frac{e^{-4.866+0.083(30)+0.351(0)}}{1+e^{-4.866+0.083(30)+0.351(0)}} = 0.09$$

To make predictions for the logistic probabilities in R:

```
> new_x <- data.frame(BMI = c(20, 30, 40), Smoke100 = "Yes")
> predict(glm2, newdata = new_x, type = "response")
         1          2          3
0.05492645 0.11812142 0.23587759
```

```
> new_x <- data.frame(BMI = c(20, 30, 40), Smoke100 = "No")
> predict(glm2, newdata = new_x, type = "response")
         1          2          3
0.03930259 0.08616053 0.17850395
```

Your Turn

(a) Fit a mutiple logistic regression model for `Diabetes` using `BMI` and `Age` as predictors. Are both predictors significant in the model?

(b) What is predicted probability that a 30 year-old person with a BMI $= 25$ has diabetes? What is the predicted probability that a 60 year-old person with a BMI $= 25$ has diabetes?