# Logistic Regression

Professor: Hammou El Barmi
Baruch College, CUNY

- At this point we have covered:
  - Simple linear regression: Relationship between numerical response and a numerical or categorical predictor
  - Multiple regression: Relationship between numerical response and multiple numerical and/or categorical predictors
- What we have not seen is what to when the response is categorical
- Logistic regression is a method used to model a binary categorical variable ( (Yes, No), (0, 1), (approve, does not approve) ) using numerical and categorical variables
- Example: ) For the 23 space shuttle flights that occurred before the Challenger mission disaster in 1986, the data below shows the temperature in fahrenheit at the time of the flight and whether at least one primary O-ring suffered thermal distress (Yes=1 and 0=No)

| Flight | Temperature (x) | ThermalDistress (y) |
|---|---|---|
| 1 | 66 | 0 |
| 2 | 70 | 1 |
| 3 | 69 | 0 |
| 4 | 68 | 0 |
| 5 | 67 | 0 |
| 6 | 72 | 0 |
| 7 | 73 | 0 |
| 8 | 70 | 0 |
| 9 | 57 | 1 |
| 10 | 63 | 1 |
| 11 | 70 | 1 |
| 12 | 78 | 0 |
| 13 | 67 | 0 |
| 14 | 53 | 1 |
| 15 | 67 | 0 |
| 16 | 75 | 0 |
| 17 | 70 | 0 |
| 18 | 81 | 0 |
| 19 | 76 | 0 |
| 20 | 79 | 0 |
| 21 | 75 | 1 |
| 22 | 76 | 0 |
| 23 | 58 | 1 |

- Clearly we can not use simple linear

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

since y is yes or no. In stead we model the odds of the event (y = 1).

- What are the Odds? The odds are another way of quantifying the probability of an event (commonly used in gambling and logistic regression)
- For some event E,

$$odds(E) = P(E)/(1 - P(E)) = P(E)/P(E^c)$$

- Usually we are told that the odds of E are x to y, then

$$odds(E) = x/y = \frac{x/(x+y)}{y/(x+y)}$$

which implies that

$$P(E) = \frac{x}{x+y} \quad \text{and} \quad P(E^c) = \frac{y}{x+y}$$

- Assume that we have only one predictor x and let

$$\pi(x) = P(y = 1|x)$$

in which case

$$odds(y = 1|x) = \frac{P(y = 1|x)}{1 - P(y = 1|x)} = \frac{\pi(x)}{1 - \pi(x)}$$

- In logistic regression we assume that the odds of the event (y=1) satisfy

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

- call

$$log\left(\frac{\pi(x)}{1 - \pi(x)}\right)$$

$logit(\pi(y = 1|x))$.

- The model implies that
  1. $$odds(y = 1|x) = e^{\beta_0+\beta_1 x}$$
  2. $$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$
- Interpretation of the coefficients;
  1. $\beta_0 = odds(y = 1|x = 0)$ or equivalently
     $$\pi(0) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$$
  2. $\beta_1$: when we increase x by 1, the odds of (y=1) change by a multiplicative factor $e^{\beta_1}$.
     Why
     The odds of (y=1) at is
     $$odds(y = 1|x) = e^{\beta_0+\beta_1 x}$$
     and when we increase by one, the odds of (y=1) is
     $$odds(y = 1|x + 1) = e^{\beta_0+\beta_1(x+1)} = e^{\beta_1} e^{\beta_0+\beta_1 x} = e^{\beta_1} odd(y = 1|x)$$

## Simple logistic regression

- To estimate $\beta_0$ and $\beta_1$, we use a method called maximum likelihood estimation. Basically we seek the values of $\beta_0$ and $\beta_1$ that maximize the likelihood of observing that the data that we have observed.
- The estimates of $\beta_0$ and $\beta_1$ are denoted by $b_0$ and $b_1$ and fitted model is

$$logit(\hat{\pi}(x)) = b_0 + b_1 x$$

- $e^{b_0}$ is the estimated odds of (y=1) when x=0
- If we increase x by 1, the odds of (y=1) change by a multiplicative factor of about $e^{b_1}$.
- In R we fit logistic regression in the same way as we did in linear regression except that we use glm instead of lm.

# Example (continued

```
> fit<- glm(ThermalDistress~Temperature, family = binomial)
> fit

Call:  glm(formula = ThermalDistress ~ Temperature, family = binomial)

Coefficients:
(Intercept)  Temperature
    15.0429      -0.2322

Degrees of Freedom: 22 Total (i.e. Null);  21 Residual
Null Deviance:      28.27
Residual Deviance: 20.32  AIC: 24.32
```

- The fitted model is

$$logit(\hat{\pi}(x)) = 15.0429 - 0.2322x$$

# Example(continued)

- When the temperature increases by one fahrenheit, the odds of (y=1) ( at least one primary O-ring suffered thermal distress) change by a multiplicative factor of about $e^{-0.2322} = 0.7927875$ (a decrease of about 20%).
- Suppose we asked to test the probability that least one primary O-ring suffers thermal distress when the temperature is 70 fahrenheit

$$\hat{\pi}(70) = \frac{e^{15.0429 - 0.2322(70)}}{1 + e^{15.0429 - 0.2322(70)}} = 0.2295065$$

- A $(1-\alpha)$ confidence interval for $\beta_1$ is

$$b_1 \pm Z_{\alpha/2} s_{b_1}$$

- and $(1-\alpha)$ confidence interval for $e^{\beta_1}$ is

$$[e^{b_1 - Z_{\alpha/2} s_{b_1}}, e^{b_1 + Z_{\alpha/2} s_{b_1}}].$$

- Interpretation: We are $(1-\alpha)$ confident that when we increase $x$ by 1, the odds of (y=1) change by a multiplicative factor between $e^{b_1 - Z_{\alpha/2} s_{b_1}}$ and $e^{b_1 + Z_{\alpha/2} s_{b_1}}$.

- Construct a 95% confidence interval for $\beta_1$.
- Output from R

```
> summary(fit)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  15.0429     7.3786   2.039   0.0415 *
Temperature  -0.2322     0.1082  -2.145   0.0320 *
```

- Answer: from the output we see that $s_{b_1} = 0.1082$. The confidence interval is given by

$$b_1 \pm Z_{\alpha/2} s_{b_1} = -0.2322 \pm 1.96(0.1082) = [-0.444272, -0.020128]$$

- Interpretation: We are 95% confident that when we increase the temperature by on 1 fahrenheit, the odds of at least primary O-ring suffers thermal distress change by a multiplicative factor between $e^{-0.444272} = 0.641291$ and $e^{-0.020128} = 0.9800732$

- Suppose we want to test $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$,
- The test statistic is

$$Z = \frac{b_1 - 0}{s_{b_1}}$$

- We reject $H_0$ if $|Z| > Z_{\alpha/2}$ or if p-value$< \alpha$.
- Example(Continued)
  1. Test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ at $\alpha = 0.05$.
  2. Answer: We have

  $$Z = \frac{-0.2322 - 0}{0.1082} = -2.145.$$

  Since $| - 2.145| > 1.96$, we reject $H_0$..
  3. As can be seen also in the output, the p-value $= 0.0320$. Since it is less that 0.05, we reject $H_0$.

# Multiple Logistic Regression

- Assume that $x_1, x_2, \ldots, x_k$ are the predictor variables.
- The model we use it

$$logit(\pi(x_1, x_2, \ldots, x_k)) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

- Interpretation of the coefficient $\beta_i, i = 1, 2, \ldots, k$.
- If we increase $x_i$ by one while holding the other xs fixed, the odds of (y=1) change by a multiplicative factor $e^{\beta_i}$

- We estimate $\beta_0, \beta_1, \ldots, \beta_1$ by $b_0, b_1, \ldots, b_k$ and the fitted model is

$$logit(\hat{\pi}(x_1, x_2, \ldots, x_k)) = b_0 + b_1 x_1 + \ldots + b_k x_k$$

- Interpretation of the coefficient $b_i, i = 1, 2, \ldots, k$.
  If we increase $x_i$ by one while holding the other xs fixed, the odds of (y=1) change by a multiplicative factor of about $e^{\beta_1}$
- The estimated probability of (y=1) at $x_1, x_2, \ldots, x_k$ is

$$\hat{\pi}(x_1, x_2, \ldots, x_k) = \frac{e^{b_0 + b_1 x_1 + \ldots + b_k x_k}}{1 + e^{b_0 + b_1 x_1 + \ldots + b_k x_k}}$$

- A $(1-\alpha)$ confidence interval for $\beta_i$ is

$$b_i \pm Z_{\alpha/2} s_{b_i}$$

- and $(1-\alpha)$ confidence interval for $e^{\beta_i}$ is

$$[e^{b_i - Z_{\alpha/2} s_{b_i}}, e^{b_i + Z_{\alpha/2} s_{b_i}}].$$

- Interpretation: We are $(1-\alpha)$ confident that when we increase $x_1$ by 1 while holding all the other xs fixed, the odds of (y=1) change by a multiplicative factor between $e^{b_i - Z_{\alpha/2} s_{b_i}}$ and $e^{b_i + Z_{\alpha/2} s_{b_i}}$.

The following data (described in New York Times, Feb. 15, 1191) is used to study the effect of AZT in slowing the development of AIDS symptoms. In the study 338 veterans whose immune systems we beginning to falter after infection with AIDS virus were randomly assigned wither to receive AZT immediately or to wait until their T cells showed severe immune weakness. The data is a 2x2x2 cross classification of the veterans' race, whether they received AZT immediately and whether they developed AIDS symptoms during the three year study.

```
> aids<-read.csv("C:\\Users\\helbarmi\\Desktop\\deathpenalty.csv",
header=TRUE, sep=',')
> attach(aids)
> aids
  race AZTuse yes no
1    w    yes  14 93
2    w     no  32 81
3    b    yes  11 52
4    b     no  12 43
```

The model we want to use here is

$$\text{logit}(P(yes|\text{race, AZTuse})) = \beta_0 + \beta_1\text{race} + \beta_2\text{AZTuse}$$

To fit this model in R, we use

```
> logit1<-glm(cbind(yes, no)~factor(race)+factor(AZTuse), family=binomial)
> logit1

Call:  glm(formula = cbind(yes, no) ~ factor(race) + factor(AZTuse),
    family = binomial)

Coefficients:
      (Intercept)      factor(race)w  factor(AZTuse)yes
         -1.07357            0.05548           -0.71946

Degrees of Freedom: 3 Total (i.e. Null);  1 Residual
Null Deviance:      8.35
Residual Deviance: 1.384  AIC: 24.86
```

Interpretation of the result:

1. Interpretation of $b_1$ the estimate of $\beta_1$.: If we hold AZTuse fixed (i.e controlling for AZT use), we estimate the odds that a white person develops AIDS symptoms to be $e^{0.05548} = 1.057$ times the odds that a back person does (a 5.7% increase roughly)
2. Interpretation of $b_2$ the estimate of $\beta_2$.: If we hold race fixed (i.e controlling for race), we estimate the odds that a person who takes AZT develops AIDS symptoms to be $e^{-0.71946} = 0.49$ times the odds that a person does who does not(a 50% decrease roughly)

You can compute these numbers using

```
> OR=exp(coef(logit1))
> OR
     (Intercept)      factor(race)w factor(AZTuse)yes
       0.3417849          1.0570527         0.4870152
```