## LOGISTIC REGRESSION

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

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

- Similarly, if we are told the odds of E are x to y, then

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

which implies that

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

- Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical variables
- We assume a binomial distribution produced the outcome variable and we therefore want to model $\pi$ the probability of success for a given set of predictors
- It turns out that there is a very general way of addressing this type of a problem and the resulting models are called generalized linear models. Logistic regression is just one example of this type of model
- All generalized linear models has the following three characteristics:

  1. A probability distribution describing the outcome variable

2. A linear model

$$\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

3. A link function that relates the linear model to the parameter of the outcome distribution

$$g(\pi) = \eta \quad \text{or} \quad \pi = g^{-1}(\eta)$$

- Logistic regression is a GLM used to model a binary categorical variable using numerical and categorical predictors

- We assume a binomial distribution produced the outcome variable therefore we want to model $\pi$, the probability of success, as a function of some predictors.

- There are a variety of reasonable link functions to use to connect $\pi$ and $\eta$, One such function that is commonly used is the logit function

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi)}\right), \quad 0 < \pi < 1.$$

- The logit fuction takes a value between 0 and 1 and maps it to a value between $-\infty$ and $+\infty$.

- The inverse logit (logistic) function is

$$g^{-1}(x) = \frac{e^x}{1+e^x}$$

- The inverse logit function takes a value between $-\infty$ and $\infty$ and maps it to a value between 0 and 1

- This formulation also some use when it comes to interpreting the model a logit can be interpreted as a the log odds of success

- The assumptions are

$$\begin{aligned} y|x_1, x_2, \ldots, x_p &= \begin{cases} 1 & \text{with probability } \pi(x_1, x_2, \ldots, x_p) \\ 0 & \text{with probability } 1 - \pi(x_1, x_2, \ldots, x_p) \end{cases} \\ \text{logit}(\pi(x_1, x_2, \ldots, x_p) &= \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \end{aligned}$$

- This implies that

$$\pi(x_1, x_2, \ldots, x_p) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}$$

- Also

$$\frac{\pi(x_1, x_2, \ldots, x_p)}{1-\pi(x_1, x_2, \ldots, x_p)} = e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}$$

- This implies that when we increase $x_i$ by one while holding all the other xs fixed, the odds of getting of 1 change by a multiplicative factor equal to $e_i^\beta$.

- In R we fit a GLM model in the same way as we did in linear regression except that we use glm instead of lm and we must specify the type of GLM to fit using the family argument.

- The data include the number of students admitted, the total number of applicants broken down by gender (the variable female), and whether or not they had taken AP calculus (the variable apcalc). Since the dataset is so small, we will read it in directly.

```
Gender= 0 male 1 female, AP = 1 took AP calculus, 0 did not.
 Admit =1 admitted 0 not admitted
Gender  AP  Admit
   0  0  0
   0  0  0
   0  0  0
   0  0  0
   0  0  0
   0  0  0
   0  0  0
   0  0  1
   0  1  0
   0  1  0
   0  1  0
   0  1  1
   0  1  1
   0  1  1
```

```
0  1  1
0  1  1
0  1  1
0  1  1
1  0  0
1  0  0
1  0  0
1  0  0
1  0  0
1  0  1
1  1  0
1  1  1
1  1  1
1  1  1
1  1  1
```

```
 > glm(Admit~Gender+AP, family = binomial("logit"))

Call:  glm(formula = Admit ~ Gender + AP, family = binomial("logit"))

Coefficients:
(Intercept)       Gender           AP
    -2.0043       0.4537       2.8755

Degrees of Freedom: 28 Total (i.e. Null);  26 Residual
Null Deviance:     39.89
Residual Deviance: 28.67  AIC: 34.67


 > summary(glm(Admit~Gender+AP, family = binomial("logit")))

Call:
glm(formula = Admit ~ Gender + AP, family = binomial("logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7667  -0.6203  -0.5028   0.8361   2.0643
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.0043     0.9170  -2.186  0.02884 *
Gender        0.4537     0.9908   0.458  0.64700
AP            2.8755     0.9898   2.905  0.00367 **
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 39.892  on 28  degrees of freedom
Residual deviance: 28.666  on 26  degrees of freedom
AIC: 34.666
Number of Fisher Scoring iterations: 4
```

To test $H_0 : \beta_i = 0$ against $H_a : \beta_i \neq 0$, the test statistics is

$$Z = \frac{b_i - 0}{SE(b_i}$$

and we reject $H_0$ if $|Z| > Z_{\alpha/2}$ or if $p - value < \alpha$. Example: Test $H_0 : \beta_{Gender} = 0$ against $H_a : \beta_{Gender} \neq 0$.. The test statistic is

$$Z = \frac{0.4537 - 0}{0.9908} = 0.458$$

If $\alpha = 0.05$ then $Z_{0.025} = 1.96$. Since $|0.458| < 1.96$, we fail to fail to reject $H_0$.

```
> confint(glm(Admit~Gender+AP, family = binomial("logit")))
                2.5 %    97.5 %
(Intercept) -4.206356 -0.450995
Gender      -1.456204  2.605742
AP           1.115573  5.130797
```

```
Probit Model
```

```
> glm(Admit~Gender+AP, family = binomial("probit"))

Call:  glm(formula = Admit ~ Gender + AP, family = binomial("probit"))

Coefficients:
(Intercept)       Gender           AP
    -1.1848       0.2561       1.7276

Degrees of Freedom: 28 Total (i.e. Null);  26 Residual
Null Deviance:     39.89
Residual Deviance: 28.67  AIC: 34.67
```