Eco311 Optional Reading: Logistic Regression

(Jing Li, Miami University)

1. This note focuses on explaining a **binary outcome** variable that can only take two values of 1 and 0. For example, $y = 1$ if a student is obese, and $y = 0$ if not obese.

2. Part of the obesity data looks like

```
> library("readxl")
> data = read_excel("obesity.xlsx")
> attach(data)
> cbind(y,x1,x14,x15)[1:6,]
     y x1 x14 x15
[1,] 0  0   1   0
[2,] 0  0   1   0
[3,] 0  0   0   0
[4,] 1  0   1   0
[5,] 1  1   0   1
[6,] 1  0   0   0
```

   All those variables shown here are binary.

3. First, let's examine how y and x_1 are related using a **two-way table**

```
> table(y, x1)
   x1
y       0     1
  0 13071  5699
  1  1663   699
> odds_x0 = 1663/13071
> odds_x1 = 699/5699
> odds_x0
[1] 0.1272282
> odds_x1
[1] 0.1226531
```

13071 students have $y = 0, x_1 = 0$; while 1663 students have $y = 1, x_1 = 0$. So for those students with $x_1 = 0$, the odds of $y = 1$ is $1663/13071 = 0.1272282$. For those students with $x_1 = 1$, the odds of $y = 1$ is $699/5699 = 0.1226531$. The two odds are very *close*, indicating that x_1 may not matter for y.

4. We can formally test the **null hypothesis** that x_1 and y are **independent** using the **chi-squared test**

```
> chisq.test(x1,y,correct=FALSE)
Pearson's Chi-squared test
data:  x1 and y
X-squared = 0.58729, df = 1, p-value = 0.4435
```

   We cannot reject the independence hypothesis since p value 0.4435 exceeds 0.05.

5. Now we switch to x_{14} and y :

```
> table(y, x14)
   x14
y       0     1
  0  4274 14496
  1   443  1919
> chisq.test(x14,y,correct=FALSE)
Pearson's Chi-squared test
data:  x14 and y
X-squared = 19.506, df = 1, p-value = 1.003e-05
> cor(x14, y)
[1] 0.0303818
```

   We see that the two odds $443/4274 = 0.10365$ and $1919/14496 = 0.1323813$ are not that close, and the p value of chi-squared test is less than 0.05. Actually the odds of $y = 1$ rises from 0.10365 to 0.1323813 when x_{14} changes from 0 to 1. So the two variables are *positively* related. Here x_{14} equals one if a student is a single child in a family. So a single child is *more* likely to be obese than a non-single child.

6. The limitation of two-way table and chi-squared test is that they can only look at two variables, or in other words, they cannot control for other variables. A more general and better approach is running a logistic regression.

7. Let y be the outcome variable and x be the covariate or independent variable. **Logistic regression** (or logit model) specifies the two probabilities for y as

$$P(y=1) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{1}$$

$$P(y=0) = 1 - P(y=1) = \frac{1}{1+e^{\beta_0+\beta_1 x}} \tag{2}$$

We can verify that $0 < \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} < 1$, so this particular function imposes the 0-1 boundary for probability. Actually, the function $\frac{e^z}{1+e^z}$ is the cumulative distribution function for the logistic distribution.

8. Based on those two probabilities we can show the **odds** of $y = 1$ is defined as

$$\texttt{odds} \equiv \frac{P(y=1)}{P(y=0)} = e^{\beta_0+\beta_1 x} \tag{3}$$

We get **log odds** (called logit) after taking natural log

$$\texttt{log odds} = \beta_0 + \beta_1 x \tag{4}$$

This is an example of **generalized linear model** (GLM) since the term on the right hand side $\beta_0 + \beta_1 x$ is a linear model.

9. Based on (4), one interpretation is that β_1 measures the **change in log odds** when x changes from 0 to 1

$$\beta_1 = \frac{\Delta\,\texttt{log odds}}{\Delta\, x} \tag{5}$$

For instance, when x_1 changes from 0 to 1, the change in log odds is

```
> log(odds_x1)-log(odds_x0)
[1] -0.03662242
```

That value is the same as the $\hat{\beta}_1$ reported by R **glm** function

```
> logit_model_1 = glm(formula = y ~ x1, family = "binomial")
> summary(logit_model_1)$coef
               Estimate Std. Error     z value  Pr(>|z|)
(Intercept) -2.06177283 0.02603462 -79.1934954 0.0000000
x1          -0.03662242 0.04778901  -0.7663355 0.4434767
```

Thus we conclude that the log odds of being obese *falls* by 0.03662242 when x_1 changes from 0 to 1. However, that change is statistically insignificant since z value -0.766335 is less than 1.96 in absolute value, or p value 0.4434767 is greater than 0.05. This finding is consistent with the chi-squared test. They all indicate that x_1 does not matter for y.

10. We get interpretation for β_0 by setting $x = 0$ in (4)

$$\beta_0 = \texttt{log odds when x=0} \tag{6}$$

11. There is an alternative interpretation for β_1, which involves **odds ratio**

$$\texttt{odds ratio} \equiv \frac{\texttt{odds when x=1}}{\texttt{odds when x=0}} = \frac{e^{\beta_0+\beta_1(1)}}{e^{\beta_0+\beta_1(0)}} = e^{\beta_1} \tag{7}$$

where the second equality is due to (3). To summarize, computing e^{β_1} provides the odds ratio.

12. Go back to our example. The odds ratio is 0.9640401, which is the same as $e^{-0.03662242}$

```
> odds_x1/odds_x0
[1] 0.9640401
> exp(-0.03662242)
[1] 0.9640401
```

Recall that $\frac{A}{B} < 1$ implies $A < B$. This odds ratio 0.9640401 is *less* than 1, so odds of $y = 1$ when $x_1 = 1$ is *less* than the odds when $x_1 = 0$. In other words, changing x_1 from 0 to 1 *reduces* the odds of $y = 1$. The two variables are negatively (but not significantly) related.

13. Finally, let's run a logistic regression using covariates x_1, x_{14} and x_{15}

```
> logit_model_final = glm(formula = y ~ x1+x14+x15, family = "binomial")
> summary(logit_model_final)$coef
               Estimate Std. Error     z value     Pr(>|z|)
(Intercept) -2.20432918 0.05361788 -41.1118278 0.000000e+00
x1          -0.03052552 0.04784421  -0.6380191 5.234612e-01
x14          0.24589127 0.05554281   4.4270587 9.552674e-06
```

```
x15          -0.20465873 0.05048499  -4.0538531 5.038086e-05
> exp(0.24589127)
[1] 1.278761
> exp(-0.20465873)
[1] 0.8149254
```

x_1 does not matter for being obese (z value -0.6380191 is insignificant). A single child is *more* likely to be obese (z value 4.4270587 is significant, odds ratio $e^{0.24589127} = 1.278761$ is *greater* than one). A child from a family with more than 100k income is *less* likely to be obese (z value -4.0538531 is significant, odds ratio $e^{-0.20465873} = 0.8149254$ is *less* than one)

14. The logistic regression is estimated by **maximum likelihood method**. Google "logistic regression" to learn more.