Logistic Regression Model

Note on required packages: The following code requires the package `erer` to compute marginal effects statistics for a logistic regression. If you have not already done so, download, install, and load the library with the following code:

```
install.packages("erer")
```

```
library("erer")
```

---

1. Logistic Regression Structure

Recall from a previous tutorial that binary variables can be used to estimate *proportions* or *probabilities* that an event will occur. If a binary variable is equal to 1 for when the event occurs, and 0 otherwise, estimates for the mean can be interpreted as the probability that the event occurs.

A **logistic regression** or **logit** is a regression model where the *outcome* is *related to* the probability that a binary variable takes on a value equal to 1.0. The probability for the outcome variable is predicted by one or more explanatory variables, which themselves can be binary or continuous.

A logistic regression takes the form,

$$\log(Odds_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_k x_{k,i} + \epsilon_i$$

where $Odds_i$ is called an **odds ratio** and is given by,

$$Odds_i = \frac{P(y_i = 1)}{P(y_i = 0)} = \frac{P(y_i = 1)}{1 - P(y_i = 1)}$$

The odds ratio has a range between 0 and $+\infty$, and increases as the probability that y_i takes on a value of 1.0 increases. The log odds ratio, $\log(Odds_i)$, can take on any value between $-\infty$ and $+\infty$. The construction of the odds ratio and logarithmic operation transform our binary outcome variable to a variable with an infinite range.

2. Example: Mortgage loan applications

The data set, `loanapp.RData`, includes actual data from 1,989 mortgage loan applications, including whether or not a loan was approved, and a number of possible explanatory variables including variables related to the applicant's ability to pay the loan such as the applicant's income and employment information, value of the mortgaged property, and credit history. Also included in the data set are variables measuring the applicant's race and ethnicity.

The code below loads the `R` data set, which creates a data set called `data`, and a list of descriptions for the variables called `desc`.

```
download.file("http://murraylax.org/datasets/loanapp.RData", "loanapp.RData")
load("loanapp.RData")
```

3. Estimating a Logistic Regression

Let us estimate a logistic with loan approval status as the outcome variable (`approve`) and total housing expenditure relative to total income as the sole explanatory variable (`hrat`). We call the function, `glm()`, which stands for general linear model, which is a large class of models that includes the logistic regression model.

```
lmapp <- glm(approve ~ hrat, data=data, family=binomial(link="logit"), x=TRUE)
```

The first parameter, `approve ~ hrat`, is a formula specifying a linear model with `approve` as the outcome variable and `hrat` as the sole explanatory variable. The parameter `data=data` tells `glm()` that these two variables can be found in the object called `data`. The parameter, `family=binomial(link="logit")` specifies a logistic regression model. Finally, `x=TRUE`, tells `glm()` to include in its return value intermediate calculations used to produce the regression results. We will not directly use these intermediate results, but we need them stored for functions that we run later in this tutorial.

We can see a summary of the regression results with a call to `summary()`.

```
summary(lmapp)
```

```
##
## Call:
## glm(formula = approve ~ hrat, family = binomial(link = "logit"),
##     data = data, x = TRUE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3256   0.4634   0.5025   0.5302   0.9232
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.692180   0.253449  10.622  < 2e-16 ***
## hrat        -0.028609   0.009432  -3.033  0.00242 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1480.7  on 1988  degrees of freedom
## Residual deviance: 1471.6  on 1987  degrees of freedom
## AIC: 1475.6
##
## Number of Fisher Scoring iterations: 4
```

The negative sign on the coefficient for `hrat` indicates that as housing expenses increase relative to income, the probability of being accepted for a loan decreases. The low p-value (`0.00242`) indicates the result is statistically significant.

The magnitude of the coefficient has little or no intuitive meaning. It is the magnitude that the log odds ratio increases when the explanatory variable increases by one meaning.

4. Marginal Effects

The *marginal effects* of a regression are estimates of the impact that one-unit increases in the explanatory variables have on the outcome variable.

In a regular bivariate or multiple regression, the coefficients are equal to the marginal effects.

In a logistic regression, the marginal effects are defined as the impact that one-unit increases in the explanatory variables have on *the probability that the outcome variable is equal to 1.0*. The outcome variable is a log-odds ratio, which is a mathematical transformation of the probability that the outcome variable is equal to 1.0. Therefore, we must take a transformation of the coefficients to obtain an estimate for the marginal effects.

The calculation of the marginal effects is further complicated in that the estimate depends on the value of the explanatory variables in the model. Unlike a simple regression where the estimate of the marginal effects is always equal to the coefficient, the estimate of the marginal effect of a logistic regression changes when the explanatory variable (housing expenditures in our example) changes. Rather than giving a range of marginal effects for a range of each explanatory variable, it is common to evaluate the marginal effect at the mean for every explanatory variable. The function call below does this by default.

The function `maBina()` from the package `erer` can be used to compute the marginal effects of the logistic regression as follows.

```
maBina(lmapp)
```

```
##             effect error t.value p.value
## (Intercept)  0.286 0.025  11.488   0.000
## hrat        -0.003 0.001  -3.066   0.002
```

The output above indicates the estimate for the marginal effect of housing expenditures is equal to `-0.003`. This implies for every one percentage point increase in housing expenditures as a percentage of income, the probability of mortgage approval decreases by 0.003 or 0.3%.