Logistic Regression Interpretations and Examples

In the example below, Y is an indicator saying whether a person was in the hospital to receive general medcial care (Y=1) or surgical care (Y=0). We have two predictor variables. X_1 is the person's age in years and X_2 is gender ($X_2 = 1$ fopr male and $X_2 = 0$ for female). We are interested in predicting how these factors influence how likely the person is to be in the hospital for acute medical care as opposed to surgery. The basic formula for a logistic regression is below. Here $p = P(Y = 1)$ is the probability the person is in the hospital for medical care. Of course you can have as many X's as you want.

$$ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

The quantity $ln(\frac{p}{1-p})$ is called the "log odds". It is easier intuitively to think about the odds of a person needing a particular kind of care, which you can get by exponentiating the log odds, or of the probability of needing a particular kind of care which you get by solving for p. The basic relationship is

$$p = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}$$

Logistic regression is basically like regular regression except that the formulas are incredibly messy (we will let the computer do all the work), there is the extra conversion to do to go from the regression equation to the predicted probability, and there are some interpretations in terms of odds and odds ratios. STATA gives two kinds of printouts. The first, using the logit command, gives the coefficients (b's) and their corresponding standard errors, test statistics and confidence intervals. The second type gives the corresponding odds ratios with their tests and confidence intervals. Printouts for the hospital data are given on the next page.

Start by noting that there is the equivalent of an overall F test in logitstic regression called the likelihood ratio chi-square test (LR chi2 on the STATA printouts). It's hypotheses are the same as in regression:

$H_0 : \beta_1 = \beta_2 = 0$–neither age nor gender affects the likelihood that a person is in the hospitcal for medical (as opposed to surgical) care.
H_A : AT least one of the β's is not 0. At least one of the variables is useful for predicting the likelihood the patient is hospitalized for medical care.

In this example (see the printouts on the next page) the test statistic is 10.04 and the corresponding p-value is .0066. We reject the null hypothesis and conclude at least one of the variables is useful.

STATA also gives an equivalent of R^2-adjusted called the "Pseudo R-squared". In this example it is .3073 meaning that age and gender account for about 30% of the variability in the reason for a person's hospitalization. This suggests that there are many other factors that are also important which is hardly surprising!

Next let's interpret the coefficient. As in all multiple regressions we need to take into accont the presence of the other variables and the difference in meaning between indicator variables and quantitative variables. The only tricky part here is that we are working on the log odds scale. The intercept gives the log odds for a subject all of whose X values are 0. here $b_0 = -2.52$ means that the log odds for a 0 year old (newborn) girl ($X_2 = 0$) needing medical as opposed to surgical care is -2.52. If we convert this into a probability we get

$$\frac{e^{-2.52}}{1 + e^{-2.52}} = .074$$

There is just over a 7% chance that a newborn girl would be staying in the hospital for medical (as opposed to surgical) care. This is reliable only if we had any newborns in our study. This probability seems rather low since not many newborns need surgery but some do stay in the hospital for a while.

```
. logit Service Age Gender

Logistic regression                                     Number of obs   =         25
                                                        LR chi2(2)      =      10.04
                                                        Prob > chi2     =     0.0066
Log likelihood =  -11.31536                             Pseudo R2       =     0.3073

------------------------------------------------------------------------------
     Service |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         Age |   .0636751   .0300868     2.12   0.034     .0047062    .1226441
      Gender |  -2.055256   1.195394    -1.72   0.086    -4.398185    .2876724
       _cons |  -2.524607   1.399137    -1.80   0.071    -5.266866    .2176522
------------------------------------------------------------------------------

. logistic Service Age Gender

Logistic regression                                     Number of obs   =         25
                                                        LR chi2(2)      =      10.04
                                                        Prob > chi2     =     0.0066
Log likelihood =  -11.31536                             Pseudo R2       =     0.3073

------------------------------------------------------------------------------
     Service | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         Age |   1.065746   .0320648     2.12   0.034     1.004717    1.130482
      Gender |     .12806   .1530821    -1.72   0.086     .0122996     1.33332
------------------------------------------------------------------------------
```

The coefficient for the age variable means that, assuming gender is fixed, for every additional year of age the log odds of needing medical care (as opposed to surgery) goes up $b_1 = .0637$. This is a little difficult to interpret unless we convert it to an odds ratio. We do this by exponentiating

$$\hat{OR} = e^{.0637} = 1.065$$

This value can be read off the second STATA printout in the Odds Ratio column. It means for people with the same gender, one of whom is a year older than the other, the older one has odds of needing medical care 1.065 times higher than that of the younger one. Another way to say this is that the odds go up by 6.5% each year. Note that an odds ratio of 1 means no difference. It corresponds to $\beta_1 = 0$ on the coefficient scale.

For the gender variable we are dealing with an indicator so our interpretation is that assuming age is held fixed the log odds of needing medical care is -2.055 lower for a man than a woman. It seems men are more likely, all things equal to need surgery than women and women all things equal are more likely to need medical care than men. This is also more interpretable on the odds ratio scale. The value of .128 for the odds ratio means that a man's odds of needing medical care (as opposed to surgery) is only .128 tims what a woman's odds would be. Note that to get the odds ratio for a woman we could just invert this and get $1/.128 = 7.8125$. A woman's odds of needing medical (as opposed to surgical) care is nearly 8 times has a high as a man's.

The next step is to determine whether the variables individually are useful predictors and to get confidence intervals for the coefficients and odds ratios. Looking at the STATA printout we see that the p-value for the age variable is .034 so using $\alpha = .05$ we reject the null hypothesis (that $\beta_1 = 0$) and conclude that age IS a useful variable even after accounting for gender, in predicting what kind of care a person will need. Note that the test statistic here is a Z rather than t statistic because we are really dealing with proportions–the fraction of people who will need medical or surgical care–our Y variable is qualitative! The p-value for the gender variable is .086. Thus at the $\alpha = .05$ significance level gender is not sigiificant. After accounting for a eprson's age the gender variable does not provide any new information. This may seem surprising given that the odds ratio for gender was so much more dramatic than that for age. The reason is that there is much more variability within gender so even though the effect looks large wwe are less confident about it.

We could also get these results using confidence intervals. The basic confidence interval for a coefficient in a logistic regression is the same as in a regular regression except that we use Z instead of t. Thus our interval is

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

For a 95% interval we use Z = 1.96. The STATA printout gives the b's and their standard errors. It also gives the confidence intervals. You can verify that if you plug in $b_1 = .0636751$ and $s_{b_1} = .0300868$ the resulting interval is [.0047062, .1226441]. This says that if gender is held fixed, the log odds of needing medical care goes up between .0047 and .1226 for every extra year of age. This sort of makes sense. Young people typically don't need acute medical care but an emergency, broken bone, etc. might make them need surgery. However as people age chronic illness is more likely to put them in the hospital. We can convert the CI for β_1 into a confidence interval for the odds ratio by exponentiating:

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

For the age variable the CI for the odds ratio is [1.004717, 1.130482] which we can also get from the second STATA printout. This says that for people of the same gender the odds of needing medical as opposed to surgical care goes up by between .47 and 13% per year. Note that you can get an odds ratio for any difference in age, δ, by multiplying by delta in all the exponentiation formulas:

$$e^{b_1 \Delta}$$

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

Since the gender variable is not significant I will not repeat the confidence interval calculations. However they work exactly the same way. In fact they are simpler because you don't need to worry about delta for doing different degrees of difference. Since gender is an indicator there is only one odds ratio of interest.