WWS 509 Generalized Linear Models: Precept 1
Section 3.1 through 3.4.3

Kristin E. Bietsch

Office of Population Research, Princeton University

October 2012

1 Introducing the Data

Here is some data looking at childbirth and HIV status. We will be looking at data that tells us if a woman had a birth in the follow up period after HIV testing. There are 3307 women in this data set. Do not site this data.

2 Binary Data

Binary data can only take two possible values. We want to code these as 0 and 1.
For example:

1. You can be pregnant or not pregnant
2. You can be approved or not for a mortgage
3. What kind of binary data have you encountered in your own research?

2.1 The Binomial Distribution

$$y_i = \begin{cases} 1 & if\, the\, i-th\, woman\, gave\, birth \\ 0 & otherwise \end{cases}$$

1. The probability of a woman giving birth is π_i
2. The probability of a woman not giving birth is $1 - \pi_i$

2.2 Expected Value and Variance

- $E(Y_i) = \mu_i = \pi_i$
- $var(Y_i) = \sigma_i^2 = \pi_i(1 - \pi_i)$

If you have grouped data (maybe I give you data by HIV status)

- $E(Y_i) = \mu_i = \eta_i \pi_i$
- $var(Y_i) = \sigma_i^2 = \eta_i \pi_i(1 - \pi_i)$

Note: any factor that affects the probability will affect both the mean and the variance of the observations

3 Logit Transformation

The problem with binary data is that it is confined to zero or 1. We need to transform the data so it can range from $-\infty$ to ∞.

- Step one: transform the probablity into odds: $odds_i = \frac{\pi_i}{1-\pi_i}$
  - What are odds: the ratio of a probability to its complement, ratio of favorable to unfavorable cases
  - What if you have a probability of 50%- odds are 1 to 1
  - What if the probability is 1/3: the odds are 1 to 2
  - Sometimes, I like to report probabilities, not odds, because that is how my mind works, you can choose whatever
  - This makes it so you can go from 0 to infinity, but what about negative infinity?
- Step 2: transform the odd into the logit by taking the logarithm: $\eta_i = logit(\pi_i) = log\frac{\pi_i}{1-\pi_i}$
  - Now if there is a 50% probability, or even odds, the log is 0
  - Negative logits represent probabilities below 1/2

If you want to transform the logit into the probability: $\pi_i = logit^{-1}(\eta_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}}$

Lets use the our data to illustrate. 1114 women gave birth in the follow-up period. Calculate:

- The probability of giving birth:
  - 1114/3307=.3368612
- The odds of giving birth:
  - (.3368612)/(1-.3368612)=.50797993

- The logit is:
  - log(.50797993)=-.67731334
- And now take the logit and transform it back into the probability: $\frac{e^{-}}{1+e^{-}}$
  - $\frac{e^{.67731334}}{1+e^{.67731334}}$

4 Logistic Regression

4.1 Interpreting β

β_j represents the change in the logit of the probability associated with a unit change in the j-th predictor holding all other predictors constant.

- You could also exponentiate β which represents the odds ratio (and this is multiplicative)
  - Stata can give you β or e^{β}

```
. logit chi_dum hiv

Iteration 0:   log likelihood = -2112.9466
Iteration 1:   log likelihood = -2106.3007
Iteration 2:   log likelihood = -2106.2914
Iteration 3:   log likelihood = -2106.2914

Logistic regression                               Number of obs   =       3307
                                                  LR chi2(1)      =      13.31
                                                  Prob > chi2     =     0.0003
Log likelihood = -2106.2914                       Pseudo R2       =     0.0031

------------------------------------------------------------------------------
     chi_dum |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        hiv5 |  -.5553665   .1584862    -3.50   0.000    -.8659937   -.2447392
       _cons |   -.641307   .0379678   -16.89   0.000    -.7157225   -.5668916
------------------------------------------------------------------------------
```

Figure 1: Regression of HIV on Childbearing

- How would you interpret the constant? What about the coefficient of HIV5?
  - The constant is the log odds of a woman without HIV having a child in the follow-up. If I convert this into a probability, I find that 35% of these women had a child in the follow up.
  - The coefficent of HIV5 tells us that the log odds of having a child in the follow-up are .555 less on average for HIV positive women than non-infected women. This translates into an odds ratio of 0.57, which means the odds of having a child are 0.57, the odds of having a child are nearly half as low as a non-infected women. We can translate this into a probability of HIV-postive women: $\frac{e^{(-.641307+-.5553665)}}{1+e^{(-.641307+-.5553665)}} =$

0.23. The probability of an HIV positive women having a child in the follow up period is 23%.

```
. logit chi_dum hiv, or

Iteration 0:   log likelihood = -2112.9466
Iteration 1:   log likelihood = -2106.3007
Iteration 2:   log likelihood = -2106.2914
Iteration 3:   log likelihood = -2106.2914

Logistic regression                               Number of obs   =       3307
                                                  LR chi2(1)      =      13.31
                                                  Prob > chi2     =     0.0003
Log likelihood = -2106.2914                       Pseudo R2       =     0.0031

------------------------------------------------------------------------------
     chi_dum | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        hiv5 |   .5738619   .0909492    -3.50   0.000     .4206334    .7829087
------------------------------------------------------------------------------
```

Figure 2: Regression of HIV on Childbearing

- How would you interpret the coefficient of HIV5?
  - The odds of having a child are lower for HIV positive women.
- Which regression is easier to interpret?
  - I find this one easier to interpret, since I don't have to exponentiate things.
  - But the first regression allowed me to calculate probabilities for different groups.

Here is a regression with age as categorical variables.

```
. do "C:\DOCUME~1\kbietsch\LOCALS~1\Temp\4d\STD00000000.tmp"

. logit chi_dum age20 age25 age30 age35 age40

Iteration 0:   log likelihood = -2112.9466
Iteration 1:   log likelihood = -1952.6956
Iteration 2:   log likelihood = -1949.1539
Iteration 3:   log likelihood =  -1949.142
Iteration 4:   log likelihood =  -1949.142

Logistic regression                               Number of obs   =       3307
                                                  LR chi2(5)      =     327.61
                                                  Prob > chi2     =     0.0000
Log likelihood =  -1949.142                       Pseudo R2       =     0.0775

------------------------------------------------------------------------------
     chi_dum |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       age20 |   1.174999   .1260839     9.32   0.000     .9278791    1.422119
       age25 |   1.625005   .1255021    12.95   0.000     1.379025    1.870984
       age30 |   1.528567   .1297104    11.78   0.000      1.27434    1.782795
       age35 |   1.109389   .1393674     7.96   0.000     .8362339    1.382544
       age40 |  -.1650128   .1627317    -1.01   0.311    -.4839611     .1539355
       _cons |  -1.593367   .0933763   -17.06   0.000    -1.776382   -1.410353
------------------------------------------------------------------------------
```

Figure 3: Regression of HIV on Childbearing

- How would you interpret the coefficients for each age?
  - The log odds of having a child in the period are higher for all categories compared to age 15 except for the oldest category.
  - We can see that age is not linear in its relationship with fertility, so using categorical variables or perhaps a quadratic term is more relatistic for this data.
  - The odds of having a child for the women 20-25 are over 3 times as high as for the youngest age group: $e^{1.174999} = 3.2381397$
  - There is no statistically significant difference between the log odds of having a child for the 15-19 group and the 40-44 group.