# 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  $\pi_i$
- 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  $\infty$ .

- Step one: transform the probability into odds:  $odds_i = \frac{\pi_i}{1-\pi_i}$
- Step 2: transform the odd into the logit by taking the logarithm:  $\eta_i = logit(\pi_i) = log \frac{\pi_i}{1-\pi_i}$

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:
- The odds of giving birth:
- The logit is:
- And now take the logit and transform it back into the probability:  $\frac{e^-}{1+e^-}$

# 4 Logistic Regression

## 4.1 Interpreting $\beta$

 $\beta_j$  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  $\beta$  which represents the odds ratio (and this is multiplicative)
  - Stata can give you  $\beta$  or  $e^{\beta}$

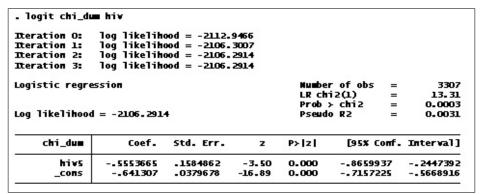


Figure 1: Regression of HIV on Childbearing

 How would you interpret the constant? What about the coefficient of HIV5?

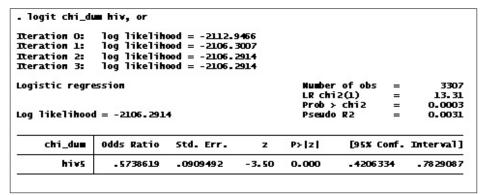


Figure 2: Regression of HIV on Childbearing

- How would you interpret the coefficient of HIV5?
- Which regression is easier to interpret?

Here is a regression with age as categorical variables.

WWS509 Precept 3 K.Bietsch

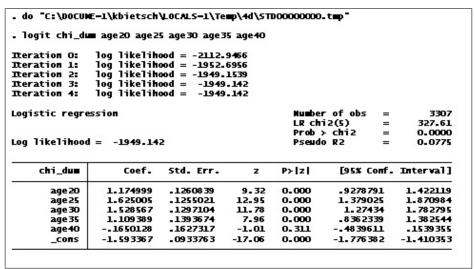


Figure 3: Regression of HIV on Childbearing

• How would you interpret the coefficients for each age?