Module 9: Probit Regression

Rebecca C. Steorts

Agenda

- Ordinal, numeric, and continous variables
- Probit versus linear regression
- Full conditionals
- An application to the 1994 General Social Survey

Generalized Linear Regression

- Many datasets include variables whose distributions cannot be represented by the normal, binomial or Poisson distributions we have studied so far.
- Distributions of common survey variables such as age, education level and income generally cannot be accurately described the above sampling models.
- In this module, we will use the probit regression model to handle such cases.

Terminology

- We use the term ordinal to refer to any variable for which there is a logical ordering of the sample space.
- We use the term numeric to refer to variables that have meaningful numerical scales.
- We use the term continuous if a variable can have a value that is (roughly) any real number in an interval.

Data

- The 1994 General Social Survey provides data on many variables in households (United States).
- There are variables such as the following:

1. DEG (highest degree obtained by individual)
2. CHILD (the number of children in a household)
3. PDEG (binary indicator of whether or not either parent obtained a college degree)

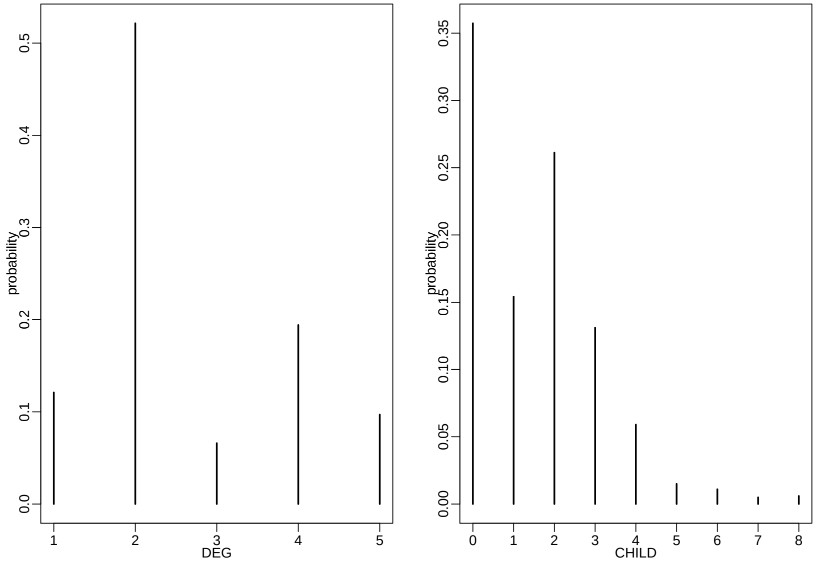
- Using these data, we might be tempted to investigate the relationship between the variables with a linear regression model.
- This is problematic due to the ordinal nature of the data.

Data

```r
dat<-read.table("http://lib.stat.cmu.edu/aoas/107/data.txt",
                header=TRUE)
head(dat)
```

```
##   INCOME DEGREE CHILDREN PINCOME PDEGREE PCHILDREN AGE
## 1     NA      1        3       3       1         5  59
## 2     11      0        3      NA       0         7  59
## 3      8      1        1      NA       0         9  25
## 4     25      3        2      NA       0         5  55
## 5    100      3        2       4       3         2  56
## 6     40      4        0      NA       4         5  36
```

Two ordinal variables having non-normal distributions



Probit regression

- Linear or generalized linear regression models, which assume a numeric scale to the data, may be appropriate for variables like height or weight, but are not appropriate for non-numeric ordinal variables like DEG or CHILD.
- This idea motivates a modeling technique known as ordered probit regression.
- We relate the response Y to a vector of predictors x via a regression model using a latent variable Z.

Probit regression model

The model can be written as

$$
\begin{align}
Y_i &= g(Z_i) \tag{1}\\
Z_i &= \beta^T x_i + \epsilon_i \tag{2}\\
\epsilon_i &\overset{iid}{\sim} Normal(0,1) \tag{3}\\
\beta &\sim MVN(0, n(X^T X)^{-1}), \tag{4}
\end{align}
$$

where g is any non-decreasing function.

Notation

- $X_{n \times p}$: regression features or covariates (design matrix)
- $Z_{n \times 1}$: latent variable
- $\boldsymbol{y}_{n \times 1}$: response variable (vector)
- $\boldsymbol{\beta}_{p \times 1}$: vector of regression coefficients

The role of g

$$Y_{n\times 1} = g(Z) \tag{5}$$
$$Z_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1} \tag{6}$$
$$\epsilon_{n\times 1} \overset{iid}{\sim} Normal(0, I) \tag{7}$$
$$\beta_{p\times 1} \sim MVN(0, n(X^T X)^{-1}) \tag{8}$$

Suppose the sample space for Y takes on K values $\{1, 2, \ldots, K\}$, then g can be described with $K-1$ ordered parameters.

You can think of the values of $g_1, \ldots g_{K-1}$ as thresholds so that moving past z will move y into the next (highest) category.

Full conditional of β

$$
\begin{aligned}
Y_{n\times 1} \mid Z &= g(Z) && (9)\\
Z_{n\times 1} \mid \boldsymbol{\beta} &\sim MVN(X\boldsymbol{\beta}, I) && (10)\\
\beta_{p\times 1} &\sim MVN(0, n(X^T X)^{-1}) && (11)
\end{aligned}
$$

$$p(\beta \mid y, z, g) \propto p(\beta)p(z \mid \beta)$$

Using the MVN conjugacy that we looked at before, $\beta \mid y, z, g$ will be MVN where

$$E[\beta \mid z] = \frac{n}{n+1}(X^T X)^{-1} X^T z$$

$$Var[\beta \mid z] = \frac{n}{n+1}(X^T X)^{-1}$$

Full conditional of Z

Under the sampling model, the conditional distribution of

$$Z_i \mid \beta \sim \text{Normal}(\beta^T x_i, 1)$$

Given g and observing $Y_i = y_i$, we know that Z_i lies in the interval (g_{i-1}, g_i).

Let $a = g_{i-1}, b = g_i$.

Then

$$p(z_i \mid \beta, z, y, g) \propto \text{dnorm}(z_i, \beta^T x_i, 1) \times I_{a,b}(z_i)$$

This is simply a density of a constrained normal distribution. How to sample? Apply the inverse CDF trick that we have done previously!

Full conditional of g

Suppose the prior distribution is $p(g)$.

Given $Y = y$ and $Z = z$, then we know:

- g_k must be higher than all z_i's for which $y_i = k$ and
- g_k must be lower than all z_i's for which $y_i = k + 1$

Let $a_k = \max\{z_i : y_i = k\}$ and $b_k = \min\{z_i : y_i = k + 1\}$.

Then the full conditional distribution of g is then proportional to $p(g)$ but constrained to the set $\{g : a_k < g_k < b_k\}$.

Application to the General Social Survey

Some researchers suggest that having children reduces opportunities for educational attainment (Moore and Waite, 1977).

Here we examine this hypothesis in a sample of males in the labor force (meaning not retired, not in school, and not in an institution), obtained from the 1994 General Social Survey.

For 959 of the 1,002 survey respondents we have complete data on the variables DEG, CHILD and PDEG described above.

We have the following variables:

- $Y_i = DEG_i$
- $x_i = (CHILD_i, PDEG_i, CHILD_i \times PDEG_i)$

Model

Our model specification is the following:

$$Y_{n\times 1} = g(Z) \tag{12}$$
$$Z_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon \tag{13}$$
$$\epsilon_{n\times 1} \overset{iid}{\sim} Normal(0, I) \tag{14}$$
$$\beta_{p\times 1} \sim MVN(0, n(X^T X)^{-1}) \tag{15}$$

$$p(g) \propto \prod_{k=1}^{K-1} (g_k, 0, 100)$$

Application to the General Social Survey

```r
X<-cbind(ychild,ypdeg,ychild*ypdeg)
head(X)
```

```
##      ychild ypdeg
## [1,]      3     0 0
## [2,]      3     0 0
## [3,]      1     0 0
## [4,]      2     0 0
## [5,]      2     1 2
## [6,]      0     1 0
```

```r
head(y<-ydegr)
```

```
## [1] 2 1 2 4 4 5
```

Application to the General Social Survey

```r
# replacing missing values with the mean
keep<-(1:length(y))[ !is.na( apply( cbind(X,y),1,mean) ) ]
X<-X[keep,]
y<-y[keep]
```

Application to the General Social Survey

```r
## data without missing values
head(X)
```

```
##      ychild ypdeg
## [1,]      3     0 0
## [2,]      3     0 0
## [3,]      1     0 0
## [4,]      2     0 0
## [5,]      2     1 2
## [6,]      0     1 0
```

```r
## response without missing values
head(y)
```

```
## [1] 2 1 2 4 4 5
```

Application to the General Social Survey

```
## short calculations
n<-dim(X)[1]
p<-dim(X)[2]
iXX<-solve(t(X)%*%X)
V<-iXX*(n/(n+1))
## review the cholesk decomposition
## if you have forgotten this from linear algebra
cholV <- chol(V)
```

Application to the General Social Survey

```
# find the unique y
# then sort them
# then return the first occurence of
# y in sort(unique(y))
ranks<-match(y,sort(unique(y)))
head(ranks)
```

```
## [1] 2 1 2 4 4 5
```

```
# sort the ranks
uranks<-sort(unique(ranks))
head(uranks)
```

```
## [1] 1 2 3 4 5
```

Application to the General Social Survey

```
###starting values
set.seed(1)
(beta<-rep(0,p))
```

```
## [1] 0 0 0
```

Initializing z

We know that Z is Gaussian. We can rank the y's (breaking ties at random). Then we can evaluation this with the qnorm function.

```
z<-qnorm(rank(y,ties.method="random")/(n+1))
head(z)
```

```
## [1] -0.7347365 -1.6348365 -0.2398885  1.1554244  0.61975
```

Other initializions

```r
(g<-rep(NA,length(uranks)-1))
```

```
## [1] NA NA NA NA
```

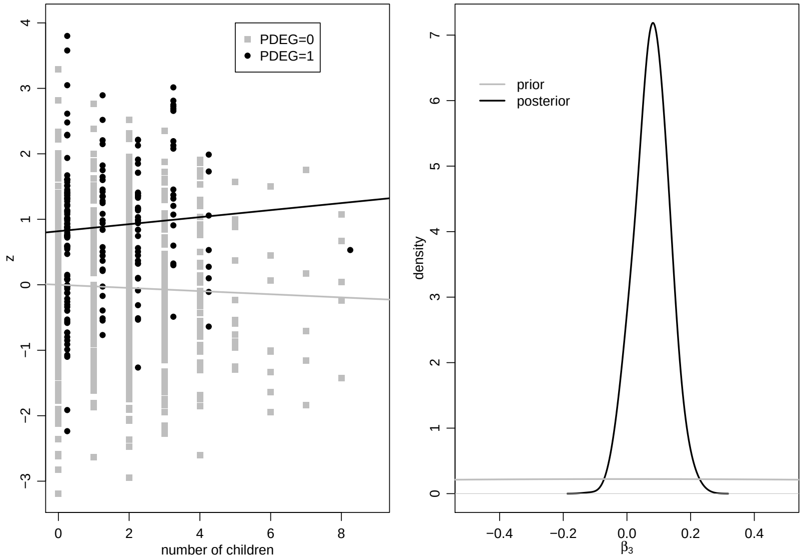
```r
K<-length(uranks)
BETA<-matrix(NA,1000,p)
Z<-matrix(NA,1000,n)
ac<-0
mu<-rep(0,K-1)
sigma<-rep(100,K-1)
S<-25000
```

Gibbs sampler

```
for(s in 1:S) {
  #update g
  for(k in 1:(K-1)){
  a<-max(z[y==k])
  b<-min(z[y==k+1])
  u<-runif(1, pnorm( (a-mu[k])/sigma[k] ),
              pnorm( (b-mu[k])/sigma[k] ) )
  g[k]<- mu[k] + sigma[k]*qnorm(u)
  }

  #update beta
  E<- V%*%( t(X)%*%z )
  beta<- cholV%*%rnorm(p) + E

  #update z
  ez<-X%*%beta
  a<-c(-Inf,g)[ match( y-1, 0:K) ]
  b<-c(g,Inf)[y]
  u<-runif(n, pnorm(a-ez),pnorm(b-ez) )
  z<- ez + qnorm(u)

  #help mixing
  c<-rnorm(1,0,n^(-1/3))
  zp<-z+c ; gp<-g+c
  lhr<-  sum(dnorm(zp,ez,1,log=T) - dnorm(z,ez,1,log=T) ) +
         sum(dnorm(gp,mu,sigma,log=T) - dnorm(g,mu,sigma,log=T) )
  if(log(runif(1))<lhr) { z<-zp ; g<-gp ; ac<-ac+1 }

  if(s%%(S/1000)==0){
    cat(s/S,ac/s,"\n")
    BETA[s/(S/1000),]<-  beta
    Z[s/(S/1000),]<- z
    }}
```

```
## 0.001 0.4
```

Plot


PDEG=0
PDEG=1
z
number of children
prior
posterior
density
β_3

Commentary for Left Plot

The posterior mean regression line for people without a college-educated parent ($x_{i,2} = 0$) is

$$E[Z \mid y, x_1, x_2 = 0] = -0.024x_1.$$

The posterior mean regression line for people with a college-educated parent

$$E[Z \mid y, x_1, x_2 = 0] = 0.818 + 0.054x_1.$$

In the above figure (left), we see that for people whose parents did not go to college, the number of children they have is indeed weakly negatively associated with their educational outcome. (The opposite is true for people whose parents did go to college).

Commentary for Right Plot

Next we give the posterior distribution of β_3 along with the prior distribution for comparison.

The 95% quantile-based posterior confidence interval for β_3 is $(-0.026, 0.178)$ which contains zero but still represents a reasonable amount of evidence that the slope for the $x_2 = 1$ is larger than the $x_2 = 0$ group.