Lecture 25 EEP118

Limited Dependent Variable

1. Logit Model Parameters/ Marginal Effects

   Parameters

   Marginal effects for continuous x

   Marginal effects for discrete x

2. Estimation Maximum Likelihood

3. Tests. Goodness of Fit. Likelihood Ratio Test

   The chi square distribution

Guest speaker: Law school

Study all of chapter 17.1

Posted all remaining DA and solutions, Practice final also

Limited Depedent Variable Y

The basic context of this set of lectures is when Y is not continuous

Y=0 or 1, Y is binary. YES/NO

Use a Data set on Women labor force participation

Source: MROZ.RAW in Wooldridge. T.A. Mroz (1987), “The Sensitivity of an Empirical Model of Married Women’s Hours of Work to Economic and Statistical Assumptions,” Econometrica 55, 765-799.

Y= 1 or 0 column called inlf (short for in labor force)

Obs: N=753

inlf byte %9.0g inlf=1 if in labor force, 1975, inlf=0 otherwise

age byte %9.0g woman's age in years

educ byte %9.0g years of schooling

kidslt6 byte %9.0g # kids < 6 years

kidsge6 byte %9.0g # kids 6-18

nwifeinc float %9.0g (faminc - wage*hours)/1000

hushrs int %9.0g hours worked by husband, 1975

husage byte %9.0g husband's age

huseduc byte %9.0g husband's years of schooling

huswage float %9.0g husband's hourly wage, 1975

city byte %9.0g =1 if live in SMSA

In [1]:

```
# Load the 'pacman' package
library(pacman)
#packages to use load them now using the pacman "manager"
p_load(dplyr, haven, readr)
#Another great feature of p_load(): if you try to load a package that is not
p_load(ggplot2)


pacman::p_load(lfe, lmtest, haven, sandwich, tidyverse)
# lfe for running fixed effects regression
# lmtest for displaying robust SE in output table
# haven for loading in dta files
# sandwich for producing robust Var-Cov matrix
# tidyverse for manipulating data and producing plots


#The big difference with Stata that appears here is lm() by default
#doesn't compute robust SE - we have to use additional packages/functions
#to compute it. felm does allow for multi-way clustering by default though
#which is nice.

#I added an alternate version of the first plots to show that we can
#change the color of the points according to whether the prediction
#is in [0,1] or outside of it. You can also specify factor(inlf) for
#the latter plots of actual vs. predicted to only have the values 0 or 1 on
#the x-axis.


pacman::p_load(lfe, lmtest, margins, haven, sandwich, tidyverse)
# lfe for running fixed effects regression
# lmtest for displaying robust SE in output table
# haven for loading in dta files
# sandwich for producing robust Var-Cov matrix
# tidyverse for manipulating data and producing plots
install.packages(sandwich)
install.packages(lfe)
install.packages(lmtest)
install.packages(tidyverse)
library(sandwich)
```

Loading [MathJax]/extensions/Safe.js

```
library(lmtest)
library(tidyverse)


# alternate plot theme for ggplot
theme_ed <- theme(
  legend.position = "bottom",
  panel.background = element_rect(fill = NA),
  # panel.border = element_rect(fill = NA, color = "grey75"),
  axis.ticks = element_line(color = "grey95", size = 0.3),
  panel.grid.major = element_line(color = "grey95", size = 0.3),
  panel.grid.minor = element_line(color = "grey95", size = 0.3),
  legend.key = element_blank())
```

```
Installing package into ‘/srv/r’
(as ‘lib’ is unspecified)

also installing the dependency ‘prediction’



margins installed

Installing package into ‘/srv/r’
(as ‘lib’ is unspecified)

Error in as.character(x): cannot coerce type 'closure' to vector of type 'ch
aracter'
Traceback:

1. install.packages(sandwich)
2. grepl("[.]tar[.](gz|bz2|xz)$", pkgs)
```

In [2]:
```
#load data
mydata<- read_dta("Lecture24MROZ.DTA")
#Summary stats inlf age educ kidslt6 kidsge6 nwifeinc hushrs husage huseduc
summary(mydata)
```

```
      inlf            hours           kidslt6         kidsge6     
 Min.   :0.0000   Min.   :   0.0   Min.   :0.0000   Min.   :0.000  
 1st Qu.:0.0000   1st Qu.:   0.0   1st Qu.:0.0000   1st Qu.:0.000  
 Median :1.0000   Median : 288.0   Median :0.0000   Median :1.000  
 Mean   :0.5684   Mean   : 740.6   Mean   :0.2377   Mean   :1.353  
 3rd Qu.:1.0000   3rd Qu.:1516.0   3rd Qu.:0.0000   3rd Qu.:2.000  
 Max.   :1.0000   Max.   :4950.0   Max.   :3.0000   Max.   :8.000  
                                                                   
      age             educ            wage            repwage    
 Min.   :30.00   Min.   : 5.00   Min.   : 0.1282   Min.   :0.00  
 1st Qu.:36.00   1st Qu.:12.00   1st Qu.: 2.2626   1st Qu.:0.00  
 Median :43.00   Median :12.00   Median : 3.4819   Median :0.00  
 Mean   :42.54   Mean   :12.29   Mean   : 4.1777   Mean   :1.85  
 3rd Qu.:49.00   3rd Qu.:13.00   3rd Qu.: 4.9708   3rd Qu.:3.58  
 Max.   :60.00   Max.   :17.00   Max.   :25.0000   Max.   :9.98  
                                 NA's   :325                     
     hushrs         husage         huseduc         huswage       
 Min.   : 175   Min.   :30.00   Min.   : 3.00   Min.   : 0.4121  
 1st Qu.:1928   1st Qu.:38.00   1st Qu.:11.00   1st Qu.: 4.7883  
 Median :2164   Median :46.00   Median :12.00   Median : 6.9758  
 Mean   :2267   Mean   :45.12   Mean   :12.49   Mean   : 7.4822  
 3rd Qu.:2553   3rd Qu.:52.00   3rd Qu.:15.00   3rd Qu.: 9.1667  
 Max.   :5010   Max.   :60.00   Max.   :17.00   Max.   :40.5090  
                                                                 
     faminc           mtr            motheduc         fatheduc     
 Min.   : 1500   Min.   :0.4415   Min.   : 0.000   Min.   : 0.000  
 1st Qu.:15428   1st Qu.:0.6215   1st Qu.: 7.000   1st Qu.: 7.000  
 Median :20880   Median :0.6915   Median :10.000   Median : 7.000  
 Mean   :23081   Mean   :0.6789   Mean   : 9.251   Mean   : 8.809  
 3rd Qu.:28200   3rd Qu.:0.7215   3rd Qu.:12.000   3rd Qu.:12.000  
 Max.   :96000   Max.   :0.9415   Max.   :17.000   Max.   :17.000  
                                                                   
      unem             city            exper          nwifeinc       
 Min.   : 3.000   Min.   :0.0000   Min.   : 0.00   Min.   :-0.02906  
 1st Qu.: 7.500   1st Qu.:0.0000   1st Qu.: 4.00   1st Qu.:13.02504  
 Median : 7.500   Median :1.0000   Median : 9.00   Median :17.70000  
 Mean   : 8.624   Mean   :0.6428   Mean   :10.63   Mean   :20.12896  
 3rd Qu.:11.000   3rd Qu.:1.0000   3rd Qu.:15.00   3rd Qu.:24.46600  
 Max.   :14.000   Max.   :1.0000   Max.   :45.00   Max.   :96.00000  
                                                                     
     lwage            expersq    
 Min.   :-2.0542   Min.   :   0  
 1st Qu.: 0.8165   1st Qu.:  16  
 Median : 1.2476   Median :  81  
 Mean   : 1.1902   Mean   : 178  
 3rd Qu.: 1.6036   3rd Qu.: 225  
 Max.   : 3.2189   Max.   :2025  
 NA's   :325                     
```

Fixing Problem 2, make sure predictions are between 0 and 1

Solution Problem 1 –

use a functional for for the probability as a function G( ) of the xs that stays between 0 and 1

e.g., the Logit Model!

the ratio of exponents in the logit below is always between 0 and 1

Prob [Y=1 | x] = G($\beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \cdots + \beta_3\, x_k$)

Get a G that stays between 0 and 1, and **the Logit is**

Prob [Y=1 | x] = G($\beta_0 + \ldots + \beta_k\;\; X_k$) = $\frac{e^{\beta_0+\beta_1\, x_1+\beta_2\, x_2+\cdots+\beta_3\, x_k}}{1+e^{\beta_0+\beta_1\, x_1+\beta_2\, x_2+\cdots+\beta_3\, x_k}}$

This ratio of exponentials is always between 0 and 1 no matter the betas and xs

Prob[Y=1| X]= $\Lambda\,(\beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \cdots + \beta_3\, x_k)$

24

In [3]:

```
###### Fixing Problem 2 so that predicted Y hats are less than 1 and greater
# In R, use the glm(formula, data, family = binomial(link = "logit")) functi

logit <- glm(inlf ~ nwifeinc + educ + age + kidslt6 + kidsge6, mydata, famil
summary(logit)
```

```
Call:
glm(formula = inlf ~ nwifeinc + educ + age + kidslt6 + kidsge6,
    family = binomial(link = "logit"), data = mydata)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.722993   0.788698   0.917    0.359
nwifeinc    -0.034891   0.007884  -4.426 9.62e-06 ***
educ         0.257965   0.040744   6.331 2.43e-10 ***
age         -0.057553   0.012737  -4.519 6.23e-06 ***
kidslt6     -1.484437   0.198013  -7.497 6.55e-14 ***
kidsge6     -0.066363   0.067856  -0.978    0.328
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  908.37  on 747  degrees of freedom
AIC: 920.37

Number of Fisher Scoring iterations: 4
```

Cannot easily interpret parameters here,

next class estimate implied marginal effects given the above estimated Logit parameters

Parameters not very meaningful here. (they enter two exponentials to get Phat)

What we want is if say education changes by one, how does the Prob(y=1) change?

Logit Model Marginal Effects

<u>For a continuous variable x_1 education for example:</u>

Given that $P(\boldsymbol{y}=\mathbf{1}) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}} = \frac{e^z}{1+e^z}$ where

$$z = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Then $\frac{\partial P(\boldsymbol{y}=\mathbf{1})}{\partial x_1} = \frac{\partial P(\boldsymbol{y}=\mathbf{1})}{\partial z}\frac{\partial z}{\partial x_1} = \frac{e^z}{(1+e^z)^2}\frac{\partial z}{\partial x_1}$

$$<=> \frac{\partial P(\boldsymbol{y}=\mathbf{1})}{\partial x_1} = \frac{e^z}{(1+e^z)^2}\beta_1$$

Logit Model Marginal Effects (ME)

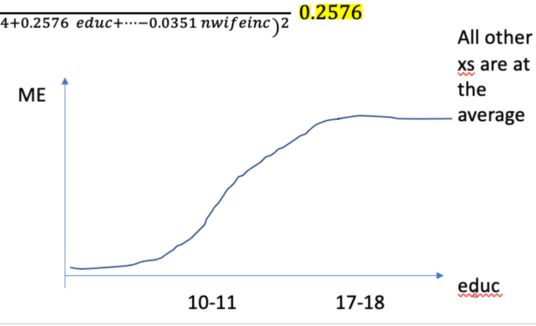
<u>For a continuous variable x_1 education for example:</u>

For education, given the estimates above, if education changes by one, then the ME on the Prob(y=1) is given by

$$\frac{e^{\widehat{\beta_0}+\widehat{\beta_1}educ+\cdots+\widehat{\beta_k}x_k}}{\left(1+e^{\widehat{\beta_0}+\widehat{\beta}_1 educ+\cdots+\widehat{\beta_k}x_k}\right)^2}\widehat{\beta_1} = \frac{e^{0.72544+0.2576\ educ+\cdots-0.0351\ nwifeinc}}{(1+e^{0.72544+0.2576\ educ+\cdots-0.0351\ nwifeinc})^2}\ 0.2576$$

Where we substitute the estimated beta hats.

Note that the ME depends on the starting point of educ and also on all the other x's.



Loading [MathJax]/extensions/Safe.js

Logit Model Marginal Effects (ME)

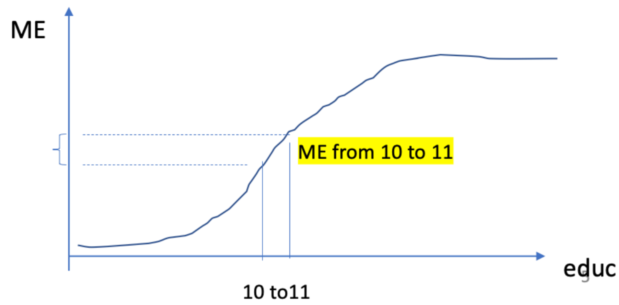
For a continuous variable x_1 education for example:

$$\frac{e^{\widehat{\beta_0}+\widehat{\beta_1}educ+\cdots+\widehat{\beta_k}x_k}}{\left(1+e^{\widehat{\beta_0}+\widehat{\beta_1}educ+\cdots+\widehat{\beta_k}x_k}\right)^2}\,\widehat{\beta_1} = \frac{e^{0.72544+0.2576\ educ+\cdots-0.0351\ nwifeinc}}{(1+e^{0.72544+0.2576\ educ+\cdots-0.0351\ nwifeinc})^2}\,0.2576$$

Where we substitute the estimated beta hats.

Note that the ME depends on the starting point of educ

and also on all the other x's.



For a continuous variable x1 education for example: How does one report the marginal effects (ME) then given that it depends on xs and starting point?

Report it for a fictitious person that would have all x's at the average, that is, for $\overline{(educ)}$=12.2, $\overline{(kids)}$=0.238 etc etc, all average of all x's, in this case, ME education is 0.0537, see next cell on how to get estimated ME

In [4]:

```
# replicate R's margins, dydx(*) command:

margins <- margins(logit)
summary(margins)
```

A summary.margins: 5 × 7

	factor <chr>	AME <dbl>	SE <dbl>	z <dbl>	p <dbl>	lower <dbl>	
1	age	-0.011991809	0.002524595	-4.7499936	2.034231e-06	-0.01693992	-0.007
2	educ	0.053749589	0.007646993	7.0288530	2.082382e-12	0.03876176	0.068
3	kidsge6	-0.013827441	0.014110502	-0.9799397	3.271159e-01	-0.04148352	0.013
4	kidslt6	-0.309296805	0.035369678	-8.7446882	2.236323e-18	-0.37862010	-0.239
5	nwifeinc	-0.007269922	0.001564987	-4.6453559	3.394907e-06	-0.01033724	-0.004

In [5]:

```
# create dataframe of mean data (i.e. one obs of X bar values)
meandata <- mydata %>%
  select(nwifeinc, educ, age, kidslt6, kidsge6) %>%
          all(mean)
```

```
# replicate Stata's mfx command:
mfx25 <- margins(logit, data = meandata)
summary(mfx25)
```

A summary.margins: 5 × 7

	factor	AME	SE	z	p	lower	
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
1	age	-0.014070720	0.003109436	-4.5251680	6.034755e-06	-0.02016510	-0.007
2	educ	0.063067667	0.009951639	6.3374153	2.336516e-10	0.04356281	0.082
3	kidsge6	-0.016224578	0.016588956	-0.9780349	3.280571e-01	-0.04873834	0.016
4	kidslt6	-0.362916807	0.048603820	-7.4668371	8.214534e-14	-0.45817854	-0.267
5	nwifeinc	-0.008530243	0.001929778	-4.4203228	9.855356e-06	-0.01231254	-0.004

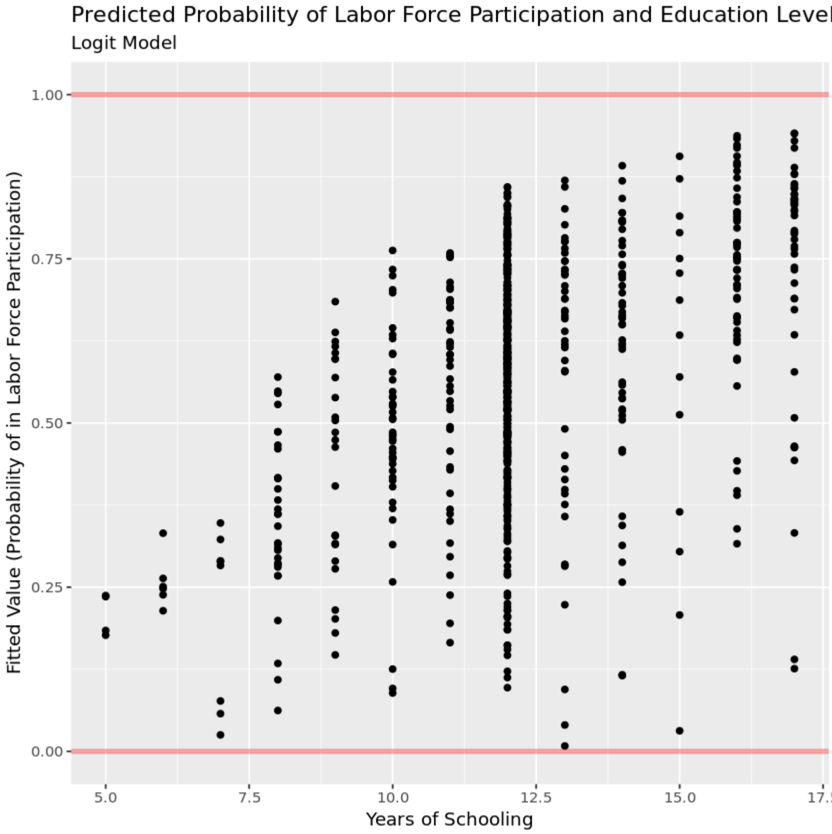
In [6]:
```
#generate predictions

mydata <- mutate(mydata, log_fit = logit$fitted.values) # add in the logit f


#Reproduce figures for logit
# no need to use the second approach as we're always within [0,1] with logit
# set data and aesthetics (x and y vars here since the same for all elements
ggplot(mydata, aes(x = educ, y = log_fit)) +
  # First add points, color determined by whether in or out of [0,1]
  geom_point() + # add points
  # add horizontal lines, width slightly wider, making partially transparent
  geom_hline(yintercept=0, size = 1.4, alpha = 0.35, color = "red") + # add
  geom_hline(yintercept=1, size = 1.4, alpha = 0.35, color = "red") + # add
  # generate labels
  labs(title = "Predicted Probability of Labor Force Participation and Educa
       subtitle = "Logit Model",
       x = "Years of Schooling",
       y = "Fitted Value (Probability of in Labor Force Participation)")

# actual vs predicted
ggplot(mydata, aes(x = factor(inlf), y = log_fit)) +
  # First add points, color determined by whether in or out of [0,1]
  geom_point() +
  # add horizontal lines, width slightly wider, making partially transparent
  geom_hline(yintercept=0, size = 1.4, alpha = 0.35, color = "red") + # add
  geom_hline(yintercept=1, size = 1.4, alpha = 0.35, color = "red") + # add
  # generate labels
  labs(title = "Predicted vs Actual Probability of Labor Force Participatior
       subtitle = "Logit Model",
       x = "Actual Labor Force Participation, 1975",
       y = "Estimated Labor Force Participation")
```

```
Warning message:
“Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
ℹ Please use `linewidth` instead.”
```



Marginal Effect for a Discrete variable X

For a discrete variable x1 city for example:

We need to compute the difference in probability, that is ME city= Prob(y=1| x, city=1) – Prob(y=1| x, city=0)

And once again we evaluate all at the average of all other x's

(*) dy/dx is for discrete change of dummy variable from 0 to 1

In [7]:
```
#run a logit with a city dummy variable
logit2 <- glm(inlf ~ nwifeinc + educ + age + kidslt6 + kidsge6+city, mydata,
summary(logit2)

# create dataframe of mean data (i.e. one obs of X bar values)
meandata2 <- mydata %>%
  select(nwifeinc, educ, age, kidslt6, kidsge6, city) %>%
  summarise_all(mean)



# replicate Stata's margins, dydx(*) command:
margins2 <- margins(logit2)
summary(margins2)
```

```
Call:
glm(formula = inlf ~ nwifeinc + educ + age + kidslt6 + kidsge6 +
    city, family = binomial(link = "logit"), data = mydata)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.725440   0.789091   0.919    0.358
nwifeinc    -0.035075   0.008067  -4.348 1.37e-05 ***
educ         0.257560   0.040910   6.296 3.06e-10 ***
age         -0.057689   0.012800  -4.507 6.58e-06 ***
kidslt6     -1.484777   0.198075  -7.496 6.58e-14 ***
kidsge6     -0.066625   0.067901  -0.981    0.326
city         0.019103   0.174730   0.109    0.913
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  908.36  on 746  degrees of freedom
AIC: 922.36

Number of Fisher Scoring iterations: 4
```

A summary.margins: 6 × 7

	factor	AME	SE	z	p	lower	
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	
1	age	-0.012019966	0.002538022	-4.7359580	2.180227e-06	-0.01699440	-0.007
2	city	0.003980240	0.036406176	0.1093287	9.129418e-01	-0.06737455	0.075
3	educ	0.053665101	0.007687514	6.9808135	2.934757e-12	0.03859785	0.068
4	kidsge6	-0.013881927	0.014119766	-0.9831556	3.255309e-01	-0.04155616	0.013
5	kidslt6	-0.309367325	0.035382599	-8.7434879	2.260221e-18	-0.37871594	-0.240
6	nwifeinc	-0.007308291	0.001604323	-4.5553738	5.229250e-06	-0.01045271	-0.004

For a city relative to not a city the probability of a woman being in the labor force increases by 0.004,

but not significantly because the p value of the marginal effect is 0.913

and confidence interval for city Marginal effect covers zero : lower= -0.06737453 upper=0.075335006

Estimation of Logit - by Maximum Likelihood

Maximum Likelihood

Derivation: for each observation of a woman

Suppose woman i working Yi=1, then, the prob is Pr(Yi=1|xi]= $\Lambda(\beta_0+\beta_1 x_{1i})$

Suppose woman j is not working, Yi=0, then the prob of that is Pr(Yj=0|xj]= 1-$\Lambda(\beta_0+\beta_1 x_{1j})$

Maximum Likelihood

The Probability of observing i working and j not is equal to the product below which is the

Likelihood

= $(\Lambda \beta_0+\beta_1 x_{1i}) * [1- \Lambda(\beta_0+\beta_1 x_{1j})]$

= Pr(Yi=1|xi] times Pr(Yj=0|xj]

- Put all the working in data together and all the non working
- The prob to see what we see in the sample is the product of the prob of all the working i's

Likelihood = $\Pi_i (\Lambda(\beta_0 + \beta_1 \, x_{1i}))\, \Pi_j [1 - \Lambda(\beta_0 + \beta_1 \, x_{1j})]$

all Y_i = inlf$_i$ =1 if women in labor market

Y_i = inlf$_i$ =0 if not in labor market

and the product of the prob of all the non working j's.

$$L = \prod_j [\frac{e^{X_i \beta}}{1+e^{X_i \beta}}]^{y_i} \prod_i [1 - \frac{e^{X_i \beta}}{1+e^{X_i \beta}}]^{1-y_i}$$

21

- Logging all that

log Likelihood =

LL = $\sum_i \boldsymbol{ln}[\Lambda(\boldsymbol{\beta_o} + \boldsymbol{X_i}\, \boldsymbol{\beta})] \quad + \quad \sum_j \boldsymbol{ln}\, [\mathbf{1} - \Lambda(\boldsymbol{\beta_o} + \boldsymbol{X_j}\, \boldsymbol{\beta})]$

all Y_i = inlf$_i$ =1
if women in labor market

Y_i = inlf$_i$ =0 if
not in labor market

$$logL = \sum_j y_i * log[\frac{e^{X_i \beta}}{1+e^{X_i \beta}}] + \sum_i (1 - y_i) log[1 - \frac{e^{X_i \beta}}{1+e^{X_i \beta}}]$$

Loading [MathJax]/extensions/Safe.js

Estimation Logit, Max Likelihood

- Put all the working in data together and all the non working
- The prob to see what we see in the sample is the product of the prob of all the working i's and the product of the prob of all the non working j's.
- If we log all of that we get
- **log Likelihood =**

LL = $\sum_i \boldsymbol{ln}[\Lambda(\boldsymbol{\beta_o} + \boldsymbol{X_i}\,\boldsymbol{\beta})] \quad + \quad \sum_j \boldsymbol{ln}\,[\mathbf{1} - \Lambda(\boldsymbol{\beta_o} + \boldsymbol{X_j}\,\boldsymbol{\beta})]$

for all Y_i = $inlf_i$ =1 if women in labor market

for all Y_i = $inlf_i$ =0 if not in labor market

$$logL = \sum_j y_i * log[\frac{e^{X_i\beta}}{1+e^{X_i\beta}}] + \sum_i (1-y_i) log[1 - \frac{e^{X_i\beta}}{1+e^{X_i\beta}}]$$

In [8]:
```
#estimate a model with lots of X's

logit_u <- glm(inlf ~ nwifeinc + educ + age + kidslt6 + kidsge6+city+hushrs+
summary(logit_u)
```

```
Call:
glm(formula = inlf ~ nwifeinc + educ + age + kidslt6 + kidsge6 + 
    city + hushrs + husage + huseduc + huswage, family = binomial(link = "lo
git"), 
    data = mydata)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept)  2.1072318  0.9407917   2.240   0.0251 *  
nwifeinc    -0.0182788  0.0128726  -1.420   0.1556    
educ         0.2893468  0.0478669   6.045 1.50e-09 ***
age         -0.0383568  0.0224972  -1.705   0.0882 .  
kidslt6     -1.5370349  0.2009480  -7.649 2.03e-14 ***
kidsge6     -0.0648634  0.0684488  -0.948   0.3433    
city         0.0147352  0.1809473   0.081   0.9351    
hushrs      -0.0003818  0.0001706  -2.238   0.0252 *  
husage      -0.0283468  0.0224390  -1.263   0.2065    
huseduc     -0.0354425  0.0365281  -0.970   0.3319    
huswage     -0.0434876  0.0372837  -1.166   0.2435    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  900.47  on 742  degrees of freedom
AIC: 922.47

Number of Fisher Scoring iterations: 4
```

What do you see in the output above?

Loading [MathJax]/extensions/Safe.js

AIC reported, a good measure of fit that is also used for model comparison

Akaike information Criterion (AIC) , not R squared any more, no more minimizing SSR

now we are maximizing log Likelihood as the estimation criterion, what are the parameters that make the sample we see the most likely?

AIC: 922.36

obtained by

Akaike Information Criterion

AIC=ln(ei2/n)+(2k/n)=ln(SSR/n)+(2k/n)

Hypothesis testing for one coefficient?

In [9]:
```
#Hypothesis testing for one coefficient

#Single parameter test-  use normal  z below


#Coefficients:
#                             Estimate  Std. Error    z value  Pr(>|z|)

#(Intercept)                  0.725440   0.789091    0.919    0.358

#nwifeinc                     -0.035075   0.008067   -4.348   1.37e-05 ***

#educ                          0.257560   0.040910    6.296    3.06e-10 ***

#age                          -0.057689   0.012800   -4.507   6.58e-06 ***

#kidslt6                      -1.484777   0.198075   -7.496   6.58e-14 ***

#kidsge6                        -0.066625   0.067901   -0.981   0.326

#city                            0.019103   0.174730    0.109    0.913
```

For example, reject that educaition coefficient is zero. z stat is 6.29 p value 3.06e-10 ***

Hypothesis Testing for multiple coefficients?

likelihood ratio test in step 2

and critical values of a chi squared distribution in step 3

Hypothesis Testing for multiple betas

LIKELIHOOD RATIO TEST

Example: H_o $beta_1 = beta_2 = beta_3 = beta_4 = \ldots = 0$ (q = number of restrictions)

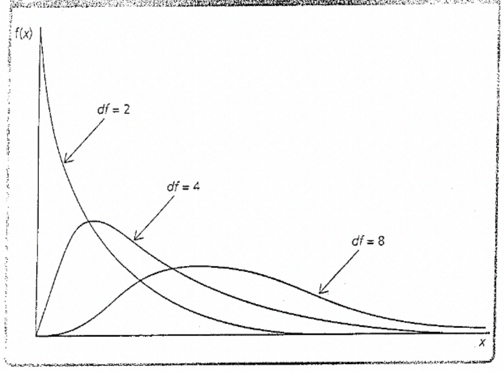
Under the null hypothesis:

LR=2 [Log likelihood unrestricted – Log Likelihood restricted]

is distributed χ^2 (q) **Chi square with q= degrees of freedom**

The chi -square distribution and table

Figure B.9
The chi-square distribution with various degrees of freedom



χ^2 (q)

TABLE G.4
Critical Values of the Chi-Square Distribution

Degrees of Freedom	Significance Level .10	.05	.01
1	2.71	3.84	6.63
2	4.61	5.99	9.21
3	6.25	7.81	11.34
4	7.78	9.49	13.28
5	9.24	11.07	15.09
6	10.64	12.59	16.81
7	12.02	14.07	18.48
8	13.36	15.51	20.09
9	14.68	16.92	21.67
10	15.99	18.31	23.21
11	17.28	19.68	24.72
12	18.55	21.03	26.22
13	19.81	22.36	27.69
14	21.06	23.68	29.14
15	22.31	25.00	30.58
16	23.54	26.30	32.00
17	24.77	27.59	33.41
18	25.99	28.87	34.81
19	27.20	30.14	36.19
20	28.41	31.41	37.57
21	29.62	32.67	38.93
22	30.81	33.92	40.29
23	32.01	35.17	41.64
24	33.20	36.42	42.98
25	34.38	37.65	44.31
26	35.56	38.89	45.64
27	36.74	40.11	46.96
28	37.92	41.34	48.28
29	39.09	42.56	49.59
30	40.26	43.77	50.89

Example: The 5% critical value with $df = 8$ is 15.51.
Source: This table was generated using the Stata® function invchi.

5 Steps as usual in hypothesis Testing

STEPS in Hypothesis testing

- Specify the null and the alternative hypothesis
- Run logit with all xs on the right = unrestricted model
  - Get the Log Likelihood value for the unrestricted L_{UR}
- Then run logit omitting 4 x's, we are testing whether those betas for those x's are zero – this is the restricted model
  - Get the Log Likelihood value for the restricted L_R
- Compute Likelihood Ratio Test Statistic= **LR**=2 (L_{UR}-L_R)
- Compare with critical value of χ^2 with 4 degrees of freedom for significance level chosen
- If critical value less than **LR** then we reject the null. Otherwise cannot reject the null

30

now coding and computing and doing the actual test

In [10]:

```
#step 1 Null that coefficients on the four husbands charactetistics, all fou

#step 2


#likelihood testing

#run unrestricted model
logit_u <- glm(inlf ~ nwifeinc + educ + age + kidslt6 + kidsge6+city+hushrs+
summary(logit_u)

#get the log likelihood of the unrestricted model
```

Loading [MathJax]/extensions/Safe.js

```
Call:
glm(formula = inlf ~ nwifeinc + educ + age + kidslt6 + kidsge6 +
    city + hushrs + husage + huseduc + huswage, family = binomial(link = "lo
git"),
    data = mydata)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.1072318  0.9407917   2.240   0.0251 *
nwifeinc    -0.0182788  0.0128726  -1.420   0.1556
educ         0.2893468  0.0478669   6.045 1.50e-09 ***
age         -0.0383568  0.0224972  -1.705   0.0882 .
kidslt6     -1.5370349  0.2009480  -7.649 2.03e-14 ***
kidsge6     -0.0648634  0.0684488  -0.948   0.3433
city         0.0147352  0.1809473   0.081   0.9351
hushrs      -0.0003818  0.0001706  -2.238   0.0252 *
husage      -0.0283468  0.0224390  -1.263   0.2065
huseduc     -0.0354425  0.0365281  -0.970   0.3319
huswage     -0.0434876  0.0372837  -1.166   0.2435
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  900.47  on 742  degrees of freedom
AIC: 922.47

Number of Fisher Scoring iterations: 4
```

In [11]:

```
#run the restricted model
#no husband charct as regressors

logit_r <- glm(inlf ~ nwifeinc + educ + age + kidslt6 + kidsge6+city, mydata
summary(logit_r)



#get the log likelihood of restricted model
```

Loading [MathJax]/extensions/Safe.js

```
Call:
glm(formula = inlf ~ nwifeinc + educ + age + kidslt6 + kidsge6 +
    city, family = binomial(link = "logit"), data = mydata)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.725440   0.789091   0.919    0.358
nwifeinc    -0.035075   0.008067  -4.348 1.37e-05 ***
educ         0.257560   0.040910   6.296 3.06e-10 ***
age         -0.057689   0.012800  -4.507 6.58e-06 ***
kidslt6     -1.484777   0.198075  -7.496 6.58e-14 ***
kidsge6     -0.066625   0.067901  -0.981    0.326
city         0.019103   0.174730   0.109    0.913
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  908.36  on 746  degrees of freedom
AIC: 922.36

Number of Fisher Scoring iterations: 4
```

In [12]:

```
#get both log likelihood values for the test statistics we will compute to e

#get log likelihood value unrestricted
logLik(logit_u)
```

```
'log Lik.' -450.2368 (df=11)
```

In [13]:

```
#get log likelihood value restricted
logLik(logit_r)
```

```
'log Lik.' -454.1793 (df=7)
```

compute the chi square stat

By hand, you will do this in Pset 5:

LR = 2 (loglikelihood UR – loglikelihood R) = 2 *(-450.237- + 454.179)=2 *3.94

So LR = chi2(4) = 7.89

step 3 go to the table and get the critical value for a certain significance level

see below

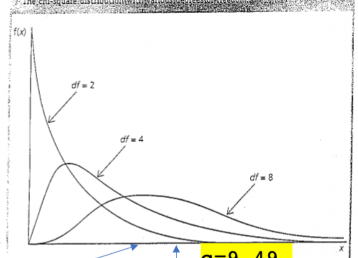
Step 3: get critical value from Chi squared Table:

at 10% c=7.78

At 5% c=9.49

Figure B.9

The chi-square distribution with various degrees of freedom



c=9.49.

LR = chi2(4)
= 7.89

Step 4: get critical value from Chi squared Table: at 10% c=7.78 At 5% c=9.49. reject at 10% cannot reject at 5%.

Step 5: conclude with a sentence. At 5%, there is no statistical evidence that husbands characteristics matter for prob woman in labor force controlling for non wife income, kids, woman educ, etc

TABLE G.4

Critical Values of the Chi-Square Distribution

Degrees of Freedom	Significance Level .10	.05	.01
1	2.71	3.84	6.63
2	4.61	5.99	9.21
3	6.25	7.81	11.34
4	7.78	9.49	13.28
5	9.24	11.07	15.09
6	10.64	12.59	16.81
7	12.02	14.07	18.48
8	13.36	15.51	20.09
9	14.68	16.92	21.67
10	15.99	18.31	23.21
11	17.28	19.68	24.72
12	18.55	21.03	26.22
13	19.81	22.36	27.69
14	21.06	23.68	29.14
15	22.31	25.00	30.58
16	23.54	26.30	32.00
17	24.77	27.59	33.41
18	25.99	28.87	34.81
19	27.20	30.14	36.19
20	28.41	31.41	37.57
21	29.62	32.67	38.93
22	30.81	33.92	40.29
23	32.01	35.17	41.64
24	33.20	36.42	42.98
25	34.38	37.65	44.31
26	35.56	38.89	45.64
27	36.74	40.11	46.96
28	37.92	41.34	48.28
29	39.09	42.56	49.59
30	40.26	43.77	50.89

Example: The 5% critical value with $df = 8$ is 15.51.
Source: This table was generated using the Stata® function invchi.

34

step 4

at 10% c=7.78 < LR=7.89 so we reject the null at 10%

at 5% c=9.49 > LR = 7.89, so we cannot reject the null at 5%

Loading [MathJax]/extensions/Safe.js

Step5: conclude with a sentence. At 5%, there is no statistical evidence that husbands characteristics matter for prob woman in labor force controlling for non wife income, kids, woman educ, etc

all together

Step 1: H0 Beta_hushrs=Beta_husage=Beta_huseduc=Beta_huswage=0

```
H1   not H0
```

Step 1: under the null 2 (loglikelihood UR – loglikelihood R) follows a Chi Square with q degrees of freedom

Step 2:

By hand, you will do this in Pset 5:

LR = 2 (loglikelihood UR – loglikelihood R) = 2 *(-450.237- + 454.179)=2 *3.94

So LR = chi2(4) = 7.89

Step 3: get critical value from Chi squared Table: at 10% c=7.78 At 5% c=9.49. reject at 10% cannot reject at 5%.

Step 4/5: conclude with a sentence. At 5%, there is no statistical evidence that husbands charct matter for prob woman in labor force controlling for non wife income, kids, woman educ, etc

In [14]:
```
#in your career you can use a canned command, not in this class though...

##in R: various equivalent specifications of the LR test
lrtest(logit_u, logit_r)
```

A anova: 2 × 5

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	11	-450.2368	NA	NA	NA
2	7	-454.1793	-4	7.885107	0.09587869

Loading [MathJax]/extensions/Safe.js

In R- for your future work in Metrics in life ☺

##in R: various equivalent specifications of the LR test

lrtest(logit_u, logit_r)

You get the output in R then:

Likelihood ratio test

Model 1: inlf ~ nwifeinc + educ + age + kidslt6 + kidsge6 + city + hushrs + husage + huseduc + huswage
Model 2: inlf ~ nwifeinc + educ + age + kidslt6 + kidsge6 + city

```
  #Df   LogLik    Df      Chisq    Pr(>Chisq)
1  11  -450.24
2   7  -454.18  -3.94     7.8851    0.09588 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here would not even reject at 10% because p value 0.0958

the end