Week 12: Logistic and Probit regression

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2019

These slides are part of a forthcoming book to be published by Cambridge University Press. For more information, go to perraillon.com/PLH. ©This material is copyrighted. Please see the entire copyright notice on the book's website.

Updated notes are here: https://clas.ucdenver.edu/marcelo-perraillon/teaching/health-services-research-methods-i-hsmp-7607

Outline

- Logistic regression once again
- Parameter interpretation
- Log odds, odds ratios, probability scale
- Goodness of fit
- Marginal effects preview

Review of MLE

- At the risk of being repetitive, recall the log-likelihood of the logistic model:

  $lnL(p) = \sum_{i=1}^{n} y_i ln(p) + \sum_{i=1}^{n}(1 - y_i)ln(1 - p)$
- When we use the logistic transformation to ensure that p is bounded between 0 and 1, we plug-in $p = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$
- When estimating the betas, the estimated model is in the log-odds scale:

  $log(\frac{p_i}{1-P_i}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

Latent variable approach

- Suppose that there is a **latent** (unobserved) and **continuous** variable y^* that take values from $-\infty$ to $+\infty$
- We also assume that the latent variable is a function of covariates **X**. For simplicity, let's just assume a linear relationship and just one covariate: $y_i^* = \beta_0 + \beta_1 x_i + u$
- u plays the same role as ϵ in the linear model: a source of random error
- We do not observe the latent variable y, we only observe if an event happens or not but whether the event happens depends on the value of the latent variable. We use y_i to denote the **observed** variable, which we assume is a 1 or 0 variable
- If $y_i * > 0$ then $y_i = 0$. If $y_i * \leq 0$ then $y_i = 1$. Note that in this case 0 is a **threshold**
- Think of $y*$ as intelligence and y is whether a person answers a question correctly or not. Or more relevant to something like it's done in economics, $y*$ is preference over some good and $y = 1$ if a person buys the good

Latent variable approach

- Because of the way we set up the problem, we can write the probability of $y = 1$ conditional on the covariate x as:

  $P(y = 1|x) = P(y* > 0|x)$
- Since we assumed that $y_i^* = \beta_0 + \beta_1 x_i + u$ the above equation becomes

  $P(y = 1|x) = P(\beta_0 + \beta_1 x_i + u > 0|x) = P(u < [\beta_0 + \beta_1 x_i]|X) = F([\beta_0 + \beta_1 x_i]|x)$
- From the above equation you get the insight that the probability of **observing** the $y = 1$ depends on the distribution of u and we can calculate it if you know the **cumulative distribution function** $F()$
- This one is not so obvious but it's apparent that one must make a strong assumption about the underlying form of $y*$ to be able to solve the problem
- Also note that $P(y = 0|x) = 1 - P(y = 1|x)$

Latent variable approach

- In econometrics this type of models are often called **index function models**
- There are two assumptions about u that are used: u distributes either **standard logistic** or **standard normal**
- Both distributions have a mean of 0 and constant variance. In the standard logistic the variance $var(u) = \frac{\pi^2}{3}$. In the standard normal $var(u) = 1$
- The idea of fixing the variance is not that trivial in the sense that if we don't fix it, then we can't estimate it because we only observe a 0 or 1 and the probability of 1 depends on the sign on $y*$ but not the scale (variance)
- In other words, we don't have information to estimate $var(u)$ yet **we lose nothing** by fixing it because $P(y = 1|x)$, and therefore, $P(y = 0|x)$, does not depend on $var(u)$

Latent variable approach

- Remember that the **cumulative distribution function** (cdf) gives you $P(X < a)$. Remember too that to get the probability you need to integrate the density f(t) from $-\infty$ to a: $\int_{-\infty}^{a} f(t)dt$
- If we assume **standard normal cdf**, our model then becomes $P(y = 1|x) = \int_{-\infty}^{\beta_0+\beta_1 x} \frac{1}{2\pi} e^{(-\frac{t^2}{2})} dt$
- And that's the **probit** model. Note that because we use the cdf, the probability will obviously be constrained between 0 and 1 because, well, it's a cdf
- If we assume that u distributes **standard logistic** then our model becomes $P(y = 1|x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$
- Remember that there are two different concepts: **logistic response function** and **logistic distribution**. The **standard logistic cdf** happens to have the above formula (the pdf is different)

Estimation

- Estimation is straighforward with MLE. We did it for the logistic model already
- For probit, the likelihood is just like writing $P(y = 1|x)$ above because that's the probability of seeing the data. We need to multiply n times and also consider that the probability of 0 is $1 - P(y = 1|x)$. If we take the log, it's a sum
- This is often a source of confusion but remember that the likelihood function is the probability of seeing the data given assumptions about the distribution of the data
- So what is the probability of observing a data point $y = 1$? It's $P(y = 1|x) = \int_{-\infty}^{\beta_0+\beta_1 x} \frac{1}{2\pi} e^{(-\frac{t^2}{2})} dt$
- What is the probability of observing a data point $y = 0$? It's $P(y = 0|x) = 1 - P(y = 1|x)$

Estimation

- Sometimes is easier to see how you could program Stata to maximize the log-likelihood
- I have more examples on my site `http://tinyurl.com/mcperraillon`
- Note below that writing the likelihood makes it obvious that the betas are shifts in the standard normal cdf scale

```
program probit_lf
        version 12
        args todo b lnf
        tempvar xb lj
        mleval `xb' = `b'
        * latent variable assumed cumm standard normal
        qui gen double `lj'  = normal( `xb')  if $ML_y1 == 1
        qui replace    `lj'  = normal(-`xb')  if $ML_y1 == 0
        qui mlsum `lnf' = ln(`lj')
end
```

Digression

- Assuming standard normal cdf or logistic are not the only options
- There is the complementary log-log model commonly used in discrete time survival because the exponent of coefficients are hazard rates
- Or the Gumbel model used to model extreme values
- Or the Burr model. Or the Scobit model
- Statistics and econometrics are large fields... Papers must be written, dissertations must be completed
- Sometimes a proposed new method goes to the Journal Article Graveyard. Sometimes they are resurrected 30 years later when somebody discovers that they are perfect for a particular application
- So many ideas and clever people out there...
- See Greene (2018) for the gory details

Example

- Women's labor force participation (inlf); main predictor is "extra" money in family

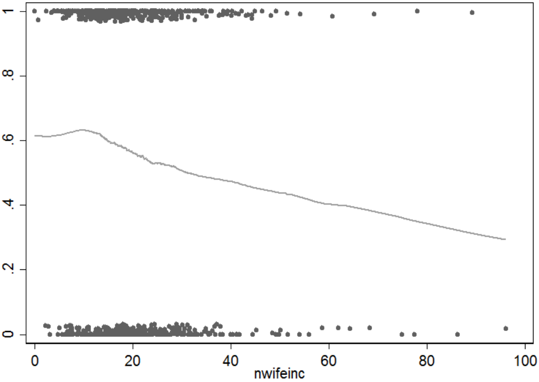
```
 bcuse mroz, clear

inlf      =1 if in labor force, 1975
nwifeinc  (faminc - wage*hours)/1000
educ      years of schooling
exper     actual labor mkt exper
age       woman's age in yrs
kidslt6   # kids < 6 years
kidsge6   # kids 6-18
```

Variable	Obs	Mean	Std. Dev.	Min	Max
inlf	753	.5683931	.4956295	0	1
nwifeinc	753	20.12896	11.6348	-.0290575	96
educ	753	12.28685	2.280246	5	17
exper	753	10.63081	8.06913	0	45
age	753	42.53785	8.072574	30	60
kidslt6	753	.2377158	.523959	0	3
kidsge6	753	1.353254	1.319874	0	8

Labor force participation

- The probability of working is decreasing as a function of "extra" income

```
lowess inlf nwifeinc, gen(lflow) nograph
scatter inlf nwifeinc, jitter(5) msize(small) || line lflow nwifeinc, sort ///
        legend(off) saving(lblow.gph, replace)
graph export lblow.png, replace
```



Writing down the model

- We want to estimate the following model:
  $P(inlf_i = 1|nwifeinc_i) = \Lambda(\beta_0 + \beta_1 nwifeinc_i)$
- **By convention** (in economics and health economics), when we write capital lambda, $\Lambda()$, we imply a logistic model (Λ is not a non-linear function). When we write phi, $\phi()$, we imply a probit model
- As I told you last class, write the logistic model this way:
  $log(\frac{inlf_i}{1-inlf_i}) = \beta_0 + \beta_1 nwifeinc_i$
- Or

  $logit(inlf_i) = \beta_0 + \beta_1 nwifeinc_i$
- Again, write it like this: $log(\frac{inlf_i}{1-inlf_i}) = \beta_0 + \beta_1 nwifeinc_i$ because this will **match Stata's (or any other statistical package) output**. Remember, we are not directly estimating $P(inlf_i = 1|nwifeinc_i)$

On pet peeves...

- For the love of everything you hold dear, please do not write logistic of probit models like this. Please, please, please, **don't do this**
- $P(y = 1|x) = \beta_0 + \beta_1 x$
- $P(y = 1|x) = \beta_0 + \beta_1 x + \epsilon$
- $P(y) = \beta_0 + \beta_1 x + \epsilon$
- $logit(y) = \beta_0 + \beta_1 x + \epsilon$
- $log(\frac{p}{1-p}) = \beta_0 + \beta_1 x + \epsilon$
- $P(y = 1|x) = f(\beta_0 + \beta_1 x1 + \epsilon)$
- Worse: $p = \beta_0 + \beta_1 x$ or $p = \beta_0 + \beta_1 x + \epsilon$

Estimating the model

- So, we will estimate $log(\frac{inlf_i}{1-inlf_i}) = \beta_0 + \beta_1 nwifeinc_i$

```
 logit inlf nwifeinc, nolog

Logistic regression                             Number of obs     =        753
                                                LR chi2(1)        =      10.44
                                                Prob > chi2       =     0.0012
Log likelihood = -509.65435                     Pseudo R2         =     0.0101

------------------------------------------------------------------------------
        inlf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -.0207569   .0065907    -3.15   0.002    -.0336744   -.0078394
       _cons |   .6946059   .1521569     4.57   0.000      .396384    .9928279
------------------------------------------------------------------------------
```

- A one thousand increase in “extra” income decreases the log-odds of participating in the labor force by 0.021. And it’s statistically significant (p-value $= 0.002$). Same Wald test as before: $-.0207569/.0065907 = -3.1494227$. The difference is that the it’s not t-student distributed but normally distributed

Overall significance

- The χ^2 (chi-square) test of the overall significance should look familiar. It compares the current model to the null model (without covariates); the null hypothesis is that all the cofficients in current model are zero
- It's the **likelihood ratio test** that we have seen before; the equivalent of ANOVA:

```
* LRT
qui logit inlf nwifeinc, nolog
est sto full

qui logit inlf, nolog
est sto redu

lrtest full redu

Likelihood-ratio test                                 LR chi2(1)  =     10.44
(Assumption: redu nested in full)                     Prob > chi2 =    0.0012
```

What about that Pseudo R^2?

- We can't partition the variance into explained and unexplained as before so we don't have a nice R^2 that goes from 0 to 1
- But one way to come up with a measure of fit is to use the (log) likelihood function to compare the **current model** to the **model without any explanatory variable** (the null model)
- The formula is: $1 - \frac{ll_{cm}}{ll_{nul}}$, where ll_{cm} is the log-likelihood of the current model and ll_{nul} is the log-likelihood of the null model
- If the current model is as good as the null model, then $\frac{ll_{cm}}{ll_{nul}}$ is going to close to 1 and the $pseudo - R^2$ is going to be close to zero
- In other words, adding variables doesn't improve the likelihood. If adding variables improves the likelihood, then the pseudo R^2 will be greater than zero

Pseudo-R^2

- Replicate Pseudo R^2

```
qui logit inlf nwifeinc, nolog
scalar ll_cm = e(ll)

qui logit inlf, nolog
scalar ll_n = e(ll)

di 1 - (ll_cm/ll_n)

.0101362

di  "cm: " ll_cm  " "  "null: "  ll_n  "  " "(ll_cm/ll_n): "  (ll_cm/ll_n)
cm: -509.65435 null: -514.8732  (ll_cm/ll_n): .9898638
```

- Psuedo R^2 is **not** a measure of how good the model is at prediction; just how better it fits compared to null model. I don't think that calling it pseudo R^2 is a good idea
- **Big picture**: comparing the log-likelihood of models is a way of comparing goodness of fit. If nested, we have the a test (LRT); if not nested, we have BIC or AIC

Not the only pseudo R^2?

- Stata uses one version of pseudo R^2 but there are plenty more. Other software may use different metrics
- Long and Freese (2014) have a laundry list of different pseudo R^2 (it's an excellent book, by the way)
- There is the McFadden one, MLE, Cragg and Uhler (also known as Nagelkerke), Efron's, Tjur's... (page 127)
- In any case, none of them have the same meaning as the R^2 in linear regression
- In particular, they **don't mean that predictions are good**. Recall that in linear regression the R^2 is also the square of the correlation between observed and predicted values
- **See, context matters a lot**

Let's try a different predictor

- We will estimate $log(\frac{inlf_i}{1-inlf_i}) = \beta_0 + \beta_1 hsp_i$, where *hsp* if education > 12

```
gen hsp = 0
replace hsp = 1 if educ > 12 & educ ~= .

logit inlf hsp, nolog

Logistic regression                             Number of obs     =        753
                                                LR chi2(1)        =      15.08
                                                Prob > chi2       =     0.0001
Log likelihood = -507.33524                     Pseudo R2         =     0.0146

------------------------------------------------------------------------------
        inlf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hsp |   .6504074   .1704773     3.82   0.000     .3162781    .9845368
       _cons |   .0998982    .086094     1.16   0.246    -.068843    .2686393
------------------------------------------------------------------------------
```

- The log-odds of entering the labor force is 0.65 higher for those with more than high school education compared to those with high-school completed or less than high-school

Odds ratios

- Let's do our usual math to make sense of coefficients. We just estimated the model $log(\frac{inlf_i}{1-inlf_i}) = \beta_0 + \beta_1 hsp_i$
- For those with $hsp = 1$, the model is $log(\frac{inlf_{hsp}}{1-inlf_{hsp}}) = \beta_0 + \beta_1$
- For those with $hsp = 0$, the model is $log(\frac{inlf_{nohsp}}{1-inlf_{nohsp}}) = \beta_0$
- The difference of the two is $log(\frac{inlf_{hsp}}{1-inlf_{hsp}}) - log(\frac{inlf_{nohsp}}{1-inlf_{nohsp}}) = \beta_1$
- Applying the rules of logs: $log(\frac{\frac{inlf_{hsp}}{1-inlf_{hsp}}}{\frac{inlf_{nohsp}}{1-inlf_{nohsp}}}) = \beta_1$
- Taking $e^{()}$: $\frac{\frac{inlf_{hsp}}{1-inlf_{hsp}}}{\frac{inlf_{nohsp}}{1-inlf_{nohsp}}} = e^{\beta_1}$

Odds ratios

$$\frac{\frac{inlf_{hsp}}{1-inlf_{hsp}}}{\frac{inlf_{nohsp}}{1-inlf_{nohsp}}} = e^{\beta_1}$$

- And that's the **(in)famous odds-ratio**
- In our example, $e^{.6504074} = 1.92$. So the odds of entering the labor force is almost twice as high for those with more than high school education compare to those without
- That's the way careful reporters would report this finding. **And it's correct**. The problem is that we would then interpret this as saying that the **probability** of entering the labor force is twice as high for those with more than high school
- **That interpretation is wrong**. A ratio of odds is more often than not far away from the ratio of probabilities

Odds ratios **are NOT** relative risks or relative probabilities

- One quick way to see this is by doing some algebra
- Changing the notation to make it easier:
  $\frac{\frac{P_A}{1-P_A}}{\frac{P_B}{1-P_B}} = e^{\beta_1}$
- After some simple algebra:
  $\frac{P_A}{P_B} = \frac{1-P_A}{1-P_B} e^{\beta_1}$
- Only when rare events (both P_A and P_B are small) or the ratio close to 1 are odds ratios close to relative probabilities ($\frac{1-P_A}{1-P_B}$ will be close to 1)
- For a more epi explanation, see `http://www.mdedge.com/jfponline/article/65515/relative-risks-and-odds-ratios-whats-difference`

Relative probabilities

- With only a dummy variable as predictor we can very easily calculate the **probabilities**
- Remember, we are modeling $log(\frac{p}{1-p}) = \beta_0 + \beta_1 X$. We also know that we can solve for p:
- $p = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$
- So we can calculate the probability for those with more than high school education and the probability for those with less

Probabilities, odds, relative risks, differences **(pay attention!!!**)

```
qui logit inlf hsp, nolog
* hsp = 1
di exp(_b[_cons] + _b[hsp]) / (1 + exp(_b[_cons] + _b[hsp]))
.67924528
* hsp = 0
di exp(_b[_cons]) / (1 + exp(_b[_cons]))
.52495379

* Odds
di (.67924528/(1-.67924528)) / (.52495379/(1-.52495379))
1.9163214

* Relative probabilities
di .67924528/ .52495379
1.2939144

 * Difference
di .67924528 - .52495379
.15429149
```

- Odds ratios are **confusing, misleading, evil**: Before, we said that the odds were doubled, or 100% higher. Now in, the scale that matters, we say that the *probability* is only 30% higher. Or 15% percent points different

Using GLM

```
 glm inlf hsp, family(binomial) link(logit) nolog

Generalized linear models                          No. of obs      =        753
Optimization     : ML                              Residual df     =        751
                                                   Scale parameter =          1
Deviance         =  1014.670487                    (1/df) Deviance =   1.351093
Pearson          =          753                    (1/df) Pearson  =   1.002663

Variance function: V(u) = u*(1-u)                  [Bernoulli]
Link function    : g(u) = ln(u/(1-u))              [Logit]

                                                   AIC             =   1.352816
Log likelihood   = -507.3352435                    BIC             =  -3960.002

------------------------------------------------------------------------------
             |                 OIM
        inlf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hsp |   .6504074   .1704773     3.82   0.000     .3162781    .9845368
       _cons |   .0998982    .086094     1.16   0.246     -.068843    .2686393
------------------------------------------------------------------------------
```

- Same coefficients, in log odds scale. Link is logit

Using GLM to get relative risk

```
  glm inlf hsp, family(binomial) link(log) nolog

Generalized linear models                          No. of obs      =        753
Optimization     : ML                              Residual df     =        751
                                                   Scale parameter =          1
Deviance         =  1014.670487                    (1/df) Deviance =   1.351093
Pearson          =          753                    (1/df) Pearson  =   1.002663

Variance function: V(u) = u*(1-u)                  [Bernoulli]
Link function    : g(u) = ln(u)                    [Log]

                                                   AIC             =   1.352816
Log likelihood   = -507.3352435                    BIC             =  -3960.002

------------------------------------------------------------------------------
             |                 OIM
        inlf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hsp |   .2576721   .0624513     4.13   0.000     .1352698    .3800743
       _cons |   -.644445   .0408986   -15.76   0.000    -.7246049   -.5642852
------------------------------------------------------------------------------
. di exp(0.2576721)
1.2939145
```

- Note that link is now **log**, not logit. The exponent of the coefficient is the relative risk. Check it matches our result by "hand," 1.29. No that the value of the log-likelihood is the same, **SEs are different!**

Why does it work?

- Because estimated model is now $log(p_i) = \beta_1 + \beta_1 hsp_i$
- Note that we are not taking the log of the outcome variable. We are still assuming that the outcome comes from a Bernoulli/Binomial distribution; in the likelihood $p = e^{(\beta_1 + \beta_1 hsp_i)}$
- So the difference between those with $hsp = 1$ and those with $hsp = 0$ is $log(p_{hsp}) - log(p_{nohsp}) = \beta_1$
- We can rewrite as $log(\frac{p_{hsp}}{p_{nohsp}}) = \beta_1$. Take exponent on both sides and we have $\frac{p_{hsp}}{p_{nohsp}} = e^{\beta_1}$
- Neat trick; GLM keeps on giving **but for inference stick to logistic**

Big picture

- A ratio of odds is hard to interpret at best. At worse, it is misleading
- We tend to think of them as a ratio of probabilities, but they are NOT
- Often there is little resemblance between relative probabilities and odds ratios (unless events are rare)
- They tend to be often misreported and confusing; same with ratio of probabilities
- For example, it sounds bad that event A is 10 times more likely to make you sick than event B, but that could be because $P_A = 0.001$ and $P_B = 0.0001$; their difference is 0.0009
- My personal opinion: A ratio of probabilities can be confusing, **a ratio of odds is EVIL**

Back to the continuous case

- Let's go back to the model $log(\frac{inlf_i}{1-inlf_i}) = \beta_0 + \beta_1 nwifeinc_i$
- We can also take $exp(\beta_1)$. In this case, $exp(-.0207569) = .97945704$
- A thousand dollars of extra income decreases the odds of participating in the labor force by a factor of 0.98
- Again, same issue. We can also solve for p or inlf in this case but not as easy as before because nwifeinc is *continous*
- We could take, as with the linear model, the derivative of p with respect to nwifeinc, but we know that it's non-linear so there is not a single effect; it depends on the values of nwifeinc
- Solution: **We will do it numerically**

Average prediction sketch

- We will do something that is conceptually very simple to numerically get the derivative
  1. Estimate the model
  2. For each observation, calculate predictions in the probability scale
  3. Increase the nwifeinc by a "small" amount and calculate predictions again
  4. Calculate the the change in the two predictions as a fraction of the change in nwifeinc. In other words, calculate $\frac{\Delta Y}{\Delta X}$, which is the **definition of the derivative**
  5. Take the **average** of the change in previous step across observations
- That's it

Numerical derivative

```
preserve
   qui logit inlf nwifeinc, nolog

   predict inlf_0 if e(sample)
   replace nwifeinc = nwifeinc + 0.011
   predict inlf_1 if e(sample)

   gen dydx = (inlf_1 - inlf_0) / 0.011

   sum dydx
restore

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        dydx |        753   -.0050217    .0001554   -.005191  -.0034977
```

- A small increase in extra income decreases the probability of entering the labor force by 0.005

That's what Stata calls **marginal effects**

```
qui logit inlf nwifeinc, nolog

margins, dydx(nwifeinc)

Average marginal effects                        Number of obs     =        753
Model VCE    : OIM

Expression   : Pr(inlf), predict()
dy/dx w.r.t. : nwifeinc

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    nwifeinc |  -.0050217   .0015533    -3.23   0.001     -.008066   -.0019773
------------------------------------------------------------------------------
```

- See, piece of cake! **We will cover in detail exactly how Stata does it (not same as my code)**

Margins for indicator variables

```
qui logit inlf i.hsp, nolog
margin, dydx(hsp)

Conditional marginal effects                      Number of obs     =        753
Model VCE    : OIM
Expression   : Pr(inlf), predict()
dy/dx w.r.t. : 1.hsp
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.hsp |   .1542915    .038583     4.00   0.000     .0786701    .2299128
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

- Same as what we found before doing it by hand. If we have covariates, we need to hold them constant at some value
- **Always use factor variable notation** with margins to avoid mistakes

Please be **fearful** of the margin command; it's healthy

```
margin, dydx(hsp)
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.hsp |   .1542915    .038583     4.00   0.000     .0786701    .2299128
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.

 margin i.hsp
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hsp |
          0  |   .5249538   .0214699    24.45   0.000     .4828736     .567034
          1  |   .6792453   .0320577    21.19   0.000     .6164134    .7420772
------------------------------------------------------------------------------
. margin
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .5683931   .0178717    31.80   0.000     .5333652     .603421
------------------------------------------------------------------------------
```

- **Small syntax changes make a big difference**. The third version is just the average prediction; same as observed proportion

Note on predictions and Stata and odds ratios

- By default, Stata calculates predictions in the probability scale
- You can also request predictions in the log-odds or logit scale
- By default, Stata shows you the coefficients in the estimation scale (that is, log-odds)
- You can also request coefficients in the odds-ration scale
- But since you know they are evil, don't do it

Sata things

```
qui logit inlf i.hsp nwifeinc, nolog

* Default, probability scale
predict hatp if e(sample)
(option pr assumed; Pr(inlf))

* Logit scale
predict hatp_l, xb

* Request odds ratios
logit inlf i.hsp nwifeinc, or nolog
------------------------------------------------------------------------------
        inlf | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.hsp |   2.461153   .4532018     4.89   0.000     1.715523    3.530861
    nwifeinc |   .9689898   .0069954    -4.36   0.000     .9553756    .9827981
       _cons |    1.95093    .303736     4.29   0.000     1.437872    2.647058
------------------------------------------------------------------------------
* That 2.46? 0.20 in probability scale, 39% more in relative probability:
margins hsp
-----------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         hsp |
          0  |   .5100289   .0213527    23.89   0.000     .4681784    .5518794
          1  |   .7137439   .0306774    23.27   0.000     .6536173    .7738705
------------------------------------------------------------------------------
di .7137439 /.5100289
1.3994185
di .7137439 -.5100289
.203715
```

Let's use probit now

```
  probit inlf i.hsp nwifeinc

Iteration 0:   log likelihood =  -514.8732
Iteration 1:   log likelihood = -496.87387
Iteration 2:   log likelihood = -496.81531
Iteration 3:   log likelihood = -496.81531

Probit regression                                 Number of obs     =        753
                                                  LR chi2(2)        =      36.12
                                                  Prob > chi2       =     0.0000
Log likelihood = -496.81531                       Pseudo R2         =     0.0351

------------------------------------------------------------------------------
        inlf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.hsp |   .5585346   .1118311     4.99   0.000     .3393497    .7777195
    nwifeinc |  -.0194555   .0043249    -4.50   0.000    -.0279322   -.0109787
       _cons |   .4140307   .0947847     4.37   0.000      .228256    .5998053
------------------------------------------------------------------------------
```

- Note that pseudo R^2 is close to logit, 0.0349
- No way to interpret coefficients other than with marginal effects. But how do we predict using probit?

Prediction in Probit models

- If coefficients are shifts in the cumulative standard, how do we make predictions?
- Well, calculating the probability given the index function
- Similar to using the inverse of the logistic response function

```
probit inlf i.hsp nwifeinc

* "By hand"
gen phat2 = normal(_b[_cons] + _b[1.hsp]*hsp + _b[nwifeinc]*nwifeinc)

* Using predict
predict phatprobit

sum phat phatprobit
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        phat |        753    .5688501    .1071053   .0971352   .8347558
  phatprobit |        753    .5688501    .1071053   .0971352   .8347558
```

Prediction in Probit models

- You can get some additional insight comparing the predicted index function $\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$ to the predicted probability
- In Stata, you can use the xb option in the predict command. Below, I do it by hand
- If the index function is postive, then predicted probabilities are greater than 0.5. Why, because that's how we define the threshold...

```
* Probability
gen phat2 = normal(_b[_cons] + _b[1.hsp]*hsp + _b[nwifeinc]*nwifeinc)

* Index function
gen xb = _b[_cons] + _b[1.hsp]*hsp + _b[nwifeinc]*nwifeinc

* Compare them
 sum phat2 if xb >0
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       phat2 |        590    .6066335    .0807844   .5004731   .8347558

sum phat2 if xb <0
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       phat2 |        163    .4320882    .0741921   .0971352   .4998522
```

Summary

- Main difficulty with logistic and probit models is to interpret parameters
- We estimate models in log-odds scale, we can easily convert coefficients into odds ratios but we really care about probabilities because a ratio of odds is not that informative (they are EVIL)
- **All effects in the probability scale are nonlinear in both models so the effect one variable depends on the value of that variable and the value of all other variables in the model**
- We can use numerical "derivatives" to come up with average predicted differences, what economists and Stata call marginal effects
- With more covariates, we just add our usual "holding other factors constant" or "taking into account other factors"
- We will do more of that next class