# ECON-C4200 - Econometrics II

## Lecture 6: Maximum likelihood approach to estimation

Otto Toivanen

# 1. Coin tosses

- Think of a tossing a coin that is potentially weighted, i.e., does not give the outcomes with 50% probability.
- Your task is to find out what the weight is.
- How to do this? Well, toss the coin lots and lots of times, record the outcomes.
- What then? Calculate the **share** of tails and heads, i.e., the **average** of tails / heads, i.e., the **probability** of getting tails / heads.

# 2. Bernoulli distribution

- More formally, you can think of what you did as a **stochastic process** with two possible outcomes, coded 0 and 1.
- Such a process is called a **Bernoulli process**.

# 2. Bernoulli distribution

- More formally, you can think of what you did as a **stochastic process** with two possible outcomes, coded 0 and 1.
- Such a process is called a **Bernoulli process**.
- It yields a sequence of 0s and 1s...

# 2. Bernoulli distribution

- More formally, you can think of what you did as a **stochastic process** with two possible outcomes, coded 0 and 1.
- Such a process is called a **Bernoulli process**.
- It yields a sequence of 0s and 1s...
- How to estimate the probability of 1 occuring?

# 3. Constructing the likelihood function

- How could we formalize this?

  1. Let's denote the probability of heads for any given coin toss with $P$. Then the probability of tails is $1 - P$.
  2. Let us toss the coin $N$ times, and index the coin tosses by $i$.
  3. Let us further denote the outcome of coin toss $i$ by $y_i$ which takes value $y_i = 1$ if heads, $y_i = 0$ if tails; $i = 1, ..., N$.

- Given $N$ coin tosses, our data are the outcomes $y_i$, and the unknown parameter is $P$.

- How can we estimate $P$?

# 3. Constructing the likelihood function

- Let's start by applying the tool we know, i.e., Least Squares (LS):

$$\min_P \sum_i (y_i - P)^2 \tag{1}$$

- We recall from Econometrics I that the answer LS gives is

$$\begin{aligned} \hat{P}^{LS} &= \frac{1}{N} \sum y_i \\ &= \frac{1}{N} (\underbrace{1 + 1 + ... + 1}_{n_H} + \underbrace{0 + 0 + ... + 0}_{N - n_h}) \\ &= \frac{n_h}{N} = \bar{y} \end{aligned} \tag{2}$$

- In other words, LS gives the answer we would have calculated without knowledge of econometrics.

# 3. Constructing the likelihood function

- Let's take another approach and ask ourselves: With $N$ coin tosses, what is the **likelihood** of getting $n_H$ heads and $N - n_H = n_T$ tails, given $P$?

# 3. Constructing the likelihood function

- Let's take another approach and ask ourselves: With $N$ coin tosses, what is the **likelihood** of getting $n_H$ heads and $N - n_H = n_T$ tails, given $P$?
- Answer:

$$L = P^{n_H}(1 - P)^{N - n_H} \tag{3}$$

- Equation (3) is the **Likelihood function** (uskottavuusfunktio) for our data, and also our problem (of finding the best estimate of $P$).

# 3. Constructing the likelihood function

- What is the next step?
- Let's find the value for $P$ that maximizes the likelihood of observing exactly $n_H$ heads and $N - n_H$ tails.
- How to do this? By maximizing the likelihood function with respect to the unknown parameter $P$, i.e., by (recall $y_i = 1$ if coin toss $i$ gives heads, $y_i = 0$ if tails):

$$\begin{aligned} \max_P \ L = & \prod_i P^{y_i}(1-P)^{1-y_i} \\ = & \underbrace{P \times P \times ... \times P}_{n_H} \times \underbrace{(1-P) \times (1-P) ... \times (1-P)}_{N-n_H} \\ = & P^{n_H}(1-P)^{N-n_H} \end{aligned} \tag{4}$$

- This can obviously be done, but often the likelihood function is difficult to work with.

# 3. Constructing the likelihood function

- Trick: let's use a monotonic transformation, i.e., let's take logs:

$$\begin{aligned}\max_P \ln L &= \sum_i [y_i \ln P + (1-y_i)\ln(1-P)] \\ &= \sum_{n_H} \ln P + \sum_{N-n_H} \ln(1-P) \\ &= n_H \ln P + (N-n_H)\ln(1-P)\end{aligned} \tag{5}$$

- Now do the differentiation and solve for $P$.

# 3. Constructing the likelihood function

- The ML estimate of $P$, $\hat{P}^{ML}$, is:

$$\hat{P}^{ML} = \frac{n_H}{N} = \hat{P}^{LS} \tag{6}$$

- Note: the ML estimate is not always equal to the LS estimate.

# 3. Constructing the likelihood function

- The idea underlying ML: construct the likelihood function.
- Ask: what parameter values are the likeliest to have lead to the data we observe?

# 4. ML estimation with observable characteristics

- Thus far we did not have any explanatory variables, i.e., observable characteristics of the observation units.
- To extend our coin example, assume that instead of tossing a single coin $N$ times, you toss $N$ different coins once each.
- Assume further that you observe some characteristics of each coin $i$. Denote the characteristics with $\boldsymbol{X}$.
- Let suppose you want to study how characteristics $\boldsymbol{X}$ affect the probability of getting heads.

# 4. ML estimation with observable characteristics

- By now you know how to build a linear probability model for this setting.
- How could you introduce the explanatory variable into our ML setup?

# 4. ML estimation with observable characteristics

- By building on what we studied in the previous lecture.
- Step #1: Assume that

$$y_i = 1 \Leftrightarrow \boldsymbol{X}_i \beta + \epsilon_i \geq 0$$

$$y_i = 0 \Leftrightarrow \boldsymbol{X}_i \beta + \epsilon_i < 0$$

- Step #2: assume a distribution for $\epsilon$. Let's denote the CDF of $\epsilon$ with $F(.)$. Let's further assume it is symmetric.

# 4. ML estimation with observable characteristics

- Step #3: Now (due to the symmetry of $F(.)$) the probability of observing $y_i = 1$ is

$$1 - F(-\boldsymbol{X}_i\beta) = F(\boldsymbol{X}_i\beta)$$

- Notice that this is not that different from assuming the probability of observing $y_i = 1$ is $P$.
- Indeed, I can replace $P$ with $F(\boldsymbol{X}_i\beta)$ in the likelihood function we just worked with.
- The difference is that the unknown parameters are now $\boldsymbol{\beta}$, not $P$.

# 4. ML estimation with observable characteristics

- We can now write the likelihood and the log likelihood functions as:

$$L = Pr(Y_1 = y_1, ..., Y_N = y_N) = \prod_i F(\boldsymbol{X}_i \beta)^{y_i} [1 - F(\boldsymbol{X}_i \beta)]^{1-y_i} \quad (7)$$

$$\ln L = \sum_i \{ y_i \ln F(\boldsymbol{X_i}\beta) + (1 - y_{\boldsymbol{i}}) \ln [1 - F(\boldsymbol{X_i}\beta)] \} \quad (8)$$

- The marginal effect (wrt. to the $k^{th}$ expl. variable $X_k$) is now given by:

$$\frac{\partial F(\boldsymbol{X_i}\beta)}{\partial X_k} = f(\boldsymbol{X_i}\beta)\beta_k \quad (9)$$

# 4. ML estimation with observable characteristics

- Key question: What is $F()$?
- Obviously, $F()$ is a cdf and hence $[0, 1]$.
- $F()$ need not be symmetric (around 0), but most of the time is.

# 4. ML estimation with observable characteristics

- $F()$ could come from:

1. Theory (= assumptions).
2. Data (non- / semi-parametric regression).
3. Past practice.

# 4. ML estimation with observable characteristics

- Does the choice matter $F()$ empirically?
- Experience shows that in most ("well-behaved") data sets and as long as $F(.)$ symmetric, makes essentially no difference to marginal effects.
- Key for being "well-behaved"; mean of the dependent variable neither "very" large nor "very" small.

# 5. Estimation

- If we assume that the error term has a normal distribution, then we are estimating a **probit** model.
- Another popular model is the **logit** model where error term has an extreme value distribution. This yields the following expression for the probability that $y_i = 1$:

$$Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \frac{\exp(\boldsymbol{x}\beta)}{\exp(0) + \exp(\boldsymbol{x}\beta)} = \frac{\exp(\boldsymbol{x}\beta)}{1 + \exp(\boldsymbol{x}\beta)}$$

- Note that the exp(0) in the denominator is the exponential of the utility from choosing the outside good, which has been normalized to be zero.

# 5. Estimation

- One cannot estimate probit or logit with OLS.
- One needs either
  1. maximum likelihood (this is what the Stata probit / logit functions do).
  2. nonlinear least squares (usually not used)
  3. generalized method of moments (sometimes used).
- Let's estimate the VI decision of cinema's in Gil's data with OLS, probit and logit.
- Unlike OLS, where we can solve for the coefficients using matrix algebra, ML models require (numerical) optimization.

# How to calculate the ME?

1. The derivative is going to depend on X.
2. Different ME for each possible value of X.
3. How to average?

# How to calculate the ME?

- Many solutions:

1. Only at the mean of X (and other variables).
2. At some interesting value of X.
3. Some avg example: weighted avg.

# Stata commands for OLS, probit and logit

## Stata code

```
regr vi_ever capacity_1000 , robust
probit vi_ever capacity_1000
margins
logit vi_ever capacity_1000
margins
```

# OLS results

```
. regr vi_ever capacity_1000, robust

Linear regression                               Number of obs     =        393
                                                F(1, 391)         =     108.07
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1844
                                                Root MSE          =     .44887
```

| vi_ever | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| capacity_1000 | .1841115 | .0177105 | 10.40 | 0.000 | .1492917 | .2189313 |
| _cons | .1281997 | .0355462 | 3.61 | 0.000 | .0583142 | .1980853 |

# Probit results

```
. probit vi_ever capacity_1000

Iteration 0:   log likelihood = -269.08833
Iteration 1:   log likelihood = -229.07358
Iteration 2:   log likelihood = -228.75776
Iteration 3:   log likelihood = -228.75752
Iteration 4:   log likelihood = -228.75752

Probit regression                               Number of obs     =        393
                                                LR chi2(1)        =      80.66
                                                Prob > chi2       =     0.0000
Log likelihood = -228.75752                     Pseudo R2         =     0.1499

------------------------------------------------------------------------------
     vi_ever |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
capacity_1000 |   .5689945   .0698458     8.15   0.000     .4320993    .7058897
        _cons |  -1.108613   .1320205    -8.40   0.000    -1.367369   -.8498576
------------------------------------------------------------------------------

. margins

Predictive margins                              Number of obs     =        393
Model VCE    : OIM

Expression   : Pr(vi_ever), predict()

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .4314227   .0225493    19.13   0.000     .3872268    .4756185
------------------------------------------------------------------------------
```

# Logit results

```
. logit vi_ever capacity_1000

Iteration 0:   log likelihood = -269.08833
Iteration 1:   log likelihood = -228.60832
Iteration 2:   log likelihood = -228.44027
Iteration 3:   log likelihood = -228.44013
Iteration 4:   log likelihood = -228.44013

Logistic regression                             Number of obs     =        393
                                                LR chi2(1)        =      81.30
                                                Prob > chi2       =     0.0000
Log likelihood = -228.44013                     Pseudo R2         =     0.1511

------------------------------------------------------------------------------
     vi_ever |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
capacity_1000 |   .9644566   .1268482     7.60   0.000     .7158387    1.213075
       _cons |  -1.843564   .2296731    -8.03   0.000    -2.293715   -1.393413
------------------------------------------------------------------------------

. margins

Predictive margins                              Number of obs     =        393
Model VCE    : OIM

Expression   : Pr(vi_ever), predict()

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .4351145   .0224736    19.36   0.000     .3910671    .4791619
------------------------------------------------------------------------------
```

# Stata commands different marginal effects

## Stata code

```
probit vi_ever capacity_1000
margins
margins , atmeans
logit vi_ever capacity_1000
margins
margins , atmeans
```

# Probit results

```
. margins

Predictive margins                              Number of obs     =        393
Model VCE    : OIM

Expression   : Pr(vi_ever), predict()
```

| | Margin | Delta-method Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| _cons | .4314227 | .0225493 | 19.13 | 0.000 | .3872268 | .4756185 |

```
. margins , atmeans

Adjusted predictions                            Number of obs     =        393
Model VCE    : OIM

Expression   : Pr(vi_ever), predict()
at           : capacit~1000    =    1.667005 (mean)
```

| | Margin | Delta-method Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| _cons | .4364026 | .026794 | 16.29 | 0.000 | .3838872 | .4889179 |

# Logit results

```
. margins

Predictive margins                              Number of obs     =        393
Model VCE    : OIM

Expression   : Pr(vi_ever), predict()

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .4351145   .0224736    19.36   0.000     .3910671    .4791619
------------------------------------------------------------------------------

. margins , atmeans

Adjusted predictions                            Number of obs     =        393
Model VCE    : OIM

Expression   : Pr(vi_ever), predict()
at           : capacit~1000    =    1.667005 (mean)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .4413192   .0280628    15.73   0.000     .3863171    .4963214
------------------------------------------------------------------------------
```

# Probit, Logit, ...?

- One can use any cumulative density function (cdf).
- Most popular are probit and logit.
- Differences in ME between probit and logit small. If you only are interested in ME (and especially with large data), OLS works OK.
- Choice may depend on convenience / prior practice.

# Why not LPM?

- Sometimes you are interested in the actual parameters, not only the ME.
- Example: estimating the demand for a good in order to understand substitution patterns.