ECON207 Session 10

Maximum Likelihood Estimation / Limited Dependent Variables

Anthony Tay

This Version: 05 Nov 2024

Agenda and Quick Review

- Introduce Maximum Likelihood Estimation using a Simple Example
- General Principles
- Hypothesis Testing / Model Selection
- Application to Limited Dependent Variables

Roughly speaking, ML estimation takes as estimators the values of the parameters that maximize the probability of obtaining the observed sample

R packages for this session:

```
library(tidyverse); library(patchwork); library(stargazer)
library(sandwich); library(lmtest); library(margins);
library(wooldridge); library(AER); library(truncreg)
```


Quick Probability Review

If $Y \sim N(\mu, \sigma^2)$, then its pdf is

$$f_Y(y; \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$$

Purpose of pdf is to describe probability of events involving Y

e.g., If $Y \sim N(\mu, \sigma^2)$, then what is $Pr(Y > 1)$?

Ans:

$$\Pr(Y > 1) = \int_1^\infty (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} dy$$

Quick Probability Review

Multivariate Normal Distribution

If $X_K \sim N_K(\mu, \Sigma)$ where μ is $(K \times 1)$ and Σ is $(K \times K)$, then

$$f(x) = (2\pi)^{-K/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right\}$$

For X, Y bivariate normal distribution, can be shown that

$$Y \mid X \sim \text{Normal}(\mu_{y|x}, \sigma^2_{y|x}) \quad \text{and } X \sim \text{Normal}(\mu_x, \sigma^2_x)$$

where

$$\mu_{y|x} = \mu_y + \frac{\sigma_{xy}}{\sigma^2_x}(x - \mu_x) \quad \text{and} \quad \sigma^2_{y|x} = \sigma^2_y - \frac{\sigma_{xy}}{\sigma^2_x}$$

Quick Probability Review

In general,

- For two random variables X and Y: $f_{X,Y}(x,y) = f_{Y|X}(y|x)f_X(x)$
- Can extend to three or more variables:

$$f_{X,Y,Z}(x,y,z) = f_{Z|Y,X}(z \mid y,x) f_{Y|X}(y \mid x) f_X(x)$$

- For multivariate normal, all conditionals are normal
- I'll drop the subscripts from now: $f(x,y,z)$ will mean $f_{X,Y,Z}(x,y,z)$, etc.
- If X, Y, Z are independent, then $f(x,y,z) = f(z)f(y)f(z)$

Intro: A Simple Coin Toss Example

Objective: Estimate probability p of observing heads for a certain coin

Observations: $\{Head, Tail, Tail\}$ coded as $\{1, 0, 0\}$

$$Y_i \overset{iid}{\sim} \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

Prior to observing outcomes, probability of obtaining $\{1,0,0\}$ is

$$\Pr(Y_1 = 1, Y_2 = 0, Y_3 = 0 \mid p) = p(1-p)^2$$

What is the "most likely" value for p?

- If $p = 0.1$, then $\Pr(Y_1 = 1, Y_2 = 0, Y_3 = 0) = 0.1(1-0.1)^2 = 0.081$
- If $p = 0.9$, then $\Pr(Y_1 = 1, Y_2 = 0, Y_3 = 0) = 0.9(1-0.9)^2 = 0.009$

Both seem "unlikely" (though certainly possible)

Intro: A Simple Coin Toss Example

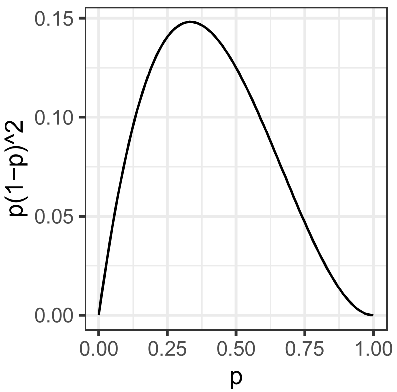
Plot of

$$\Pr(Y_1 = 1, Y_2 = 0, Y_3 = 0 \mid p) = p(1-p)^2$$

A reasonable estimate would seem to be in the vicinity of 0.3

In fact $p(1-p)^2$ is maximized at $p = 1/3$

Our maximum likelihood estimate for p is $\hat{p} = 1/3$



Maximum Likelihood Estimation

Maximum Likelihood choose as estimates those parameter values that make your sample observations the most likely, prior to seeing the sample

Maximum likelihood estimators are (subject to certain conditions holding)

- Consistent
- Asymptotically efficient
- Asymptotically normal
- may be biased in finite samples

Maximum Likelihood Estimation

Step 1 Write down a "complete model" describing the data generating process

- Y_i iid with mean $E(Y_i) = \mu$ and $Var(Y_i) = \sigma^2$, $i = 1, 2, \dots, n$ is not complete
- $Y_i \overset{iid}{\sim}$ Normal(μ, σ^2) , $i = 1, 2, \dots, n$ is a complete model

With a complete model you can write down the p.d.f. for your data

$$p(Y_1, Y_2, \dots, Y_n \mid \theta)$$

where θ is a vector containing your parameters

E.g., if $Y_i \overset{iid}{\sim}$ Normal(μ, σ^2) , $i = 1, 2, \dots, n$, then

$$p(y_1, y_2, \dots, y_n | \mu, \sigma^2) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{1}{2} \frac{(y_i - \mu)^2}{\sigma^2} \right\}$$

Maximum Likelihood Estimation

Step 2 Reinterpret your pdf as a function of your parameters given the data, i.e.,

$$L(\theta \mid Y_1, Y_2, \dots, Y_n)$$

and call this the "Likelihood Function"

E.g., if $Y_i \overset{iid}{\sim}$ Normal(μ, σ^2) , $i = 1, 2, \dots, n$, then

$$L(\mu, \sigma^2 \mid Y_1, Y_2, \dots, Y_n) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{1}{2} \frac{(Y_i - \mu)^2}{\sigma^2} \right\}$$

The likelihood function and the pdf have the same form, but different interpretation

Maximum Likelihood Estimation

Step 3 The maximum likelihood estimator is the value of the parameters that maximize the Likelihood Function

E.g., for the $Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$, $i = 1, 2, \dots, n$ example, we have

$$\hat{\mu}, \widehat{\sigma^2} = \underset{\mu, \sigma^2}{\arg\max} \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{1}{2} \frac{(Y_i - \mu)^2}{\sigma^2} \right\}$$

Maximum Likelihood Estimation

We often maximize the log-likelihood instead of the likelihood

e.g., for the $Y_i \overset{iid}{\sim}$ Normal(μ, σ^2), example,

$$L(\mu, \sigma^2 \mid Y_1, Y_2, \dots, Y_n) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{1}{2} \frac{(Y_i - \mu)^2}{\sigma^2} \right\}$$

$$\ln L(\mu, \sigma^2 \mid Y_1, Y_2, \dots, Y_n) = \sum_{i=1}^{n} -\frac{1}{2} \ln(2\pi\sigma^2) + \sum_{i=1}^{n} \left\{ -\frac{1}{2} \frac{(Y_i - \mu)^2}{\sigma^2} \right\}$$

$$= -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \mu)^2$$

Maximum Likelihood Estimation

(Example continued) Straightforward to show

$$\hat{\mu}_{ml} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}$$
$$\widehat{\sigma^2}_{ml} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

Notice that the ML estimator for μ is unbiased and consistent, but the ML estimator for σ^2 is biased (but consistent)

ML estimator is also asymptotically normal, but what is the estimator variance?

Maximum Likelihood Estimation

Suppose the likelihood function is $L(\theta \mid Y_1, Y_2, \dots, Y_n)$ (θ may be a vector)

Then

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \text{Normal}\left(0, -nE\left(\frac{\partial^2 \ln L(\theta \mid Y_1, Y_2, \dots, Y_n)}{\partial\theta\partial\theta^{\mathrm{T}}}\right)^{-1}\right)$$

i.e., we can use the asymptotic approximation

$$\hat{\theta} \overset{a}{\sim} \text{Normal}\left(\theta, -E\left(\frac{\partial^2 \ln L(\theta \mid Y_1, Y_2, \dots, Y_n)}{\partial\theta\partial\theta^{\mathrm{T}}}\right)^{-1}\right)$$

Maximum Likelihood Estimation

In practice, we further approximate

$$-E\left(\frac{\partial^2 \ln L(\theta \mid Y_1, Y_2, \dots, Y_n)}{\partial\theta\partial\theta^{\mathrm{T}}}\right)^{-1} \quad \text{with} \quad -E\left(\frac{\partial^2 \ln L(\hat{\theta} \mid Y_1, Y_2, \dots, Y_n)}{\partial\theta\partial\theta^{\mathrm{T}}}\right)^{-1}$$

The expression

$$I(\theta) = -E\left(\frac{\partial^2 \ln L(\theta \mid Y_1, Y_2, \dots, Y_n)}{\partial\theta\partial\theta^{\mathrm{T}}}\right)$$

is called the "information matrix"

Maximum Likelihood Estimation

For the $Y_i \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2)$, $i = 1, 2, \dots, n$ example, we can show

$$-E\left(\frac{\partial^2 \ln L(\theta \mid Y_1, \dots, Y_n)}{\partial\theta\partial\theta^{\mathrm{T}}}\right) = -E\begin{bmatrix} \frac{\partial^2 \ln L(\theta \mid Y_1, \dots, Y_n)}{(\partial\mu)^2} & \frac{\partial^2 \ln L(\theta \mid Y_1, \dots, Y_n)}{\partial\mu\partial\sigma^2} \\ \frac{\partial^2 \ln L(\theta \mid Y_1, \dots, Y_n)}{\partial\mu\partial\sigma^2} & \frac{\partial^2 \ln L(\theta \mid Y_1, \dots, Y_n)}{(\partial\sigma^2)^2} \end{bmatrix}$$
$$= \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{bmatrix}$$

$$\text{therefore} \quad \begin{bmatrix} \hat{\mu} \\ \widehat{\sigma^2} \end{bmatrix} \overset{a}{\sim} \text{Normal}\left(\begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}, \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}\right)$$

General Coin Toss Example

$$Y_i = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases} \qquad \text{iid} \qquad i = 1, 2, \ldots, n$$

PDF: $p(y_1, y_2, \ldots, y_n \mid p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$

Likelihood: $L(p \mid Y_1, Y_2, \ldots, Y_n) = \prod_{i=1}^{n} p^{Y_i}(1-p)^{1-Y_i}$

log-Likelihood: $\ln L(p \mid Y_1, Y_2, \ldots, Y_n) = \ln p \sum_{i=1}^{n} Y_i + \ln(1-p) \sum_{i=1}^{n}(1-Y_i)$

Maximize:

$$\frac{\partial \ln L(\hat{p})}{\partial p} = \frac{1}{\hat{p}} \sum_{i=1}^{n} Y_i - \frac{1}{1-\hat{p}} \sum_{i=1}^{n}(1-Y_i) = 0 \;\; \Rightarrow \;\; \hat{p} = \overline{Y}$$

General Coin Toss Example

We also have

$$\frac{\partial^2 \ln L(p)}{\partial p^2} = -\frac{1}{p^2} \sum_{i=1}^{n} Y_i - \frac{1}{(1-p)^2} \sum_{i=1}^{n}(1-Y_i)$$

$$E\left(\frac{\partial^2 \ln L(p)}{\partial p^2}\right) = -\frac{1}{p^2} np - \frac{1}{(1-p)^2} n(1-p) = -\frac{n}{p(1-p)}$$

$$-E\left(\frac{\partial^2 \ln L(p)}{\partial p^2}\right)^{-1} = \frac{p(1-p)}{n}$$

That is,

$$\hat{p} \overset{a}{\sim} \text{Normal}\left(p\,,\, \frac{p(1-p)}{n}\right)$$

OLS

Suppose

$$y = X\beta + \varepsilon\,,\; \varepsilon \sim N(0, \sigma^2 I_n)$$

where X is $n \times K$ and β is $K \times 1$, i.e.,

$$\varepsilon \sim \text{Normal}_n(X\beta, \sigma^2 I_n)$$

$$f(y \mid X; \beta, \sigma^2) = (2\pi)^{-n/2} |\sigma^2 I_n|^{-1/2} \exp\left\{-\frac{1}{2}\varepsilon^{\mathrm{T}}(\sigma^2 I_n)^{-1}\varepsilon\right\}$$

$$= (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)^{\mathrm{T}}(y - X\beta)\right\}$$

$$\ln L(\beta, \sigma^2 \mid y, X) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(y - X\beta)^{\mathrm{T}}(y - X\beta)$$

OLS

Can show ML estimators for β and σ^2 are

$$\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y \quad \text{and} \quad \widehat{\sigma^2} = \frac{\hat{\varepsilon}^{\mathrm{T}}\hat{\varepsilon}}{n}$$

where $\hat{\varepsilon} = y - X\hat{\beta}$, and

$$\begin{bmatrix} \hat{\beta} \\ \widehat{\sigma^2} \end{bmatrix} \overset{a}{\sim} \text{Normal}\left(\begin{bmatrix} \beta \\ \sigma^2 \end{bmatrix}, \begin{bmatrix} \sigma^2(X^{\mathrm{T}}X)^{-1} & 0 \\ 0 & \dfrac{2\sigma^4}{n} \end{bmatrix}\right)$$

Remark

In most cases, we are unable to derive a formula for our ML estimators

It seems more common to have to maximize the log-likelihood numerically

Likewise, the information matrix is often also found numerically

We will not concern ourselves with the details

- In the applications of interest to us, the numerical maximization routine is included in the packages we use

Likelihood Ratio Test

Suppose $\ln L(\theta)$ where θ is $(K \times 1)$ vector of parameters

You want to test $R\theta = r_0$ where R is $J \times K$

1. Maximize $\ln L(\theta)$, get $\ln L_{UR} = \ln L(\hat{\theta})$
2. Maximize $\ln L(\theta)$ subject to $R\theta = r_0$, get $\ln L_R = \ln L(\tilde{\theta})$ where $\tilde{\theta}$ satisfies $R\tilde{\theta} = r_0$
3. Under the null hypothesis

$$LR = 2(\ln L_{UR} - \ln L_R) \overset{a}{\sim} \chi^2_{(J)}$$

Likelihood Ratio Test

Example: Jointly test $\beta_1 = 0$ and $\beta_2 = \beta_3$ in:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_K X_{iK} + \epsilon_i\,,\ \epsilon_i \sim N(0, \sigma^2)$$

Get $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_K, \widehat{\sigma^2}$ (unrestricted parameter estimates)

Get $\tilde{\beta}_0, \tilde{\beta}_2 = \tilde{\beta}_3, \tilde{\beta}_4, \dots, \tilde{\beta}_K, \widetilde{\sigma^2}$ (restricted parameter estimates)

Note that

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \dots - \hat{\beta}_K X_{iK})^2$$

$$\widetilde{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \tilde{\beta}_0 - \tilde{\beta}_2(X_{i2} + X_{i3}) - \tilde{\beta}_4 X_{i4} - \dots - \tilde{\beta}_K X_{iK})^2$$

Likelihood Ratio Test

$$\begin{aligned}
\ln L_{UR} &= -\frac{1}{2}\sum_{i=1}^{n}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\ln\widehat{\sigma^2} - \frac{1}{2\widehat{\sigma^2}}\sum_{i=1}^{n}(Y_t - \widehat{\beta}_0 - \widehat{\beta}_1 X_{i1} - \dots - \widehat{\beta}_K X_{iK})^2 \\
&= -\frac{1}{2}\sum_{i=1}^{n}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\ln\widehat{\sigma^2} - \frac{n}{2} \\
\ln L_{R} &= -\frac{1}{2}\sum_{i=1}^{n}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\ln\widetilde{\sigma^2} - \frac{1}{2\widetilde{\sigma^2}}\sum_{i=1}^{n}(Y_i - \widetilde{\beta}_0 - \widetilde{\beta}_2(X_{i2} + X_{i3}) - \widetilde{\beta}_4 X_{i4} - \dots \\
&\qquad - \widetilde{\beta}_K X_{iK})^2 \\
&= -\frac{1}{2}\sum_{i=1}^{n}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\ln\widetilde{\sigma^2} - \frac{n}{2}
\end{aligned}$$

$$LR = 2(\ln L_{UR} - \ln L_R) = n(\ln\widetilde{\sigma^2} - \ln\widehat{\sigma^2}) \sim \chi^2_{(2)}$$

Information Criteria

Information Criterion Model Selection

- A higher likelihood indicates a better fit, or a "more likely" model
- max value of log-likelihood decreases when restrictions are imposed (like R^2)
- if restrictions are very wrong, log-likelihood will decrease substantially

Information Criteria: choose model with a higher log-likelihood, but with a penalty term to control for excessive parameterization (similar to adjusted-R^2)

Information Criteria

Akaike Information Criterion (AIC)

$$AIC = -2\ln L + 2q$$

where q is number of parameters in the model

- log-likelihood is negated, try to minimize AIC
- attempting to lower AIC by using "larger" model will succeed only if fall in $\ln L$ is greater than increase in q

Bayes Information Criterion (BIC) / Schwarz Information Criterion (SIC):

$$SIC = -2\ln L + q\ln n$$

- Again, choose model with lower SIC

Information Criteria

Remarks:

- Some implementations of AIC/ SIC divide expression throughout by n
  - not a problem if comparing models with same software package
- Models compared must be estimated over the same sample period
- AIC and SIC can be used to compare non-nested models (unlike LR test)
- Dependent variable must be the same across models
  - e.g., cannot compare AIC / SIC of model for Y_i vs those of model for $\ln Y_i$

Information Criteria

The AIC / SIC differ in their asymptotic properties

The SIC is "consistent"

- SIC chooses correct model asymp. if true model is in class of candidate models

AIC is not "consistent"

- positive prob. that AIC does not choose correct model even when true model is in class of candidate models
- However, AIC may perform better than SIC when true model is not in class of candidate models, and in finite samples even when it is

SIC contains a stronger penalty term for additional parameters, and will choose more parsimonious models (fewer parameters) than the AIC

Limited Dependent Variables

Limited Dependent Variables

Situations where dependent variable Y has some special structure

- $Y = 0, 1$
- $Y = 1, 2, 3, \ldots$
- Y is truncated or censored, or there are "corner solutions"

Linear Probability Model

Recall that if Y takes values 1 or 0 only, with probability p and $1-p$ respectively, then $E(Y) = p$ and $Var(Y) = p(1-p)$

This also applies to conditional expectations, e.g.,

$$E(Y \mid X_1, \ldots, X_K) = \Pr(Y = 1 \mid X_1, \ldots, X_K)$$

Regression of Y on $X_1, \ldots, X_K$ estimates $E(Y \mid X_1, \ldots, X_K)$ assuming $E(Y \mid X_1 \ldots, X_K) = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K$

If Y is binary, then

$$\widehat{\Pr}(Y = 1 \mid X_1, \ldots, X_K) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_K X_K$$

"Linear Probability Model (LPM)"

Logit / Probit Models

Difficulty of LPM is that predicted probabilities may be below 0 or above 1

Logit and Probit models solve this issue by assuming

$$\Pr(Y = 1 \mid X_1, \ldots, X_K) = F(\beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K)$$

where $F(.)$ is a Cumulative Distribution Function

- F is a function such that $\lim_{x \to -\infty} F(z) = 0$ and $\lim_{x \to +\infty} F(z) = 1$
- F is sometimes called the "link" function

Logit / Probit Models

Then

$$\widehat{\Pr}(Y = 1 \mid X_1, \ldots, X_K) = F(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_K X_K) \in (0, 1)$$

Logit: $F(z)$ is the CDF of the *logistic distribution* $\Lambda(z) = \dfrac{e^Z}{1 + e^Z}$, i.e.

$$\Pr(Y = 1 \mid X_1, \ldots, X_K) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K}}$$

Probit: $F(z)$ is the CDF of the *standard normal distribution* $\Phi(z)$, i.e.

$$\Pr(Y = 1 \mid X_1, \ldots, X_K) = \int_{-\infty}^{Z} (2\pi)^{-1/2} \exp(-u^2/2)\, du$$

where $Z = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K$

Logit / Probit Models

A justification for this approach

Suppose $Y^* = \beta_0 + \beta_1 X_1 + \epsilon$, $Y = 1$ if $Y^* > 0$, where ϵ has cdf $F(.)$, then

$$\begin{aligned}
\Pr(Y = 1 \mid X_1) &= \Pr(Y^* > 0 \mid X_1) \\
&= \Pr(\beta_0 + \beta_1 X_1 + \epsilon > 0 \mid X_1) \\
&= \Pr(\epsilon > -\beta_0 - \beta_1 X_1 \mid X_1) \\
&= 1 - \Pr(\epsilon \leq -\beta_0 - \beta_1 X_1 \mid X_1) \\
&= 1 - F(-\beta_0 - \beta_1 X_1 \mid X_1) \\
&= F(\beta_0 + \beta_1 X_1 \mid X_1)
\end{aligned}$$

if pdf is symmetric

Logit / Probit Models

Estimation by Maximum Likelihood (maximized numerically)

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} F(Z_i)^{Y_i}(1 - F(Z_i))^{1-Y_i} \quad \text{where} \quad Z_i = \beta_0 + \beta_1 X_{i1} + \dots \beta_K X_{iK}$$

or $$\ln \mathcal{L}(\beta) = \sum_{i=1}^{n} Y_i \ln F(Z_i) + \sum_{i=1}^{n} (1 - Y_i) \ln(1 - F(Z_i))$$

Together with coefficients, often report AIC, SIC, and Pseudo-$R^2 = 1 - \dfrac{\ln \mathcal{L}_{ur}}{\ln \mathcal{L}_0}$

- note that $\ln \mathcal{L}_0 \leq \ln \mathcal{L}_{ur} \leq 0$
- $\ln \mathcal{L}_0$ is log-Likelihood from model with intercept only

LPM, Logit, Probit Example

Dataset: mroz

inlf = in labor force
nwifeinc = non-wife income
educ = years of education

```
library(stargazer)
library(wooldridge)
library(sandwich)
library(lmtest)

data(mroz)
probit_mdl0 <- glm(inlf ~ 1, data=mroz, family=binomial(link="probit"))
probit_mdl1 <- glm(inlf ~ educ, data=mroz, family=binomial(link="probit"))
probit_mdl2 <- glm(inlf ~ nwifeinc + educ + age + kidslt6 + kidsge6, data=mroz,
                   family=binomial(link="probit"))
stargazer(probit_mdl0, probit_mdl1, probit_mdl2,
  type="text",
  add.lines = list(
    c("Pseudo R-sqr",
      round(1 - logLik(probit_mdl0)[1]/logLik(probit_mdl0)[1], 3),
      round(1 - logLik(probit_mdl1)[1]/logLik(probit_mdl0)[1], 3),
      round(1 - logLik(probit_mdl2)[1]/logLik(logit_mdl0)[1], 3))),
  align=TRUE, no.space=TRUE)
```

	Dependent variable:		
		inlf	
	(1)	(2)	(3)
nwifeinc			-0.021***
			(0.005)
educ		0.108***	0.156***
		(0.021)	(0.024)
age			-0.034***
			(0.008)
kidslt6			-0.892***
			(0.115)
kidsge6			-0.038
			(0.041)
Constant	0.172***	-1.148***	0.422
	(0.046)	(0.261)	(0.472)
Pseudo R-sqr	0	0.026	0.118
Observations	753	753	753
Log Likelihood	-514.873	-501.302	-454.225
Akaike Inf. Crit.	1,031.746	1,006.604	920.451
Note:	*p<0.1;	**p<0.05;	***p<0.01

Linear Probability Model and Logit / Probit Models

For LPM: $\dfrac{\partial \Pr(Y = 1 \mid X_1, \dots, X_K)}{\partial X_k} = \beta_k$

For Logit / Probit models

$$\begin{aligned}
\frac{\partial \Pr(Y = 1 \mid X_1, \dots, X_K)}{\partial X_k} &= \beta_k \frac{\partial F(\beta_0 + \beta_1 X_1 + \dots + \beta_K X_K)}{\partial X_k} \\
&= \beta_k f(\beta_0 + \beta_1 X_1 + \dots + \beta_K X_K)
\end{aligned}$$

where f is the p.d.f. corresponding to F

- This is the main difference between LPM and Logit/Probit (apart from the fact that LPM can give you probability estimates less than zero or greater than one)
- Care must be taken when comparing LPM / Logt / Probit coefficient estimates

LPM, Logit, Probit Example

Usually compare β_k from LPM with

- Partial Effect at the Average: $\hat{\beta}_k f(\hat{\beta}_0 + \hat{\beta}_1 \overline{X_1} + \cdots + \hat{\beta}_K \overline{X_K})$
- Average Partial Effect: $\hat{\beta}_k \left[\frac{1}{n} f(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_K X_{iK}) \right]$

from Logit / Probit

If $f(\hat{\beta}_0 + \hat{\beta}_1 \overline{X_1} + \cdots + \hat{\beta}_K \overline{X_K})$ or $\frac{1}{n} f(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_K X_{iK})$ is unavailable, can use $\lambda(0) \approx 0.25$ for logit or $\phi(0) \approx 0.4$ for probit

LPM, Logit, Probit Example

```
# Estimate Models
lpm_mdl <- lm(inlf ~ nwifeinc + educ + age + kidslt6 + kidsge6, data=mroz)  # LPM Model
robust_se_lpm <- sqrt(diag(vcovHC(lpm_mdl, type="HC0")))  # Always use HC Standard Errors for LPM
probit_mdl <- glm(inlf ~ nwifeinc + educ + age + kidslt6 + kidsge6, data=mroz, family=binomial(link="probit")) # Probit
logit_mdl <- glm(inlf ~ nwifeinc + educ + age + kidslt6 + kidsge6, data=mroz, family=binomial(link="logit"))   # Logit

# Show LPM, Probit and Logit Output
stargazer(lpm_mdl, probit_mdl, logit_mdl, type="text", se = list(robust_se_lpm, NULL, NULL),
          omit.stat = "all")

# Show APE for Probit and Logit
cat("\n APE Probit")
margins(probit_mdl) %>% summary()
cat("\n APE Logit")
margins(logit_mdl) %>% summary()

# Show PEA for Probit and Logit
at_list <- list(nwifeinc=mean(mroz$nwifeinc), educ=mean(mroz$educ),
  age=mean(mroz$age), kidslt6=mean(mroz$kidslt6), kidsge6=mean(mroz$kidsge6))
cat("\n PEA Probit")
margins(probit_mdl, at = at_list) %>% summary()
cat("\n PEA Logit")
margins(logit_mdl, at = at_list) %>% summary()
```


LPM, Logit, Probit Example

```
=======================================
            Dependent variable:
         ------------------------------
                     inlf
           OLS      probit   logistic
           (1)       (2)        (3)
---------------------------------------
nwifeinc -0.007*** -0.021*** -0.035***
          (0.002)   (0.005)   (0.008)

educ     0.052***  0.156***  0.258***
          (0.007)   (0.024)   (0.041)

age      -0.012*** -0.034*** -0.058***
          (0.002)   (0.008)   (0.013)

kidslt6  -0.297*** -0.892*** -1.484***
          (0.033)   (0.115)   (0.198)

kidsge6   -0.012    -0.038    -0.066
          (0.014)   (0.041)   (0.068)

Constant 0.645***    0.422     0.723
          (0.155)   (0.472)   (0.789)

=======================================
=======================================
```

Very roughly, $LPM \approx 0.4 Probit \approx 0.25 Logit$

```
APE Probit:
   factor     AME     SE       z      p   lower   upper
      age -0.0118 0.0025 -4.7315 0.0000 -0.0167 -0.0069
     educ  0.0536 0.0075  7.1019 0.0000  0.0388  0.0684
  kidsge6 -0.0130 0.0140 -0.9240 0.3555 -0.0404  0.0145
  kidslt6 -0.3067 0.0348 -8.8029 0.0000 -0.3750 -0.2384
 nwifeinc -0.0072 0.0015 -4.6620 0.0000 -0.0102 -0.0042

APE Logit:
   factor     AME     SE       z      p   lower   upper
      age -0.0120 0.0025 -4.7500 0.0000 -0.0169 -0.0070
     educ  0.0537 0.0076  7.0289 0.0000  0.0388  0.0687
  kidsge6 -0.0138 0.0141 -0.9799 0.3271 -0.0415  0.0138
  kidslt6 -0.3093 0.0354 -8.7447 0.0000 -0.3786 -0.2400
 nwifeinc -0.0073 0.0016 -4.6454 0.0000 -0.0103 -0.0042
```


LPM, Logit, Probit Example

```
PEA Probit:
   factor nwifeinc    educ     age kidslt6 kidsge6     AME     SE       z      p   lower   upper
      age  20.1290 12.2869 42.5378  0.2377  1.3533 -0.0135 0.0030 -4.5458 0.0000 -0.0193 -0.0077
     educ  20.1290 12.2869 42.5378  0.2377  1.3533  0.0611 0.0094  6.5063 0.0000  0.0427  0.0795
  kidsge6  20.1290 12.2869 42.5378  0.2377  1.3533 -0.0148 0.0160 -0.9227 0.3562 -0.0462  0.0166
  kidslt6  20.1290 12.2869 42.5378  0.2377  1.3533 -0.3497 0.0453 -7.7115 0.0000 -0.4386 -0.2608
 nwifeinc  20.1290 12.2869 42.5378  0.2377  1.3533 -0.0082 0.0018 -4.4763 0.0000 -0.0118 -0.0046

PEA Logit:
   factor nwifeinc    educ     age kidslt6 kidsge6     AME     SE       z      p   lower   upper
      age  20.1290 12.2869 42.5378  0.2377  1.3533 -0.0141 0.0031 -4.5252 0.0000 -0.0202 -0.0080
     educ  20.1290 12.2869 42.5378  0.2377  1.3533  0.0631 0.0100  6.3374 0.0000  0.0436  0.0826
  kidsge6  20.1290 12.2869 42.5378  0.2377  1.3533 -0.0162 0.0166 -0.9780 0.3281 -0.0487  0.0163
  kidslt6  20.1290 12.2869 42.5378  0.2377  1.3533 -0.3629 0.0486 -7.4668 0.0000 -0.4582 -0.2677
 nwifeinc  20.1290 12.2869 42.5378  0.2377  1.3533 -0.0085 0.0019 -4.4203 0.0000 -0.0123 -0.0047
```

LPM / Logit / Probit Summary

- Linear Probability Model, Logit and Probit Models used for Binary Dependent Variables
- Fitted values are predicted probabilities $\widehat{\Pr}(Y = 1 \mid X_1, \ldots, X_K)$
  - Can use for classification: $\widehat{Y} = 1$ if $\widehat{\Pr}(Y = 1 \mid X_1, \ldots, X_K) \geq 0.5$
- Predictions from LPM, Logit and Probit are often similar
  - But LPM can sometimes give predicted probabilities greater than 1 or less than 0
- Partial effects from Probit / Logit depend on regressor values, partial effects from LPM are constant (main difference)
- Don't compare coefficients across LPM, Logit and Probit. Compare Average Partial Effects (APE), or Partial Effects at the Average (PEA)

Poisson Regressions

For **Count Dependent Variable** $Y = 0, 1, 2, \ldots$ where most outcomes are low integers

- No. of children, no. of patents filed in a year, etc.

Count data usually modelled using Poisson Distribution with parameter $\lambda > 0$

$$\Pr(Y = h) = \frac{\exp(-\lambda)\lambda^h}{h!}, \quad h = 0, 1, 2, \ldots,$$

If $Y \sim$ Poisson, then $E(Y) = \lambda$ and $Var(Y) = \lambda$

Proof omitted, but uses the fact that $\exp(x) = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \sum_{h=0}^{\infty} \frac{x^h}{h!}$

Poisson Regressions

A Poisson regression assumes

$$\Pr(Y = h \mid X_1, \ldots, X_K) = \frac{\exp(-\lambda)\lambda^h}{h!}$$

where λ now is a *conditional* expectation, and specifies

$$E(Y \mid X_1 \ldots, X_K) = \lambda = \exp(\beta_0 + \beta_1 X_1 + \ldots \beta_K X_K)$$

Interpretations:

- $\dfrac{\partial E(Y \mid X_1, \ldots, X_K)}{\partial X_j} = \exp(\beta_0 + \beta_1 X_1 + \ldots \beta_K X_K)\beta_j$
- $\dfrac{\partial \ln E(Y \mid X_1, \ldots, X_K)}{\partial X_j} = \beta_j$

Poisson Regressions

Estimate by maximum likelihood (numerical maximization)

$$\mathcal{L} = \prod_{i=1}^{n} \frac{\exp(-\exp(\beta_0 + \beta_1 X_{i1} + \ldots \beta_K X_{iK}))\exp(\beta_0 + \beta_1 X_{i1} + \ldots \beta_K X_{iK})^{Y_i}}{Y_i!}$$

$$\ln \mathcal{L} = \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_{i1} + \ldots \beta_K X_{iK}) - \sum_{i=1}^{n} \exp(\beta_0 + \beta_1 X_{i1} + \ldots \beta_K X_{iK}) - \sum_{i=1}^{n} \ln(Y_i!)$$

Can calculate goodness of fit in the usual way

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{u}_i^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \quad \text{where} \quad \hat{u}_i = Y_i - \widehat{Y}_i = Y_i - (\exp(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \ldots \hat{\beta}_K X_{iK}))$$

Incidentally, can show that $\overline{\hat{u}} = 0$

Poisson Regressions

We can continue to use count Y in usual linear regression model

- but noise term obviously not normally distributed
- Coefficients in linear regression not directly comparable to Poisson regression

$$\text{Poisson: } \frac{\partial E(Y \mid X_1, \dots, X_K)}{\partial X_j} = \exp(\beta_0 + \beta_1 X_1 + \dots \beta_K X_K)\beta_j$$

$$\text{Linear Regression: } \frac{\partial E(Y \mid X_1, \dots, X_K)}{\partial X_j} = \beta_j$$

Can compare OLS estimates with "Average Partial Effects" from Poisson

$$(1/n)\sum_{i=1}^{n} \exp(\hat{\beta}_0^{pois} + \hat{\beta}_1 X_{i1}^{pois} + \dots + \hat{\beta}_K^{pois} X_{iK})\hat{\beta}_j^{pois} = \overline{Y}\hat{\beta}_j^{pois}$$

Poisson Regressions

One issue with Poisson regression is the assumption that

$$E(Y \mid X_1, \dots, X_k) = Var(Y \mid X_1, \dots, X_K)$$

Does not hold in many applications (there is often "overdispersion")

Nonetheless, we often proceed with Poisson regression anyway (we call it "Quasi-Maximum Likelihood"), but assume

$$Var(Y \mid X_1, \dots, X_K) = \sigma^2 E(Y \mid X_1, \dots, X_K)$$

- $\widehat{\sigma^2} = \frac{1}{n-K+1}\sum_{i=1}^{n} \frac{\hat{u}_i^2}{\widehat{Y}_i}$
- Adjust estimator standard errors accordingly

Poisson Regressions (Example)

data: crime1

variables: narr86 (times arrested in 86), pcnv (proportion of prior arrests that led to conviction), qemp86 (quarters employed in 86), inc86 (legal income in 86), black (= 1 if black), hispan (=1 if hispanic)

```
data(crime1)
ols_count <- lm(narr86 ~ pcnv + qemp86 + inc86 + black + hispan, data = crime1)
robust_se_ols <- sqrt(diag(vcovHC(ols_count, type="HC0")))

pois_mle <- glm(narr86 ~ pcnv + qemp86 + inc86 + black + hispan, data = crime1, family = poisson())
pois_qmle <- glm(narr86 ~ pcnv + qemp86 + inc86 + black + hispan , data = crime1,
             family = quasipoisson())
```

```
stargazer(ols_count, pois_mle, pois_qmle, type="text",
         se = list(robust_se_ols, NULL, NULL), omit.stat = "all", no.space = TRUE)
```


Poisson Regressions (Example)

```
=============================================
                 Dependent variable:
         ------------------------------------
                       narr86
            OLS     Poisson  glm: quasipoisson
                                link = log
            (1)       (2)          (3)
---------------------------------------------
pcnv     -0.143*** -0.433***     -0.433***
          (0.033)   (0.085)       (0.106)
qemp86   -0.037***  -0.010        -0.010
          (0.013)   (0.029)       (0.035)
inc86    -0.002*** -0.008***     -0.008***
         (0.0002)   (0.001)       (0.001)
black     0.319***  0.643***      0.643***
          (0.058)   (0.073)       (0.091)
hispan    0.182***  0.472***      0.472***
          (0.041)   (0.074)       (0.091)
Constant  0.536*** -0.667***     -0.667***
          (0.038)   (0.064)       (0.079)
```

```
summary(margins(pois_mle))[,1:5]
```

```
 factor     AME     SE       z      p
  black  0.2599 0.0307  8.4588 0.0000
 hispan  0.1910 0.0304  6.2946 0.0000
  inc86 -0.0034 0.0004 -7.8449 0.0000
   pcnv -0.1752 0.0349 -5.0191 0.0000
 qemp86 -0.0039 0.0115 -0.3385 0.7350
```

Corner Solutions

E.g. Y "essentially continuous" over strictly positive values, 0 with positive probability

- $Y \sim$ amount spent on alcohol
- $Y \sim$ no. of hours worked

Can use Linear Regression Model, but

- Can get negative predictions
- Estimate conditional expectation might be misleading

Tobit Model (Corner Solutions)

$$Y^* = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K + \epsilon, \;\; u \mid X \sim N(0, \sigma^2)$$
$$Y = \max\{0, Y^*\}$$

Can show (proof omitted) that

$$E(Y \mid Y > 0, X_1, \dots, X_K) = Z + \sigma \underbrace{\frac{\phi(Z/\sigma)}{\Phi(Z/\sigma)}}_{\lambda(Z/\sigma)}$$

$$\frac{\partial E(Y \mid Y > 0, X_1, \dots, X_K)}{\partial X_j} = \beta_j \left[1 - \lambda\left(\frac{Z}{\sigma}\right)\left\{\frac{Z}{\sigma} + \lambda\left(\frac{Z}{\sigma}\right)\right\}\right]$$

where $Z = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K$, $\lambda(Z/\sigma)$ is the "inverse Mills ratio", $\phi()$ and $\Phi()$ are the pdf and cdf of the standard normal distribution

Tobit Model (Corner Solutions)

Also:

$$E(Y \mid X_1, \dots, X_K) = \Phi(Z/\sigma) Z + \sigma \phi(Z/\sigma)$$

$$\frac{\partial E(Y \mid X_1, \dots, X_K)}{\partial X_j} = \beta_j \Phi(Z/\sigma)$$

Example

data: mroz

```r
data(mroz)
tobit_mdl <- tobit(hours ~ nwifeinc + educ + exper + expersq + age + kidslt6 + kidsge6,
                   data=mroz, left=0) # from AER package
```

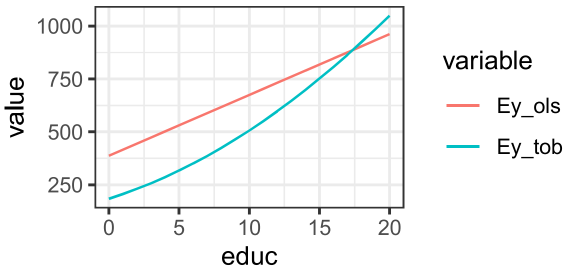

Tobit Model (Corner Solutions)

```r
summary(tobit_mdl)
```

```
Call:
tobit(formula = hours ~ nwifeinc + educ + exper + expersq + age +
    kidslt6 + kidsge6, left = 0, data = mroz)

Observations:
         Total  Left-censored     Uncensored Right-censored
           753            325            428              0

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  965.30528  446.43614   2.162 0.030599 *
nwifeinc      -8.81424    4.45910  -1.977 0.048077 *
educ          80.64561   21.58324   3.736 0.000187 ***
exper        131.56430   17.27939   7.614 2.66e-14 ***
expersq       -1.86416    0.53766  -3.467 0.000526 ***
age          -54.40501    7.41850  -7.334 2.24e-13 ***
kidslt6     -894.02174  111.87804  -7.991 1.34e-15 ***
kidsge6      -16.21800   38.64139  -0.420 0.674701
Log(scale)     7.02289    0.03706 189.514  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Scale: 1122

Gaussian distribution
```

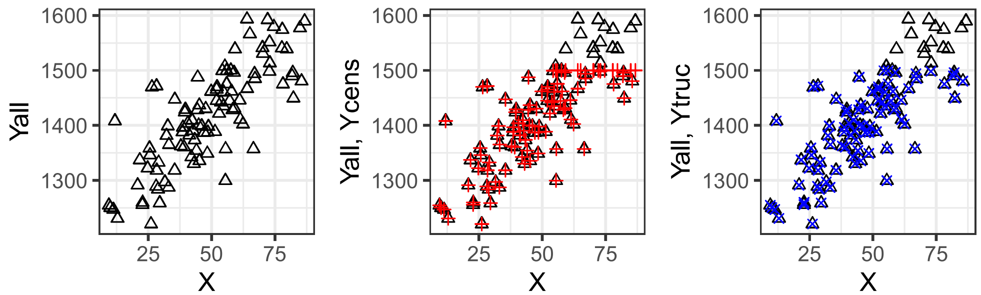
Tobit Model (Corner Solutions)

```r
newdata <- data.frame(educ=0:20, nwifeinc=mean(mroz$nwifeinc), exper=mean(mroz$exper), expersq=mean(mroz$expersq),
            age=mean(mroz$age), kidslt6=mean(mroz$kidslt6), kidsge6=mean(mroz$kidsge6))
# Compare predictions Ehat(y \mid educ) between OLS and Tobit, other predictors set at average
ols_corner <- lm(hours ~ nwifeinc + educ + exper + expersq + age + kidslt6 + kidsge6, data=mroz)
Ey_ols <- predict(ols_corner, newdata)
Z <- predict(tobit_mdl, newdata)
Ey_tob <- pnorm(Z/tobit_mdl$scale)*Z + tobit_mdl$scale*dnorm(Z/tobit_mdl$scale)
plotdat <- tibble(educ = 0:20, Ey_ols=Ey_ols, Ey_tob=Ey_tob)
plotdat %>% pivot_longer(cols = c(Ey_ols, Ey_tob), names_to="variable", values_to="value") %>%
  ggplot(aes(x=educ, y=value, color=variable)) + geom_line() + theme_bw()
```



Censored and Truncated Samples

```r
df <- read_csv("./data/trunc_censored.csv", show_col_types=FALSE)
p1 <- ggplot(data=df) + geom_point(aes(x=X, y=Yall), pch=2) + theme_bw()
p2 <- p1 + geom_point(aes(x=X, y=Ycens), pch=3, color="red") + ylab("Yall, Ycens")
p3 <- p1 + geom_point(aes(x=X, y=Ytruc), pch=4, color="blue") + ylab("Yall, Ytruc")
p1 | p2 | p3
```



Censored and Truncated Samples

Censored Samples:

$$Y^* = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K + \epsilon\,,\ \epsilon \sim N(0, \sigma^2)\,,\ Y = \min\{Y^*, c\}$$

Likelihood similar to Tobit

Truncated Samples

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K + \epsilon\,,\ \epsilon \sim \text{Truncated Normal}(0, \sigma^2)$$

Both estimate by Maximum Likelihood (Numerical Maximization)

In both cases $E(Y \mid X_1, \ldots, X_K) = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K$

Censored and Truncated Samples

```r
ols_full <- lm(Yall ~ X, data=df)
ols_cens <- lm(Ycens ~ X, data=df)
ols_truc <- lm(Ytruc ~ X, data=df)

# from AER package
tob_cens <- tobit(Ycens ~ X,
  data=df, right=1500)

# from truncreg package
nor_truc <- truncreg(Ytruc ~ X,
  data=df, point=1500,
  direction="right")
```

```r
stargazer(ols_full, ols_cens, ols_truc, tob_cens, type="text",
  omit.stat = "all", no.space = TRUE)
```

	Dependent variable:			
	Yall	Ycens	Ytruc	Ycens
	OLS	OLS	OLS	Tobit
	(1)	(2)	(3)	(4)
X	3.862***	3.275***	3.134***	3.789***
	(0.290)	(0.262)	(0.329)	(0.311)
Constant	1,228.326***	1,248.594***	1,251.424***	1,230.884***
	(15.144)	(13.686)	(15.668)	(15.607)

Note: *p<0.1; **p<0.05; ***p<0.01

Censored and Truncated Samples

```
summary(nor_truc)
```

```
Call:
truncreg(formula = Ytruc ~ X, data = df, point = 1500, direction = "right")

BFGS maximization method
56 iterations, 0h:0m:0s
g'(-H)^-1g = 9.96E-05



Coefficients :
              Estimate Std. Error t-value  Pr(>|t|)
(Intercept) 1212.52967   21.55490 56.2531 < 2.2e-16 ***
X              4.39835    0.55319  7.9509 1.776e-15 ***
sigma         57.75881    5.86850  9.8422 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -437.36 on 3 Df
```


Roadmap

- (Previous) Session 1: Statistics Review
- (Previous) Session 2: Simple Linear Regression
- (Previous) Session 3: Estimator Standard Errors; Multiple Linear Regression
- (Previous) Session 4: Matrix Algebra
- (Previous) Session 5: OLS using Matrix Algebra
- (Previous) Session 6: Hypothesis Testing
- (Previous) Session 7: Prediction
- (Previous) Session 8: Instrumental Variable Regression
- (Previous) Session 9: Generalized Least Squares / Panel Data Regressions
- *This Session 10: MLE / Limited Dependent Variable Models*
- **Next Session 11-12: Introduction to Time Series / Time Series Regressions**