USC, Fall 2016, Economics 513

Lecture 6: Inference in the multiple linear regression model: Asymptotics.

In lecture 5 we were able to derive the exact sampling distribution of the OLS estimator $\hat{\beta}$ and related statistics, so that we obtained confidence intervals with an exact coverage probability and test statistics with an exact distribution under the null hypothesis.

This was possible because we made the strong CLR assumptions: $\mathrm{E}(\varepsilon|X) = 0$ and $\mathrm{Var}(\varepsilon|X) = \sigma^2 I$. In addition we fully specified the conditional distribution $\varepsilon|X \sim N(0, \sigma^2 I)$.

In many cases these assumptions are too strong. We consider (i) failure of $\mathrm{E}(\varepsilon|X) = 0$, (ii) failure of conditional normality of the random errors.

Failure of $\mathrm{E}(\varepsilon|X) = 0$

To estimate partial effects we need that ε_i and $x_i = (x_{i1} \cdots x_{iK})'$ are unrelated. We expressed this as $\mathrm{E}(\varepsilon_i|x_i) = 0$, i.e. mean independence. With cross-sectional data this implies $\mathrm{E}(\varepsilon_i|x_1, \ldots x_n) = \mathrm{E}(\varepsilon_i|X) = 0$, but not with time-series data (see autoregressive example).

In time-series (and panel data) often we can only assume

Assumption 1': $\mathrm{E}(\varepsilon_i|x_i) = 0$

We may want to express the assumption that ε_i and x_i are unrelated in even weaker form:

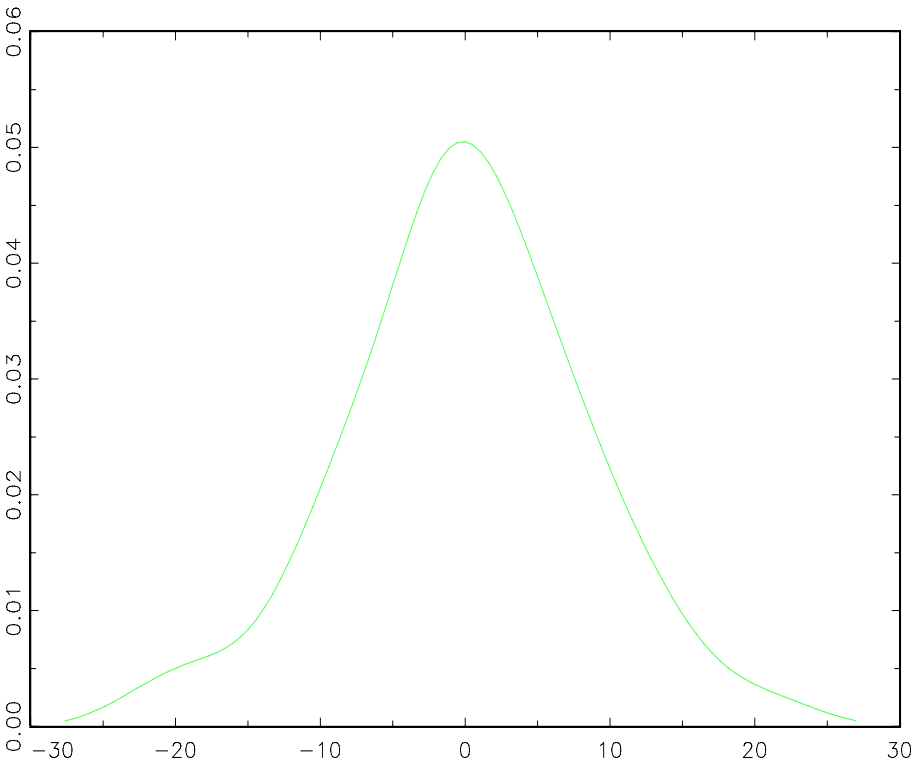
Assumption 1": $\mathrm{E}(\varepsilon_i x_i) = 0$, i.e. ε_i and x_i are uncorrelated.

In this lecture we consider the properties of the OLS estimator $\hat{\beta}$ under the weaker assumption 1". Under this assumption the OLS estimator is in general not unbiased, nor can we derive its sampling variance as we did in lecture 5.

Failure of normality, i.e. of $\varepsilon|X \sim N(0, \sigma^2 I)$

The figures plot nonparametric density estimates of the distribution of the OLS residuals for the lottery data.

Figure 1: Density of OLS residuals: Specification II of lecture 3



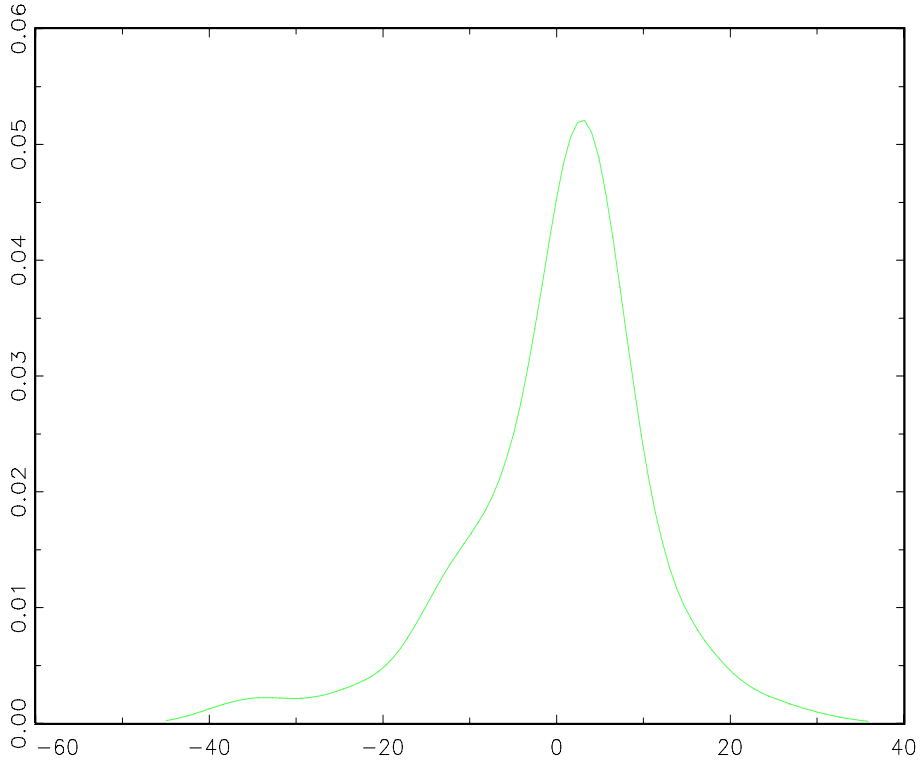
For specification III (dependent variable is change in earnings) the distribution does not seem normal.

Asymptotic approximations

We can derive an approximation to the sampling distribution of the OLS estimator $\hat{\beta}$ if the sample size n is large.

The procedure is to derive a limiting result for $n \to \infty$ and use this in a finite sample as an approximation.

Figure 2: Density of OLS residuals: Specification III of lecture 3



Consider linear regression model (without intercept)

$$y_i = \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n$$

For the OLS estimator

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} = \beta + \frac{\frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i}{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

Note

$$\frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i \qquad \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

are both sample averages. In probability theory Law of Large Numbers says that a sample average for $n \to \infty$ 'converges' to the population mean, e.g.

$$\frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i \xrightarrow{p} \mathrm{E}(x\varepsilon) = 0$$

if Assumption 1" holds.

Also consider

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i \varepsilon_i}{\frac{1}{n} \sum_{i=1}^{n} x_i^2}$$

The numerator

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} x_i \varepsilon_i$$

reminds of the type of sum in a Central Limit Theorem. The distribution of this type of sum 'converges' to a normal distribution.

Note that we deal with ratios of sample averages and we need results for such ratios.

Limits of sequences of random variables

A random variable X is a function that assigns a real number to the outcome of a random experiment. A random variable has a probability distribution that can be characterized by the (cumulative) distribution funnction (cdf) $F(x) = \Pr(X \leq x)$.

Let $X_1, X_2, \ldots$ be a sequence of random variables. We are interested in the limit of this sequence if $n \to \infty$. Note that this is a sequence of functions, not of real numbers.

There is no unique way to define the limit $X_n \to X$. We consider two definitions that focus on

(i) the probability that X_n differs from X, i.e. $\Pr(|X_n - X| > \varepsilon)$.

(ii) the limit of the corresponding sequence of cdf's, $\{F_n\}$.

The various possibilities lead to different modes of convergence and the limit depends on which mode we adopt.

The limit of a sequence of random variables is a random variable. A special limit is a random variable with a degenerate distribution: $\Pr(X = c) = 1$, i.e. X is a deterministic constant c.

Convergence in probability

The first definition is *convergence in probability* (also known as *plim*). A sequence of random variables $\{X_n\}$ converges in probability to a random variable

X if, for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr(|X_n - X| > \varepsilon) = 0 \tag{1}$$

Notation:

$$X_n \xrightarrow{p} X \tag{2}$$

Why is convergence in probability a useful concept?

Consider a sequence of random variables $\{X_n\}$ that are independent and all have the same distribution, i.e. are identically distributed. Such a sequence is called an i.i.d. sequence or a random sample.

For random samples there is a famous limit theorem, the *Law of Large Numbers* (LLN): If $X_1, \ldots, X_n, \ldots$ is a random sample and $\mathrm{E}(|X_1|) < \infty$, then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mathrm{E}(X_1) \tag{3}$$

In statistics and econometrics, we usually require that a sequence of estimators $\hat{\beta}_n$ of a parameter β converges in probability to the population value of that parameter. Such an estimator (sequence) is called *(weakly) consistent*.

Convergence in distribution

The sequence of random variables $X_1, \ldots, X_n, \ldots$ with corresponding c.d.f.'s $F_1, \ldots, F_n, \ldots$ converges in distribution to a random variable X with c.d.f. F if

$$\lim_{n\to\infty} F_n(x) = F(x) \tag{4}$$

for all x where F is continuous.

Notation:

$$X_n \xrightarrow{d} X \tag{5}$$

It can be shown that

$$X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{d} X \tag{6}$$

but

$$X_n \xrightarrow{d} X \text{ not} \Rightarrow X_n \xrightarrow{p} X \tag{7}$$

Hence convergence in distribution is weaker. However, if $X_n \xrightarrow{d} c$, i.e. the limit distribution is degenerate in c, then $X_n \xrightarrow{p} c$, i.e. if the limit is a constant (degenerate distribution), then convergence in probability and in distribution are equivalent.

To see that convergence in distribution in general does not imply convergence in probability let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables all with

cdf $F = \Phi$. Also let X be a random variable that is independent of these and has the same standard normal distribution. Obviously

$$X_n \xrightarrow{d} X$$

but for all $\varepsilon > 0$,

$$\Pr\left(|X_n - X| < \varepsilon\right) = 2\Phi\left(\frac{\varepsilon}{\sqrt{2}}\right) - 1$$

and this does not converge to 1.

Why is convergence in distribution useful?

Let $X_1, \ldots, X_n, \ldots$ be a random sample with

$$\begin{aligned} \mathrm{E}(X_1) &= \mu \\ \mathrm{Var}(X_1) &= \sigma^2 < \infty \end{aligned}$$

Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} \xrightarrow{d} N(0,1)$$

This is the famous *Central Limit Theorem* (C.L.T.).

Useful properties of limits of sequences of random variables

The first property shows that the limit of a function is the function of the limit, provided that the function is continuous. This is an important simplification that does not hold for expectations.

Continuous Mapping Theorem: Let $\{X_n\}$ be a sequence of r.v.'s with (i) $X_n \xrightarrow{p} X$, (ii) $X_n \xrightarrow{d} X$, and let $g(x)$ be continuous in x. Then (i) $g(X_n) \xrightarrow{p} g(X)$, (ii) $g(X_n) \xrightarrow{d} g(X)$.

The next property summarizes the relation between the modes of convergence.

Modes of convergence: Let $\{X_n\}$ and $\{Y_n\}$ be sequences of random variables. Then

(i) $X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{d} X$

(ii) $X_n \xrightarrow{p} c \Leftrightarrow X_n \xrightarrow{d} c$

(iii) If $X_n \xrightarrow{d} X$ and $|X_n - Y_n| \xrightarrow{p} 0$, then $Y_n \xrightarrow{d} X$

(iv) If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then $(X_n, Y_n) \xrightarrow{d} (X, c)$

(v) If $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$, then $(X_n, Y_n) \xrightarrow{p} (X, Y)$

The last two statements indicate what can be concluded on the joint convergence of random vectors from the marginal convergence of their components. Note, in particular that marginal convergence in distribution is not enough to have joint convergence of the random vector (why?).

If we combine the Continuous Mapping Theorem with (iv) of the previous theorem, we have

Slutsky's Theorem: Let $\{X_n\}$ and $\{Y_n\}$ be sequences of random variables, such that

(i) $X_n \xrightarrow{d} X$

(ii) $Y_n \xrightarrow{p} c$ (a non − stochastic constant)

Then

$$X_n Y_n \xrightarrow{d} cX,\ X_n + Y_n \xrightarrow{d} X + c,\ \frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c} \text{ provided } c \neq 0$$

We often need results on the convergence in distribution of random vectors. If the limit random variable has a normal distribution, then

Cramér-Wold device: If for all vectors (of deterministic constants) $c \neq 0$

$$c' X_n \xrightarrow{d} N(0, c'\Sigma c)$$

then $X_n \xrightarrow{d} N(0, \Sigma)$.

Application of these results

Let $\{X_n\}$ be a random sample with

$$\mathrm{E}(X_1) = \mu \qquad \mathrm{Var}(X_1) = \sigma^2 < \infty$$

Hence the conditions of CLT are satisfied. Further, let $S_1, \ldots, S_n, \ldots$ be a sequence of r.v.'s with

$$S_n \xrightarrow{p} \sigma^2$$

Then

$$\frac{\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)\right)^2}{S_n} = \frac{\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}\right)^2}{\frac{S_n}{\sigma^2}} \xrightarrow{d} \chi^2(1)$$

because

$$\frac{S_n}{\sigma^2} \xrightarrow{p} 1$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i-\mu}{\sigma}\xrightarrow{d}N(0,1)$$

The asymptotic distribution of the OLS estimator

Consider a linear regression model with an intercept. We do not make (all) CLR assumptions, and we shall see which are needed. In the regression model the $i-$th observation satisfies

$$y_i=\beta_0+x_i'\beta_1+\varepsilon_i$$

with $x_i=(x_{i2}\cdots x_{iK})'$. We assume

- ε_i are independent and identically distributed or i.i.d. with mean 0 and variance σ^2.
- x_i are i.i.d. (μ_x,Σ_x) with Σ_x non-singular and finite.
- These assumptions hold automatically if $y_i,x_i,i=1,\ldots,n$ is a random sample from a population.
- $\mathrm{E}(\varepsilon_i x_i)=0$.

By partial regression the OLS estimator of β_1

$$\hat{\beta}_{1n}=(X'M_1X)^{-1}X'M_1y=$$

$$=\left(\sum_{i=1}^{n}(x_i-\bar{x}_n)(x_i-\bar{x}_n)'\right)^{-1}\left(\sum_{i=1}^{n}(x_i-\bar{x}_n)y_i\right)$$

with $M_1=I-\frac{1}{n}\iota\iota'$ and $\bar{x}_n$ the vector of sample means of the non-constant independent variables. Here we use the subscript n to indicate that the OLS estimator uses a sample of size n.

Upon substitution of $y_i=\beta_0+x_i'\beta_1+\varepsilon_i$ we have

$$\hat{\beta}_{1n}-\beta_1=\left(\sum_{i=1}^{n}(x_i-\bar{x}_n)(x_i-\bar{x}_n)'\right)^{-1}\left(\sum_{i=1}^{n}(x_i-\bar{x}_n)\varepsilon_i\right)$$

Now

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}_n)(x_i-\bar{x}_n)'=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_x)(x_i-\mu_x)'+$$

with the rhs converging in probability to $\mathrm{E}[(x_i-\mu_x)(x_i-\mu_x)']=\Sigma_x<\infty$ by the WLLN and joint convergence (because we are dealing with a matrix of averages and we can apply the WLLN to each of them separately and then invoke joint convergence, i.e. (v) of the modes of convergence result)

$$+(\bar{x}_n-\mu_x)(\bar{x}_n-\mu_x)'$$

which converges in probability to 0 by the WLLN (that implies that $\bar{x}_n - \mu_x \xrightarrow{p} 0$), joint convergence and continuous mapping (because the function z z' is continuous in z). Therefore the limit is Σ_x (in probability) by continuous mapping (because $z_1 + z_2$ is a continuous function of z_1 and z_2).

Also

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)\varepsilon_i = \frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i - \bar{x}_n\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i$$

with the first term on the rhs converging to $\mathrm{E}(x_i\varepsilon_i) = 0$ in probability by the WLLN and joint convergence and the first factor of the second term on the rhs converging to μ_x and the second factor converging to 0 in probability by the WLLN. Therefore by continuous mapping the limit (in probability) is 0.

Because the inverse A^{-1} is a continuous function of A , if A is non-singular, i.e. can be inverted, repeated application of continuous mapping gives

$$\hat{\beta}_{1n} - \beta_1 \xrightarrow{p} 0 \qquad \text{or} \quad \hat{\beta}_{1n} \xrightarrow{p} \beta_1$$

Hence (the sequence of) OLS estimator(s) $\hat{\beta}_{1n}$ is (weakly) consistent for β_1.

Next, consider

$$\sqrt{n}\left(\hat{\beta}_{1n} - \beta_1\right) = \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)(x_i - \bar{x}_n)'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i - \bar{x}_n)\varepsilon_i\right)$$

with

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i - \bar{x}_n)\varepsilon_i = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i - \mu_x)\varepsilon_i - (\bar{x}_n - \mu_x)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i$$

with the first factor of the second term on the rhs, $\bar{x}_n - \mu_x$, converging to 0 in probability by the WLLN and joint convergence and the second factor by the CLT converging in distribution to $N(0, \sigma^2)$. Hence by Slutsky theorem (product of a sequence that converges in probability and a sequence that converges in distribution) second term on RHS converges to 0 in probability (and therefore also in distribution).

For first term let $c \neq 0$ be a vector of constants. Then using the Cramer-Wold device and CLT

$$c'\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i - \mu_x)\varepsilon_i\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}c'(x_i - \mu_x)\varepsilon_i \xrightarrow{d} N(0, c'\Sigma c)$$

because for all $c \neq 0$

$$\mathrm{E}[c'(x_i - \mu_x)\varepsilon_i] = 0 \qquad \mathrm{Var}[c'(x_i - \mu_x)\varepsilon_i] = c'\mathrm{E}[\varepsilon_i^2(x_i - \mu_x)(x_i - \mu_x)']c = c'\Sigma c$$

with

$$\Sigma = \mathrm{E}\left(\varepsilon_i^2 \left(x_i - \mu_x\right)\left(x_i - \mu_x\right)'\right) < \infty$$

and if x_i and ε_i are independent, then $\Sigma = \sigma^2 \Sigma_x$. We conclude that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (x_i - \mu_x)\varepsilon_i \xrightarrow{p} N(0, \Sigma)$$

Applications of continuous mapping and the Slutsky theorem gives (note the difference with the usual formula $\sigma^2 \Sigma_x$)

$$\sqrt{n}\left(\hat{\beta}_{1n} - \beta_1\right) \xrightarrow{d} N\left(0, \Sigma_x^{-1} \Sigma \Sigma_x^{-1}\right)$$

We obtain the usual formula for the variance of the OLS estimator if

$$\mathrm{E}[\varepsilon_i^2 | x_i] = \sigma^2$$

i.e. the random errors are homoskedastic. With this assumption

$$\mathrm{E}\left(\varepsilon_i^2 \left(x_i - \mu_x\right)\left(x_i - \mu_x\right)'\right) = \sigma^2 \Sigma_x$$

and we get the usual expression for the variance.

The general formula allows for arbitrary heteroskedasticity. The variance matrix is called the heteroskedasticity-consistent variance matrix of the OLS estimator. Note that all results were obtained under weaker conditions than in the CLR model.

The heteroskedasticity-consistent variance matrix can be estimated by

$$\hat{\Sigma}_x^{-1} \hat{\Sigma} \hat{\Sigma}_x^{-1}$$

with

$$\hat{\Sigma}_x = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x}_n)(x_i - \overline{x}_n)'$$

and

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} e_i^2 (x_i - \overline{x}_n)(x_i - \overline{x}_n)'$$

Note that this is a consistent estimator (by WLLN and continuous mapping).

The asymptotic distribution result is used to obtain a finite sample approximate sampling distribution

$$\hat{\beta}_{1n} \approx N(\beta_1, \frac{1}{n} \hat{\Sigma}_x^{-1} \hat{\Sigma} \hat{\Sigma}_x^{-1})$$

Note that the approximate variance goes to 0 if the sample size increases. Why should this be true?

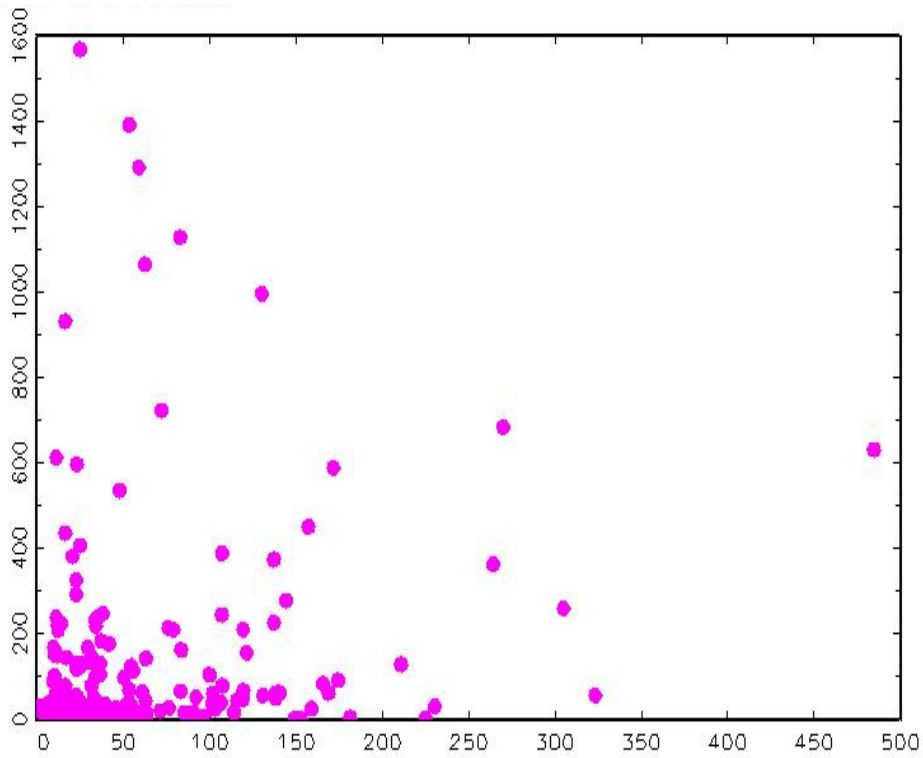
If $\text{se}(\hat{\beta}_{k,1n})$ denotes the standard error of $\hat{\beta}_{k,1n}$, i.e. the k-th component of the main diagonal, then an approximate confidence interval with coverage probability α is

$$\left[\hat{\beta}_{k,1n} - n_\alpha \text{se}(\hat{\beta}_{k,1n}), \hat{\beta}_{k,1n} + n_\alpha \text{se}(\hat{\beta}_{k,1n})\right]$$

with n_α the $(1+\alpha)/2$-th quantile of the standard normal distribution.

In the lottery data if we consider the model with the change in earnings as the dependent variable and PRIZE as the independent variable the OLS estimate of the coefficient on PRIZE is -0.0503 with regular standard error 0.0113 and heteroskedasticity-consistent standard error .0171 which is 51% larger.

To check on heteroskedasticity we plot the squared OLS residuals against the independent variable PRIZE.



The coefficient of PRIZE in a regression of the squared residuals on PRIZE is 0.762 with a standard error of 0.241. This is not really a test, but it is an indication that the errors are heteroskedastic.

Omitted variable bias

We found earlier that

$$\hat{\beta}_{1n} = \beta_1 + \left(\sum_{i=1}^{n} (x_i - \bar{x}_n)(x_i - \bar{x}_n)' \right)^{-1} \left(\sum_{i=1}^{n} (x_i - \bar{x}_n)\varepsilon_i \right) \qquad (1)$$

Now assume that there is an omitted variable z so that the correct regression model is

$$y_i = \beta_0 + x_i'\beta_1 + \gamma z_i + \eta_i$$

with $\mathrm{E}(\eta_i | x_i, z_i) = 0$ (or uncorrelated). This implies that

$$\varepsilon_i = \gamma z_i + \eta_i$$

Substitution in (1) gives

$$\hat{\beta}_{1n} = \beta_1 + \gamma \left(\sum_{i=1}^{n} (x_i - \bar{x}_n)(x_i - \bar{x}_n)' \right)^{-1} \left(\sum_{i=1}^{n} (x_i - \bar{x}_n) z_i \right) +$$

$$\left(\sum_{i=1}^{n} (x_i - \bar{x}_n)(x_i - \bar{x}_n)' \right)^{-1} \left(\sum_{i=1}^{n} (x_i - \bar{x}_n)\eta_i \right)$$

By an analogous argument as used above the last term converges to 0 in probability. The second term converges in probability to the vector δ_1 with

$$z_i = \delta_0 + \delta_1' x_i + \nu_i$$

with $\mathrm{E}(x_i \nu_i) = 0$ and we do not worry whether δ_1 is a vector of partial effects.

We conclude that for the OLS estimator

$$\hat{\beta}_{1n} \xrightarrow{p} \beta_1 + \gamma\delta_1$$

so that this estimator is inconsistent.

The asymptotic bias is $\gamma\delta_1$ and this is the product of the effect (not necessarily causal) of x_1 on z, i.e. δ_1 and the (causal) effect of z on y, i.e. γ.

In empirical research researchers often speculate about the sign and size of this bias. Example: Ability omitted in earnings equation (positive effect on earnings; $\gamma > 0$). Education likely to be positively related to ability ($\delta_1 > 0$). Therefore the ability bias for the returns to education is likely to be positive. In assignment 2 you will use this formula.