Principles of Econometrics

Gianfranco Piras

THE SIMPLE REGRESSION MODEL: STATISTICAL PROPERTIES OF OLS I

Expected Value of OLS I

- We motivated simple regression using a **population model**.
- But our analysis so far has been purely algebraic!
- Now our job gets harder!
- We have to study **statistical properties of the OLS estimator**.

Expected Value of OLS II

- How do our estimators behave **across different samples**?
- **On average**, would we get the right answer if we could repeatedly sample?
- We need to find the **expected value** of the OLS estimators and determine if we are right on average.
- Leads to the notion of **unbiasedness**.

Expected Value of OLS III

Assumption SLR.1 (Linear in Parameters)

The **population model** can be written as

$$y = \beta_0 + \beta_1 x + u$$

where β_0 and β_1 are the (unknown) population parameters.

Expected Value of OLS IV

Assumption SLR.2 (Random Sampling)

We have a **random sample** of size n, $\{(x_i, y_i) : i = 1, ..., n\}$, following the population model.

Expected Value of OLS V

Assumption SLR.3 (Sample Variation in the Explanatory Variable)

The sample outcomes on x_i are **not all the same value**.

Expected Value of OLS VI

Assumption SLR.4 (Zero Conditional Mean)

In the population, the **error term has zero (conditional) mean** given any value of the explanatory variable:

$$E(u|x) = 0 \text{ for all } x.$$

- This is the key assumption for showing that OLS is unbiased!

Expected Value of OLS VII

- How do we show $\hat{\beta}_1$ is **unbiased** for β_1? What we need to show is

$$E(\hat{\beta}_1) = \beta_1$$

  where the expected value means **averaging across random samples**.
- To prove the result above, we will treat the x_i as nonrandom (or fixed in repeated samples).
- There are a few steps involved.

Expected Value of OLS VIII

1. Write down a formula for $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

It is convenient to define $SST_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$, the total variation in the x_i, and write

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{SST_x}$$

Remember, SST_x is just some positive number. The existence of $\hat{\beta}_1$ follows from SLR.3.

Expected Value of OLS IX

2. Replace each y_i with $y_i = \beta_0 + \beta_1 x_i + u_i$

Expected Value of OLS X

The numerator becomes

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})y_i &= \sum_{i=1}^{n}(x_i-\bar{x})(\beta_0+\beta_1 x_i+u_i)\\ &= \beta_0\sum_{i=1}^{n}(x_i-\bar{x})+\beta_1\sum_{i=1}^{n}(x_i-\bar{x})x_i+\sum_{i=1}^{n}(x_i-\bar{x})u_i\\ &= 0+\beta_1\sum_{i=1}^{n}(x_i-\bar{x})^2+\sum_{i=1}^{n}(x_i-\bar{x})u_i\\ &= \beta_1 SST_x+\sum_{i=1}^{n}(x_i-\bar{x})u_i\end{aligned}$$

Expected Value of OLS XI

We used two results:

a) $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$
b) $\sum_{i=1}^{n}(x_i - \bar{x})x_i = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

We have shown

$$\hat{\beta}_1 = \frac{\beta_1 SST_x + \sum_{i=1}^{n}(x_i - \bar{x})u_i}{SST_x} = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{SST_x}$$

Note how the last piece is the slope coefficient from the OLS regression of u_i on x_i, $i = 1, ..., n$.
We cannot do this regression because the u_i are not observed.

Expected Value of OLS XII

Now define

$$w_i = \frac{(x_i - \bar{x})}{SST_x}$$

so we have

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} w_i u_i$$

- $\hat{\beta}_1$ is a linear function of the unobserved errors, u_i. The w_i are all functions of $\{x_1, x_2, ..., x_n\}$.
- The (random) difference between $\hat{\beta}_1$ and β_1 is due to this linear function of the unobservables.

Expected Value of OLS XIII

3. Find $E(\hat{\beta}_1)$.
   Under Assumptions SLR.2 and SLR.4, $E(u_i|x_1, x_2, ..., x_n) = 0$. That means, conditional on $\{x_1, x_2, ..., x_n\}$ (and using SLR.3),

$$E(w_i u_i) = w_i E(u_i) = 0$$

   because w_i is a function of $\{x_1, x_2, ..., x_n\}$.
   This would not be true if, in the population, u and x are correlated.

Expected Value of OLS XIV

Now we can complete the proof: Conditional on $\{x_1, x_2, ..., x_n\}$,

$$\begin{aligned} E(\hat{\beta}_1) &= E\left(\beta_1 + \sum_{i=1}^{n} w_i u_i\right) \\ &= \beta_1 + \sum_{i=1}^{n} E(w_i u_i) = \beta_1 + \sum_{i=1}^{n} w_i E(u_i) \\ &= \beta_1, \end{aligned}$$

where we used two important properties of expected values:

- the expected value of a sum is the sum of the expected values;
- the expected value of a constant, β_1 in this case, is just itself.

Expected Value of OLS XV

THEOREM (Unbiasedness of OLS)

Under Assumptions SLR.1 through SLR.4 and conditional on the outcomes $\{x_1, x_2, ..., x_n\}$,

$$E(\hat{\beta}_0) = \beta_0 \text{ and } E(\hat{\beta}_1) = \beta_1.$$

Expected Value of OLS XVI

Now simulate data in R:

$$y = 3 + 2x + u$$

$x \sim Normal(0, 9)$, $u \sim Normal(0, 36)$, and they are independent.

Expected Value of OLS XVII

```
R> x <- rnorm(250, mean = 0, sd = 3)
R> u <- rnorm(250, mean = 0, sd = 6)
R> y <- 3 + 2*x + u
R> coefficients(lm(y ~ x))
(Intercept)           x
   2.576036    1.929405
R> coefficients(lm(y ~ x)) - c(3, 2)
(Intercept)           x
-0.42396411 -0.07059531
```

Expected Value of OLS XVIII

```
R> u <- rnorm(250, mean = 0, sd = 6)
R> y <- 3 + 2*x + u
R> coefficients(lm(y ~ x))
(Intercept)           x
   2.666639    2.050404
R> coefficients(lm(y ~ x)) - c(3, 2)
(Intercept)           x
-0.33336096  0.05040401
```

Expected Value of OLS XIX

```
R> u <- rnorm(250, mean = 0, sd = 6)
R> y <- 3 + 2*x + u
R> coefficients(lm(y ~ x))
(Intercept)           x
   3.085537    1.807165
R> coefficients(lm(y ~ x)) - c(3, 2)
(Intercept)           x
 0.08553664 -0.19283505
```

Expected Value of OLS XX

```
R> u <- rnorm(250, mean = 0, sd = 6)
R> y <- 3 + 2*x + u
R> coefficients(lm(y ~ x))
(Intercept)           x
   3.603080    1.786789
R> coefficients(lm(y ~ x)) - c(3, 2)
(Intercept)           x
  0.6030804  -0.2132105
```

Expected Value of OLS XXI

- The second generated data set gets us very close to $\beta_1 = 2$, with $\hat{\beta}_1 \approx 2.050$. The third data set gets us closest to $\beta_0 = 3$.
- If we repeat the experiment again and again, and average the $\hat{\beta}_1$, we would get very close to 2.
- The problem is, we do not know which kind of sample we have. We can never know whether we are close to the population value.
- Generally, we hope that our sample is "typical" and produces a slope estimate close to β_1, but we can never know!!

Expected Value of OLS XXII

```
R> set.seed(1)
R> mc_samples <- 5000
R> beta_0 <- 3
R> beta_1 <- 2
R> n <- 250
R> strg <- matrix( , nrow = mc_samples, ncol = 2)
R> x <- rnorm(n, mean = 0, sd = 3)
R> i <- 1
R> while(i <= mc_samples){
+    u <- rnorm(n, mean = 0, sd = 6)
+    y <- beta_0 + beta_1*x + u
+    strg[i,] <- coefficients(lm(y ~ x))
+    i <- i + 1
+  }
R> apply(strg, 2, mean)
[1] 3.001956 1.997946
R> apply(strg, 2, mean) - c(3, 2)
[1]  0.001956029 -0.002054387
```

Expected Value of OLS XXIII

IMPORTANT

Unbiasedness is a property of the *procedure*. After estimating an equation like

$$\begin{aligned} \widehat{lwage} &= 0.583 + .083\ educ \\ n &= 526,\ R^2 = .186 \end{aligned}$$

it is tempting to say 8.3% is an "unbiased estimate" of the return to education. Technically, this statement is incorrect!!! We can only say that the rule used to get $\hat{\beta}_0 = 0.583$ and $\hat{\beta}_1 = .083$ is unbiased.

Expected Value of OLS XXIV

- The four assumptions:

SLR.1: $y = \beta_0 + \beta_1 x + u$

SLR.2: random sampling from the population

SLR.3: some sample variation in the x_i

SLR.4: $E(u|x) = 0$

- The focus should mainly be on the last of these. What are the omitted factors? Are they likely to be correlated with x? If so, SLR.4 fails and OLS will be biased.

Expected Value of OLS XXV

EXAMPLE: Student Performance and Student-Teacher Ratios

Using data from mathpnl,

$$\begin{aligned} \widehat{math4} &= 76.01 - 0.064\ ptr \\ n &= 550,\ R^2 = .00017 \end{aligned}$$

Notice the minus sign for $\hat{\beta}_1 = -0.064$. The estimate implies that one more student per teacher decreases the estimated pass rate by about .064 percentage points (*math*4 is a percent). A *decrease* of one standard deviation (about 2.7) in *ptr* is predicted to *increase* the pass rate by about $0.064(2.7) = 0.17$ percentage points.

Expected Value of OLS XXVI

Is the OLS estimator likely to be unbiased in this setting?

- Students from advantaged backgrounds tend to go to schools with smaller *ptr* and likely would perform better, on average, without small classes!
- We may just be picking up the negative correlation between "ability" and class size, rather than a causal effect of class size.

Expected Value of OLS XXVII

- Notice the extremely small R-squared. Basically none of the variation in $math4$ across schools is explained by ptr (in this sample).
- The low R-squared means that using this equation for predicting $math4$ likely will produce poor results.
- But a high variance for u (leading to a low R-squared) is separate from whether u and ptr are correlated. We must judge that based on introspection or external evidence.

Expected Value of OLS XXVIII

```
R> library(wooldridge)
R> data("mathpnl")
R> meap98 <- mathpnl[mathpnl$y98 == 1,]
R> summary(meap98$ptr)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.40   19.70   21.30   20.96   22.60   28.80
R> sd(meap98$ptr)
[1] 2.688068
```

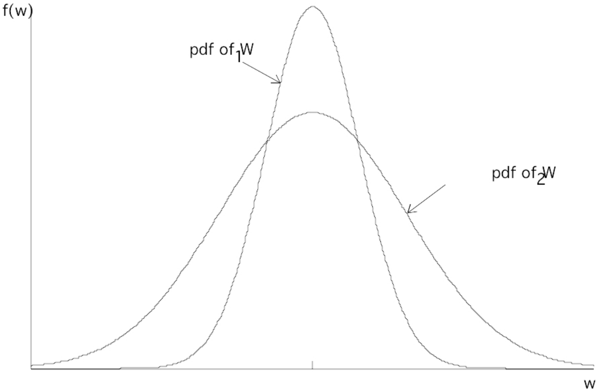
Expected Value of OLS XXIX

```
R> mod <- lm(math4 ~ ptr, data = meap98)
R> summary(mod)

Call:
lm(formula = math4 ~ ptr, data = meap98)

Residuals:
    Min      1Q  Median      3Q     Max
-48.638  -6.817   1.343   9.317  25.291

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 76.00955    4.43162  17.152   <2e-16 ***
ptr         -0.06439    0.20971  -0.307    0.759
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.21 on 548 degrees of freedom
Multiple R-squared:  0.000172,       Adjusted R-squared:  -0.001653
F-statistic: 0.09426 on 1 and 548 DF,  p-value: 0.7589
```

Variance of the OLS Estimators I

- Under SLR.1 to SLR.4, the OLS estimators are **unbiased**.
- But we need a measure of dispersion in the sampling distribution of the estimators: this measure is the variance!

Variance of the OLS Estimators II



Variance of the OLS Estimators III

Assumption SLR.5 (Homoskedasticity, or Constant Variance)

The error has the same variance given any value of the explanatory variable x:

$$Var(u|x) = \sigma^2 > 0 \text{ for all } x,$$

where σ^2 is unknown.

Variance of the OLS Estimators IV

Since $E(u|x) = 0$ and SLR.5, we can also write

$$E(u^2|x) = \sigma^2 = E(u^2)$$

Also SLR.1, SRL.4 and SLR.5 imply that:

$$\begin{aligned} E(y|x) &= \beta_0 + \beta_1 x \\ Var(y|x) &= \sigma^2 \end{aligned}$$

Variance of the OLS Estimators V

Is the assumption SLR.5 always reasonable?

Variance of the OLS Estimators VI

EXAMPLE:

Suppose $y = sav$, $x = inc$ and we think

$$E(sav|inc) = \beta_0 + \beta_1 inc$$

with $\beta_1 > 0$. This means average family saving increases with income. If we impose SLR.5 then

$$Var(sav|inc) = \sigma^2$$

which means the variability in saving does not change with income. There are reasons to think saving would be more variable as income increases.

Variance of the OLS Estimators VII

THEOREM (Sampling Variances of OLS)

Under Assumptions SLR.1 to SLR.5:

$$
\begin{aligned}
Var(\hat{\beta}_1) &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - x)^2} = \frac{\sigma^2}{SST_x} \\
Var(\hat{\beta}_0) &= \frac{\sigma^2 \left(n^{-1} \sum_{i=1}^n x_i^2\right)}{SST_x}
\end{aligned}
$$

Variance of the OLS Estimators VIII

To show this result, write, as before,

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} w_i u_i$$

where $w_i = (x_i - \bar{x})/SST_x$.

- We are treating the w_i as nonrandom in the derivation.
- Because β_1 is a constant, it does not affect $Var(\hat{\beta}_1)$.
- For uncorrelated random variables, the variance of the sum is the sum of the variances.

Variance of the OLS Estimators IX

The $\{u_i : i = 1, 2, ..., n\}$ are actually independent across i, and so they are uncorrelated. Therefore,

$$\begin{aligned} Var(\hat{\beta}_1) &= Var\left(\sum_{i=1}^{n} w_i u_i\right) = \sum_{i=1}^{n} Var(w_i u_i) \\ &= \sum_{i=1}^{n} w_i^2 Var(u_i) = \sum_{i=1}^{n} w_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^{n} w_i^2 \end{aligned}$$

Variance of the OLS Estimators X

Now we have

$$\begin{aligned}\sum_{i=1}^{n} w_i^2 &= \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{(SST_x)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(SST_x)^2} \\ &= \frac{SST_x}{(SST_x)^2} = \frac{1}{SST_x}.\end{aligned}$$

We have shown

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_x}$$

Variance of the OLS Estimators XI

Two things to note:

1. This is the "standard" formula for the **variance of the OLS slope** estimator. It is **not** valid if Assumption SLR.5 does not hold.
2. The homoskedasticity assumption was **not** used to show unbiasedness of the OLS estimators.

Variance of the OLS Estimators XII

Usually we are interested in β_1. We can easily study the two factors that affect its variance.

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_x}$$

1. As the error variance increases, that is, as σ^2 increases, so does $Var(\hat{\beta}_1)$.
2. More variation in $\{x_i\}$ is desirable!

$$SST_x \uparrow \text{ implies } Var(\hat{\beta}_1) \downarrow$$

Variance of the OLS Estimators XIII

The standard deviation of $\hat{\beta}_1$ is the square root of the variance. So

$$sd(\hat{\beta}_1) = \frac{\sigma}{\sqrt{SST_x}}$$

This turns out to be the measure of variation that appears in **confidence intervals** and **test statistics**.

Variance of the OLS Estimators XIV

Estimating the Error Variance

In the formula

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_x}$$

we can compute SST_x from the observed data $\{x_i : i = 1, ..., n\}$.

We need a way to estimate σ^2 Recall that

$$\sigma^2 = E(u^2).$$

Variance of the OLS Estimators XV

Therefore, if we could observe a sample on the errors, an unbiased estimator of σ^2 would be the sample average of the squared errors,

$$n^{-1} \sum_{i=1}^{n} u_i^2$$

But this not an estimator because we cannot compute it from the data we observe!

How about replacing each u_i with its "estimate," the OLS residual $\hat{u}_i$?

$$\begin{aligned} u_i &= y_i - \beta_0 - \beta_1 x_i \\ \hat{u}_i &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \end{aligned}$$

Variance of the OLS Estimators XVI

$\hat{u}_i$ can be computed from the data because it depends on the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$.

$$\hat{u}_i \neq u_i$$

for any i.

In fact, simple algebra gives

$$\begin{aligned}
\hat{u}_i &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = (\beta_0 + \beta_1 x_i + u_i) - \hat{\beta}_0 - \hat{\beta}_1 x_i \\
&= u_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1) x_i
\end{aligned}$$

$E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$, but the estimators almost always differ from the population values in a sample.

Variance of the OLS Estimators XVII

What about this as an estimator of σ^2?

$$n^{-1}\sum_{i=1}^{n}\hat{u}_i^2 = SSR/n$$

It is a true estimator and easily computed from the data after OLS.

As it turns out, this estimator is slightly biased: Its expected value is less than σ^2.

Variance of the OLS Estimators XVIII

The unbiased estimator of σ^2 uses a **degrees-of-freedom** adjustment.

The estimator used universally is

$$\hat{\sigma}^2 = SSR/(n-2) = (n-2)^{-1}\sum_{i=1}^{n}\hat{u}_i^2.$$

Variance of the OLS Estimators XIX

THEOREM (Unbiased Estimator of σ^2)

Under Assumptions SLR.1 to SLR.5, and conditional on $\{x_1, ..., x_n\}$,

$$E(\hat{\sigma}^2) = \sigma^2.$$

Variance of the OLS Estimators XX

In regression output, it is

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{SSR/(n-2)}$$

that is usually reported. This is an estimator of $sd(u)$, the standard deviation of the population error.

Variance of the OLS Estimators XXI

$\hat{\sigma}$ is called the **standard error of the regression**, which means it is an estimate of the standard deviation of the error in the regression.

Given $\hat{\sigma}$, we can now estimate $sd(\hat{\beta}_1)$ and $sd(\hat{\beta}_0)$. The estimates of these are called the **standard errors** of the $\hat{\beta}_j$. We will use these a lot.

Almost all regression packages report the standard errors in a column next to the coefficient estimates.

Variance of the OLS Estimators XXII

We just plug $\hat{\sigma}$ in for σ:

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{SST_x}}$$

where both the numerator and denominator are easily computed from the data.

For reasons we will see, it is useful to report the standard errors below the corresponding coefficient, usually in parentheses.

Variance of the OLS Estimators XXIII

EXAMPLE: Return to Education Using WAGE2

$$\begin{aligned} \widehat{lwage} &= \underset{(.081)}{5.973} + \underset{(.0059)}{.0598}\, educ \\ n &= 935,\ R^2 = .097 \end{aligned}$$

In this regression, $\hat{\sigma} = .4003$

Variance of the OLS Estimators XXIV

```
R> data("wage2")
R> summary(lm(lwage ~ educ, data = wage2))
Call:
lm(formula = lwage ~ educ, data = wage2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.94620 -0.24832  0.03507  0.27440  1.28106

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.973063   0.081374   73.40   <2e-16 ***
educ        0.059839   0.005963   10.04   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4003 on 933 degrees of freedom
Multiple R-squared:  0.09742,       Adjusted R-squared:  0.09645
F-statistic: 100.7 on 1 and 933 DF,  p-value: < 2.2e-16
```