EEP/IAS 118 - Introductory Applied Econometrics, Section 3

Pierre Biscaye and Jed Silver

September 2021

Quick review (1/2)

- Variance and covariance
  1. (Population) variance: $\frac{\sum_{i=1}^{n}(x_i - \mu_x)^2}{n}$
  2. (Population) Covariance: $\frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{n}$
- Punchline of SLRs 1 through 4:
  1. $E(\hat{\beta}_0) = \beta_0$
  2. $E(\hat{\beta}_1) = \beta_1$

Quick review (2/2): Goodness of fit (R^2)

Three main terms to define to understand the R^2 and how to calculate it:

1. Sum of Squares Total (SST) $= \sum_i^n (y_i - \bar{y})^2$
2. Sum of Squares Explained (SSE) $=\sum_i^n (\hat{y}_i - \bar{y})^2$
3. Sum of Squared Residuals (SSR) $= \sum_i^n (y_i - \hat{y})^2$

- Note that $SST = SSE + SSR$
- The R^2 is defined: $R^2 = \frac{SSE}{SST} = \frac{SSE}{SSE+SSR} = 1 - \frac{SSR}{SST}$
- You can think of the R^2 as how much of the *total* sample variation in y is explained by our model
- R^2 is always less than 1. Being closer to one indicates a better model fit

Agenda

1. Population parameters and sample estimators
2. Multiple linear regression

Population Parameters

If X is a random variable, the expected value (or expectation) of X, is the weighted average of all possible values of X.

$$E(X) = \mu = \sum_{j=1}^{k} x_j f(x_j)$$

If X is a random variable, the variance tells us the expected distance from X to its mean:

$$Var(X) = \sigma^2 = E[(X - E(X))^2]$$

Both of these are **population parameters**.

Sample Estimator

Since we don't observe the population, we can calculate the average and variance in a sample. This is the best *estimate* for the average and variance in the population.

Definition: An **estimator** is a pre-specified rule (function) that assigns a value to some unknown population parameter θ for any possible outcome of the sample.

Sample Estimator

Recall our population has mean μ and variance σ_X^2. Then

- An **estimator** of μ is the sample mean $\bar{X} = \dfrac{1}{n} \sum_i X_i$
- An **estimator** of σ_X^2 is $s_X^2 = \dfrac{1}{n-1} \sum_i (X_i - \bar{X})^2$

When we collect a specific sample from this population, we can get a particular **estimate** for $\bar{X}$ and s_X^2

Note: Prof. Magruder will sometimes write $\hat{\sigma}_X^2$ or s_X^2, but they mean the same thing.

Standard Errors of Estimators

Remember that estimators themselves are **random variables** because they depend on a random sample: as we obtain different random samples from the population, the values of $\bar{X}$ can change. Hence they have a certain probability distribution, with a certain mean and a certain variance/ standard deviation.

Standard Errors of Estimators

Just like our underlying variable X has an expected value and variance, so does the estimator $\bar{X}$.

$$E[\bar{X}] = E[\frac{1}{n}\sum_i X_i] = \frac{1}{n}E[\sum_i X_i] = \frac{1}{n}nE[X_i] = \frac{1}{n}n(\mu) = \mu$$

$$Var[\bar{X}] = Var\left[\frac{1}{n}\sum_i X_i\right] = \frac{1}{n^2}Var\left[\sum_i X_i\right] = \frac{1}{n^2}nVar[X_i] = \frac{\sigma_X^2}{n}$$

$$Sd[\bar{X}] = \sqrt{(Var[\bar{X}])} = \frac{\sigma_X}{\sqrt{n}}$$

BUT we don't know σ_X because this is a *population* parameter! So how can get the standard deviation of our estimator?

Standard Errors of Estimators

So, instead we use our estimator for σ_X, $s_X = \dfrac{1}{n-1}\sum_i^n (x_i - \bar{x})^2$.

We call this term the **standard error** - essentially the standard deviation of our estimator once we replaced the population σ_X with the sample estimator s_X

$$Se[\bar{X}] = \frac{s_X}{\sqrt{n}}$$

- **Note:** It may seem unituitive to divide by $n-1$ rather than n but basically this is a way of accounting for fact that the sample mean in this formula is also an estimator with its own variance

Summary: X as continuous variable

	Symbol	Formula
Population parameters	μ	$\sum_{j=1}^{k} x_j f(x_j)$
	σ_X^2	$E[(X - E(X))^2]$
	σ_X	$\sqrt{E[(X - E(X))^2]}$
Sample estimators	$\bar{X}$	$\frac{1}{n} \sum_i X_i$
	s_X^2	$\frac{1}{n-1} \sum_i (X_i - \bar{X})^2$
	s_X	$\sqrt{\frac{1}{n-1} \sum_i (X_i - \bar{X})^2}$
Estimator parameters	$E(\bar{X})$	μ
	$Var(\bar{X})$	$\frac{\sigma_X^2}{n}$
	$Sd(\bar{X})$	$\frac{\sigma_X}{\sqrt{n}}$
SE of estimator	$Se(\bar{X})$	$\frac{s_X}{\sqrt{n}}$

Summary: X as binary variable

	Symbol	Formula
Population parameters	μ	p
	σ_X^2	$p(1-p)$
	σ_X	$\sqrt{p(1-p)}$
Sample estimators	$\bar{X}$	$\hat{p}$
	s_X^2	$\hat{p}(1-\hat{p})$
	s_X	$\sqrt{\hat{p}(1-\hat{p})}$
Estimator parameters	$E(\bar{X})$	p
	$Var(\bar{X})$	$p(1-p)$
	$Sd(\bar{X})$	$\sqrt{p(1-p)}$
SE of estimator	$Se(\bar{X})$	$\frac{s_X}{\sqrt{n}}$

The estimator $\hat{\beta}$

Transitioning back to the population model we discussed previously:

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

$\hat{\beta}_0$ and $\hat{\beta}_1$ are **estimators** for the parameters β_0 and β_1. Indeed, we derived a formula for our βs. This was a rule that assigns each possible outcome of the sample a value of β. Then, for the given sample of data we work with, we obtain particular intercept and slope **estimates**, $\hat{\beta}_0$ and $\hat{\beta}_1$.

Recall that because $\hat{\beta}_1 \equiv \text{cov}(x, y)/\text{var}(x)$ is an estimator based off a random sample, it has a **standard error** of its own.

Summary: Regression

	Symbol	Formula
Population estimators	β_0 β_1	
Sample estimators	$\hat{\beta}_0$ $\hat{\beta}_1$	$\bar{y} - \hat{\beta}_1 \bar{x}$ $\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
Estimator parameters*	$E(\hat{\beta}_0)$ $E(\hat{\beta}_1)$ $Var(\hat{\beta}_1)$ $Sd(\hat{\beta}_1)$	β_0 β_1 $\frac{\sigma_u^2}{SST_x}$ $\frac{\sigma_u}{\sqrt{SST_x}}$
SE of estimator	$Se(\hat{\beta}_1)$	$\frac{\hat{\sigma}_u}{\sqrt{SST_x}}$

*We don't show $Var(\hat{\beta}_0)$, $Sd(\hat{\beta}_0)$, or $Se(\hat{\beta}_0)$ because we rarely care

Review: Assumptions of (Simple) Linear Regression

We make these assumption about the “true data generating process”

Model	Simple
SLR.1	The population model is linear in parameters $y = \beta_0 + \beta_1 x_1 + u$
SLR.2	$\{(x_i, y_i), \quad i = 1 \cdots N\}$ is a random sample from the population
SLR.3	The observed explanatory variable (x) is not constant: $Var(x) \neq 0$
SLR.4	No matter what we observe x to be, we expect the unobserved u to be zero $E[u\|x] = 0$
SLR.5	The “error term” has the same variance for any value of x : $Var(u\|x) = \sigma^2$

Why Multiple Linear Regression?

If we think other variables besides our variable interest x_1 belong in our model, then we want to include them in our model

- Otherwise they'd enter in u, violating SLR4.

The following two equations illustrate these benefits explicitly:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u \tag{1}$$

$$consumption = \beta + 0 + \beta_1 inc + \beta_2 inc^2 + u \tag{2}$$

In equation (1) we want to know the direct effect of education on wages. Here we explicitly control for experience.

- Compared to SLR, we have effectively taken *experience* out of the error term and put it explicitly in the equation.
- Otherwise we would have had to unrealistically assume that experience is uncorrelated with education to make SLR 4 hold.

In equation (2) the model falls outside simple regression because it contains two functions of income, *income* and $income^2$

What Changes With Multiple Linear Regression

Our model is now $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \mu$. How do we interpret β_1 now?

- We now say it is the effect of x_1 on y *holding all else fixed*
- Can also rephrase this as *ceteris paribus* or *conditional on* or *controlling for* $x_2 \ldots x_k$

We will also need to adjust our assumptions to accomodate the additional x variables...

Assumptions for Multiple Linear Regression

Up to now, we have dealt with regressions with only one explanatory variable. How do the necessary assumptions change when we have multiple X ?

Model	Multiple
MLR.1	The population model is linear in parameters $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \mu$
MLR.2	$\{(x_{i1}, \cdots, x_{ik}, y_i), \quad i = 1 \cdots N\}$ is a random sample from the population
MLR.3	No perfect colinearity among observed variables and $Var(x_j) \neq 0$, $j = 1 \cdots k$
MLR.4	No matter what we observe $(x_{i1}, \cdots, x_{ik})$ to be, we expect the unobserved u to be zero $E[u\|x_1, \cdots, x_k] = 0$
MLR.5	The "error term" has the same variance for any value of$(x_1, \cdots x_k)$: $Var(u\|x_1, \cdots x_k) = \sigma^2$

What do we get from these assumptions?

Using only assumptions 1 - 4, we can prove that:

1. $E(\hat{\beta}_j) = \beta_j$

This means that the mean of our estimators $\hat{\beta}_j$ are our true population parameters β_j

If we add assumption 5, we can also show that:

2. $Var(\hat{\beta}_j) = \frac{\sigma_u^2}{SST_j(1-R_j^2)}$

where $SST = \sum_j (x_{ij} - \bar{x})^2$ is the total sample variation in x_j, and R_j^2 is the R squared from regressing x_j on all other independent variables, and

$$\hat{\sigma}_u^2 = \frac{\sum_i \hat{u}_i^2}{n-2}$$

Exercise

Suppose we estimated the following equation:

$$\widehat{educ} = 10.36 - 0.094 sibs + 0.131 meduc + 0.210 feduc$$

where $educ$ is years of schooling, $sibs$ is number of siblings, $meduc$ is mother's years of schooling, and $feduc$ is father's years of schooling.

1. Does sibs have the expected effect?
2. Holding meduc and feduc fixed, by how much does sibs have to increase to reduce predicted years of education by one year?
3. Discuss the interpretation of the coefficient on $meduc$
4. Suppose that Alice has no siblings, and her mother and father each have 12 years of education. Bob has no siblings, and his mother and father each have 16 years of education. What is the predicted difference in years of education between Alice and Bob?

Exercise Solutions

1 Does sibs have the expected effect?
Having one additional sibling is associated with a 0.094 decrease in predicted years of education, holding $meduc$ *and* $feduc$ *fixed. Because of budget constraints, it makes sense that, the more siblings there are in a family, the less education any one child in the family has.*

2 Holding meduc and feduc fixed, by how much does sibs have to increase to reduce predicted years of education by one year?
To find the increase in the number of siblings that reduces predicted education by one year, we solve:

$$\begin{aligned} 1 &= 0.094(\Delta sibs) \\ \Delta sibs &= \frac{1}{0.094} \\ &= 10.6 \end{aligned}$$

Group Exercise Solutions

3 Discuss the interpretation of the coefficient on $meduc$
One more year of mother's education is associated with 0.131 years more of predicted education holding $sibs$ *and* $feduc$ *fixed. So if a mother has four more years of education, her child is predicted to have about a half a year (.524) more years of education.*

Exercise Solutions

4 Suppose that Alice has no siblings, and her mother and father each have 12 years of education. Bob has no siblings, and his mother and father each have 16 years of education. What is the predicted difference in years of education between Bob and Alice?

For Alice we have:

$$\widehat{educ} = 10.36 - 0.094(0) + 0.131(12) + 0.210(12)$$

For Bob we have:

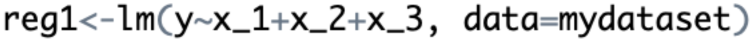
$$\widehat{edu} = 10.36 - 0.094(0) + 0.131(16) + 0.210(16)$$

So the predicted difference in education between Bob and Alice is

$$0.131(4) + 0.210(4) = 1.364$$

How to run MLR models in R

- What happens if you want to run a model with multiple x's in R?
- It's very similar to single regression, we still use `lm()`
- Just now we have:

  ```
  reg1<-lm(y~x_1+x_2+x_3, data=mydataset)
  ```

  for a data set object we've named "mydataset" that contains dependent variable 'y' and independent variables 'x_1', 'x_2' and 'x_3'
- Click here for an example!