Econometrics
Chapter 4: Linear Regression with One Regressor

Basic Model

Lets look at the relationship between class size (measured as student teacher ratio) and student test scores.

What are we interested in is important. There are two possibilities:

Model 1: The statistical model - *Model 2*: The structural model

What we want to know is $\frac{\Delta TestScore}{\Delta ClassSize}$ - but what do we mean by this?

Model 1: What is the relationship in the world - what is observed - a property of the joint distribution function (used for prediction)

Model 2: What is the causal effect of class size on test scores - what would happen if we intervened in class size holding all else constant (ceteris paribus)? This is not necessarily in the data/joint distribution function. (used for policy)

Keep this distinction in mind, but for now lets talk about how we will answer either question. Then we will discuss under what assumptions will we get the answer we want.

Lets assume a linear model:

$TestScore = \beta_0 + \beta_{ClassSize} \times ClassSize$

Now this cannot be exactly right for every students' test score.

But we could say this relationship holds *on average* - so it is modeling how the average test score changes as class size changes (again it will become important what we mean by 'as class size changes')

How about we add up all the other things that affect test score for each student and use a more common notation:

$Y_i = \beta_0 + \beta_1 X_i + u_i$ for $i = 1, 2, 3...n$

These are linear models with a single regressor - note without the u_i it is the equation of a line

Y is the *dependent variable* (explained)
X is the *independent variable* (explanatory)

the first part: $Y_i = \beta_0 + \beta_1 X_i$ is the population model/function. It is the part that holds on average for the population.

β_0 is the intercept
β_1 is the slope, or more generally, the parameter of interest

u_i is the disturbance/error term.

Think about u_i catching all other things that effect test scores when the values of X are *fixed*: family background, age, race, sex, teacher quality etc.

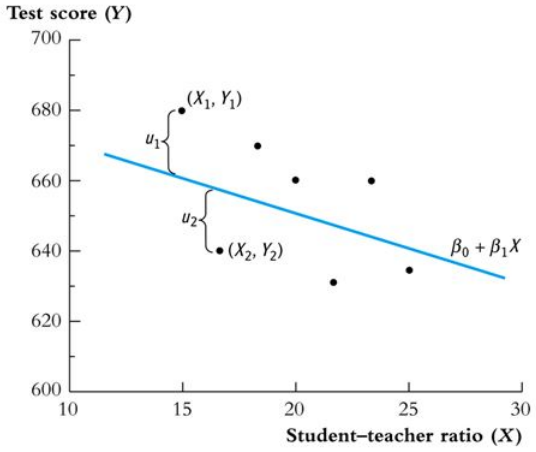
Of course β is not known, so we must use some data to estimate them.

So how might we go about 'estimating' what is a 'good' value of or β?

First lets look at the mechanics of estimation that we will use, and then discuss if this approach (or under what assumptions will this approach) give us the answer we have in mind.

Estimation

Figure 1: 'Best Fit' Line



Main idea: How would we fit a 'best' line through the dots?

Thats what we want to do right - fit a line that sort of 'catches' the overall trend in the scatterplot.

But, how to pick what is the 'best' line? Well the standard is the 'least square error' line.

Ordinary Least Squares (OLS) Estimator

We choose the line that is 'closest' to the data in the sense of the least sum of the square errors in predicting Y given X

Recall the sample average $\overline{Y}$ was the *least squares estimator* of the population mean (section 3.1 Eq. 3.2), well the OLS estimator extends this notion.

Let b_0 and b_1 be some estimators of β_0 and β_1, we want a pair that minimizes:

$\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$

Whichever b's that do this are called the Ordinary Least Squares (OLS) estimators.

We usually denote the OLS estimate of with a 'hat' $\hat{\beta}$.

The predicted Y_i given the OLS estimates and X_i is $\hat{Y}_i$.

The *residual* (different than error term) is $\hat{u}_i = Y_i - \hat{Y}_i$.

So what is the OLS estimator?

$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2}$ = (sample cov(X,Y))/(sample var(X))

Also $\hat{\beta}_1 = Corr(Y, X)\frac{s_y}{s_x}$ so though the sample correlation is unit free, β_1 has units of Y/X

$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

Derivation not important (in appendix 4.2)

Say get results of:

$\hat{TestScore} = 698.9 - 2.28 \times STR$

So our model says 'a decrease' of the student-teacher ratio by 2 will lead to increase expected (average) test score of 4.56 points. What do we mean by 'a decrease'? This again depends on our interpretation of the model.

What about decreasing or increasing by large amounts? Be careful when trying to use results to predict very large changes - ones outside of the sample area.

Why Should We Use this OLS Estimator?

The OLS estimator has 'good' properties. If certain assumptions hold then the OLS estimator is unbiased and consistent. And under some additional assumptions (put off till later) it is also the most efficient among a certain class of estimators.

Measures of Fit

How well does your regression line describe your data?

There are two measures: R^2 and the standard error of the regression (SER)

The R^2

The regressions R^2 is the fraction of the variance of Y that is described/predicted by your X

Note that we can write our observations of Y as a sum of their predictions (based on the regression) and the resulting residual:

$Y_i = \hat{Y}_i + \hat{\epsilon}_i$

So the R^2 is the ratio of the sample variance of $\hat{Y}$ to that of Y

We can write these as:

Explained Sum of Squares (ESS): $\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$

Total Sum of Squares (TSS): $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$

(Note the ESS uses $\overline{Y}$ instead of $\overline{\hat{Y}}$ because OLS average of $\hat{Y}$ is the average of Y - see appendix 4.3)

So $R^2 = \frac{ESS}{TSS}$

Or, since there are only two parts to the actual Y (that explained and that not), can think of R^2 in terms of the fraction of the variance of Y not explained by the regression:

Sum of Squared Residuals (SSR): $\sum_{i=1}^{n}\hat{u}_i^2$

Can be shown (appendix 4.3) that TSS = ESS + SSR. So could also write:

$R^2 = 1 - \frac{SSR}{TSS}$

In the simple case with only one X the R^2 is also the square of the correlation coefficient between Y and X.

Lets see:

$$
\begin{aligned}
R^2 = \frac{ESS}{TSS} &= \frac{\sum(\hat{Y} - \overline{Y})^2}{\sum(Y - \overline{Y})^2} \quad \text{By def. of ESS and TSS} \\
&= \frac{\sum(\hat{\beta}_0 + \hat{\beta}_1 X - \overline{Y})^2}{\sum(Y - \overline{Y})^2} \quad \text{By def of } \hat{Y} \\
&= \frac{\sum(\overline{Y} - \hat{\beta}_1\overline{X} + \hat{\beta}_1 X - \overline{Y})^2}{\sum(Y - \overline{Y})^2} \quad \text{By def. of } \hat{\beta}_0 \\
&= \frac{\sum(\hat{\beta}_1 X - \hat{\beta}_1\overline{X})^2}{\sum(Y - \overline{Y})^2} \\
&= \frac{{\hat{\beta}_1}^2\sum(X - \overline{X})^2}{\sum(Y - \overline{Y})^2} \\
&= \Big(\frac{cov(X,Y)}{var(X)}\Big)^2 \cdot \frac{\sum(X - \overline{X})^2}{\sum(Y - \overline{Y})^2} \quad \text{By def. of } \hat{\beta}_1 \\
&= \Big(\frac{cov(X,Y)}{var(X)}\Big)^2 \cdot \frac{var(X)}{var(Y)} \\
&= \frac{cov(X,Y)^2}{var(X)var(Y)} = corr(X,Y)^2
\end{aligned}
$$

Also note R^2 ranges between 0 and 1.

If $\hat{\beta}_1$ is 0 then so is R squared. This should make sense - X does not explain any of Y because it does not affect it. (Try it yourself. Start with the definition of R^2 - either one - and substitute in like we did above. The key part will be when you substitute in the definition of $\hat{\beta}_0$).

One the other hand, if your predicted model happened to hit every observation exactly then your R^2 would be 1.

Standard Error of the Regression - SER

SER ($= s_{\hat{u}}$) is a *estimator* of the standard deviation of the regression error (u). And it is in same units as Y.

This tells us on average how much off is the regression line in terms of units of Y.

Of course the regression errors are not observed, so we estimate using the OLS residuals.

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

Why (n-2) again degrees of freedom correction - beyond our scope to really discuss with justice. This uses fact that average $\hat{u}$ is zero (see appendix).

Lets get back to test score example:
The R^2 is 0.051 and SER is 18.6.

So 5.1% of the variance in test scores is predicted by class size. And the standard deviation of the true Y from the predicted is 18.6

Assumptions of OLS

Start with our regression line: $Y_i = \beta_0 + \beta_1 X_i + u_i$

Now before we move on we have to decide what we are trying to do. This matters. For example do you just want to know how education and earnings are related? Or do you want to know how education causes earnings?

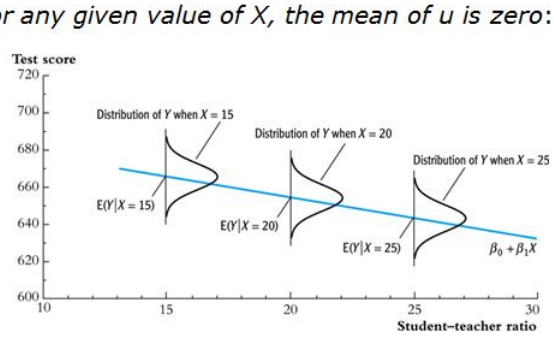
To simplify things for the time being assume that things are linear. That is X and Y are related in a linear way

#1. The Conditional Distribution of u_i Given X_i has a mean of zero. $E[u_i|X_i] = 0$

This is a mathematical way of saying the 'things' in u should not be related to the X's in the sense that given a value of X the mean of the distribution of these other things is zero.

This is the assumption that depends on what your goal is - ie. what question are you asking. I come back to this below

Figure 2: Errors centered at zero for all values of X

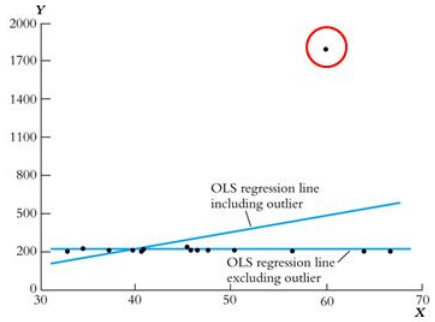


#2. (X_i, Y_i), are i.i.d.: This is automatic if drawn from random simple sample. Identical: if from same population, Independent: if randomly drawn.

Won't hold for things like time series, earnings for someone this year is probably not independent of their earnings last year.

#3. Large Outliers are Unlikely: Technically we mean X and Y have nonzero finite fourth moments: $0 < E(X_i^4) < \infty$, and for Y too. It is used to justify the large sample approximations to the distributions of the OLS test statistics.

Graphically can see how large outliers can hurt our OLS est.



What role do these assumption play??

Role 1. Under these assumptions (2 & 3), in large samples, the OLS estimators have sampling distributions that are normal - thus we can form test statistics/confidence intervals (based on large sample properties).

Role 2. And (#1) gives guidelines for when OLS gives useful (unbiased) estimates.

For what model do we 'need' assumption #1?

Again, recall for now we are assuming the world is linear.

For model #1 (statistical model), we only need to worry about assumptions 2 and 3.

So if we only care about the *association* between test scores and class size we do not need to worry about assumption 1. Why? - because it will hold by *definition*.

We *define* the model that we are interested in as the 'average in the population given a value of X' and the error as the difference between what we observed for individuals and that average.

Note that what we care about in this case is the Conditional Expectation Function (CEF): $E(Y|X)$.

So the model we want is $Y_i = E(Y|X_i) + u_i$. And $u_i = Y_i - E(Y|X_i)$.

So: $E(u_i|X_i) = E(Y_i|X_i) - E[E(Y|X_i)] = 0$

And so OLS will give unbiased estimates of β.

It may be worth noting that the CEF is the best predictor of Y given X. And even when the CEF is not linear OLS gives the best linear predictor of the CEF, and if the CEF is linear (as we assumed by linearity here) then OLS gives the CEF.

For model #2 (structural model) we **do** need to worry about assumption #1. And normally what we care about is a structural model.

Because we are no longer trying to estimate the CEF ($E(Y|X)$) but rather a model with some *causal interpretation*. Remember the question you are asking (and so model you are trying to estimate) is an idea, OLS is algebra. The algebra is not guaranteed to give you what you want.

Some people refer to this model as a 'data generating process', and some denote it as $E(Y|do(X))$ because it involves thinking about some possible *intervention*.

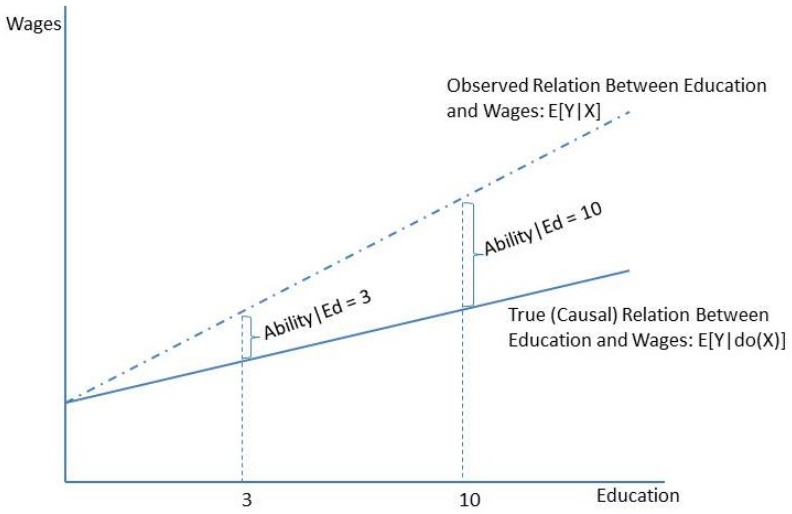
In this case we need to think about whether the things in the error term are systematically related to our X. In other words, if we look at the real world, do the other things that affect Y (ie. the error term) shift in mean when we change X?

This is commonly called an 'exogeneity assumption'. Do we believe this?

The classic case is the effect of education on wages. Are there other things that affect wages that are maybe related to education levels? What about ability/IQ?

What does it look like if assumption 1 does not hold?

Figure 3: Omitted Variable Bias.



IQ is likely related to education levels and also wages. So if we leave it out OLS does not give us the 'causal' relationship between education and wages - this is the Omitted Variable Bias (OVB).

This latter model is typically what we care about - and the one we will discuss from here on out - and so we need to worry about assumption #1. And *if* it holds then OLS will also give us unbiased estimates of the β we care about.

Sampling Distribution of the OLS Estimators

REMEMBER: since our estimates (of β_0 and β_1) are computed using random variables then they too are random and thus also have a distribution - their sampling distribution.

Under the Assumptions above, the means of the sampling distributions of our β's are the true values (that is: they are *unbiased* estimators):

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

Also, if the sample is 'large' then the distributions of our β's are well approximated by a bivariate normal distribution. This implies that the marginal distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ are also normal.

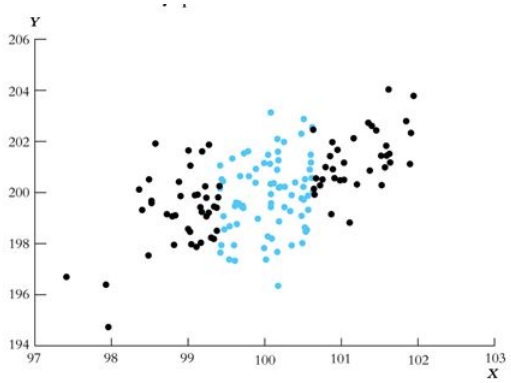
The large sample normal distribution of $\hat{\beta}_1$ is $N(\beta_1, \sigma^2_{\hat{\beta}_1})$ where:
$\sigma^2_{\hat{\beta}_1} = \frac{1}{n}\frac{var[(X_i - \mu_X)u_i]}{[var(X_i)]^2}$

The large sample normal distribution of $\hat{\beta}_0$ is $N(\beta_0, \sigma^2_{\hat{\beta}_0})$ where:
$\sigma^2_{\hat{\beta}_0} = \frac{1}{n}\frac{var(H_i u_i)}{[E(H_i^2)]^2}$ where $H_i = 1 - [\frac{\mu_X}{E(X_i^2)}]X_i$

This draws on the CLT. This allows us, since we know the sampling distribution, to conduct tests etc.

And you can see above, as n grows the sampling distribution collapses around the true value - the estimators are *consistent* also.

Also note that the larger the variance of X then the lower the variance of $\hat{\beta}_1$ and the smaller the variance of u_i then the smaller the variance of $\hat{\beta}_1$. To understand the first note the following picture, which would you rather have? the full data (larger variance in X) or just the part in the middle (smaller variance of X).



Brief History of 'Regression'

The term 'regression' comes from a paper by Sir Francis Galton (1886)

In it he was measuring the relation between parental height and children's height.