OLS and the Classical Regression Model

Devesh Raval

1 Projection- Geometry of OLS

In the previous classes, I went over three different analogy estimators for OLS. Now- will go to the matrix algebra projection interpretation of OLS. Basically, OLS is a projection in the X space in the y direction.

Our dependent variable is Y and independent variables $X = [X_1 X_2 ... X_k]$. I then define $S(X)$ as a subspace spanned by the $x_1, ..., x_k$ vectors:

$$S(X) \quad = \quad \{z | z = \sum_{i=1}^{k} b_i x_i, b_i \in \Re\}$$

Thus, $S(X)$ consists of every vector that can be formed as a linear combination of the X_i.

Now put up graph example- remember this is all in the population!!

Can always do this- define y as combination of $x\beta$ and u- not necessarily making any assumptions- BLP!!!

$x\beta$ line is the "shadow" line of y- closest one can get to y in $S(X)$.

Now to estimation in sample:

Recall how we found $\hat{\beta}$ from normal equations (i.e. moment conditions):

$$\begin{aligned} x'\hat{u} &= 0 \\ \hat{u} &= y - x\hat{\beta} \\ y &= x\hat{\beta} + \hat{u} \end{aligned}$$

In sample of 2 Xs:

$$\begin{matrix} X_1' \\ X_2' \end{matrix} \quad \hat{u} = \quad 0$$

X_1', X_2' both 1 by N, $\hat{u}$ is N by 1. Thus, the vector $\hat{u}$ is orthogonal to each x_1 and x_2. But then its orthogonal to every vector in $S(X)$. That is,

$$\begin{aligned} X\beta &\in S(X) \\ (X\beta)'\hat{u} = \beta' X' \hat{u} &= 0 \end{aligned}$$

Thus, we know that, for $\hat{\beta}$:

1. $\hat{u}$ is orthogonal to $S(X)$.

2. $y = X\hat{\beta} + \hat{u}$

where by definition $X\hat{\beta} \in S(X)$ also.

See graphic.

Notes on graph- remember that any other β will give a larger $||u||$- and so will be worse. The best projection is one that gives a right angle so $\hat{u}$ is perpendicular.

We know that $\hat{\beta}$ is chosen such that $\hat{u}$ minimizes $||u||$, where $u = y - x\beta$, so depends on β. But minimizing $||u||$ is the same as min $(||u||^2)$- OLS procedure!!

By Pythagorean Thm:

$$\begin{aligned} ||y||^2 &= ||x\hat{\beta}||^2 + ||\hat{u}||^2 \\ y'y &= \hat{\beta}'XX\hat{\beta} + \hat{u}'\hat{u} \\ TSS &= ESS + RSS \end{aligned}$$

Here is a decomposition that tells us about the goodness of fit of the regression line:

$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$

Since $y = X\hat{\beta} + \hat{u}$, and $\hat{\beta} = (X'X)^{-1}X'Y$ and $\hat{u} = Y - X(X'X)^{-1}X'Y$

$$\begin{aligned} Y &= X\hat{\beta} + \hat{u} \\ &= X(X'X)^{-1}X'Y + (I - X(X'X)^{-1}X')Y \\ Y &= PY + MY \end{aligned}$$

where $M = I - P$.

Thus, $\hat{\beta}$ decomposes Y into two vectors- P projects Y into $S(X)$. M projects Y into a space orthogonal to $S(X)$.

Some properties of these matrices, given that $rank(X) = k$:

i) $P = P^1 = P^2$- P is symmetric and idempotent.

ii) $rank(P) = k$

iii) N eigenvalues of P- k are 1s, N-K 0s

iv) M is idempotent, $rank(M) = N - k$

v) of N eigenvalues, $N - K$ 1s, K 0s

We call P the projection matrix, M the annihalator matrix. Why?

P will take any vector and give the part that is projected onto S(X). M will take out the part that can be projected onto S(X) and give the part orthogonal to S(X). Thus PY gives the fitted values of the regression and MY the residuals.

What if we take X and find PX? Each column of X projected on $S(X)$ but each column of $X \in S(X)$- so what do we obtain?

$PX = X$ or $X(X'X)^{-1}X'X = X$.

Get fitted value of X if start with X- predict perfectly!!

Here

If we take X and find MX- there is no residual.

i.e. $(I - X(X'X)^{-1}X')X = 0$.

Since X is in S(X) nothing can be orthogonal to it- so no residuals.

2 Asymptotic Properties of LS Estimator

Remember that we showed that in the population, the BLP is:

$$\beta = (E(X'X)^{-1}E(X'Y)$$

either by solving the LS problem or by constructing U s.t. $E(X'U) = 0$ and $Y = X\beta + U$. No assumptions were made here except the rank condition on X.

We can then define the sample estimator $b_N = E_N(X'X)^{-1}E_N(X'Y)$ through the analogy principle. I will now show that it's consistent and asymptotically normal- nice properties!!

Consistency:

By the Law of Large Numbers,

$$\begin{aligned} E_N(X'X) &\to_p E(X'X) \\ E_N(X'Y) &\to_p E(X'Y) \end{aligned}$$

By the Slutsky and Mann Wald Thms:

$$\begin{aligned} E_N(X'X)^{-1}E_N(X'Y) &\to_p (E(X'X)^{-1}E(X'Y) \\ &= \beta \end{aligned}$$

Asymptotic Normality:

Note the following identity:

$$\beta = E_N(X'X)^{-1}E_N(X'X)\beta$$

I then use it as follows:

$$\begin{aligned} \sqrt{N}(b_N - \beta) &= \sqrt{N}(E_N(X'X)^{-1}E_N(X'Y) - \beta) \\ &= \sqrt{N}(E_N(X'X)^{-1}E_N(X'Y) - E_N(X'X)^{-1}E_N(X'X)\beta) \\ &= E_N(X'X)^{-1}\sqrt{N}(\frac{1}{N}\sum_i X_i'Y_i - \frac{1}{N}\sum_i X_i'X_i\beta) \\ &= E_N(X'X)^{-1}\sqrt{N}(\frac{1}{N}\sum_i X_i'(Y_i - X_i\beta)) \\ &= E_N(X'X)^{-1}\sqrt{N}(E_N(X'U)) \end{aligned}$$

Remember we know by the first order conditions to construct the BLP in the population that

$$E(X'U) \quad = \quad 0$$

We then have that:

$$\begin{aligned} E_N(X'X)^{-1} &\rightarrow_p E(X'X)^{-1} \\ \sqrt{N}(E_N(X'U)) &= \sqrt{N}(E_N(X'U) - E(X'U)) \\ &\rightarrow_d N(0, E(X'UU'X)) \end{aligned}$$

This is using the CLT- impt assuming iid sampling!!!- Cant have correlation between variables.

Putting these together, and using the CMT:

$$\begin{aligned} \sqrt{N}(b_N - \beta) &\rightarrow_d N(0, V) \\ V &= E(X'X)^{-1}E(X'UU'X)E(X'X)^{-1} \end{aligned}$$

This matrix is k by k show why given X is n by k.

A Special Case:

Conditional Homoskedasticity, No Autocorrelation

No autocorrelation- since CLT depends on iid sampling everything above rested on this.

$$E(X'UU'X) = E(X'XU^2)$$

Lets do this in sums:

$$\frac{1}{N}\sum_i X_i'U_iU_i'X_i \quad = \quad \frac{1}{N}\sum_i X_i'X_iU_i^2$$

Conditional Homoskedasticity:

$$E(U^2|X) \quad = \quad E(U^2)\forall X$$

Observe then that

$$\begin{aligned} E(X'XU^2|X) &= X'XE(U^2|X) \\ &= X'XE(U^2) \end{aligned}$$

By the LIE,

$$\begin{aligned} E(X'XU^2) &= E_X(E(X'XU^2|X)) \\ &= E(X'XE(U^2)) = E(X'X)E(U^2) \end{aligned}$$

If we put this in the formula for V:

$$\begin{aligned} V &= E(X'X)^{-1}E(X'UU'X)E(X'X)^{-1} \\ &= E(X'X)^{-1}E(X'X)E(U^2)E(X'X)^{-1} \\ &= E(X'X)^{-1}E(U^2) \end{aligned}$$

In the HW you have to derive what the bivariate version of this looks like for the Linear Max Lik case.

3 Classical Regression Model

Here we start putting some assumptions- to get finite sample results. Already derived the asymptotic results without any assumptions!!

Assumption 1. $E(Y|X) = X\beta$- this states that the CEF is linear.

Notice that this linearity is linear in terms of parameters β, not X.
Another way to put this is $Y = X\beta + \epsilon$ or $E(\epsilon|X) = 0$.

Assumption 2. $V(Y|X) = \sigma^2 I$ or $V(\epsilon|X) = \sigma^2 I$.

This assumes homoskedasticity and no autocorrelation- will get to this a minute.

Assumption 3. $Rank(X) = k$

This is no perfect multicollinearity.
Give an example with constant, male, and female.
Fix constant at 2- can put male at 10, female at 2 or constant at 1, male at 11, female at 3.
So no one parameter solving everything!!

Assumption 4. X is non-stochastic. In repeated sampling, X always takes on the same values.

Do not need this but simplifies some proofs. Can think of doing an experiment- on a farm and set X as amt of fertilizer- always set X same way in different random samples of experiment.
An alternative to this assumption is the following:

Assumption 4′. Neoclassical Regression Model: X is stochastic (random sampling from the joint distribution $\{Y, X\} \sim P$.

Assumption 5. Classical Normal Regression Model: $Y|X \sim N(X\beta, \sigma^2 I)$

If take the maximum likelihood estimate of this equation- get the LS estimator- we already saw this!!
We then have:

$E(Y|X) = \begin{array}{c} X_1\beta \\ X_2\beta \\ \\ \\ X_n\beta \end{array}$ where X_i is a 1 by k row of X.

$V(Y|X)$ =matrix with diagonal σ^2, offdiagonals zero. n by n matrix. Explain what that means.

Homoskedastic vs. heteroskedastic results:

Given an example with x axis schooling, y axis wage, z axis distn. Some LS line in X Y plane, distn around it.

Classic normal - same distribution and variance at each point on LS line.

Classic- same variance at each point on LS line.

3.1 Finite Sample Properties

Note that we didnt need any of these assumptions for the asymptotic results- only the assumption 2 for a simpler form for the variance.

But for small sample properties do need!

What are the properties of $b_N = (X'X)^{-1}X'Y$?

We start with the rank assumption on X- only assumption on X not Y|X!!.

1. No assumptions on $P(Y|X)$- distribution of $Y|X$.

$$\begin{aligned} E(b_N|X) &= E((X'X)^{-1}X'Y|X) \\ &= (X'X)^{-1}X'E(Y|X) \end{aligned}$$

Since no assumption on $P(Y|X)$- cant say more! This is because the expectation of a ratio is not equal to the ratio of an expectation. Cant just break up big expectation into multiple of smaller ones!!

$$\begin{aligned} V(b_N|X) &= V((X'X)^{-1}X'Y|X) \\ &= (X'X)^{-1}X'V(Y|X)X(X'X)^{-1} \end{aligned}$$

This is similar to variance of limiting distn.

2. $E(Y|X) = X\beta$

$$\begin{aligned} E(b_N|X) &= E((X'X)^{-1}X'Y|X) \\ &= (X'X)^{-1}X'E(Y|X) \\ &= (X'X)^{-1}X'X\beta = \beta \end{aligned}$$

If $E(b_N|X) = \beta$ for all values of X, also true unconditionally- this is the LIE!! We then have an unbiased estimator.

3. $V(Y|X) = \sigma^2 I$

$$\begin{aligned} V(b_N|X) &= (X'X)^{-1}X'(\sigma^2 I)X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

$$\begin{aligned} V(b_N) &= E(V(b_N|X)) + V(E(b_N|X)) \\ &= E(V(b_N|X)) = \sigma^2 E(X'X)^{-1} \end{aligned}$$

The first inequality uses a conditional variance equality and the second the fact that $E(b_N|X)$ is a constant as we showed.

Example for intuition:

Let $X = [1\ X_1]$ This is the bivariate regression case. Note: if you are ever confused, look at the bivariate regression case!!

Then we have (you should have to do this for HW??):

$$\begin{aligned} V(b_N|X) &= \sigma^2(X'X)^{-1} \\ &= \sigma^2[\begin{array}{cc} \sum_i 1*1 & \sum_i 1*X_{i,1} \\ \sum_i 1*X_{i,1} & \sum_i X_{i,1}*X_{i,1} \end{array}]^{-1} \\ & \sigma^2[\begin{array}{cc} n & \sum_i 1*\bar{X}_{i,1} \\ n\bar{X}_1 & \sum_i X^2_{i,1} \end{array}]^{-1} \\ &= \frac{\sigma^2}{NS^2_{x1}}[\begin{array}{cc} \bar{X_1}^2 & -\bar{X}_1 \\ -\bar{X}_1 & 1 \end{array}] = [\begin{array}{cc} Var(b_1|x) & Cov(b_1,b_2|x) \\ Cov(b_1,b_2|x) & Var(b_2|x) \end{array}] \end{aligned}$$

Here:

$$S^2_{x1} = \frac{1}{N-1}\sum_i (X_{i,1} - \bar{X}_i)^2$$

Then:

$$V(b_{2N}|X) = \frac{\sigma^2}{NS^2_{x1}}$$

We can see it increases with σ^2- as y has more variance conditional on X we get less accurate results. Decreases with N and decreases with s^2_{x1}- as have more variance in X- can get better estimate of β- have to think of variance in X variables!!

4. If we are in the CNLRM:

$$\begin{aligned} Y|X &\sim N(x\beta, \sigma^2 I) \\ P(b_N|X) &\sim N(\beta, \sigma^2(X'X)^{-1}) \end{aligned}$$

5.

Theorem 1. *Gauss Markov Theorem In the class of linear unbiased estimators, OLS estimator attains minimum variance.*

Proof:

Note: Hopefully this helps with the proof. Let $\hat{\theta}$ and $\tilde{\theta}$be estimators of a vector parameter θ. Then define $A = E((\hat{\theta}-\theta)(\hat{\theta}-\theta)')$ and $B = E((\tilde{\theta}-\theta)(\tilde{\theta}-\theta)')$.

Both of these are mean squared error type calculations- but in vector form. Then $\hat{\theta}$ is better than $\tilde{\theta}$ if:

$$C'(B-A)C \geq 0$$

for every vector C and every para value, and true without equality for at least one value of C, one value of parameter.

This is the same as saying $B-A$ non negative definite, $B \neq A$.

Now the proof.

Proof. Let $A = (X'X)^{-1}X'$ and so $\hat{\beta} = AY$. Then an alternative estimator can be defined as, WLOG, $\tilde{\beta} = (A+C)Y$- also linear.

Now we calculate the expectation and variance:

The expectation is:

$$\begin{aligned} E(\tilde{\beta}|X) &= (A+C)E(Y|X) \\ &= AX\beta + CX\beta \\ &= \beta + CX\beta \end{aligned}$$

where the second equality comes from OLS matrix algebra. For $\tilde{\beta}$ to be unbiased, we need that $CX = 0$- this will prove handy later.

$$\begin{aligned} V(\tilde{\beta}|X) &= E((\tilde{\beta} - E(\tilde{\beta}|X))((\tilde{\beta} - E(\tilde{\beta}|X))'|X) \\ &= E((\tilde{\beta} - \beta)((\tilde{\beta} - \beta)'|X) \\ &= E((A+C)UU'(A+C)|X) \\ &= \end{aligned}$$

Here we use

$$\begin{aligned} \tilde{\beta} &= (A+C)(X\beta + U) \\ &= AX\beta + AU + CU \\ \tilde{\beta} - \beta &= (A+C)U \end{aligned}$$

Earlier $AX = I$ b/c $PX = X$.

Continuing:

$$\begin{aligned} &= E((A+C)UU'(A+C)|X) \\ &= (A+C)E(UU'|X)(A+C)' \\ &= (A+C)(A+C)'\sigma^2 I \\ &= (A+C)(A'+C')\sigma^2 I \\ &= (AA' + AC' + CA' + CC')\sigma^2 \end{aligned}$$

Since $CA' = CX(X'X)^{-1} = 0$ for the estimator to be unbiased.

Then we have:

$$\begin{aligned} V(\tilde{\beta}|X) &= (AA' + CC')\sigma^2 \\ &= \sigma^2(X'X)^{-1} + \sigma^2 CC' \end{aligned}$$

Thus, $V(\tilde{\beta}|X) = V(\hat{\beta}|X) + \sigma^2 CC'$

Since CC' is a positive def matrix, $\sigma^2 > 0$ this is strictly bigger

This fullfills the previous criterion as for any h, $h'CC'h = (C'h)'(C'H) = v'v \geq 0$.

Thus we have shown that $\hat{\beta}$ is BLUE- best in MSE!! □

Some comments:

1. Still only applies to linear unbiased estimators.

2. If relax homoskedasticity- no longer BLUE- used this assumption in 1 step.

3. Both homoskedastic and normal- MLE and can use Cramer-Rao lower bound.

4 Model Structure in the CRM

So far we have assumed that we know the correct CEF and use the correct X variables in the regression. But in real life- may not be able to do this, because:

1. Some variables we can not observe or are not in our dataset

2. There are lots of variables in our dataset- which to use? We do not necessarily know the true CEF

3. Even if we know which variables are important, how do we put them in the regression?

4.1 Model Structure

- Dummy Variables and Interactions:

Dummy Variable is just variable that is 1 or 0. For example, male is 1 if someone is male otherwise 0.

Can use log wage example here.

Draw graph- dummy variables in regression affect intercepts!!

But can also include interactions- affect slope and intercept.

For ex: investigating discrimination:

wage= schooling+female+schooling*female

Does slope of schooling change with sex? Or just intercept?

Can have interactions with continuous variable also- try schooling and ability.

- Logarithms

Two reasons we in general use logarithms-

Sometimes variables are on a log scale so linear regressions work better in logs. For example, wages- Mincer regressions!

Second- want to measure elasticities or percentage changes (which are unitless)

Remember what an elasticity is-

$$\frac{dlogN}{dlogw} = \frac{dN}{dw} * \frac{w}{N}$$

Measures the percentage change in N for a percentage change in w- dont have to worry about the units of N or w.

vs.

$$\frac{dN}{dw}$$

only makes sense given units of both quantities. We can estimate an elasticity by running the following type of regression:

$$logN = \alpha + \beta \log w + \epsilon$$

Production function - in logs get elasticities also.

Nonlinearities:

Draw graph- quadratic CEF- what do residuals look like? what does line look like?

Can add quadratic, cubic terms to approximate some nonlinearity. If just have $\alpha + \beta X$, the marginal effect is linear- for example- schooling- same return from going from 8th to 9th grade as 15th to 16th.

If quadratic, marginal effect now depends on X.

Another way to do this- put in dummy variables. X* 1(in range). Or schooling- instead of having no of years- HS, college, etc.

4.2 Omitted Variables

What happens if we omit a variable that is in the true CEF?

Example 2. Ability Bias

Y=wage

X_1 =years of schooling

X_2 = ability

What is the return from going to school?

The true CEF includes ability:

$$E(Y|X_1, X_2) \quad = \quad \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Thus, β_1 measures the gain from extra schooling controlling for ability. If years of schooling increases by 1 year, Y will increase by β_1, after accounting for ability.

But- in our dataset we don't have IQ or ability- what can we do?

Just do OLS on years of schooling- what happens to the coefficient on the years of schooling? Does it really measure the partial effect of schooling on wages?

4.3 Omitted Variable Rule

Lets start by partitioning X into two sets of variables, $X = [X_1 X_2]$, X is n by k, X_1 n by k_1, X_2 n by k_2.

In our experiment here X belongs in the true CEF- but we only do OLS with X_1. Now, if we could do OLS with the full dataset, we would estimate b_N, where:

$$\begin{aligned} e_N &= Y - Xb_N \\ Y &= Xb_N + e_N \end{aligned}$$

This expression is just what Y equals in the sample- no assumptions here.

Now- what happens if we just run the regession of Y on X_1? We can think of the full regression as the long regression and the X_1 regression as the short regression.

We will get b_{1N}^* k_1 by 1:

$$\begin{aligned} b_{1N}^* &= (X_1'X_1)^{-1}X_1'Y \\ &= (X_1'X_1)^{-1}X_1'(X_1 b_{1N} + X_2 b_{2N} + e_N) \\ &= b_{1N} + (X_1'X_1)^{-1}X_1'X_2 b_{2N} + X_1'e_N \\ &= b_{1N} + (X_1'X_1)^{-1}X_1'X_2 b_{2N} \end{aligned}$$

Here $X_1'e_N = 0$ by the foc of the LS problem.

Thus, we have that the short regression coefficient is:

$$\begin{aligned} b_{1N}^* &= b_{1N} + Fb_{2N} \\ F &= (X_1'X_1)^{-1}X_1'X_2 \end{aligned}$$

Here X_1 is n by k1, X2 is n by k2- F is then k1 by k2. What does F look like? Its just X_2 is replacing Y- or coefficients of regression of X_2 on X_1.

We can call this regression the Auxiliary Regression:
Regress X_2 on X_1:

$$X_2 = X_1F + X_2^*$$

where X_2^* is the residual of the regression.

If we look at the bigger picture, the difference between the short regression coefficient and long regression coefficient is about the total derivative of Y wrt X_1 or the partial derivative of Y wrt X_1:

$$\frac{dY}{dX_1} = \frac{\partial Y}{\partial X_1} + \frac{\partial Y}{\partial X_2}\frac{\partial X_2}{\partial X_1}$$

What are these partial derivatives in terms of the CRM? How can we derive them?

Also, which derivative are we interested in? This depends on our economic question. Go back to the wage example.

If our question is- what is the difference in wages between someone who has a college degree and someone who has a high school degree, we want the total derivative.

If our question is- what is the rise in someone's wage if I make him go to college, its the partial derivative.

The total derivative is including the fact that those who go to college are higher ability people and so earn more anyway.

Now- if we omit a variable- is our estimate unbiased? consistent?

Consistency:

$$\begin{aligned} b_{1N}^* &= b_{1N} + E_N(X_1'X_1)^{-1}E_N(X_1'X_2)b_{2N} \\ &\to_p \beta_1 + E(X_1'X_1)^{-1}E(X_1'X_2)\beta_2 \end{aligned}$$

So inconsistent unless:
1. $\beta_2 = 0$- so omitted term does not belong in CEF
2. or X_1 orthogonal to X_2
Bias:

$$\begin{aligned} E(b_{1N}^*|X) &= E(b_{1N}|X) + (X_1'X_1)^{-1}(X_1'X_2)E(b_{2N}|X) \\ &= \beta_1 + (X_1'X_1)^{-1}(X_1'X_2)\beta_2 \end{aligned}$$

where the second step relies on the true CEF being $E(Y|X) = X\beta$.

Thus, we will have the same conditions for biasedness- Omitted Variable Bias!!

Lets look at this in the case where X_1 and X_2 are scalars and there is a constant:

$$\begin{aligned} plim\ b^*_{1N} &= \frac{Cov(x_1, y)}{Var(x_1)} \\ &= \frac{Cov(x_1, \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u)}{Var(x_1)} \\ &= \beta_1 + \frac{Cov(x_1, x_2)}{Var(x_1)} \beta_2 \end{aligned}$$

where $\frac{Cov(x_1,x_2)}{Var(x_1)}$ is the plim of the coefficient of a regression of x_2 on x_1.

This should be the same basic formula as before.

In the schooling example, the bias should be positive- as β_2 is positive (why?) and schooling and ability are positively correlated (why?). Usually we are interested in both the sign and magnitude of the bias!

4.4 Residual Regression Rule (or Frisch Waugh Thm)

The following are equivalent:

1. Get b_{2N} from regressing Y on X_1 and X_2.

2. Regress X_2 on X_1- the auxiliary regression- take the residuals X_2^* and get regress Y on X_2^*

3. Regress X_2 on X_1- the auxiliary regression- take the residuals X_2^*. Regress Y on X_1 and get residuals Y^*. Then regress Y^* on X_2^*.

Here by equivalent I mean we get the same estimates.

Why is this interesting?

1. Sometimes it can be nice to use the residuals instead of the variables themselves- I will give some examples below.

2. This "trick" will help us derive results on omitted and irrelevant variable properties.

Example 3. Demeaning variables:

If we regress a variable on a constant- the value of the constant is the mean and the residuals are the demeaned variables. (This is b/c $E(U) = 0$). Thus, instead of including a constant in the regression we can simply demean all the variables.

Example 4. Trend Removal

If we are doing time series analysis we may be concerned with a time trend in the data- some variables are just moving up over time. For ex- imagine effect of new law on crime- if crime rising over time want to account for that before seeing how law affects crime- otherwise may find law inc crime just b/c crime increases.

So- can add a time trend to the regression- or detrend all variables first (regressing against time trend) and then run regression on detrended variables.

Example 5. Seasonal Adjustment

Looking for BC patterns, but 1st quarter always has low sales, 4th quarter high sales. Can put in quarter dummies- or substract quarter means from x variables and then run regression.

Proof. I wish to show that $c_{2N} = (X_2^{*'} X_2^*)^{-1} X_2^{*'} Y = b_{2N}$

Lets first consider what X_2^* is- its the residual of a regression of X_2 on X_1. Thus,

$$\begin{aligned} X_2^* &= X_2 - X_1(X_1' X_1)^{-1} X_1' X_2 \\ &= X_2 - X_1 F \\ &= (I - X_1(X_1' X_1)^{-1} X_1') X_2 \\ &= M_1 X_2 \end{aligned}$$

Here M_1 is the Annihilator Matrix from a regression of X_1- so it gives the residuals of a regression with X_1- exactly what we have!

Then:

$$\begin{aligned} c_{2N} &= (X_2^{*'} X_2^*)^{-1} X_2^{*'} Y \\ &= (X_2^{*'} X_2^*)^{-1} X_2' M_1 Y \\ &= (X_2^{*'} X_2^*)^{-1} X_2' M_1 (X_1 b_{1N} + X_2 b_{2N} + e_N) \end{aligned}$$

Here $(X_1 b_{1N} + X_2 b_{2N} + e_N)$ come from a regression of Y on X- can always do this.

Now, we know $M_1 X_1$ is zero as a regression of X_1 on itself delivers a perfect fit.

$$\begin{aligned} c_{2N} &= (X_2^{*'} X_2^*)^{-1} X_2' M_1 (X_1 b_{1N} + X_2 b_{2N} + e_N) \\ &= (X_2^{*'} X_2^*)^{-1} X_2' M_1 (X_2 b_{2N} + e_N) \\ &= \end{aligned}$$

Again, $X_2' M_1 e_N = 0$ since $M_1 e_N = e_N$ as e_N is constructed to be orthogonal to X_1 and X_2. But then $X_2' e_N = 0$.

$$\begin{aligned} c_{2N} &= (X_2^{*'} X_2^*)^{-1} X_2' M_1 (X_1 b_{1N} + X_2 b_{2N} + e_N) \\ &= (X_2^{*'} X_2^*)^{-1} X_2' M_1 (X_2 b_{2N} + e_N) \\ &= (X_2^{*'} X_2^*)^{-1} X_2' M_1 (X_2 b_{2N}) \\ &= (X_2^{*'} X_2^*)^{-1} X_2' M_1' M_1 (X_2 b_{2N}) \\ &= (X_2^{*'} X_2^*)^{-1} X_2^{*'} X_2^* b_{2N} = b_{2N} \end{aligned}$$

If we also took the residuals from Y- would be equivalent to $M_1 Y$ above- we can always do this b/c of symmetric idempotent properties of M_1. □

Now I will show:

In Omitted Variable Case, short regression coefficient has a lower variance than long regression coefficient. Thus, even if short coeff (omitting a var) is biased, can have lower variance- so bias-variance tradeoff!

I will show that:

$$Var(b_{1N}^*|X) \quad = \quad Var(b_{1N}|X) - FVar(b_{2N}|X)F'$$

Thus its smaller generically unless F=0.

Proof. We know that:

$$\begin{aligned} b_{1N} &= b_{1N}^* - Fb_{2N} \\ Var(b_{1N}|X) &= Var(b_{1N}^*|X) + FVar(b_{2N}|X)F' - Cov(b_{1N}^*, b_{2N}|X)F' \end{aligned}$$

I will now show that $Cov(b_{1N}^*, b_{2N}|X) = 0$.

Define $A_1 = (X_1'X_1)^{-1}X_1'$ and $A_2^* = (X_2^{*'}X_2^*)^{-1}X_2^{*'}$. Then, using $V(Y|X) = \sigma^2 I$, we have:

$$\begin{aligned} Cov(b_{1N}^*, b_{2N}|X) &= Cov(A_1Y, A_2^*Y|X)) \\ &= A_1Var(Y|X)A_2^{*'} \\ &= \sigma^2(X_1'X_1)^{-1}X_1'X_2^*(X_2^{*'}X_2^*)^{-1} \end{aligned}$$

Now, $X_1'X_2^*$ is the correlation between X_1 and the residuals of a regression of X_2 on X_1- so these should be zero by definition (the resids are the orthogonal part.) So the cov term is equal to zero.

Thus-

$$Var(b_{1N}^*|X) \quad = \quad Var(b_{1N}|X) - FVar(b_{2N}|X)F'$$

□

I have shown that if we omit a variable from the regression, then the other variables will be biased but will have lower variance- we can try to trade off bias and variance.

4.5 Irrelevant Variables:

What happens if we put in variables which dont exist in the true CEF? Are the estimates of the other variables biased?

In our experiment here X_2 belongs in the true CEF- but we do OLS with X_1 and X_2.

Thus the long regression is the incorrect regression and the short regression is the true regression.

There are two ways to see that the OLS coefficients in the long regression are unbiased and consistent. First, the long CEF is just the short CEF with $\beta_1 = 0$- we never said that β had to be nonzero when defining the CEF. Thus, ests of β_2 should be fine, and b_{1N} should be unbiased and consistent for zero.

The algebra is as follows:

$$\begin{aligned} c_{2N} &= (X_2^{*'} X_2^*)^{-1} X_2' M_1 Y \\ E(c_{2N}|X) &= (X_2^{*'} X_2^*)^{-1} X_2' M_1 E(Y|X) \\ &= (X_2^{*'} X_2^*)^{-1} X_2' M_1 X_2 \beta_2 \\ &= (X_2^{*'} X_2^*)^{-1} X_2' M_1 M_1 X_2 \beta_2 \\ &= (X_2^{*'} X_2^*)^{-1} X_2^{*'} X_2^{*'} \beta_2 = \beta_2 \end{aligned}$$

However, the variance of the estimator in the long regression will be higher- as we showed earlier- the intuition here is that part of what matters in the variance is the variation in X- if we do the long regression the variation in our X1 is the residual variation after accounting for other Xs- will be smaller in general- so larger variances of the estimators.

So- if omit a variable risk biasing the coefficient- but if dont omit then have a larger variance- bias variance tradeoff!!

What to do in practice- just include the variables that theory specifies, and do robustness checks!

5 Variance of the LS Estimator in the CRM

5.1 R^2

R^2 is a measure of the goodness of fit of the regression- how well does a line fit the data relative to a constant.

Draw the picture.

We have the following identity, which I proved before and will prove again now:

$$\begin{aligned} \sum_i (y_i - \bar{y})^2 &= \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i e_i^2 \\ SST &= SSR + SSE \end{aligned}$$

Basic idea- line should have less error than a constant.

Proof:

$$\begin{aligned} Y &= \hat{Y} + e \\ \hat{Y}' e &= (X b_N)' e \\ &= b_N' X' e = 0 \end{aligned}$$

1.

$$\begin{aligned} y'y &= \sum_i y_i^2 = (\hat{y}+e)'(\hat{y}+e) \\ &= \hat{y}'\hat{y} + e'e + e'\hat{y} + \hat{y}'e \\ &= \hat{y}'\hat{y} + e'e \end{aligned}$$

2. Mean of y is mean of $\hat{y}$:

$$\begin{aligned} \sum_i y_i &= \sum_i \hat{y}_i + \sum_i e_i \\ \bar{y} &= \bar{\hat{y}} \\ \bar{e} &= 0 \end{aligned}$$

If there is an intercept in the regression $\bar{e} = 0$.

3. Then we can subtract out the mean and everything is fine:

$$\begin{aligned} \sum_i (y_i - \bar{y})^2 &= \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i e_i^2 \\ SST &= SSR + SSE \end{aligned}$$

Cross terms and squared mean term are the same on both sides.

Then we define R^2 as:

$$\begin{aligned} R^2 &= \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \\ &= \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{e'e}{\sum_i (y_i - \bar{y})^2} \end{aligned}$$

$R^2 \in [0,1]$

Suppose $R^2 =$ 1- $e'e = 0$

$e_i = 0$ for all i. Then $Y_i = X_i b$. This means there are no residuals, there is a perfect linear fit!!

Suppose $R^2 = 0$. Then $\hat{y}_i = \bar{y}$ for all i- then the line is the constant, no additional info over a constant.

Some notes:

1. R^2 only makes sense when there is a constant in the regression- we used this earlier.

2. R^2 tends to increase with the number of X variables- so adjusted R^2- put a penalty for number of variables.

In general in model selection, dont just want to put in all the X variables- want to penalize fit for more vars.

3. R^2 does not mean the model is better!!

5.2 Variance of the OLS Estimator in Multivariate Case

In the CRM, we have that $V(Y|X) = \sigma^2 I$. Earlier we found that:

$$V(b_1|X) = \frac{\sigma^2}{NS^2_{X_1}}$$

for the bivariate regression model.

Now- consider a single slope parameter from a multivariate regression model.

$$b_{2N} = (X_2^{*'} X_2^*)^{-1} X_2^{*'} Y$$

should be 1 by 1-show this!!

Here again X_2^* is the residual of a regression of X_2 on all other Xs.

Then we can calculate the variance of the estimator b_{2N}:

$$\begin{aligned} V(b_{2N}|X) &= \sigma^2 (X_2^{*'} X_2^*)^{-1} \\ &= \frac{\sigma^2}{\sum_i X_{2i}^{*2}} \end{aligned}$$

I will now use the R^2 from the auxiliary regression to make a nice form for this:

$$\begin{aligned} R_2^2 &= 1 - \frac{X_2^{*'} X_2^*}{\sum_i (X_{2i} - \bar{X}_2)^2} \\ \frac{X_2^{*'} X_2^*}{\sum_i (X_{2i} - \bar{X}_2)^2} &= 1 - R_2^2 \\ X_2^{*'} X_2^* &= (1 - R_2^2) \sum_i (X_{2i} - \bar{X}_2)^2 \\ &= (1 - R_2^2) NS^2_{X_2} \end{aligned}$$

We then have:

$$V(b_{2N}|X) = \frac{\sigma^2}{(1 - R_2^2) NS^2_{X_2}}$$

So- what determines the precision of the estimator?

1. $V(Y|X)$
2. N

3. Variance of covariate of X

4. Collinearity between X_2 and other variables

In a sense, collinearity between X_2 and other variables is stripping out variation in X_2.

If R_2^2 is zero- no linear relationship between X_2 and rest of Xs- just like bivariate regression- can get same estimate by short or long regression- nothing will change (on estimate or its variance).

If perfect lin relationship between X_2 and all other Xs, R_2^2 goes to 1- variance goes to infinity (identification problem)

So multicollinearity problem- increases the variance of estimators- cant precisely estimate the parameter.

Bias-Variance tradeoff- throw out variables to reduce this R_2^2, lower variance- but will bias coefficient.

Now- how do we estimate this variance?

We should observe everything except σ^2.

If $E(U^2|X) = \sigma^2$ or $V(Y|X) = \sigma^2 I$, then

$$\begin{aligned} V &= \sigma^2 E(X'X)^{-1} \\ V_N &= \frac{e'e}{n-k} E_N(X'X)^{-1} = E_N(e_N^2)E_N(X'X)^{-1} \end{aligned}$$

Analogy principle at work- but why n-k? b/c $E(\frac{e'e}{n-k}) = \sigma^2$- get unbiased estimate.

$V_N \rightarrow_p V$ (why?). Can use divide by n and get biased estimate.

The idea here $\frac{1}{n}e'e = \frac{1}{n}\sum_i e_i^2$- since the errors in the sample are mean zero this is the sample variance of the estimator. If the CRM holds, σ^2 is the true population variance of the errors- by the analogy principle use the sample variance as an estimator!