Linear Regression - I

Professor Ji-Woong Chung
Korea University

This lecture note is based on Todd Gormley's.

Outline

The CEF and causality

- The CEF and causality (very brief)
- BLP of the CEF
- Causality and linear regression

Linear OLS model

- Basic interpretation
- Rescaling & shifting of variables
- Incorporating non-linearities

Multivariate estimation

- Properties & Interpretation
- Partial regression interpretation
- R^2 bias and consistency

Motivation

Conditional expectations, $E[Y_i \mid D_i]$, are pivotal in causal analysis.

How can we estimate conditional expectation functions (CEF)?

Linear regression is arguably the most popular modeling approach in empirical research

- Transparent and intuitive
- Very robust technique; easy to build on
- Even if not interested in causality, it is useful for describing the data

Motivation continued

Can be used to answer descriptive, predictive, and causal questions.

Linear regression is easy to compute but very difficult to interpret.

- Linear regression does not estimate the CEF directly!
- Linear regression estimates the *best linear approximation* of the CEF.

Outline

The CEF and causality

- The CEF and causality (very brief)
- BLP of the CEF
- Causality and linear regression

Linear OLS model

- Basic interpretation
- Rescaling & shifting of variables
- Incorporating non-linearities

Multivariate estimation

- Properties & Interpretation
- Partial regression interpretation
- R^2 bias and consistency

A bit about random variables

It is useful to write any random variable Y as

$$Y = E(Y \mid X) + \epsilon$$

where (Y, X, ϵ) are random variables and $E(\epsilon \mid X) = 0$[1]

- $E(Y \mid X)$ is the expected value of Y given X
- In words, Y can be broken down into the part 'explained' by X, $E(Y \mid X)$, and a piece that is mean independent of X, ϵ.[2]

[1] $E(\epsilon \mid X) = E(Y - E(Y \mid X) \mid X) = E(Y \mid X) - E[E(Y \mid X) \mid X] = 0$

[2] ϵ is independent of any functions of X. Let $h(X)$ be any function of X. $E(h(X)\epsilon) = E[E(h(X)\epsilon \mid X)] = E[h(X)E(\epsilon \mid X)] = 0$

Conditional expectation function (CEF)

$E(Y \mid X)$ is what we call the CEF and it has very desirable properties:

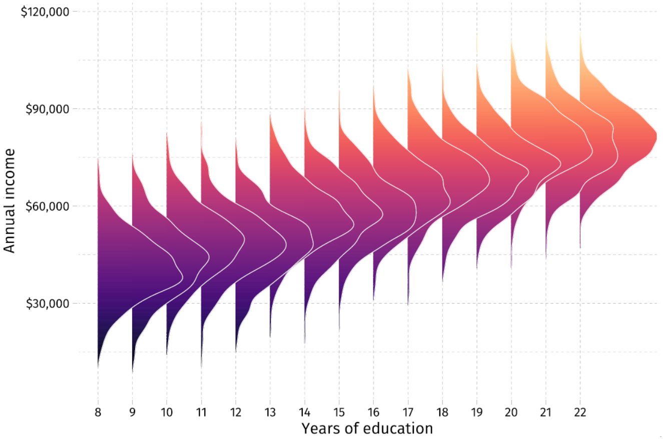
- Natural way to think about the relationship between X and Y
- And it is the best predictor of Y given X in a minimum mean-squared error sense
- I.e., $E(Y \mid X)$ minimizes $E[(Y - m(X))^2]$ where $m(X)$ can be any function of X.[3]

[3] Hint: add and subtract $E(Y \mid X)$.

CEF visually

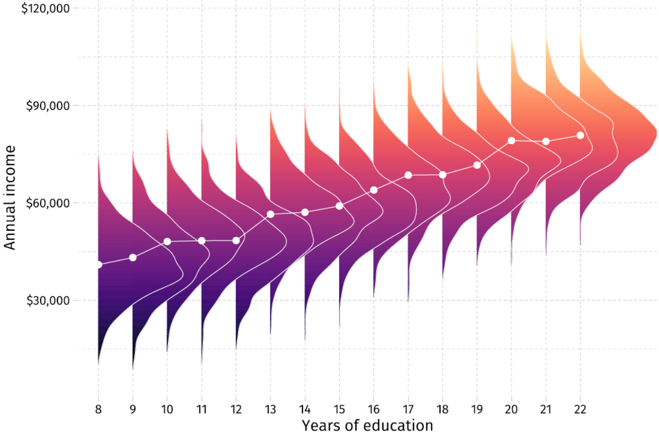
$E(Y \mid X)$ is fixed but unobservable.

- Intuition: For any value of X, the distribution of Y is centered about $E(Y \mid X)$.



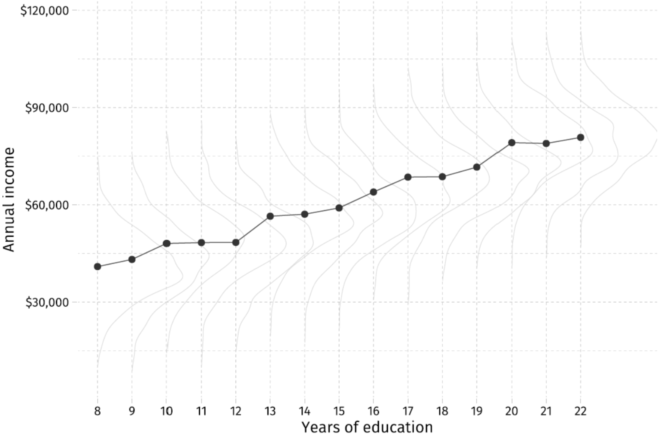
CEF visually

The CEF connects these conditional distributions' means.



CEF visually

Our goal is to learn about the CEF.



Outline

The CEF and causality

- The CEF and causality (very brief)
- BLP of the CEF
- Causality and linear regression

Linear OLS model

- Basic interpretation
- Rescaling & shifting of variables
- Incorporating non-linearities

Multivariate estimation

- Properties & Interpretation
- Partial regression interpretation
- R^2 bias and consistency

Linear regression and the CEF

If done correctly, a linear regression can help us uncover what the CEF is.

- Consider the linear regression model

$$Y = \alpha + \beta X + u$$

  - (Y, X, u) are random variables
  - (Y, X) are observable
  - (u, α, β) are unobservable
  - u captures everything that determines Y after accounting for X — This might be a lot of stuff!
  - We want to estimate β

Best Linear Predictor (BLP)

BLP is α, β that minimize the mean-squared error:

$$argmin_{\alpha,\beta} E[(Y - \alpha - \beta X)^2]^4$$

Using first order condition:

$$E[Y - \alpha - \beta X] = 0 \text{ and } E[X(Y - \alpha - \beta X)] = 0$$

Hence, $\alpha^{BLP} = E(Y) - \beta E(X)$ and $\beta^{BLP} = \frac{E(XY) - E(X)E(Y)}{E(X^2) - E(X)^2} = \frac{Cov(X,Y)}{Var(X)}$.

Note: By definition, the residual from this regression $Y - \alpha^{BLP} - \beta^{BLP} X$ is uncorrelated (not independent) with X.

Note: $u = Y - (\alpha^{BLP} + \beta^{BLP} X) = (Y - E(Y \mid X)) + (E(Y \mid X) - (\alpha^{BLP} + \beta^{BLP} X))$. Prediction error + Linear approximation error.

[4] Equivalent problem: $argmin_{\alpha,\beta} E[(E[Y \mid X] - \alpha - \beta X)^2]$

What's great about this linear regression?

It can be shown that

1. If $E(Y \mid X)$ is linear, $E(Y \mid X) = \alpha^{BLP} + \beta^{BLP} X$.[5]
2. $\alpha^{BLP} + \beta^{BLP} X$ is the best linear prediction of Y given X[6]
3. $\alpha^{BLP} + \beta^{BLP} X$ is the best linear approximation of $E(Y \mid X)$[7]

('best' in terms of minimum mean-squared error)

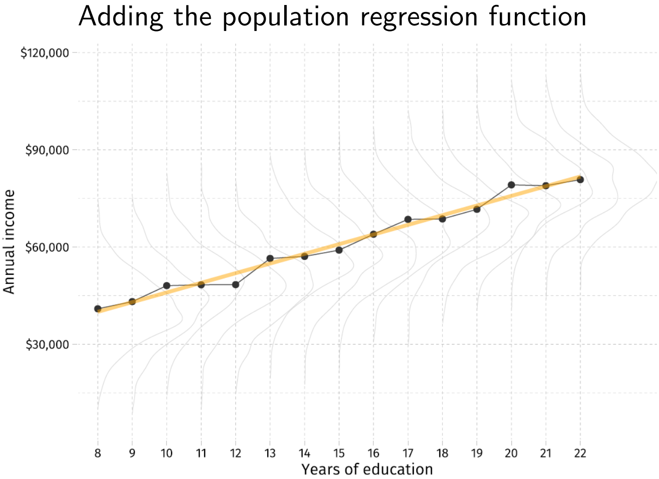
This is quite useful. I.e., even if $E(Y \mid X)$ is nonlinear, the regression gives us the best linear approximation of it.

[5] Use the fact that $E[u] = 0$ and $E[Xu] = 0$

[6] By definition, i.e., $(\alpha, \beta) \in argmin_{\alpha,\beta} E[(Y - \alpha - \beta X)^2]$

[7] I.e., $(\alpha, \beta) \in argmin_{\alpha,\beta} E[(E(Y \mid X) - \alpha - \beta X)^2]$

What's great about this linear regression? (Cont'd)



Outline

The CEF and causality

- The CEF and causality (very brief)
- BLP of the CEF
- Causality and linear regression

Linear OLS model

- Basic interpretation
- Rescaling & shifting of variables
- Incorporating non-linearities

Multivariate estimation

- Properties & Interpretation
- Partial regression interpretation
- R^2 bias and consistency

What about causality?

Need to be careful here

- How X explains Y, which this regression helps us understand is (descriptive), not the same as learning the causal effect of X on Y.
- Captures the *approximate* expected level of Y *associated* with a level of X.[8]
- For that, we need more assumptions

[8] $\beta = \frac{\partial E(Y|X)}{\partial X}$ only if $E(Y \mid X)$ is linear.

The basic assumptions [Part 1]

Assumption 1: $E(u) = 0$

- With intercept, this is totally innocuous.
- Just change the regression to $Y = \alpha + \beta X + u$ where α is the intercept term.
- Any non-zero mean is absorbed by the intercept.

The basic assumptions [Part 2]

Assumption 2: $E(u \mid X) = E(u)$

- In words, the average of u (i.e., the unexplained portion of Y) does not depend on the value of X.
- This is “conditional mean independence” (CMI)
  - True if X and u are independent of each other.
  - Implies u and X are uncorrelated.

This is the key assumption being made when people make causal inferences.

CMI Assumption

Basically, the assumption says you've got the correct CEF model for the causal effect of X on Y.

- CEF is causal if it describes differences in average outcomes for a change in X.
  - I.e., change in Y if X increases from values a to b is equal to $E(Y \mid X = b) - E(Y \mid X = a)$
- This is only true if $E(u \mid X) = E(u)$

Example of why CMI is needed

With model $Y = \alpha + \beta X + u$

- $E(Y \mid X = a) = \alpha + \beta a + E(u \mid X = a)$
- $E(Y \mid X = b) = \alpha + \beta b + E(u \mid X = b)$
- Thus, $E(Y \mid X = b) - E(Y \mid X = a) = \beta(b - a) + E(u \mid X = b) - E(u \mid X = a)$
- This only equals what we think of as the ‘causal’ effect of X changing from a to b if $E(u \mid X = b) = E(u \mid X = a)$ i.e., CMI assumption holds.

Tangent – CMI versus correlation

CMI is needed for no bias — a finite sample property[9]

However, we only need to assume a zero correlation between X and u for consistency — a large sample property[10]

We typically care about consistency which is why we often refer to correlations rather than CMI.

[9] With $y = X\beta + u$, we have $\hat{\beta} = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'u$. Hence $\hat{\beta} = \beta \Leftrightarrow E[u \mid X] = 0$

[10] $\hat{\beta} = \beta + (X'X/n)^{-1}(X'u/n)$, where $\hat{\beta} \xrightarrow{p} \beta \Leftrightarrow E[Xu] = 0$

Is it plausible?

Admittedly there are many reasons why this assumption might be violated.

- Recall u captures all the factors that affect Y other than X. It will contain a lot!
- Let's just do a couple of examples

Exaxmple – Capital structure regression

Consider the following firm-level regression:

$$\text{Leverage}_i = \alpha + \beta \text{Profitability}_i + u_i$$

- CMI implies average u is the same for each profitability.
- Easy to find a few stories why this isn't true
  1. Unprofitable firms tend to have higher bankruptcy risk which by tradeoff theory should mean lower leverage.
  2. Unprofitable firms have accumulated less cash which by pecking order means they should have more leverage.

Is there a way to test for CMI?

- Let $\hat{Y}$ be the predicted value of Y i.e. $\hat{Y} = \hat{\alpha} + \hat{\beta}X$ where $\hat{\alpha}$ and $\hat{\beta}$ are OLS estimates.
- And let $\hat{u} = Y - \hat{Y}$ be the residual.
- Can we prove CMI if residuals are $E(\hat{u}) = 0$ and if $\hat{u}$ is uncorrelated with X?

- **Answer**: No! By construction, these residuals are mean zero and uncorrelated with X.

Three reasons why CMI is violated

- Omitted variable bias
- Measurement error bias
- Simultaneity bias

We will look at each of these in much more detail in the “Causality” lecture.

Outline

The CEF and causality

- The CEF and causality (very brief)
- BLP of the CEF
- Causality and linear regression

Linear OLS model

- Basic interpretation
- Rescaling & shifting of variables
- Incorporating non-linearities

Multivariate estimation

- Properties & Interpretation
- Partial regression interpretation
- R^2 bias and consistency

Outline

The CEF and causality

- The CEF and causality (very brief)
- BLP of the CEF
- Causality and linear regression

Linear OLS model

- Basic interpretation
- Rescaling & shifting of variables
- Incorporating non-linearities

Multivariate estimation

- Properties & Interpretation
- Partial regression interpretation
- R^2 bias and consistency

Ordinary Least Square (OLS)

BLP is an approximation to $E[Y \mid X]$.

The BLP and its coefficients $(\alpha^{BLP}, \beta^{BLP})$ are theoretical concepts.

OLS estimates these coefficients using real data.

$$(\hat{\alpha}_n, \hat{\beta}_n) \in argmin_{\alpha,\beta} \frac{1}{n} \left(Y_i - (\alpha + X_i \beta)\right)^2$$

$$\hat{\alpha}_n = \frac{1}{n} \sum Y_i = \frac{1}{n} \sum X_i \hat{\beta}_n$$

$$\hat{\beta}_n = \frac{\frac{1}{n} \sum X_i Y_i - \left(\frac{1}{n} \sum X_i\right)\left(\frac{1}{n} \sum Y_i\right)}{\frac{1}{n} \sum X_i^2 - \left(\frac{1}{n} \sum X_i\right)^2}$$

Interpreting the estimates

Suppose we estimate the following model of CEO compensation

$$\text{Salary}_i = \alpha + \beta \text{ROE}_i + u_i$$

– Salary for CEO i is in 000s; ROE is a %

If you get $\text{Salary}_i = 963.2 + 18.5\text{ROE}_i + u_i$

- What do these coefficients tell us?
  - 1 percentage point increase in ROE is approximately associated with $18,500 increase in salary.
  - Average salary for CEO with ROE = 0 was equal to $963,200.
- Is CMI likely satisfied? Probably not.

Outline

The CEF and causality

- The CEF and causality (very brief)
- BLP of the CEF
- Causality and linear regression

Linear OLS model

- Basic interpretation
- Rescaling & shifting of variables
- Incorporating non-linearities

Multivariate estimation

- Properties & Interpretation
- Partial regression interpretation
- R^2 bias and consistency

Scaling the dependent variable

What if I change the measurement of salary from 000*s* to \$*s* by multiplying it by 1,000?

- Estimates were $\hat{\alpha} = 963.2$ and $\hat{\beta} = 18.50$
- Now they will be $\hat{\alpha} = 963,200$ and $\hat{\beta} = 18,500$

Scaling Y by an amount c just causes all the estimates to be scaled by the same amount

- $cy = (c\alpha) + (c\beta)X + cu$

Scaling Y continued

Notice the scaling has no effect on the relationship between ROE and salary

- I.e., because Y is expressed in \$*s* now, $\hat{\beta} = 18,500$ means that a one percentage point increase in ROE is still associated with $18,500$ increase in salary.

Scaling the independent variable

What if I instead change the measurement of ROE from percentage to decimal? (i.e., multiply ROE by 1/100)

- Estimates were $\hat{\alpha} = 963.2$ and $\hat{\beta} = 18.50$
- Now they will be $\hat{\alpha} = 963.2$ and $\hat{\beta} = 18,500$

Scaling X by an amount k causes the slope on X to be scaled by $1/k$

$$Y = \alpha + \beta X + u$$

$$Y = \alpha + k\beta \frac{X}{k} + u$$

- New slope $= k\beta$
- Will interpretation of estimates change? Answer: Again no!

Scaling both X and Y

If we scale Y by an amount c and X by amount k then we get

- Intercept scaled by c
- Slope scaled by c/k

$$Y = \alpha + \beta X + u$$

$$cy = c\alpha + kc\beta \frac{X}{k} + cu$$

When is scaling useful?

Practical application of scaling #1:

- No one wants to see a coefficient of 0.000000456 or 1234567890.
  - Just scale the variables for cosmetic purposes!
  - It will affect coefficients & SEs.
  - However, it won't affect *t*-stats or inference.

Practical application of scaling #2

To improve interpretation in terms of estimated magnitudes it's helpful to scale the variables by their sample standard deviations.

- Let σ_x and σ_y be sample standard deviations of X and Y respectively.
- Let c the scalar for Y be equal to $1/\sigma_y$.
- Let k the scalar for X be equal to $1/\sigma_x$.
- I.e., units of X and Y are now standard deviations.

Practical application of scaling #2 continued..

With the prior rescaling, how would we interpret a slope coefficient of 0.25?

- Answer = a 1 s.d. increase in X is associated with a $\frac{1}{4}$ s.d. increase in Y.
- The slope tells us how many standard deviations Y changes on average for a standard deviation change in X.
- Is 0.25 large in magnitude? What about 0.01?

Shifting the variables

Suppose we instead add c to Y and k to X (i.e., we shift Y and X up by c and k respectively)

Only the estimated intercept will change

$$\begin{aligned}
&Y = \alpha + \beta X + u \\
&Y + c = \alpha + c + \beta X + u \\
&Y + c = \alpha + c + \beta(X + k) - \beta k + u \\
&Y + c = (\alpha + c - \beta k) + \beta(X + k) + u
\end{aligned}$$

- New intercept $= \alpha + c - \beta k$
- Slope the same $= \beta$

Practical application of shifting

To improve interpretation, sometimes it is helpful to demean X by its sample mean

- Let μ_x be the sample mean of X; regress Y on $X - \mu_x$
- Intercept now reflects the expected value of Y for $X = \mu_x$

$$
\begin{aligned}
Y &= (\alpha + \beta\mu_x) + \beta(X - \mu_x) + u \\
E(Y \mid X = \mu_x) &= (\alpha + \beta\mu_x)
\end{aligned}
$$

- This will be very useful when we get to diff-in-diffs.

Outline

The CEF and causality

- The CEF and causality (very brief)
- BLP of the CEF
- Causality and linear regression

Linear OLS model

- Basic interpretation
- Rescaling & shifting of variables
- Incorporating non-linearities

Multivariate estimation

- Properties & Interpretation
- Partial regression interpretation
- R^2 bias and consistency

Incorporating nonlinearities [Part 1]

Assuming that the causal CEF is linear may not always be that realistic

- E.g., consider the following regression

$$\text{wage} = \alpha + \beta\text{education} + u$$

- Why might a linear relationship between # of years of education and level of wages be unrealistic? How can we fix it?

Incorporating nonlinearities [Part 2]

Better assumption might be that each year of education leads to a constant proportionate (i.e., percentage) increase in wages

- Approximation of this intuition captured by

$$\ln(\text{wage}) = \alpha + \beta \text{education} + u$$

- I.e., the linear specification is very flexible because it can capture linear relationships between non-linear variables.

Common nonlinear function forms

- Regressing Levels on Logs
- Regressing Logs on Levels
- Regressing Logs on Logs

Let's discuss how to interpret each of these

The usefulness of log

Log variables are useful because $100\Delta\ln(Y) \approx \%\Delta Y$[11]

- Note: When people say “Log” we really mean the natural logarithm “ln”. E.g., if you use the “log” function in Stata it assumes you meant “ln”.

[11] $\ln y_1 \approx \ln y_0 + \frac{1}{y_0}(y_1 - y_0)$. Hence. $\ln y_1 - \ln y_0 = \Delta\ln(Y) = \frac{y_1}{y_0} - 1$, by Taylor expansion.

Interpreting log-level regressions

If you estimate the ln(*wage*) equation, 100β will tell you the %Δ wage for an additional year of education. To see this

$$\begin{aligned}
&\ln(wage) = \alpha + \beta education + u \\
&\Delta \ln(wage) = \beta \Delta education \\
&100 \times \Delta \ln(wage) = (100\beta)\Delta education \\
&\%\Delta wage \approx (100\beta)\Delta education
\end{aligned}$$

Log-level interpretation continued

The proportionate change in Y for a given change in X is assumed constant.

- The change in Y is not assumed to be constant it gets larger as X increases.
- Specifically, $\ln(Y)$ is assumed to be linear in X; but Y is not a linear function of X

$$
\begin{aligned}
ln(Y) &= \alpha + \beta X + u \\
Y &= \exp(\alpha + \beta X + u)
\end{aligned}
$$

Example: interpretation

Suppose you estimated the wage equation (where wages are $/hour$) and got

$$\ln(wage) = 0.584 + 0.083 education$$

What does an additional year of education get you?

- Answer $=$ 8.3% increase in wages.

Interpreting log-log regressions

If you alternatively estimate the following

$$\ln(\mathit{salary}) = 4.822 + 0.257 \ln(\mathit{sales})$$

- β is the elasticity of Y w.r.t. X!
- i.e., β is the percentage change in Y for a percentage change in X.
- Note: regression assumes constant elasticity between Y and X regardless of the level of X.

Example: interpretation of log-log

Suppose you estimated the CEO salary model using logs and got the following:

$$\ln(salary) = 4.822 + 0.257 \ln(sales)$$

What is the interpretation of 0.257?

- Answer = For each 1% increase in sales, salary increases by 0.257%.

Interpreting level-log regressions

If estimating the following

$$Y = \alpha + \beta \ln(X) + u$$

- $\beta/100$ is the change in Y for 1% change X.

$$\begin{aligned}
Y &= \alpha + \beta \ln(X) + u \\
\Delta Y &= \beta \Delta \ln(X) \\
\Delta Y &= (\beta/100)(100 \Delta \ln(X)) \\
\Delta Y &= (\beta/100)(\% \Delta X)
\end{aligned}$$

Example: interpretation of level-log

Suppose you estimated the CEO salary model using logs and got the following, where salary is expressed in \$000s:

$$salary = 4.822 + 1812.5 ln(sales)$$

What is the interpretation of 1812.5?

- Answer = For each 1% increase in sales, salary increases by $\$18,125 (= 1,812.5 \times 1,000 \times \frac{1}{100})$.

Summary of log functional forms

Model	**Dep. Var.**	**Ind. Var.**	**Interpretation of β**
Level-Level	Y	X	$dy = \beta dx$
Log-Level	$\ln(Y)$	X	$\%dy = (100\beta)dx$
Log-Log	$\ln(Y)$	$\ln(X)$	$\%dy = \beta\%dx$
Level-Log	Y	$\ln(X)$	$dy = (\beta/100)\%dx$

Now let's talk about what happens if you change units (i.e., scale) for either Y or X in these regressions

Rescaling logs doesn't matter [Part 1]

What happens to intercept & slope if rescale (i.e., change units) of Y when in log form?

- Answer = Only intercept changes; slope unaffected because it measures proportional change in Y in Log-Level model.

$$\begin{aligned}\ln(Y) &= \alpha + \beta X + u \\ \ln(c) + \ln(Y) &= \ln(c) + \alpha + \beta X + u \\ \ln(cy) &= (\ln(c) + \alpha) + \beta X + u\end{aligned}$$

Rescaling logs doesn't matter [Part 2]

Same logic applies to changing the scale of X in level-log models only intercept changes.

$$\begin{aligned} Y &= \alpha + \beta \ln(X) + u \\ Y + \beta \ln(c) &= \alpha + \beta(\ln(X) + \ln(c)) + u \\ Y &= (\alpha - \beta \ln(c)) + \beta \ln(cx) + u \end{aligned}$$

Basic message – If you rescale a logged variable, it will not affect the slope coefficient because you are only looking at proportionate changes.

Log approximation problems

A paper argues that allowing capital inflows into the country caused -120% change in stock prices during crisis periods

- Do you see a problem with this?
- A 120% drop in stock prices isn't possible. The true percentage change was -70%. Here is where that author went wrong

Log approximation problems [Part 1]

Approximation error: as the true $\%\Delta Y$ becomes larger $100\Delta \ln(Y) \approx \%\Delta Y$ becomes a worse approximation.

- To see this consider a change from Y to Y'
- Ex. #1: $Y = 100$ and $Y' = 105$ (5%) and $100\Delta \ln(Y) = 4.9\%$
- Ex. #2: $Y = 100$ and $Y' = 175$ (75%) but $100\Delta \ln(Y) = 56\%$

Log approximation problems [Part 2]

Problem also occurs for negative changes

- Ex. #1: $Y = 100$ and $Y' = 95$ (-5%) and $100\Delta\ln(Y) = -5.1\%$
- Ex. #2: $Y = 100$ and $Y' = 25$ (-75%) but $100\Delta\ln(Y) = -139\%$

Log approximation problems [Part 3]

So if the implied percent change is large, it is better to convert it to the true % change before interpreting the estimate.

$$\begin{aligned}
\ln(Y) &= \alpha + \beta X + u \\
\ln(Y') - \ln(Y) &= \beta(X' - X) \\
\ln(Y'/Y) &= \beta(X' - X) \\
Y'/Y &= \exp(\beta(X' - X)) \\
\%\Delta Y &= 100[\exp(\beta(X' - X)) - 1]
\end{aligned}$$

Log approximation problems [Part 4]

We can now use this formula to see what the true % change in Y is for X'–$X = 1$

$$\begin{aligned}\%\Delta Y &= 100[\exp(\beta(X' - X)) - 1]\\ \%\Delta Y &= 100[\exp(\beta) - 1]\end{aligned}$$

- If $\beta = 0.56$ the percent change isn't 56% it is

$$100[\exp(0.56) - 1] = 75\%$$

Recap of last two points on logs

Two things to keep in mind about using logs

- Rescaling a logged variable doesn't affect slope coefficients; it will only affect intercept.
- Log is only an approximation for % change; it can be a very bad approximation for large changes.

Usefulness of logs – Summary

Using logs gives coefficients with appealing interpretation

Can be ignorant about the unit of measurement of log variables since they're proportionate Δs.

Logs of Y or X can mitigate the influence of outliers.

“Rules of thumb” on when to use logs

Helpful to take logs for variables with

- Positive currency amount
- Large integral values (e.g., population)

Don’t take logs for variables measured in years or for variables that can equal zero

What about using $\ln(1 + Y)$?

Because $\ln(0)$ doesn't exist, some use $\ln(1 + Y)$ for non-negative variables i.e., $Y \geq 0$.

- However, you should not do this!
- Nice interpretation no longer true, especially if a lot of zeros or many small values in Y. [Why?]
  - Ex. #1: What does it mean to go from $\ln(0)$ to $\ln(X > 0)$?
  - Ex. #2: And $\ln(X' + 1)$–$\ln(X + 1)$ is not percent change of X

See Cohn, Liu, Wardlaw (JFE 2022) for solutions & more details on why using ln(1+Y) is problematic. Use Poisson regression (with fixed effects).

Tangent – Percentage Change

What is the percent change in unemployment if it goes from 10% to 9%?

- This is a 10 percent drop.
- It is a 1-percentage point drop.
  - Percentage change is $[(X_1 - x_0)/x_0] \times 100$
  - Percentage point change is the raw change in percentages.

Please take care to get this right in the description of your empirical results.

Models with quadratic terms [Part 1]

Consider $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + u$.
Partial effect of X is given by

$$\frac{\partial Y}{\partial X} = \beta_1 + 2\beta_2 X$$

What is different about this partial effect relative to everything we've seen thus far?

- Answer = It depends on the value of X. So we will need to pick a value of X to evaluate it (e.g., $\bar{X}$).

Models with quadratic terms [Part 2]

If $\beta_1 > 0, \beta_2 < 0$ then it has a parabolic relation

- Turning point = Maximum = $|\beta_1 / 2\beta_2|$
- Know where this turning point is! Don't claim a parabolic relation if it lies outside the range of X!
- Odd values might imply misspecification or simply mean the quadratic terms are irrelevant and should be excluded from the regression.

Outline

The CEF and causality

- The CEF and causality (very brief)
- BLP of the CEF
- Causality and linear regression

Linear OLS model

- Basic interpretation
- Rescaling & shifting of variables
- Incorporating non-linearities

Multivariate estimation

- Properties & Interpretation
- Partial regression interpretation
- R^2 bias and consistency

Outline

The CEF and causality

- The CEF and causality (very brief)
- BLP of the CEF
- Causality and linear regression

Linear OLS model

- Basic interpretation
- Rescaling & shifting of variables
- Incorporating non-linearities

Multivariate estimation

- Properties & Interpretation
- Partial regression interpretation
- R^2 bias and consistency

Basic multivariable model

Example with constant and k regressors

$$Y = \alpha + \beta_1 X_1 + \ldots + \beta_k X_k + u$$

- Similar identifying assumptions as before
  - No collinearity among covariates [why?]
  - $E(u \mid X_1, \ldots, X_k) = 0$
  - Implies no correlation between any X and u which means we have the correct model of the true causal relationship between Y and $(X_1, \ldots, X_k)$.

Interpretation of estimates

Estimated intercept $\hat{\alpha}$ is the predicted value of Y when all $X = 0$; sometimes this makes sense, sometimes it doesn't.

Estimated slopes $\hat{\beta}_j$ have a more subtle interpretation now

$$Y = \hat{\alpha} + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k + \hat{u}$$

How would you interpret $\hat{\beta}_1$?

Interpretation – Answer

Estimated slopes $\hat{\beta}_j$ have partial effect interpretations

- Typically, we think about a change in just one variable e.g., ΔX_1 holding constant all other variables i.e., $(\Delta X_2, \ldots, \Delta X_k$ all equal 0).
    - This is given by $\Delta Y = \hat{\beta}_1 \Delta X_1$.
    - I.e., $\hat{\beta}_1$ is the coefficient holding all else fixed (ceteris paribus).

Interpretation continued

However, we can also look at how changes in multiple variables at once affect the predicted value of Y.

- I.e., given changes in X_1 through X_k we obtain the predicted change in Y, ΔY.

$$\Delta \hat{Y} = \hat{\beta}_1 \Delta X_1 + \ldots + \hat{\beta}_k \Delta X_k$$

Example interpretation – College GPA

Suppose we regress college GPA onto high school GPA (4-point scale) and ACT scores for $N = 141$ university students.

$$colGPA = 1.29 + 0.453 hsGPA + 0.0094 ACT$$

- What does the intercept tell us?
- What does the slope on *hsGPA* tell us?

Example – Answers

Intercept meaningless person with zero high school GPA and ACT doesn't exist

Example interpretation of slope

- Consider two students Ann and Bob with identical ACT scores but Ann's GPA is 1 point higher than Bob's. Best prediction of Ann's college GPA is that it will be 0.453 higher than Bob's.

Example continued

Now what is the effect of increasing high school GPA by 1 point and ACT by 1 point?

$$\begin{aligned}
\Delta colGPA &= 0.453\Delta hsGPA + 0.0094\Delta ACT \\
\Delta colGPA &= 0.453 + 0.0094 \\
\Delta colGPA &= 0.4624
\end{aligned}$$

Example continued

Lastly, what is the effect of increasing high school GPA by 2 points and ACT by 10 points?

$$\begin{aligned} \Delta colGPA &= 0.453\Delta hsGPA + 0.0094\Delta ACT \\ \Delta colGPA &= 0.453 \times 2 + 0.0094 \times 10 \\ \Delta colGPA &= 1 \end{aligned}$$

Fitted values and residuals

Definition of residual for observation i, $\hat{u}_i$:

$$\hat{u}_i = Y_i - \hat{Y}_i$$

- Properties of residual and fitted values:
  - Sample average of residuals $= 0$; implies that sample average of $\hat{Y}$ equals the sample average of Y.
  - Sample covariance between each independent variable and residuals $= 0$.
  - Point of means $(\bar{X}_1, \bar{X}_2, \ldots, \bar{Y})$ lies on the regression line.

Tangent about residuals

Again it bears repeating

- Looking at whether the residuals are correlated with the X's is NOT a test for causality.
- By construction, they are uncorrelated with X.
- There is no "test" of whether the CEF is the causal CEF; that justification will need to rely on economic arguments.

Outline

The CEF and causality

- The CEF and causality (very brief)
- BLP of the CEF
- Causality and linear regression

Linear OLS model

- Basic interpretation
- Rescaling & shifting of variables
- Incorporating non-linearities

Multivariate estimation

- Properties & Interpretation
- Partial regression interpretation
- R^2 bias and consistency

Question to motivate the topic

What is wrong with the following? And why?

- Researcher wants to know the effect of X on Y after controlling for z.
- So researcher removes the variation in Y that is driven by z by regressing Y on z & saves residuals.
- Then the researcher regresses these residuals on X and claims to have identified the effect of X on Y controlling for z using this regression.
- We'll answer why it's wrong in a second

Partial regression [Part 1]

The following is quite useful to know

- Suppose you want to estimate the following:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + u$$

- Is there another way to get $\hat{\beta}_1$ that doesn't involve estimating this directly?
- Answer: Yes! You can estimate it by regressing the residuals from a regression of Y on X_2 onto the residuals from a regression of X_1 onto X_2.

Partial regression [Part 2]

To be clear you get $\hat{\beta}_1$ by[12]

- $\#1$ – Regress Y on X_2; save residuals (call them $\tilde{Y}$).
- $\#2$ – Regress X_1 on X_2; save residuals (call them $\tilde{X}_1$).
- $\#3$ – Regress $\tilde{Y}$ onto $\tilde{X}_1$; the estimated coefficient will be the same as if you'd just run the original multivariate regression!!!

[12] Called the Frisch–Waugh–Lovell theorem

Partial regression – Interpretation

Multivariate estimation is basically finding the effect of each independent variable after partialing out the effects of other variables.

- I.e., the effect of X_1 on Y after controlling for X_2 (i.e., what you'd get from regressing Y on both X_1 and X_2) is the same as what you get after you partial out the effect of X_2 from both X_1 and Y and then run a regression using the residuals.

Partial regression – Generalized

This property holds more generally

- Suppose X_1 is a vector of independent variables.
- X_2 is a vector of more independent variables.
- And you want to know the coefficients on X_1 that you would get from a multivariate regression of Y onto all the variables in X_1 and X_2

Partial regression – Generalized Part 2

You can get the coefficients for each variable in X_1 by

- Regress Y and each variable in X_1 onto all the variables in X_2 (at once); save residuals from each regression.
- Do a regression of residuals; i.e., regress Y onto variables of X_1 but replace Y and X_1 with the residuals from the corresponding regression in step #1.

Practical application of partial regression

Now what is wrong with the following?

- Researcher wants to know the effect of X on Y after controlling for z.
- So the researcher removes the variation in Y that is driven by z by regressing Y on z & saves residuals.
- Then the researcher regresses these residuals on X and claims to have identified the effect of X on Y controlling for z using this regression.

Practical application – Answer

It's wrong because it didn't partial the effect of z out of X!

Therefore it is NOT the same as regressing Y onto both X and z!

Unfortunately, it was commonly done by researchers in finance [e.g., industry-adjusting].

- We will see how badly this can mess up things in a later lecture.

Outline

The CEF and causality

- The CEF and causality (very brief)
- BLP of the CEF
- Causality and linear regression

Linear OLS model

- Basic interpretation
- Rescaling & shifting of variables
- Incorporating non-linearities

Multivariate estimation

- Properties & Interpretation
- Partial regression interpretation
- R^2 bias and consistency

Goodness-of-Fit (R^2)

A lot is made of R^2; so let's quickly review exactly what it is

- Start by defining the following:
  - Sum of squares total (SST)
  - Sum of squares explained (SSE)
  - Sum of squares residual (SSR)

Definition of SST, SSE, SSR

If N is the number of observations and the regression has a constant then

$SST = \sum_{i=1}^{N}(Y_i - \bar{Y})^2$	SST is total variation in Y
$SSE = \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2$	SSE is total variation in predicted Y [mean of predicted Y=mean of Y]
$SSR = \sum_{i=1}^{N} \hat{u}_i^2$	SSR is total variation in residuals [mean of residual=0]

SSR, SST, and SSE continued

The total variation SST can be broken into two pieces the explained part SSE and unexplained part SSR.

$$SST = SSE + SSR$$

R^2 is just the share of total variation that is explained! In other words

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

More about R^2

As seen on the last slide R^2 must be between 0 and 1

It can also be shown that R^2 is equal to the square of the correlation between Y and predicted Y.

If you add an independent variable, R^2 will never go down.

Adjusted R^2

Because R^2 always goes up we often use what is called Adjusted R^2

$$\bar{R}^2 = 1 - (1 - R^2)\left(\frac{N-1}{N-k-1}\right)$$

- $k = \#$ of regressors excluding the constant.
- Basically, you get penalized for each additional regressor such that adjusted R^2 won't go up after you add another variable if it doesn't improve fit much [it can go down!].

Interpreting R^2

If I tell you the R^2 is 0.014 from a regression what does that mean? Is it bad?

- Answer $\#1 =$ It means I'm only explaining about 1.4% of the variation in Y with the regressors that I'm including in the regression.
- Answer $\#2 =$ Not necessarily! It doesn't mean the model is wrong; you might still be getting a consistent estimate of the β you care about!

Unbiasedness versus Consistency

When we say an estimate is unbiased or consistent, it means we think it has a causal interpretation

- I.e., the CMI assumption holds and the X's are all uncorrelated with the disturbance u.

Bias refers to a finite sample property; **consistency** refers to an asymptotic property.

More formally

An estimate $\hat{\beta}$ is unbiased if $E(\hat{\beta}) = \beta$

- I.e., on average the estimate is centered around the true unobserved value of β.
- Doesn't say whether you get a more precise estimate as sample size increases.

An estimate is consistent if $plim_{N \rightarrow \infty} \hat{\beta} = \beta$[13]

- I.e., as sample size increases the estimate converges (in probability limit) to the true coefficient.

[13] A sequence X_n of random variables converges in probability towards the random variable X if for all $\epsilon > 0$, $\lim_{n \rightarrow \infty} \Pr(|X_n - X| > \epsilon) = 0$

Unbiasedness of OLS

OLS will be unbiased when

- Model is linear in parameters.
- We have a random sample of X.
- No perfect collinearity between X's.
- $E(u \mid X_1, \ldots, X_k) = 0$: Earlier CMI assumptions #1 and #2 give us this.

Unbiasedness is a nice feature of OLS; but in practice, we care more about consistency.

Consistency of OLS

OLS will be consistent when

- Model is linear in parameters.
- u is not correlated with any of the X's: CMI assumptions #1 and #2 give us this; a lack of correlation is a weaker assumption than CMI. CMI precludes both linear and non-linear relationships while correlations only measure linear relationships.

Summary of Today [Part 1]

The CEF $E(Y \mid X)$ has desirable properties

- Linear OLS gives the best linear approximation of it.
- If the correlation between error u and independent variables X's is zero, it has a causal interpretation.

Scaling & shifting of variables doesn't affect inference but can be useful.

- E.g., demean to give intercepts a more meaningful interpretation or rescale for cosmetic purposes.

Summary of Today [Part 2]

Multivariate estimates are partial effects

- I.e., the effect of X_1 holding $X_2, \ldots, X_k$ constant.
- Can get the same estimates in two steps by first partialing out some variables and regressing residuals on residuals in the second step.