# Causation and the problem of choice

Professor Dan Black **PP 414:Applied Regression Analysis: Analysis of Microeconomics Data**

2018

# Overview of lecture

- **A very brief introduction to the course**
- The Roy model
- Why do we care about causation?
- How do we determine causation?
- Why is this so hard? Or, the problem of choice
- The Roy model again
- Experiments
- Lee's (inadvertent, but fundamental) critique of experiments
- "External validity" and the role of theory

# The course

- Goal: To make you a better analyst of data
- Goal: To make you a better better consumer of data analysis
- Goal: To teach you a few new techniques
- Instrument: Homework assignments with rotating groups of co-workers
- Instrument: Midterm and final, but not so useful
- Instrument: Readings
- Instrument: Discussions
- Instrument: Thought

# Overview of lecture

- *A very brief introduction to the course*
- **The Roy model**
- Why do we care about causation?
- How do we determine causation?
- Why is this so hard? Or, the problem of choice
- The Roy model again
- Experiments
- Lee's critique of experiments
- "External validity" and the role of theory

# The Roy model

- If you have taken a course from me, you have seen this example: Accountants and Economists
- Results of simulation (items in bold would be observed in the real world):

| | Accountants | Economists | Total |
|---|---|---|---|
| **Accounting Earnings** | **63,985** | | |
| **Economics Earnings** | | **72,317** | |
| | | | |
| **N** | **78,414** | **21,586** | **100,000** |

- What can you learn about the impact of becoming an economist from this data? **Nothing**

# The Roy model

- The naive among you might want to argue that we "know" the answer. It is just
$$\hat{\Delta}_N = \bar{y}_e - \bar{y}_a = 8,332$$
- But this estimates makes a bunch of "implicit assumptions" that makes use Krugman's "Accidental Theorist". We want to avoid being such assumptions
- More fundamentally, however, why do we care about causation?
- How do we make causal judgments?
- How does microdata help?

# Overview of lecture

- *A very brief introduction to the course*
- *The Roy model*
- **Why do we care about causation?**
- How do we determine causation?
- Why is this so hard? Or, the problem of choice
- The Roy model again
- Experiments
- Lee's critique of experiments
- "External validity" and the role of theory

# Why do we care about causation?

- Bertrand Russel (1913) argues: "The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm." Thus, Russel's answer to this question is: You should not. I am not fond of argument that rests on authority, but Russel was a great thinker so we should take his views seriously
- It may be that finding causal relationships is intrinsically interesting so needs no further justification
- That's a fine answer for an academic discipline, but this is a professional school
- We must believe that causal inference allows us to determine better policy. We must believe that learning about the causal impact of a policy in the past helps us understand what the impact in the future
- That is, we believe our analysis has some "external validity"

# Why do we care about causation?

- But social science should not be a religion: We should use the scientific method to justify this belief
- Key to the use of the scientific method is to have a discernible theory that disciplines our wild beliefs about the external validity of our results to other times and other populations
- This does not mean that our theories are perfectly correct, but theories guide and discipline our inference to either different times or different populations
- This is necessarily part "science" and part "art", but it will provide us with the guidance you need to make better inference
- Without the theory, we are making leaps of faith when we make projections

# Why do we care about causation?

- This makes use the data *necessarily* a theoretical exercise. This is a different view than one gets much from reading much of the current literature
- If one produces an estimate that has "internal validity" but no "external validity," who cares?
- The failure to recognize this relationship makes the empirical scientist an "accidental theorist" because we want our results to generalize
- You should be very vigilant against this accidental theorizing. It can rely on some positively goofy (implicit) assumptions

# Overview of lecture

- *A very brief introduction to the course*
- *The Roy model model*
- *Why do we care about causation?*
- **How do we determine causation?**
- Why is this so hard? Or, the problem of choice
- The Roy model again
- Experiments
- Lee's critique of experiments
- "External validity" and the role of theory

# How do we determine causality?

- Recall our Roy model example and naive estimator. Why is this not a causal estimate? A formal model of causality will indicate why. So let use develop one.
- In this case, let us call becoming an economist as the "treated" state. Let $D_i \in \{0, 1\}$ index the *ith* individual's treatment status.
- Everyone has two possible outcomes at a point in time: If treated, they received $Y_{1,i}$ and if untreated they receive $Y_{0,i}$
- This framework now allows us to define the impact of treatment for the *ith* person as $\Delta_i = Y_{1,i} - Y_{0,i}$. We now have a precise way to think about causality
- Estimating individual effects may be demanding so we might estimate the average effect or $E(\Delta_i) \equiv \Delta^{ATE} \equiv E(Y_{1,i}) - E(Y_{0,i})$. We refer to $E(\Delta^{ATE})$ as the "estimand"
- Of course, at best we get to observed either $Y_{1,i}$ or $Y_{0,i}$ but never both: We have a missing counterfactual

# How do we determine causality?

Results of simulation (items in bold would be observed in the real world):

| | **Accountants** | **Economists** | **Total** |
|---|---|---|---|
| **Accounting Earnings** | **63,985** | | |
| **Economics Earnings** | | **72,317** | |
| | | | |
| **N** | **78,414** | **21,586** | **100,000** |

$$\hat{\Delta}_N = \bar{y}_e - \bar{y}_a = 8,332$$

# How do we determine causality?

**A formal model of causality**

- In this case, let us call becoming an economist as the "treated" state. Let $D_i \in \{0,1\}$ index the *ith* individual's treatment status
- Everyone has two possible outcomes at a point in time: If treated, they received $Y_{1,i}$ and if untreated they receive $Y_{0,i}$
- This framework now allows us to define the impact of treatment for the *ith* person as $\Delta_i = Y_{1,i} - Y_{0,i}$. We now have a precise way to think about causality
- Estimating individual effects may be demanding so we might estimate the average effect or $E(\Delta_i) \equiv \Delta^{ATE} \equiv E(Y_{1,i}) - E(Y_{1,i})$. We refer to $E(\Delta^{ATE})$ as the "estimand"
- Of course, at best we get to observed either $Y_{1,i}$ or $Y_{0,i}$ but never both: We have a missing counterfactual
- How do deal with the missing counterfactual? We need to estimate it!

# How do we determine causality?

- In the case of our naive estimate of the treatment effect, we estimated the average estimate of those who chose economics to forecast the earnings of economists. This is, of course, and excellent estimate for those who chose to be economist. What about accountants?
- For accounting earnings, we estimated the average earnings of those who chose accounting to forecast the earnings of accounting. What about economists?
- For economics, the mean earnings of economists are

$$\mu_e = Pr(D_i = 1)E(Y_{1,i}|D_i = 1) + Pr(D_i = 0)E(Y_{1,i}|D_i = 0)$$

- For accounting, the mean earnings of accountants are

$$\mu_a = Pr(D_i = 0)E(Y_{0,i}|D_i = 0) + Pr(D_i = 1)E(Y_{0,i}|D_i = 1)$$

- For both $\mu_e$ and $\mu_a$, the second term of the right-hand side of the equation is missing

# How do we determine causality?

- Now, we could assume that $E(Y_{1,i}|D_i = 1) = E(Y_{1,i}|D_i = 0)$ and $E(Y_{0,i}|D_i = 0) = E(Y_{0,i}|D_i = 1)$. What would justify this assumption?
- Here is one model: Students flip a coin to determine whether to become an economist or an accountants
- This model should probably makes you uncomfortable. Do people flip coins when making important, life-changing decisions?
- Perhaps we could ask a simpler question. What if we wanted to know the impact of becoming an economists for those who become economists, or $\Delta^{ATT}$, the impact of treatment on the treated

# How do we determine causality?

- Formally, the estimand $\Delta^{ATT}$ is defined as

$$\Delta^{ATT} \equiv E(Y_{1,i}|D_i = 1) - E(Y_{0,i}|D_i = 1)$$

- The good news is that we have an excellent estimator $E(Y_{1,i}|D_i = 1)$: The observed mean of economic earnings, $\bar{y}_e$ is great
- But what about $E(Y_{0,i}|D_i = 1)$? How do we estimate it?
- Well, may be it would be easier to ask: What would be the return of accountants to becoming economists? This estimand is just

$$\Delta^{ATN} = E(Y_{1,i}|D_i = 0) - E(Y_{0,i}|D_i = 0)$$

- Again, the good news is that we have an excellent estimator $E(Y_{0,i}|D_i = 0)$: The observed mean of accounting earnings, $\bar{y}_a$ is great
- But what about $E(Y_{1,i}|D_i = 0)$? How do we estimate it?

# How do we determine causality?

- We now have four possible parameters to estimate: $\Delta_i$, $\Delta^{ATE}$, $\Delta^{ATT}$ and $\Delta^{ATN}$
- There is no reason why these four parameters should be equal. Indeed, we would be very surprised if your return to becoming an economists is the same as mine
- The last three are related by this identity

$$\Delta^{ATE} = Pr(D_i = 1)\Delta^{ATT} + Pr(D_i = 0)\Delta^{ATN}$$

- When $\Delta_i = \Delta^{ATE} = \Delta^{ATT} = \Delta^{ATN}$, we way that we have a "common effect", which is probably not particularly plausible
- We now have four parameters ($\Delta_i$, $\Delta^{ATE}$, $\Delta^{ATT}$ and $\Delta^{ATN}$), but each is afflicted with this "missing counterfactual"
- Why is this so hard?

# Overview of lecture

- *A very brief introduction to the course*
- *The Roy model*
- *Why do we care about causation?*
- *How do we determine causation?*
- **Why is this so hard? Or, the problem of choice**
- The Roy model again
- Experiments
- Lee's critique of experiments
- "External validity" and the role of theory

# Why is this so hard? Or, the problem of choice

- The problem we are facing agent's who make choices. They make those choices having different information than we (the analysts) have. Krugman provides a nice analysis:
- "An Indian born economist once explained his personal theory of reincarnation to his graduate economics class. 'If you are a good economist, a virtuous economist,' he said, 'you are reborn as a physicist. But if you are an evil, wicked economist, you are reborn as a sociologist.' A sociologist might say that this quote shows what is wrong with economists: they want a subject that is fundamentally about human beings to have the mathematical certainty of the hard sciences . . . . But good economists know that the speaker was talking about something else entirely: the sheer difficulty of the subject. Economics is harder than physics; luckily it is not quite as hard as sociology." Paul Krugman Peddling Prosperity
- This "hidden variables" problem gives rise to what we call the "selection problem"

# Why is this so hard? Or, the problem of choice

- To see how, let us write out a formal model, I will call the "canonical model". Here are the equations:

$$Y_{1,i} = g_1(X_i) + \epsilon_{1,i}$$

$$Y_{0,i} = g_0(X_i) + \epsilon_{0,i}$$

$$D_i^* = h(X_i, Z_i) + \epsilon_{D,i}$$

- We term $X_i$ the "observed variables" (or observables), $\epsilon_{j,i}$ the "unobserved variables" (or observables), and $Z_i$ are very special variables we will call "instruments". Our use of unobserved and observed variables is relative to the analyst. This model makes some relatively strong assumptions, but we can live with them for now
- $D_i^*$ is a latent variable we do not observe, but $D_i = 1$ if $D_i^* \geq 0$
- We are now able to discuss the problem of choice

# Why is this so hard? Or, the problem of choice

$$Y_{1,i} = g_1(X_i) + \epsilon_{1,i}$$
$$Y_{0,i} = g_0(X_i) + \epsilon_{0,i}$$
$$D_i^* = h(X_i, Z_i) + \epsilon_{D,i}$$

- Agents use their information contained in the variable $\epsilon_{D,i}$ to determine whether or not they choose treatment, but the analyst does not observe $\epsilon_{D,i}$
- The fear is that $E\left(\epsilon_{1,i}|\epsilon_{D,i} \geq -h(X_i, Z_i)\right) \neq E\left(\epsilon_{1,i}|\epsilon_{D,i} < -h(X_i, Z_i)\right)$ or that $E\left(\epsilon_{0,i}|\epsilon_{D,i} \geq -h(X_i, Z_i)\right) \neq E\left(\epsilon_{0,i}|\epsilon_{D,i} < -h(X_i, Z_i)\right)$
- Thus, the nonindependence of $(\epsilon_{1,i}, \epsilon_{0,i}, \epsilon_{D,i})$ is the statistical reason we find this problem so difficult
- From a theoretical perspective, we are worried that we do not observe important determinants of the outcomes that will reside in $(\epsilon_{1,i}, \epsilon_{0,i}, \epsilon_{D,i})$. How does relate to Roy model?

# Overview of lecture

- *A very brief introduction to the course*
- *The Roy model*
- *Why do we care about causation?*
- *How do we determine causation?*
- *Why is this so hard? Or, the problem of choice*
- **The Roy model again**
- Experiments
- Lee's critique of experiments
- "External validity" and the role of theory

# The Roy model again

Recall, results of simulation (items in bold would be observed in the real world):

| | **Accountants** | **Economists** | **Total** |
|---|---|---|---|
| **Accounting Earnings** | **63,985** | | |
| **Economics Earnings** | | **72,317** | |
| | | | |
| **N** | **78,414** | **21,586** | **100,000** |

# The Roy model again

- Let us fill in the details of the model that I used to simulate the model. Accounting earnings are a normal random variable with mean, or average, of $65,000 and standard deviation of $5,000
- Economic earnings are a normal random variable with mean, or average, of $60,000 and standard deviation of $10,000
- The correlation coefficient between the two is 0.86
- People simply select the field that they are best at, or $Y_i = max(Y_{E,i}, Y_{A,i})$
- We could solve this model analytically, but computers make our lives way more comfortable so I simulate this with $n = 100,00$. Here are the results:

# The Roy model again

Results of simulation (items in bold would be observed in the real world):

| | **Accountants** | **Economists** | **Total** |
|---|---|---|---|
| **Accounting Earnings** | **63,985** | 68,690 | 65,001 |
| **Economics Earnings** | 56,599 | **72,317** | 59,992 |
| | | | |
| **N** | **78,414** | **21,586** | **100,000** |

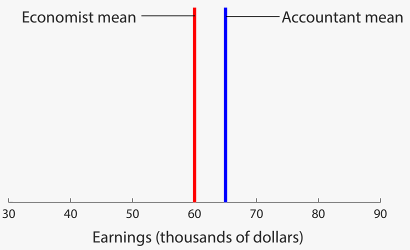
# The Roy model

- This is an extremely simple model, but is incredibly interesting
- Recall our naive $\hat{\Delta}^{N} = \$72,317 - \$63,985 = \$8,332$. Taking this estimate seriously we would recommend that accountants become economists, which is very, very wrong!
- Actually, for accountants becoming an economist would cost them **$7,386**, or $\hat{\Delta}^{ATN} = -\$7,386 = 56,599 - 63,985$
- For an economist, becoming an economist made them **$3,627**, or $\hat{\Delta}^{ATT} = \$3,627 = 72,317 - 68,690$
- For the average person, becoming an economist costs them **$5,009**, or $\hat{\Delta}^{ATE} = -5,009 = 59,992 - 65,001$
- Notice that economists are better accountants than accountants (on the average $68,690 compared to $63,985) and accountants are, well, below average accounts (on the average $63,985 compared to the mean of $65,000)

# The Roy model

- How is this possible?
- Accounting earnings are a normal random variable with mean, or average, of $65,000 and standard deviation of $5,000
- Economic earnings are a normal random variable with mean, or average, of $60,000 and standard deviation of $10,000

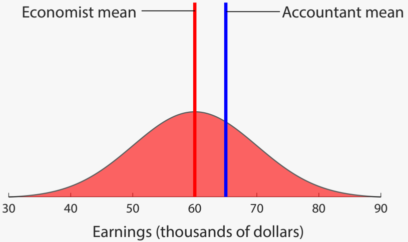
## The Roy model



# The Roy model

- How is this possible?
- Accounting earnings are a normal random variable with mean, or average, of \$65,000 and standard deviation of \$5,000
- Economic earnings are a normal random variable with mean, or average, of \$60,000 and standard deviation of \$10,000

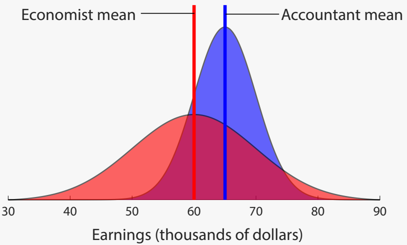
## The Roy model



# The Roy model

- How is this possible?
- Accounting earnings are a normal random variable with mean, or average, of $65,000 and standard deviation of $5,000
- Economic earnings are a normal random variable with mean, or average, of $60,000 and standard deviation of $10,000

## The Roy model



# The Roy model

- This should **terrify** you. It keeps me up at night
- Economics has the lowest average wage but the highest observed wage. Our naive policy proposal (force accountant into economics) would be a disaster!
- This example illustrates why one needs model to analyze the world: Simple models can avoid disastrous decisions. This is why the faculty of Harris have you take so much theory: You need the ability to formulate models to interpret data
- This example also puts to rest the fallacy of "letting the data speak". Data doesn't talk, it babels
- Beware of the simple and obvious: It isn't

# The Roy model

$$Y_{1,i} = g_1(X_i) + \epsilon_{1,i}$$
$$Y_{0,i} = g_0(X_i) + \epsilon_{0,i}$$
$$D_i^* = h(X_i, Z_i) + \epsilon_{D,i}$$

- Recalling our canonical model, much is missing in our example: We have no observables ($X_i$) or instruments ($Z_i$)
- If confronted with just this data, intellectual modesty would require us to say we know little about the true nature of the earnings differences between accountants and economists
- (Actually, if we knew how earnings were distributed we could solve the problem, but that never happens)
- How can we estimate a causal impact?
- Let us start with experiments

# Overview of lecture

- *A very brief introduction to the course*
- *The Roy model*
- *Why do we care about causation?*
- *How do we determine causation?*
- *Why is this so hard? Or, the problem of choice*
- *The Roy model again*
- **Experiments**
- Lee's critique of experiments
- "External validity" and the role of theory

# Experiments, idealized

$$Y_{1,i} = g_1(X_i, U_i)$$
$$Y_{0,i} = g_0(X_i, U_i)$$
$$D_i = R_i$$

- We now move to a model that Heckman calls the "all causes" model. $U_i$ is a vector of "unobserved variables, where before $\epsilon_{1,i} = k_1(U_i)$ and $\epsilon_{0,i} = k_0(U_i)$. This represents a generalization of our canonical model because the canonical model is a special case of the "all causes" model
- The important change is that $D_i = R_i$, where $R_i = 1$ when the agent is assigned to the treatment group and $R_i = 0$ when the agent is assigned to the control group
- The assignment is made by the analyst who uses a random number to determine whether $R_i = 1$ or $R_i = 0$

# Experiments, idealized

$$Y_{1,i} = g_1(X_i, U_i)$$
$$Y_{0,i} = g_0(X_i, U_i)$$
$$D_i = R_i$$

- I call this an idealized experiment because everyone honors the experimental protocol as $D_i = R_i$
- Why do experiments work? Because treatment status is randomly assigned for all $(X^0, U^0)$

$$E\Big(F(X^0, U^0|D_i = 1)\Big) = E\Big(F(X^0, U^0|D_i = 0)\Big) = E\Big(F(X^0, U^0\Big)$$

- Thus, $E(Y_{1,i}|D_i = 1) = E(Y_{1,i}|D_i = 0)$ and $E(Y_{0,i}|D_i = 1) = E(Y_{0,i}|D_i = 0)$ so we have solved our missing counterfactual problem!

# Experiments, idealized

$$Y_{1,i} = g_1(X_i, U_i)$$
$$Y_{0,i} = g_0(X_i, U_i)$$
$$D_i = R_i$$

- We may estimate $\Delta^{ATE}$ by simply comparing the means or
$$\hat{\Delta}^{ATE} = \bar{y}_1 - \bar{y}_0$$
- We would actually use the regression of the form:
$$y_i = x_i\beta + \delta R_i + e_i$$
which reduces the variance of our estimates. NB: $x_i$ must be set before random assignment
- The $E\left(\hat{\Delta}^{ATE}\right) = E\left(\hat{\delta}\right)$, but the two will differ because of random fluctuations in $x_i$. Both $\hat{\Delta}^{ATE}$ and $\hat{\delta}$ are "estimators" of the "estimand" $\Delta^{ATE}$

# Experiments, idealized

- Thus, if we subject the populations (or a random sample of the population) to the experiment we can recover an estimates $\Delta^{ATE}$
- Can we estimate $\Delta^{ATT}$, $\Delta^{ATN}$, or $\Delta_i$? Sadly, no. But we can change the design of the experiment to elicit estimates of $\Delta^{ATT}$ and $\Delta^{ATN}$
- Estimation of these parameters are equally easy to estimate
- You often hear that experiments "solve" the selection problem, but this is wrong. Experiments avoids the selection problem by replacing it with random assignment
- But are experiments ideal? Sadly, no

# Experiments, not idealized

- People sure can screw up your experiment. People in the treatment group often don't take the treatment and people in the control group somehow get treatment
- In many medical experiments this threat is avoided because the new treatments are often experimental drugs
- In social science experiments, this is much more of a problem. For instance, about 34.4% of the women in the control group in the Job Training Partnership Act Experiment received classroom training
- Similarly, only about 64.3% of the women in treatment group in the Job Training Partnership Act Experiment received classroom training
- What havoc does this wreak?

# Experiments, not idealized

- Well, we just need to define different parameters! To that end, define the impact of the intent to treat as:

  $$\Delta^{ITT} = E(Y|R_i = 1) - E(Y|R_i = 0) = E(Y|R_i = 1) - E(Y|R_i = 0)$$

  where $Y_i = D_i Y_{1,i} + (1 - D_i) Y_{0,i}$

- This parameter essentially tells you the change in the mean value of $Y$ resulting from the assignment to the treatment group
- This is of course is often the relevant parameter for policy
- This parameter, however, is not related to any of the previous parameters $(\Delta_i, \Delta^{ATE}, \Delta^{ATT}, \Delta^{ATN})$
- But what if you are actually interested in the impact of treatment per se?

# Experiments, not idealized

- Again, we simply need to define a new parameter (and add some assumptions). This estimator will be an instrumental variables estimator that we will see again in lecture 5
- The added assumption is called "Montonicity" or "Uniformity". It states:

$$D_i(R_i = 1) \geq D_i(R_i = 0)\ \forall i$$

  That is, being assigned to the treatment group does not cause people to defy the experimental protocol
- With this assumption, we can construct the "Bloom estimand":

$$\Delta^B = \frac{E(Y|R_i = 1) - E(Y|R_i = 0)}{E(D_i|R_i = 1) - E(D_i|R_i = 0)}$$

- This can be estimated with

$$\hat{\Delta}^B = \frac{\bar{y}_{R_i=1} - \bar{y}_{R_i=0}}{\bar{D}_{R_i=1} - \bar{D}_{R_i=0}}$$

# Experiments, not idealized

- But what does this estimator estimate?
- With the monotonicity assumption, we can divide people into three mutually exclusive groups: (1) people who always take the treatment regardless of assignments who occur with probability $p_a$, people who never take the treatment regardless of assignment who occur with probability $p_n$, and people who comply with the experimental protocol who occur with probability $p_c$, with $p_c + p_n + p_a = 1$
- For those assigned to treatment, we have
$$E(Y|R_i = 1) = p_n E(Y_{0,i}|n) + p_c E(Y_{1,i}|c) + p_a E(Y_{1,i}|a)$$
- For those assigned to the control group we have
$$E(Y|R_i = 0) = p_n E(Y_{0,i}|n) + p_c E(Y_{0,i}|c) + p_a E(Y_{1,i}|a)$$
- Subtracting these two equation, we get
$$E(Y|R_i = 1) - E(Y|R_i = 0) = p_c E(Y_{1,i} - Y_{0,i}|c)$$

# Experiments, not idealized

- For the denominator, we have two terms. The first is:

$$E(D_i|R_i = 1) = p_a + p_c$$

- The second term is just

$$E(D_i|R_i = 0) = p_a$$

- Thus, subtracting the two equation yields

$$E(D_i|R_i = 1) - E(D_i|R_i = 0) = p_c$$

so that

$$\Delta^B = E(Y_{1,i} - Y_{1,i}|c)$$

- Thus, the Bloom estimand identifies the impact of treatment for those who abide in the experimental protocol
- This is referred to as **L**ocal **A**verage **T**reatment **E**ffect (LATE)

# Experiments, not idealized

- LATE's play an important role in modern causal analysis thanks to the work of Imbens and Angrist (1994)
- The LATE is for those who comply with the experimental protocol, but the data do not identify who these folks are
- The LATE is not related to any of the previous parameters $(\Delta_i, \Delta^{ATE}, \Delta^{ATT}, \Delta^{ATN})$
- As such, the LATE is a bit unsatisfactory. But to quote Guido Imbens, "Better LATE than nothing"
- There are a host of other issues associated with experiments, and I will leave you to Heckman and Smith (1995) to learn about many of these
- I will consider a critique raised in Lee (2009)

# Overview of lecture

- *A very brief introduction to the course*
- *The Roy model*
- *Why do we care about causation?*
- *How do we determine causation?*
- *Why is this so hard? Or, the problem of choice*
- *The Roy model again*
- *Experiments*
- **Lee's critique of experiments**
- "External validity" and the role of theory

# Lee's critique of experiments

- Lee (2009) provides a devastating critique for those who believe that experiments eliminate the need for applied researchers to use theory
- Lee examines the impact of the US Job Corp. The Job Corp program gave the average enrollee about 400 hours of academic instructions and about 700 hours of vocational training. The goal was to raise the wages of participants
- In 1994-95, Mathematica conducted the first experimental evaluation of the program, 20 years after it's inception
- From November of 1994 to December of 1995, 15,386 selected to participate in the experiment, with 5,997 individuals assigned to the control group and 9,409 assigned to the treatment group
- In the experimental estimates, we find that 48 months after random assignment employment in increased about 4.1%
- Lee's goal is to see if the program succeeded in raising wage. That is, did the program succeed at it s goal?

# Lee's critique of experiments

- This turns out to be extremely challenging. The problem with wage is that we do not observe them unless the individual is employed
- But at 48 months, only 0.607 proportion of the treatment group and 0.566 proportion of the control group is employed. We have, therefore, a missing counterfactual: Would have been the wage of those person who are not employed at the month 48?
- We have a program that seeks to raise the wages of participants, but the data from the experiment, at least without a healthy dose of theory, cannot answer this question
- Lee conducts his analysis and finds it likely that the program had a modest positive impact on wages

# Overview of lecture

- *A very brief introduction to the course*
- *The Roy model*
- *Why do we care about causation?*
- *How do we determine causation?*
- *Why is this so hard? Or, the problem of choice*
- *The Roy model again*
- *Experiments*
- *Lee's critique of experiments*
- **"External validity" and the role of theory**

# "External validity" and the role of theory

- Even when we are fortunate enough to have our causal parameter identified and our data from the experiment allow us to estimate that parameter, our work is not done
- Almost always we wish to know whether this estimate is applicable to a different time or a different population
- This inherently involves a theoretical exercise to evaluate whether this extrapolation is likely to be successful
- Often, this exercise is left implicit. That is generally a bad idea. Formal consideration does better than "accidental theorizing"
- To quote Derek Neal (1995) "You can run from economic models, but you can't hide from them."

# Overview of lecture

- *A very brief introduction to the course*
- *The Roy model*
- *Why do we care about causation?*
- *How do we determine causation?*
- *Why is this so hard? Or, the problem of choice*
- *The Roy model again*
- *Experiments*
- *Lee's critique of experiments*
- *"External validity" and the role of theory*