Supplementary Notes 15.034

Jaume Vives-i-Bastida (MIT)

Overview:

1. What is econometric thinking?
   - Identification of causal effects.
   - The independence assumption.
   - Omitted variable bias.
2. Testing a hypothesis.
   - P-values and confidence intervals.
   - Size and power of tests.
   - Multiple hypothesis testing.
3. Improving your standard errors with ML.
   - Identification vs. estimation.
   - ML saves the day - LASSO.
   - Dealing with ML bias.
   - ML for model selection.
4. Natural experiments, DD and RDD.
   - DD.
   - RDD.

1. What is econometric thinking?

Only two things exist in the world (of econometrics): *Data* and *Models.* Data are the variables we collect and observe, models are the assumptions we make about how the data we observe was generated. Econometrics is about making precise what we can say about the world given our data and our model assumptions.

Suppose we want to understand the return of an MBA education. A more precise way of thinking about it would be asking: *if I decide to do an MBA how should I expect my career earnings to change?* It may seem that a model is not necessary to answer this question as we could simply compare earnings of people with MBAs with the earnings of people without MBAs.

Why do we need models? The answer is that we need models to *interpret* what we find from the data. Imagine that we compare earnings for a sample of individuals and we find that people with MBAs make on average \$50,000 more a year than people without MBAs. Does this mean that if I decided to do an MBA tomorrow I will increase my earnings by \$50,000 a year? Well, that depends. There are many different *stories* that could rationalize this difference. It could be that the MBA increases each person's earnings by fifty thousand dollars a year, but it could also be that the people that decide to do the MBA choose professions that have a higher pay, or that the people that are accepted into the MBAs are higher ability and therefore would have higher earnings regardless of the MBA. Without a model to structure our thinking, we can't disentangle which is the true mechanism that lead to the difference in earnings and therefore we can't answer our initial question about the return of an MBA education.
The simplest econometric model we can come up with is

$$E_i = \alpha + \beta \text{MBA}_i + \epsilon_i. \tag{1}$$

This model states that for an individual i the career earnings E_i depend on a constant α, a constant β if they did an MBA, and, importantly, some unobserved factor ϵ_i that is specific to the individual. In model (1), E_i, MBA_i and ϵ_i are random variables, meaning that they have a probability distribution over the possible values they could take. We see the sample of individuals we observe as a random sample of a larger fixed population. The model allows us to think about the parameters α and β in terms of the population and give meaning to them. For example, it allows us to answer what is the change in expected earnings?

$$\text{Expected earnings for MBAs}: \quad \mathbb{E}[E_i|MBA_i = 1] = \alpha + \beta + \mathbb{E}[\epsilon_i|MBA_i = 1], \tag{2}$$

$$\text{Expected earnings for non MBAs}: \quad \mathbb{E}[E_i|MBA_i = 0] = \alpha + \mathbb{E}[\epsilon_i|MBA_i = 0]. \tag{3}$$

From (2) and (3) we can give meaning to the parameters!

$$\begin{aligned}\beta &= \underbrace{\mathbb{E}[E_i|MBA_i = 1] - \mathbb{E}[E_i|MBA_i = 0]}_{ATE} + \underbrace{\mathbb{E}[\epsilon_i|MBA_i = 1] - \mathbb{E}[\epsilon_i|MBA_i = 0]}_{Selection},\\ \alpha &= \mathbb{E}[E_i|MBA_i = 0] - \mathbb{E}[\epsilon_i|MBA_i = 0].\end{aligned}$$

The β parameter is the *Average Treatment Effect*, the expected change in earnings between having an MBA and not having an MBA (what we wanted to answer!), plus a *Selection* term that depends on the unobserved. The selection term is the difference in unobserved characteristics between the people that have MBAs and the people that do not. Without further assumptions this means we can *not* disentangle the treatment effect from the selection effect by estimating β. As econometricians, our job is to think of what assumptions on the unobserved ϵ_i and its relation to the observed MBA_i may be reasonable and may allows us to estimate the ATE.

The independence assumption The simplest assumption that one can make is that getting an MBA is completely unrelated to the unobserved characteristics. Formally, this means that MBA_i and ϵ_i are independent. This has an immediate consequence for our interpretation of β

$$\begin{gathered}\text{MBA}_i \perp \epsilon_i \implies \mathbb{E}[\epsilon_i|MBA_i] = \mathbb{E}[\epsilon_i],\\ \beta = \underbrace{\mathbb{E}[E_i|MBA_i = 1] - \mathbb{E}[E_i|MBA_i = 0]}_{ATE} + \underbrace{\mathbb{E}[\epsilon_i] - \mathbb{E}[\epsilon_i]}_{=0}.\end{gathered}$$

Under the independence assumption the selection effect goes away! Our β parameter now has the interpretation of being the Average Treatment Effect of the MBA. This assumption would be valid if we could do an experiment and randomly assign doing an MBA to individuals with the same expected unobserved characteristics. Unfortunately, this is not a very reasonable assumption in practice as getting an MBA requires being accepted to an MBA which requires ability, and ability also affects earnings. We call ability a confounder, an omitted variable that confounds the true effect by affecting both the outcome (earnings) and the treatment (the MBA).

The omitted variable bias The presence of a confounder means that the selection effect may not be zero and so we can't interpret the β as the ATE. However, the model allows us to think about this bias in a structured way. Instead of assuming independence we now assume that the unobserved characteristics have a particular structure

$$\epsilon_i = \gamma A_i + \eta_i,$$

where A_i is the ability of individual i and η_i is the part of the unobserved characteristic that does not depend on ability that we assume is independent of ability and doing the MBA, that is $\eta_i \perp A_i, \text{MBA}_i$. Under this new assumption the β coefficient takes a new meaning

$$\begin{aligned}\beta &= \mathbb{E}[E_i|MBA_i = 1] - \mathbb{E}[E_i|MBA_i = 0] + \mathbb{E}[\gamma A_i + \eta_i|MBA_i = 1] - \mathbb{E}[\gamma A_i + \eta_i|MBA_i = 0],\\ &= \text{ATE} + \underbrace{\gamma}_{\text{Effect of ability on } E_i} \times \underbrace{(\mathbb{E}[A_i|MBA_i = 1] - \mathbb{E}[A_i|MBA_i = 0])}_{\text{Selection of ability into MBA}}\end{aligned}$$

Under the omitted variable model the β is the average treatment effect plus the product of the effect of ability on earnings and the expected ability difference between MBA and non MBA individuals. Observe that the difference in ability expectations has a familiar form, it is the same as the parameter implied by a model of ability on having done an MBA (β_a under the independence assumption for the model $A_i = \alpha_a + \beta_a MBA_i + u_i$). With this new interpretation of β, we can easily think of the potential bias of ignoring ability and interpreting β as the ATE. To do this note that

1. Ability on average should lead to higher earnings (so $\gamma > 0$).
2. Higher ability people are more likely to want to do an MBA and to be accepted into an MBA (so $\beta_a > 0$).

Therefore, we now know that in our model $\beta = \text{ATE} + \gamma \times \beta_a > \text{ATE}$, that is there is a positive bias generated by the ability confounder and the coefficient is actually larger than the true average treatment effect. That means that using just our earnings and MBA data without accounting for ability will lead us to think that the return of the MBA is larger than what it is in reality.

To summarize, econometric models allow us to precisely identify what we can say about a particular question given the data we have under different assumptions. In the next section we show how to quantify the degree of certainty with which our model answers a particular question given our data.

2. Testing a hypothesis

Now we know how to carefully think of a question through the lens of an econometric model. The next step is to bring the model to the data an *estimate* its parameters. With the estimated parameters we can finally quantify the return of the MBA. To do so we need data. Suppose, for simplicity, that we were able to do an experiment an randomly assign doing an MBA to $N/2$ people and have $N/2$ similar people as a control group. As discussed above under our simple econometric model this means that the independence assumption is likely to be true so we can interpret β as the ATE.

Our estimate of β will be the difference in the sample averages[1]

$$\hat{\beta} = \frac{1}{N/2} \sum_{i \text{ if MBA}} E_i - \frac{1}{N/2} \sum_{i \text{ if not MBA}} E_i,$$

we denote it with a hat because it is an *estimator*, a function of the sample data. Under the independence assumption $\epsilon_i \perp \text{MBA}_i$ $\hat{\beta}$ is an *unbiased* estimator of β, and therefore, of the ATE. Our estimator depends on the random sample of individuals we drew, unbiased means that if we drew many samples over and over, on average our estimator would be equal to β (that is $\mathbb{E}[\hat{\beta}] = \beta$).

Not only are we interested in the mean of the estimator, but also its distribution, as it allows us to quantify the uncertainty of the answer we get for the return of an MBA education. The easiest way to see what the distribution of our estimator is, is to think of our simple model and assume the unobserved term follows a particular distribution. In general the most reasonable assumption is that $\epsilon_i \sim_{i.i.d} N(0, \sigma^2)$. That is the unobserved characteristics are drawn from a normal distribution that is centered at zero. Under this assumption, independence and our simple model observe that

$$\begin{aligned}
\hat{\beta} &= \frac{1}{N/2} \sum_{i \text{ if MBA}} E_i - \frac{1}{N/2} \sum_{i \text{ if not MBA}} E_i \\
&= \frac{1}{N/2} \sum_{i \text{ if MBA}} \alpha + \beta + \epsilon_i - \frac{1}{N/2} \sum_{i \text{ if not MBA}} (\alpha + \epsilon_i) \\
&= \beta + \frac{1}{N/2} \sum_{i \text{ if MBA}} \epsilon_i - \frac{1}{N/2} \sum_{i \text{ if not MBA}} \epsilon_i \\
&\sim N(\beta, \sigma^2/N).
\end{aligned}$$

[1] As discussed in lectures, this estimator is equivalent to running the linear regression of E_i on the MBA indicator and a constant.

So our estimator is a random variable that follows a normal distribution centered at the true ATE β and with variance σ^2/N. This allows us to think how likely it is that $\hat{\beta}$ would be very different if we had drawn a different sample! It allows us to assess the uncertainty around our estimate of the returns of the MBA, if that uncertainty is to large then we might conclude we can't say anything about the actual return.

To make precise the quantification of the uncertainty we formulate a hypothesis and asses whether our data and model are consistent with the hypothesis or not. The most common hypothesis is that the effect of the MBA is zero:

$$H_0 : \beta = 0 \quad vs. \quad H_1 : \beta \neq 0,$$

where H_0 is called the null hypothesis and states that in the true model the MBA has no effect on earnings. We compare this hypothesis with the alternative H_1 that the MBA has some effect (positive or negative) on earnings. Given our model *and* the data we can check how likely it is that H_0 is not true. To test this we use a *test statistic*, a quantity (an observed quantity from the data) that tells us how far away our estimated $\hat{\beta}$ is from the true model (in this case $\beta = 0$). To come up with the test statistic we use the distribution of $\hat{\beta}$ imposing the restriction that $\beta = 0$:

$$\hat{\beta} \sim N(0, \sigma^2/N),$$
$$\hat{t} = \frac{\sqrt{N}\hat{\beta}}{\sigma} \sim N(0,1).$$

The $\hat{t}$ is our test statistic and it has a standard normal distribution. Our hypothesis and model say that if we sampled many times and the true β was 0 we should get a normal distribution centered at zero. If the value of $\hat{\beta}$ that we observe is very far away from zero then it is likely that the true model with $\beta = 0$ did not generate the data. So what determines how far away from zero the t-statistic needs to be so that we believe the $\beta = 0$ model did not generate the data? It is a *convention*! We choose a *significance level*: the percent of the time we are willing to conclude that the hypothesis was wrong when it was actually right. The standard significance level is $\alpha = 5\%$, which means that we want to find the values of z for which:

$$P(|\hat{t}| > z) \leq 5\%.$$

where recall that $\hat{t}$ is a random variable with a standard normal distribution. If our observed value of $\hat{t}$ falls within these values of z then we conclude that the hypothesis is likely false as less than $\alpha\%$ of the time we expect the true model to be $\beta = 0$. We call the z at which $P(|\hat{t}| > z) = \alpha\%$ the *critical value*.

Figure 1: Distribution of $\hat{t}$.

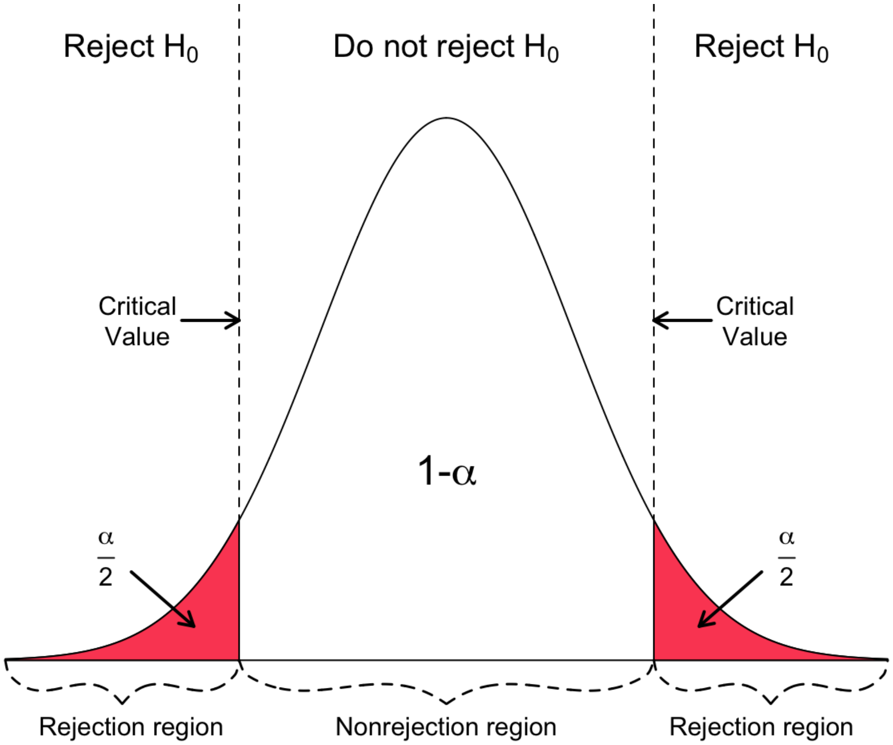

Figure 1 shows the potential values of z for which we would choose to reject the null hypothesis at the α significance level (the rejection region). If our observed t-statistic is in this region we conclude with confidence that the true model is not $\beta = 0$, if on the other hand it is in the central region then we can't say that it is likely that the true model is not $\beta = 0$. In the case of $\alpha = 5\%$ and the t-statistic is normally distributed the critical values of z are -1.96 and 1.96 as 95% of the mass of the standard normal distribution falls between these values. Therefore, if with our MBA earnings data we got a t-statistics of 3 for example, we would conclude that because $|3| > 1.96$ the ATE of an MBA education on earnings is different from zero!

p-values and confidence intervals Besides the critical value, there are two other equivalent ways of thinking about when to reject the null hypothesis. The first one, is to think of the confidence interval around your observed $\hat{\beta}$, that is the numbers that you could have gotten if you had sampled differently. As a convention we care about the 95% confidence interval, that is the we are willing to have values outside the confidence interval 5% of the time. Recall that the critical value for a 5% significance level and a normal distribution is 1.96, so

$$\begin{aligned} P(|\hat{t}| > 1.96) &= 5\% \\ P\left(\left|\frac{\sqrt{N}\hat{\beta}}{\sigma}\right| > 1.96\right) &= 5\% \\ P(\hat{\beta} - 1.96 \times \sigma/\sqrt{N} < \beta < \hat{\beta} + 1.96 \times \sigma/\sqrt{N}) &= 95\%, \end{aligned}$$

so the true value of $\beta \in [\hat{\beta} - 1.96 \times \sigma/\sqrt{N} < \beta, \hat{\beta} + 1.96 \times \sigma/\sqrt{N}]$, which we call the *95% confidence interval*. If zero is in this interval then we know we can't reject that $\beta = 0$ at the 5% significance level from the data.

The third way of deciding whether to reject the null hypothesis is to look at the $p-value$. The p-value is simply the probability that the true value of β is more extreme than the $\hat{\beta}$ we have observed. Formally, the p-value is the probability that a randomly sampled value from the distribution of $|\hat{t}|$ is greater than observed t-statistic t:

$$pvalue = P(|\hat{t}| > |t|).$$

If the p-value is 0.05 or smaller it means that our observed t-statistic falls within the rejection regions defined by the critical value in Figure 1. Therefore, we can reject that the true model includes $\beta = 0$ with confidence. Observe, that the p-value depends on your draw of the data (through the observed t-statistic). By definition of the significance level α the p-value is less than α exactly α percent of the time. That is $P(pvalue < \alpha) = \alpha$.

2.1. Size and power of a test

In the previous section we showed how to use a hypothesis test to assess how likely it is that a model with zero effect ($\beta = 0$) could have generated the data. We chose the test's *significance* level such that 5% of the time we are willing to be wrong about claiming that β is not zero when in reality $\beta = 0$. Another word for this type of mistake is *Type I* error or the *size* of the test. In fact, we can classify all the different cases we could get when testing a particular hypothesis. For simplicity, consider testing:

$$H_0 : \beta = 0 \quad vs. \quad H_1 : \beta = x,$$

that is that the return of the MBA is zero versus that it increases earnings by $x.

It is useful to recall what the intuition is behind testing.

- **Thought experiment**: We recognize that our particular value of $\hat{\beta}$ could have been different if we had drawn a different sample of the population. The question is, if we had drawn many samples and used a test to determine whether we think the true $\beta = 0$, how often are we willing to be wrong (that is claim $\beta \neq 0$ when $\beta = 0$)? We call this probability of being wrong the **significance** level, the **size** of the test or the **type I** error.

While we are often just interested in rejecting the null hypothesis. This is not the only type of mistake we could incur when using a test.

Table 1: Types of errors

	H_0 True	H_1 True
Reject H_0	type I / size	correct! (power)
Not Reject H_0	correct!	type II

Table 1 tells us everything that could happen in a test. If H_0 is true, we call the probability that we reject H_0 the size of the test. If H_1 is true we call the probability that we correctly reject H_0 the power of the test. Hence, the *type II* error is 1-power of a test and is the probability that we don't reject H_0 when the alternative is actually true. The key thing to remember about power is that it depends on the alternative hypothesis H_1!

- **Power intuition**: Thinking about the power of a test is useful because it allows us to think about what effects we can detect with our test. Suppose we consider $\$x = 10000$, given our assumptions we can pin down the probability of a type II error. For example, we could find that if we sampled over and over only 40% of the time would we reject H_0 when $\beta = 10000$, meaning that we don't have much chance of detecting effects of size 10000 or smaller.

While the size of a test is a *convention*, we choose it so that it is 5%, the power will depend on two key attributes of our setting:

1. How much data we have: with more data the standard error is smaller so it is more likely we reject H_0 when H_1 is true.

2. The alternative effect size: we choose $H_1 : \beta = x$, an x very far away from zero will make it less likely that we reject H_0 when H_1 was true.

To see these two points observe that our t-statistic also has a distribution under H_1, which is the same as under H_0 but shifted by the value of x. Under $H_1 : \beta = x$:

$$\hat{t} = \frac{\sqrt{N}(\hat{\beta}_1 - x)}{\sigma} \sim N(0, 1),$$
$$\hat{\beta}_1 \sim N(x, \sigma^2/N)$$

The power of a test is then

$$P(\text{reject } H_0 \mid H_1 \text{ is true}) = P(|\hat{\beta} - x| > z_{cr}\frac{\sigma}{\sqrt{N}}),$$

where z_{cr} is the critical value for which we reject H_0. Intuitively, that is the probability that the $\hat{\beta}$ we would observe if H_1 were true falls in the rejection region under H_0. Then increasing N, or lowering σ, has the effect of making the sampling distributions narrower so they overlap less and the power increases. Similarly, increasing x has the effect of moving the sampling distribution under H_1 further away so they also overlap less increasing power. In the formula above for the power, we can see that increasing N or lowering σ decreases the term $z_{cr}\frac{\sigma}{\sqrt{N}}$, hence making the probability of being greater (i.e. rejecting the null), larger.

2.2. Multiple Hypothesis testing

Now we know how to do one test! But in practice many times we are interested in figuring out the effect of many different policies. For example, in our MBA example if we might be interested in figuring out the return of each individual MBA course to confirm that Econometrics for Managers is the best. To do so we would need to do a test for the effect of each course separately. Suppose we choose each test such that the significance level is α.

Table 2: Types of errors but now with m tests

	H_0 True	H_1 True
Reject H_0	False Positives	True Positives
Not Reject H_0	True Negatives	False Negatives

Table 2 is also called the confusion matrix in ML classification settings. Now the question is how to be formal about our testing procedure in the same way we were with one test. In general, we are most worried about False Positives. The reason is that false positives are what we were worried about for one test, ie claiming the MBA has an effect when it does not. The question is can we still claim that our false positive rate if we sampled again and again will be α? The answer is no.

- **Multiple hypothesis test intuition**: while it is still true that for each test if we sampled over and over again only $\alpha\%$ of the time we would have a false positive, if we think of instead of the thought experiment of adding new tests, the probability that at least one of them is a false positive grows with the number of tests!

To see this, suppose we have M tests, so M null hypothesis $H_0^m : \beta^m = 0$, where β^m are the different true effects for each course m. We first focus on the probability that at least 1 test is a rejection of the null when all the nulls are true. We call this quantity the family wise error rate (FWER):

$$\begin{aligned} \text{FWER} &= P(\text{At least 1 test rejects when all } H_0^m \text{ are true }) \\ &= P(\cup_m \{H_0^m \text{ is rejected }\}|H_0^1, \ldots, H_0^M \text{ True}) \\ &= \sum_m P(H_0^m \text{ is rejected } |H_0^m \text{ True }) - \Pi \\ &= \sum_m \alpha_m M - \Pi. \end{aligned}$$

where Π is a term including the joint probability of rejecting any combination of tests (as in $P(A \cup B) = P(A) + P(B) - P(A \cap B)$), and the last equality follows from the fact that we set the significance level of each test to α_m, so $P(H_0^m \text{ is rejected } |H_0^m \text{ True }) = \alpha_m$. It follows that if we want to "control" the FWER, that is claim that $FWER \geq \alpha$ for some level α, we can't choose the significance level of each test to be $\alpha_m = \alpha$ as before. In that case,

$$FWER = \alpha M - \Pi.$$

Unfortunately, in general the Π term is likely small (for example when the tests are independent, P(A,B) = P(A)P(B), it involves products of significance levels which are small). So unless we can bound the αM term we have no hope of "controlling" the FWER. The most straightforward way of controlling the FWER is called the **Bonferroni** correction and it involves setting all $\alpha_m = \alpha/M$ such that

$$FWER \leq \sum_m \alpha_m M = \alpha M/M = \alpha.$$

While the Bonferroni correction allows you to say the probability that we reject one test by chance is at most α, it comes at the cost of increasing the precision required to reject every test by a factor of M. This means in practice we will probably discover fewer effects as our threshold for what is an effect different from zero has dropped a lot. For example, in the case of the MBA courses, implementing the bonferroni correction would likely lead to the conclusion that we can't say whether any MBA course has a non-zero effect on income.

- **How to avoid not being able to say anything?** Only two possibilities: (1) is to use knowledge of the correlation between tests (courses) to compute Π, for example if we knew all analytics courses should have similarly sized effects, we can use that information to make our α_m less strict. (2) is to focus on a less strict requirement than the FWER, the usual alternative in practice is to focus on the False Discover Rate (FDR).

The FDR is defined as the expected fraction of tests that are falsely rejected, that is:

$$FDR = \mathbb{E}[\text{\# false discoveries}/\text{\# discoveries}]$$

The expected share of false discoveries amongst discoveries is a less strict quantity than the FWER. The thought experiment is saying: we are okay with at most α% of our discoveries to be due to chance, versus saying that we are okay with the probability of finding **one** false discovery being at most α in the FWER case. As with the FWER, we would like to claim that $FDR \leq \alpha$.

When deciding how many tests to reject and which ones two things can affect the FDR. (1) if you reject more then it is more likely you make false discoveries (numerator), but you will also make more discoveries (denominator). (2) visceversa, if you reject less tests then you make less mistakes and less discoveries. The **Benjamini-Hochberg** procedure is a procedure that finds an optimal point in this trade off between making more discoveries and more false discoveries.

The reason the threshold is chosen in this way is the following. BH says the likelihood a test is a true discovery depends on the p-value rank (ie the test with smallest p value is the least likely to be a false discovery, the second smallest p value is the second least likely etc until the last test M which is the most likely to be a false discovery if we decided to reject). So as we include more tests to reject along the p value rank with each test it is more likely that we will have a false discovery, but also with each rejection we make a discovery! The assumption in BH is that this increase in prob of false discovery is linear in the rank. Then the question is where do we stop to optimize the trade off between false discovery and discoveries?

The number of discoveries if we stop at rank k will be k (as we choose to reject up to k). What will be the number of false discoveries? Well it will be bounded above by the significance level of the kth test (lets call it α_k) times the number of true null hypothesis (lets call it M_0). So we have that the $FDR \leq \alpha_k M_0/k$. Now BH tells you to choose $\alpha_k = \alpha k/M$. The reason is that if we substitute in:

$$FDR <= \alpha_k M_0/k = \alpha k M_0/(Mk) = \alpha M_0/M <= \alpha$$

So the BH successfully gets a alpha control on the FDR!

Note that this is not the only possible way of bounding the FDR, but it is one that is very used in practice bc it is simple to compute and, despite looking confusing, is simple to interpret.

3. Improving your standard errors with "ML"

Economists and statisticians often divide the steps to answering a question with data in two:

1. **Identification**: given our model of the world, identification is the meaning we assign to the parameters we are interested in if we had the whole population. For example, in the MBA example the β parameter identifies the ATE under the independence assumption. We say a model is not *identified* if given our assumptions and the data population the parameter of interest could be rationalized by different models, if that were the case we would have no hope of interpreting it!

2. **Estimation**: given our data and our model, estimation is the process of assigning a value to the parameter of interest. We do so using *estimators*, such as the $\hat{\beta}$ in linear regression. In general, we want estimators that have good properties, such as no bias (that is that on average they recover the model β) and small variance (that is if we had drawn a slightly different sample the likelihood of getting a very different parameter

estimate is not large).

While ML is not very useful for the *identification* part, as identification requires writing a model down and ML often relies on letting the data decide the functional form of your model, it can be quite useful for the *estimation* part. The reason is that the choice of estimator plays a key role in the prediction power of the estimated model and therefore in the standard errors of the estimated parameter. To see this recall our earlier conversation about the OLS estimator:

$$\hat{\beta} \sim N(0, \sigma^2/N),$$

in general the variance σ^2 is not known and therefore we need to estimate it. The simplest way of estimating it, and therefore getting an estimate of the standard error, is through the residuals

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_i \hat{\epsilon}_i^2 = \frac{1}{n-1}\sum_i (E_i - \hat{\beta}MBA_i)^2.$$

Observe that $\frac{1}{n-1}\sum_i (E_i - \hat{\beta}MBA_i)^2$ is essentially the *mean squared error* (MSE) of the regression (the error between the true earnings and your prediction $\hat{\beta}MBA_i$) . Therefore, *regressions with smaller MSE will also lead to smaller standard errors for the parameter of interest*. Now, how can ML help?

Well in the case in which you only have one variable, MBA_i, there is not much to do. But in most settings your model of the world will include other variables X_i, for example controls such as salary before the MBA, undergraduate school, age etc. In fact, in many cases the independence assumption is conditional on the controls (i.e. $\epsilon_i \perp MBA_i | X_i$), so you need to add them to your regression. In these settings ML can be helpful. The reason is that while adding more variables gives you more information, it also increases the number of parameters that you have to estimate. Every time you estimate a parameter you pay a "price" that, in non technical terms, is similar to having a smaller effective sample size.

To see this consider the regression analysis of pioneering MIT professor Victor Chernozhukov. We are now in setting in which our model of the world is given by

$$y_i = \theta' Z_i + \epsilon_i,$$

where $\theta' Z_i = \alpha + \beta D_i + \gamma' X_i$, our treatment of interest D_i plus d covariates of interest. In this setting, under some weak conditions it can be shown that

$$\mathbb{E}[MSE(\theta' Z_i, \hat{\theta}' Z_i)] \lesssim \frac{d}{n},$$

where $\hat{\theta}$ is the OLS estimate and "$\lesssim$" means less up to a constant. The result above states that the expected MSE scales at a rate of d/n. That means that when d is small relative to the sample size n OLS will work well. However, when d is large relative to n then OLS will not control the MSE and we can get overly large standard errors! In fact, if d is larger than n then we just can not use OLS!

ML saves the day - the LASSO In high dimensional settings, when d is large, ML can help. The intuition is that ML methods take into account the "cost" of estimating new parameters in the loss function through *regularization*. One of the most used ML regression methods is the LASSO, which regularizes the OLS estimate by adding an l_1 penalty term to the loss function

$$\hat{\theta}^{LASSO} \in \text{argmin}_{\theta} \frac{1}{n} \sum (y_i - \theta' Z_i)^2 + \lambda \sum_{k}^{d} |\theta_k|,$$

where the λ is a penalty parameter that controls the absolute size of the θ parameters. The more, non-zero, parameters the larger the loss. By choosing the λ optimally we can navigate the bias variance trade-off and get a lower MSE rate than in the standard OLS. Under an *approximate sparsity assumption*, that is that not all X are relevant, but only s of them are in the true model (have coefficients that are not "close" to zero), we can show that

$$\mathbb{E}[MSE(\theta' Z_i, \hat{\theta}^{LASSO\prime} Z_i)] \lesssim \frac{s \log(d)}{n},$$

so now instead of the MSE decaying at d/n rate it decays at a rate $log(d)/n$. This is *significantly* better, if before we needed $n = 1000$ to get a small SE with $d = 50$ (d/n = 0.05) now we only need n=34 ($log(50)/34 = 0.05$), an incredible drop in the sample size requirements to get precise estimates.

This improvement however, does not come for free. ML methods trade-off bias and variance when they regularize the model, therefore in general the estimators are not *unbiased*, a property that OLS has and that we would like to maintain. This means that on average if we used LASSO the $\hat{\beta}$ we would get if we sampled over and over again would not be the true β. This is not the end of the world however, we can account for this bias introduced by the ML model by cross-fitting.

Dealing with ML bias The problem with using an ML algorithm is that it biases your estimates. This bias can be quite bad, even in RCTs when the treatment is randomly assigned, as shown in a paper by Chernozhukov et al. 2016. However, one can correct for this bias through a cross-fitting procedure. A simple example of this is the **double-LASSO estimator**.

Instead of using LASSO and interpreting the coefficients $\hat{\theta}^{LASSO}$ directly, we do a two step procedure:

1. **Data split**: we split the data in two independent sets.
2. **LASSO step**: In the first split of the data we compute LASSO estimates for two regressions and store the residuals
   - LASSO of Y on X: gives us $\hat{\gamma}_{YX}$, residual $\tilde{y}_i = y_i - \hat{\gamma}'_{YX} X_i$.
   - LASSO of D on X: gives us $\hat{\gamma}_{DX}$, residual $\tilde{D}_i = D_i - \hat{\gamma}'_{DX} X_i$.
3. **OLS step**: In the second split of the data we run OLS using the residuals:

$$\hat{\beta}^{doubleLASSO} \in \text{argmin}_b \frac{1}{n} \sum (\tilde{y}_i - b\tilde{D}_i)^2.$$

It can be shown that $\hat{\beta}^{doubleLASSO}$ is an unbiased estimator of β and asymptotically normal, that is as $n \to \infty$ we have that

$$\sqrt{n}(\hat{\beta}^{doubleLASSO} - \beta) \sim N(0, V),$$

where $V = (\mathbb{E}\tilde{D}^2)^{-1}(\mathbb{E}\tilde{D}^2\epsilon^2)(\mathbb{E}\tilde{D}^2)^{-1}$ can be estimated using its sample counterpart. Furthermore, the standard error $\sqrt{\frac{\hat{V}}{n}}$, will in general be smaller than the simple OLS standard error $\sqrt{\frac{\hat{\sigma}^2}{n}}$.

- **Intuition recap**: ML methods can be useful in high dimensional settings in reducing the *variance* of our estimators, however this comes at the cost of introducing *bias*. To address this bias we have to use cross-fitting procedures such as the double LASSO.

3.1. Deciding what is important - ML for model selection

Until now we have focused on cases in which we have a clearly defined question and parameter of interest (the ATE in the MBA return example). This has let us to formulate a model and a hypothesis test to evaluate with the data we have whether some effect is different than zero ($H_0 : \beta = 0$). In many cases however, we may not know what variable is important in predicting a particular outcome, or what treatments have important effects on the outcome of interest. In these cases we might want a procedure to tell us what variables are important. As before we are in a high dimensional setting in which

$$y_i = \theta' Z_i + \epsilon_i,$$

The first approach we might think of is that testing all the possible hypothesis, called **subset selection**:

$$H_0 : \theta_1 = 0, H_0 : \theta_1 = \theta_2 = 0...$$

for d variables this would imply testing 2^d hypothesis, which might be unfeasible, and once we account for multiple hypothesis testing corrections (as we discussed before), powerless. An ML based alternative is to use a data-driven procedure to select the variables that ”matter”. Once again, the most common procedure to do is the LASSO.

A model selection/consistency result states that as $n \to \infty$,

$$P(\hat{\theta}_k = 0 \text{ if the true } \theta_k = 0) \to 1.$$

That is, if in the true model of the world a variable k has no effect on y then an estimator that correctly estimates that variable's coefficient to be zero with high probability is called *model consistent*. It turns out that LASSO is *model consistent* under two important assumptions:

1. **Sparsity**: only a few s of the d variables have non-zero effect in the true model. That is for $k = 1, \ldots, s$, $\theta_k \neq 0$, but for $k > s$ $\theta_k = 0$. Furthermore, $s/n \to 0$, that is s does not grow as fast as the sample size n.

2. **Irrepresentative condition**: this is a technical condition that states that the non-sparse variables $1, \ldots, s$ and the sparse variables $s + 1, \ldots, d$ are not very correlated. If they were, then the method would not know how to distinguish between them.

If these two conditions are met then we can use LASSO to determine what variables are importnat (non-zero) in predicting outcome y. Furthermore, we can use a **post-LASSO** procedure to select the important variables and then estimate a causal effect. Similar to the double-LASSO, the post-LASSO requires cross-fitting:

1. Split the data in two independent sets.

2. In the first split, run LASSO of Y on X, and keep the variables that LASSO does not set to zero, call these variables X^{LASSO}.

3. In the second split, run OLS of Y on X^{LASSO}.

If some technical conditions including sparsity and the irrepresentative condition are satisfied then the post-LASSO OLS is an unbiased estimator of the true parameters and we have reduced the number of parameters ot estimate from d to the set of sparse parameters!

4. Natural experiments, DD and RDD

In our discussion of using models and data to assess causal effects the key part was being able to defend the *independence assumptions* that allow us to interpret model parameters as causal effects (for example as ATEs). In a medical RCT the assumption is easy to defend because the treatment is randomly assigned and it is feasible to find a good control group, however in economics good RCTs are very hard to implement and often unfeasible. One can get around not having an RCT by making the model more complex, controlling for many variables, taking into account endogeneity concerns, signing the bias using the OVB formulate etc. However, at the core of our analysis there will always be an indenpendence assumption that we have to defend.

The *second best* in economics is to find situations in which we have a source of exogenous variation that allows us to approximate RCT conditions. We call them **Natural Experiments**. The Nobel prize in Economics in 2021 was given to three economists (David Card, Guido Imbens and MIT professor Josh Angrist) that pioneered the econometric analysis of natural experiments to study questions such as the returns of schooling, the effects of immigration on local labor outcomes or the effects of raising the minimum wage on unemployment. This body of work kick started what is called the *credibility revolution* in economics: the shift in focus from model predictions to careful causal estimates by using exogenous variation to answer economic questions.

David Card's work on the minimum wage, with Alan Krueger, offers a great example of what a natural experiment is. One of the big question in labor economics is whether raising the minimum wage leads to more unemployment. On one hand, standard economic theory suggests that higher minimum wages increase costs for firms leading to less hiring and therefore more unemployment. On the other hand, higher minimum wages could also lead to increased consumption among workers and economic growth which could lower unemployment. Hence, one might argue that the effect is theoretically ambiguous and therefore an empirical question. The problem is that minimum wages are not randomly assigned, rich regions might be able to set higher minimum wages than poorer regions, furthermore they might be on different growth paths, hence comparing minimum wages across regions suffers from serious selection/omitted variable bias concerns.

The natural experiment used in Card and Krueger 1993 exploits a policy change in New Jersey in 1992: minimum wage increased from \$4.25 to \$5.05 per hour. By comparing fast food restaurants on each side of the NJ-Pennsylvania border before and after the policy change we can estimate the causal effect of the minimum wage increase. The intuition is that given that restaurants and markets on either side of the border are very similar we can use Pennsylvania restaurants as a valid control group for the NJ restaurants after the minimum wage increase, hence approximating a real experiment. This simple method of comparing units in a natural experiment is called **Differences-in-Differences** or **DD**.

DD Consider two units $i = 1, 2$ and two time periods $t = 1, 2$. In time period 2 unit 1 gets treated with a policy, for example a minimum wage increase as in the Card and Krueger example. As before, suppose that our outcome of interest, say unemployment, is Y_{it} and our treatment variable is D_{it} with $D_{it} = 1$ if $i = 2$ and $t = 2$ and 0 otherwise. Potential outcomes given the two periods are given by $Y_{it}(0, 0)$ and $Y_{it}(0, 1)$, with $(0, 0)$ indicating no treatment in either period and $(0, 1)$ indicating treatment in the second period. As opposed to before, we are now interested in the average treatment effect on the treated (ATET), that is:

$$\tau = \mathbb{E}[Y_{i2}(0, 1) - Y_{i2}(0, 0)|D_i = 1].$$

The thought experiment is that this is the effect we would get if we picked someone at random from the set of people that received the treatment. Note that this is different than the ATE which is the effect if we picked someone at random from the whole population. The intuition of why we focus on the ATET rather than the ATE is that to say something about the ATE we would need the stronger assumption that the treatment is randomly assigned, in natural experiments we don't have that in general as we don't choose the treatment as it is often due to a policy change. The ATET can be identified from the simple DD estimator under two key assumptions:

1. **No anticipation**: the treatment does not affect the outcome in previous periods:

$$Y_{i1}(0, 0) = Y_{i1}(0, 1)$$

2. **Parallel trends**: in absence of the treatment both units would follow the same trend. That is the treated and the control unit would have had the same evolution in the

outcome

$$\delta = \mathbb{E}[Y_{i2}(0,0) - Y_{i1}(0,0)|D_i = 1] - \mathbb{E}[Y_{i2}(0,0) - Y_{i1}(0,0)|D_i = 0] = 0$$

The **DD** estimator is what you thought it was:

$$\hat{\tau}_{DD} = (Y_{22} - Y_{21}) - (Y_{12} - Y_{11}),$$

that is the difference in values post-pre treatment for the treatment unit (i=2) minus the difference in values post-pre treatment for the control unit $i = 1$. Under the no anticipation and parallel trend assumption it is easy to see that $\mathbb{E}[\hat{\tau}_{DD}] = \tau$:

$$\begin{aligned}
\mathbb{E}[\tau_{DD}] &= \mathbb{E}[(Y_{22} - Y_{21}) - (Y_{12} - Y_{11})] \\
&= \mathbb{E}[Y_{i2} - Y_{i1}|D_i = 1] - \mathbb{E}[Y_{i2} - Y_{i1}|D_i = 0] \quad \text{(since only unit 2 gets treated)} \\
&= \mathbb{E}[Y_{i2}(0,1) - Y_{i1}(0,1)|D_i = 1] - \mathbb{E}[Y_{i2}(0,0) - Y_{i1}(0,0)|D_i = 0] \quad \text{(re-writing)} \\
&= \mathbb{E}[Y_{i2}(0,1) - Y_{i1}(0,0)|D_i = 1] - \mathbb{E}[Y_{i2}(0,0) - Y_{i1}(0,0)|D_i = 0] \quad \text{(by no anticipation)} \\
&= \mathbb{E}[Y_{i2}(0,1) - Y_{i1}(0,0)|D_i = 1] - \mathbb{E}[Y_{i2}(0,0) - Y_{i1}(0,0)|D_i = 1] + \delta \quad \text{(trend definition)} \\
&= \tau \quad \text{(by parallel trend).}
\end{aligned}$$

Hence, the simple DD estimator is an unbiased estimator of the ATET! The key assumption, that often is hard to justify, is the parallel trends assumption. Researchers may show that treated and control units follow the same trends before treatment to justify this assumption in practice. Alternatively, recent methods have been proposed to create control units that are more likely to satisfy this type of assumptions, or that don't rely on this assumption, chief amongsts these new methods are Synthetic Controls (pioneered by MIT professor Alberto Abadie).

– **Recap**: **DD** designs are an example of how we can exploit weaker assumptions (no anticipation, parallel trends) than independence when a natural experiment exists to say something about a parameter like the ATET. The caviat of the method is that you still have to believe in some assumption, the parallel trends assumption, and in many settings this assumption is unlikely to be satisfied.

Regression Discontinuity Design Another great example of natural experiment variation used to identify causal effects in Economics is that of RDD. This type of research design is very popular currently to study important economic questions ranging from education to climate change. It was pioneered in part by MIT Nobel laureate Josh Angrist and MIT economist Parag Pathak in leading work on the returns of education. The idea behind RDD is also intuitive: when treatment is assigned on the basis of a threshold on some continuous variable, units on each side of the threshold should be very similar and so the ones that did not receive the treatment can be used as a control group. Let's return to the MBA example to see this. Suppose that acceptance into the Sloan MBA depended on a numerical score (say the GMAT) and that students above a certain threshold are admitted and students below are not. Let's call the score T and the threshold c, then we can define the treatment indicator as MBA:

$$MBA_i = \begin{cases} 1 & \text{if } T_i \geq c, \\ 0 & \text{if } T_i < c. \end{cases} \tag{4}$$

The regression discontinuity estimator is often motivated with the following model:

$$E_i = \alpha + \beta MBA_i + \gamma T_i + \epsilon_i.$$

Observe that

$$\begin{aligned} \lim_{t \downarrow c} \mathbb{E}[E_i | T_i = t] &= \lim_{t \downarrow c} \mathbb{E}[E_i | T_i = t, MBA_i = 1] = \alpha + \beta + \lim_{t \downarrow c} \gamma t + \mathbb{E}[\epsilon_i | T_i = t], \\ \lim_{t \uparrow c} \mathbb{E}[E_i | T_i = t] &= \lim_{t \downarrow c} \mathbb{E}[E_i | T_i = t, MBA_i = 0] = \alpha + \lim_{t \uparrow c} \gamma t + \mathbb{E}[\epsilon_i | T_i = t], \end{aligned}$$

so assuming *continuity* so that the limits are well defined it follows that:

$$\begin{aligned} \lim_{t \downarrow c} \mathbb{E}[E_i | T_i = t] - \lim_{t \uparrow c} \mathbb{E}[E_i | T_i = t] &= \lim_{t \downarrow c} \mathbb{E}[E_i | T_i = t, MBA_i = 1] - \lim_{t \uparrow c} \mathbb{E}[E_i | T_i = t, MBA_i = 0] \\ &= \mathbb{E}[E_i | T_i = c, MBA_i = 1] - \mathbb{E}[E_i | T_i = c, MBA_i = 0] \\ &= \beta + \gamma c + \mathbb{E}[\epsilon_i | T_i = c] - \gamma c - \mathbb{E}[\epsilon_i | T_i = c] \\ &= \beta. \end{aligned}$$

That is, the simple OLS estimator including the score and the threshold rule, identifies the **Average Treatment Effect** at the cutoff c!

- **RDD thought experiment**: we can identify the expected treatment effect of an MBA on a student randomly drawn from the population of students with scores exactly $T = c$ (GMAT=750 for example).

Note that this is significantly more restrictive than the treatment effects we investigated before. With the strong independence assumption we could say something about the average person (the ATE), with the DD design we could say something about the average person that was treated (i.e. everyone that is doing an MBA currently), with RDD we can say something about the *marginal* person that got into the MBA. The scope is significantly reduced, but the assumption we made is significantly weaker. Instead of assuming that the unobserved is unrelated to the MBA decision, all we are assuming is that there are no discontinuities around the cutoff. While this seems like an odd assumption it has clear economic significance. The **continuity assumption** may not be satisfied when

1. There is manipulation around the cutoff: certain type of students work harder to be on one side of the cutoff leading to different sets of students on each side.

2. Endogenous cutoff: if the school, as might be the case, selects the cutoff to screen student ability then we would expect students on each side of the cutoff to be different.

One way of ”checking” the continuity assumption is to make sure that the characteristics of the units on each side of the cutoff are similar. Like a balance check in an RCT, we would like that the treatment and control group (each side of the cutoff) do not differ meaningfully.

A note on estimation In practice, estimating the RDD parameter requires making **functional form** assumptions on when the cutoff is crossed. In the model written above we have assumed a linear functional form, meaning earnings follow a linear trend before and after the cutoff, but in many applications more flexible models might be more appropriate. Depending on the model's functional form and estimator used we will be using more or less data far away from the cutoff to estimate the point at the limit from each side. The more data we use further away from the cutoff the more worried we might that we are comparing different units and biasing our estimate, but the more data we use the better the estimate of the function at the cutoff will be. Therefore, there is a trade-off when choosing how to *estimate* regression discontinuity designs! In practice, many different solutions have been proposed, including ML driven RDD in which the threshold around the cutoff used to estimate the treatment effect is optimized to minimize the MSE. Work on this has been pioneered by Guido Imbens (winner of the Nobel prize in 2021) with Matias Cattaneo, and it involves, as in our discussion of ML methods, a cross-fitting step to account for the bias induced by the ML estimator.