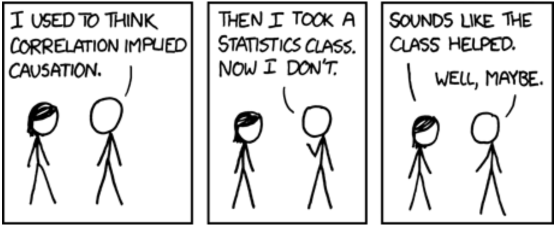
Advanced Applied Econometrics

Teacher: Felix Weinhardt



- Last week: Frish-Waugh/Regression Anatomy, OVB, AET2005 and Oster-method
- Today:
- Heterogeneous effects in OLS
- Fisher Inference

- Heterogeneous treatment effects

Heterogeneous treatment effects

- Throughout, homogeneous treatment effects are a remaining assumption
- This is not usually made explicit when OLS estimates are discussed
- In the context of IV-LATE this is well understood (we cover this in a few sessions)
- However, effect heterogeneity has consequences also for the interpreation of plain OLS (or Diff in Diff and other methods relying on least-squares estimation techniques) that extend to the interpretation of coefficient movements and the OVB forumla.
- Since heterogeneity in OLS is not well understood, there is no textbook or reading for this section (let me know if you find one!)

- A typical notation that allows for effect heterogeneity is:

$$Y = \alpha + \beta_g * X + \epsilon$$

- OLS estimates are usually interpreted as providing an *average effect* whenever such heterogeneity is not modelled explicity. (Note: I think this is what is happening, in a nutshell.)

$$\beta^P = \sum_{g=1}^{G} w_g \beta_g = \frac{\sum_{g=1}^{G} \frac{N_g}{N} VAR(X|g)}{VAR(X)} \beta_g$$

- This is only the *average effect* when each observation is weighted equally, or groups by their size.
- This is the formula that only averages over group size: $\beta^{AE} = \sum_{g=1}^{G} \frac{N_g}{N} \beta_g = \sum_{g=1}^{G} \frac{N_g}{N} \frac{COV(X_g, Y_g)}{VAR(X_g)}$.
- This is not what OLS does.

Consider the following numerical example:

- We generate a single dataset with $N = 1000$, where Y depends only on X and a normally distributed error term.
- Moreover, the variance in X is not constant across strata g, we have $VAR(X_g|1) < VAR(X_g|2)$, so the regressor values are independent but not identically distributed. Each strata g has $N = 500$.
- The outcome is defined as $Y = \beta_g X + \epsilon$ and treatment effects are heterogeneous with $\beta_1 = 1$ and $\beta_2 = 5$.

Table: Estimates with heterogeneous effects and heteroskedastic strata that are positively related

	(1) $\hat{\beta}_1$	(2) $\hat{\beta}_2$	(3) $\hat{\beta}^{AE}$	(4) $\hat{\beta}^{P}$
X	0.957***	4.979***	2.975	4.004***
	(0.0457)	(0.0253)	-	(0.0831)
Constant	0.0964*	0.0332	-	0.138
	(0.0455)	(0.0440)	-	(0.0815)
N	500	500	1000	1000

- The issue is that OLS also weights groups by their variance, not only by their size.
- This relates directly to the i.i.d. assumption. Regressors need to be independent, and identically distributed
- Angrist and Pischke write in mostly harmless that this is the case whenever samples are sufficiently large.

- But what about the i.i.d assumption when conditional independence is required?
- OLS provides the following sample-size-variance weighted average:

$$\beta^{PM} = \sum_{g=1}^{G} w_g \beta_g = \frac{\sum_{g=1}^{G} \frac{N_g}{N} VAR(\tilde{X}|g)}{VAR(\tilde{X})} \beta_g \quad (2)$$

- Are there good reasons to believe that $VAR(\tilde{X}|g)$ is constant across g?
- Recall $VAR(\tilde{X})$ comes from the auxiliary regression and so depends on the degree of multicolinearity of the RHS variables across strata.
- We do not usually make assumptions about multicolinearity (except no perfect multicolinearity)
- This will not go away in large samples...

Consider the following numerical example:

- In contrast to the previous example, we here set the variance in X as constant across strata, we have $VAR(X)|1 = VAR(X)|2$. This means that a simple OLS in this setting returns a valid estimate for the average treatment effect and $\beta^P = \beta^{AE}$ as regressors are i.i.d.
- We now want to understand what happens if control variables are added to this specification. For this, we define a single variable W that correlates with X in the following way: $COV(W, X|1) = 0$ and $COV(W, X|2) > 0$.
- Do you think that adding the "irrelevant control" W will affect the estimates?

Table: Simple OLS with heterogeneous effects and heteroskedastic strata: how "irrelevant" controls change the estimates

	(1)	(2)
	$\hat{\beta}^{P}$	$\hat{\beta}^{PM}$
X	3.000***	2.466***
	(0.113)	(0.0893)
W		1.061***
		(0.0732)
Constant	0.103	0.0960
	(0.0691)	(0.0617)
Controls included		✓
N	1000	1000

- So, controls can move the OLS estimates even if they are irrelevant - this goes against our formula for OVB.
- Since the i.i.d assumption cannot be defended in multiple regression models, the only way out is to assume homogeneous treatment effects.
- This means coefficients can move for multiple reasons, due to classical OVB and due to the way OLS is weighting observations, when effects are heterogeneous.

- In a heterogeneous world, the differences in the estimate between the short and full model is given by:

$$\delta^{diff} = \sum_{g=1}^{G} w_g^l \beta_g^l - \sum_{g=1}^{G} w_g^s \beta_g^s + \gamma * \sum_{g=1}^{G} w_g^\tau \tau_g \tag{3}$$

- The first two terms just represent the weighted average notation of the pooled estimates
- The final product assumes that the omitted variable W itself has a constant effect on Y , , but takes into account that the covariance between W and X might not be constant across groups.
- The last summation represents the variance-sample size weighted effect of W on X.

- In a heterogeneous world, the differences in the estimate between the short and full model is given by:

$$\delta^{diff} = \sum_{g=1}^{G} w_g^l \beta_g^l - \sum_{g=1}^{G} w_g^s \beta_g^s + \gamma * \sum_{g=1}^{G} w_g^\tau \tau_g \tag{3}$$

- The first two terms just represent the weighted average notation of the pooled estimates
- The final product assumes that the omitted variable W itself has a constant effect on Y , , but takes into account that the covariance between W and X might not be constant across groups.
- The last summation represents the variance-sample size weighted effect of W on X.

- **Final words:**
- Notice how the Oster-method implicitly also assumes effect homogeneity/i.i.d. regressors
- OLS is behaving "correctly". But the *averaging*-interpretation is not valid is many non-experimental settings
- We will see in a few sessions how the recent diff-in-diff literature relates to this, too. For IV, this is well understood.
- Much of this can be interpreted as specification error: effect heterogenetiy is not modelled explicitly, which generates the problem of the averaging interpretation. But it cannot be modelled explicitly without knowing the underlying groups.
- Given how poorly properties of OLS are understood –do we believe we understand fully even more compliated estimation strategies?

- **Final words:**
- Notice how the Oster-method implicitly also assumes effect homogeneity/i.i.d. regressors
- OLS is behaving "correctly". But the *averaging*-interpretation is not valid is many non-experimental settings
- We will see in a few sessions how the recent diff-in-diff literature relates to this, too. For IV, this is well understood.
- Much of this can be interpreted as specification error: effect heterogenetiy is not modelled explicitly, which generates the problem of the averaging interpretation. But it cannot be modelled explicitly without knowing the underlying groups.
- Given how poorly properties of OLS are understood –do we believe we understand fully even more compliated estimation strategies?

- Fisher inference

Lady tasting tea experiment

- Ronald Aylmer Fisher (1890-1962)
  - Two classic books on statistics: *Statistical Methods for Research Workers* (1925) and *The Design of Experiments* (1935), as well as a famous work in genetics, *The Genetical Theory of Natural Science*
  - Developed many fundamental notions of modern statistics including the theory of randomized experimental design.
- Muriel Bristol (?? - ??)
  - Worked with Fisher at the Rothamsted Experiment Station (which she established) in 1919 (*and a PhD scientist back in the days when women weren't PhD scientists*)
  - During afternoon tea, Muriel claimed she could tell from taste whether the milk was added to the cup before or after the tea
  - Scientists were incredulous, but Fisher was inspired by her strong claim
  - He devised a way to test her claim which she passed. What was the test?

Description of the tea-tasting experiment

- Original claim: Given a cup of tea with milk, Bristol claims she can discriminate the order in which the milk and tea were added to the cup
- Experiment: To test her claim, Fisher prepares 8 cups of tea – 4 **milk then tea** and 4 **tea then milk** – and presents each cup to Bristol for a taste test
- Question: How many cups must Bristol correctly identify to convince us of her unusual ability to identify the order in which the milk was poured?
- Fisher's sharp null: Assume she can't discriminate. Then what's the likelihood that random chance was responsible for her answers?

Choosing subsets

- "8 choose 4" – $\binom{8}{4}$ – ways to choose 4 cups out of 8
  - There are $8 \times 7 \times 6 \times 5 = 1,680$ ways to choose a first cup, a second cup, a third cup, and a fourth cup, in order.
  - There are $4 \times 3 \times 2 \times 1 = 24$ ways to order 4 cups.
- So there are 70 ways to choose 4 cups out of 8, and therefore a 1.4% probability of producing the correct answer by chance

$$\frac{1680}{24} = 70 = 0.014.$$

- Note: the lady performs the experiment by selecting 4 cups, say, the ones she claims to have had the tea poured first.
- For example, the probability that she would correctly identify all 4 cups is $\frac{1}{70}$

Choosing 3

- To get exactly 3 right, and, hence, 1 wrong, she would have to choose 3 from the 4 correct ones.
    1. She can do this by $4 \times 3 \times 2 = 24$ with order.
    2. Since 3 cups can be ordered in $3 \times 2 = 6$ ways, there are 4 ways for her to choose the 3 correctly.
- Since she can now choose the 1 incorrect cup 4 ways, there are a total of $4 \times 4 = 16$ ways for her to choose exactly 3 right and 1 wrong.
- Hence the probability that she chooses exactly 3 correctly is $\frac{16}{70} = \frac{8}{35}$.

Statistical significance

- Suppose the lady correctly identifies all 4 cups.
- Conclusion
  1. Either she has no ability, and has chosen the correct 4 cups purely by chance, or
  2. she has the discriminatory ability she claims.
- Since choosing correctly is highly unlikely in the first case (one chance in 70), we decide for the second.
  1. if she got 3 correct and 1 wrong, this would be evidence for her ability, but not persuasive evidence since the chance of getting 3 or more correct is $\frac{17}{70} = 0.2429$.
  2. by convention, a result is considered statistically significant if the probability of its occurrence by chance is < 0.05, or, less than 1 out of 20.

Null hypothesis

- In this example, the null hypothesis is the hypothesis that the lady has no special ability to discriminate between the cups of tea.
    - We can never prove the null hypothesis, but the data may provide evidence to reject it.
    - In most situations, rejecting the null hypothesis is what we hope to do.
- Randomization allows us to make probability calculations revealing whether the data are "statistically significant" or not.
- Randomization also takes care of all the possible causes for which we cannot control.

Example: Honey experiment

Paul et al (2007) designed a study to evaluate the effect of giving buckwheat honey or honey-flavored destromethorpan or nothing at night before bedtime on nocturnal cough frequency for a population of children with upper respiratory tract infections

- Population: 72 kids (35 received honey, 37 nothing)
- Outcome of interest: "cough frequency afterwards" (*cfa*)
- Pretreatment variable: "cough frequency prior" (*cfp*)

Notation

- Let Y_i^1 and Y_i^0 represent potential outcomes for individual i with and without honey treatment, respectively
- Let $D_i \subset \{0, 1\}$ be a binary indicator equalling 1 if the child received honey as the treatment and 0 otherwise
- Switching equation:

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$$

- X_i is a covariate/characteristic/pretreatment variable for child i. Here it is cough frequency prior, *cfp*
- Number of treatment (N_t) and control units (N_c):

$$\begin{aligned} N_t &= \Sigma_{i=1}^{N} D_i \\ N_c &= \Sigma_{i=1}^{N} (1 - D_i) \end{aligned}$$

Cough frequency for the first six units

Unit	Potential outcomes		Observed variables		
	Y_i^0	Y_i^1	D_i	X_i	Y_i^{obs}
	cfa			*cfp*	*cfa*
1	?	3	1	4	3
2	?	5	1	6	5
3	?	0	1	4	0
4	4	?	0	4	4
5	0	?	0	1	0
6	1	?	0	5	1

Sharp null

- Let $\delta = Y^1 - Y^0$ be the causal effect of the treatment.
- Assess the "sharp null" hypothesis:

$$H_0 : \delta_i = Y_i^1 - Y_i^0 = 0 \text{ for all } i = 1, \dots, N$$

  against the alternative that for some units there is some non-zero effect of the treatment ($\delta_i \neq 0$)
- **Key feature**: The null hypothesis is considered **sharp** because under the sharp null hypothesis, we know the missing potential outcomes for each observation
- How's that? If $\delta_i = 0$, then we aren't missing any data – we can replace the missing values with observed value to satisfy the null hypothesis equality, i.e., $Y^1 - Y^0 = 0$

Randomized experiment data

Cough frequency for the first six units from honey study under null of no effect

Unit	Potential outcomes		Observed variables		
	Y_i^0	Y_i^1	D_i	X_i	Y_i^{obs}
		cfa		*cfp*	*cfa*
1	(3)	3	1	4	3
2	(5)	5	1	6	5
3	(0)	0	1	4	0
4	4	(4)	0	4	4
5	0	(0)	0	1	0
6	1	(1)	0	5	1

Inference

- Consider some statistic that is a function of the observed variables, D, Y, X, such as the simple difference in means (SDO)

$$\widehat{\delta} = \overline{Y_t} - \overline{Y_c}$$

  where $\overline{Y_t} = \frac{1}{N_t}\Sigma_{i:D_i=1} Y_i$ and $\overline{Y_c} = \frac{1}{N_c}\Sigma_{i:D_i=0} Y_i$

- Given a sample of six units, the value of the statistic is

$$\widehat{\delta} = \frac{8}{3} - \frac{5}{3} = 1$$

- Fisher wants to assess how unusual would it be to estimate a 1 under the null hypothesis where there is no effect of the treatment whatsoever.
- The key insight Fisher had was that *we can derive the exact distribution* of $\widehat{\delta}(Y, X, D)$ under the randomization distribution which is the distribution induced by random assignment to the treatment units

Unit	D_1	D_2	D_3	D_4	D_5	D_6	$\widehat{\delta}$
1	0	0	0	1	1	1	-1.00
2	0	0	1	0	1	1	-3.67
3	0	0	1	1	0	1	-1.00
4	0	0	1	1	1	0	-1.67
5	0	1	0	0	1	1	-0.33
6	0	1	0	1	0	1	2.33
7	0	1	0	1	1	0	1.67
8	0	1	1	0	1	0	-0.33
9	0	1	1	0	1	0	-1.00
10	0	1	1	1	0	0	1.67
...							

Conclusion

- If we assign 3 children to the honey, and 3 to nothing, there are

$$\binom{6}{3} = \frac{6 \times 5 \times 4}{3 \times 2} = 20$$

  different assignment vectors (different values for D), and therefore at most 20 unique values for the δ (only ten are given in the table)
- Of these 20 values for δ, 16 were at least as large in absolute value as $\delta(Y, D, X) = 1$, so that the p-value is $\frac{16}{20} = 0.80$.
- At conventional levels (e.g., 0.05), we wouldn't reject the null hypothesis that there is no treatment effect.

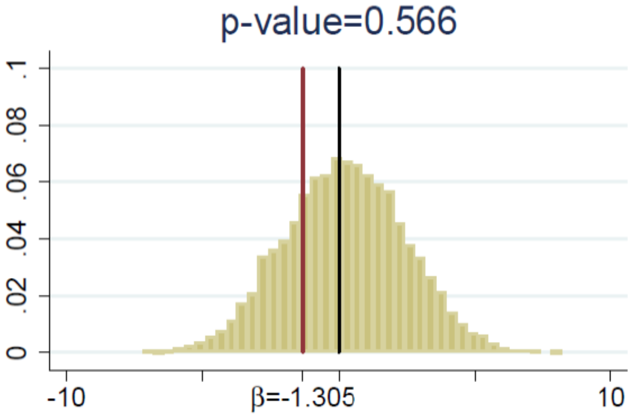
Fisher in today's work

- Useful when sharp null is hypothesis of interest
- Nice feature is that we can produce p values without making assumptions about error variance structure - and without estimating it from our sample
- As a result: preferred method of inference (espacitally in RCTs)

Fisher in practice

- Course of dimensionality
- Consider RCT in 100 schools with 50 getting a treatment
- $$\binom{100}{50} = 1.0089134454556424e + 29$$
- Cannot possible compute exact distribution of outcome under sharp null - too many possibilities
- Solution: choose a random subset of these to approximate sharp null distribution
- Implementation: Take your data and simulate random assignment to geneate outcomes under sharp null.
- Then: compare your experimental estiamte (real world sample) to this distribution

Fisher in practice: Teacher training RCT in schools in England



Example taken from: Murphy, Weinhardt and Wyness (2021) Who teaches the teachers? A RCT of peer-to-peer observation and feedback in 181 schools. *Economics of Education Review*, vol. 82. https://doi.org/10.1016/j.econedurev.2021.102091