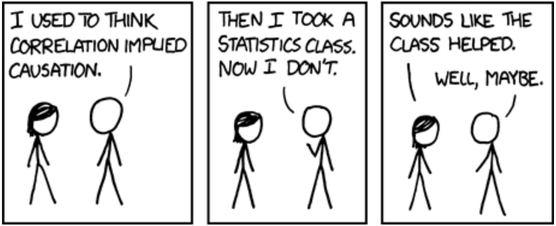
Advanced Applied Econometrics

Teacher: Felix Weinhardt


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Outline: Selected OLS topics

- Regression anatomy theorem
- Omitted variable bias
- Interpreting coefficient movements
- Heterogeneous effects

Slides are partly based on material provided by Scott Cunningham.

- Regression anatomy theorem

Theorem 3.1.7: Regression anatomy theorem

- Regression anatomy theorem is maybe more intuitive with an example and some data visualization. It concerns multiple linear regression.
- Can we estimate the causal effect of family size on labor supply by regressing labor supply (`workforpay`) on family size (`numkids`)?

$$workforpay_i = \beta_0 + \beta_1 numkids_i + u_i$$

```
.  regress workforpay numkids
```

where the first line is the causal / econometric model, and the second line is the regression command in STATA

- If family size is random, then number of kids is uncorrelated with the unobserved error term, which means we can interpret $\widehat{\beta}_1$ as the causal effect.
    - Example: if Melissa has no children in reality (i.e., $\texttt{numkids}= 0$) and we wanted to know what the effect on labor supply will be if we surgically manipulated her family size (i.e., $\texttt{numkids} = 1$) then $\widehat{\beta}_1$ would be our answer
    - Visual: Even better, we could just plot the regression coefficient in a scatter plot showing all *i* `(workforpay, numkids)` pairs and the slope coefficient would be the best fit of the data through these points, as well as tell us the average causal effect of family size on labor supply
- But how do we interpret $\widehat{\beta}_1$ if `numkids` is non-random?

- Assume that family size is random once we condition on race, age, marital status and employment. Then the model is:

$$\begin{aligned}\texttt{Workforpay}_i &= \beta_0 + \beta_1\texttt{Numkids}_i + \gamma_1\texttt{White}_i + \gamma_2\texttt{Married}_i \\ &+\gamma_3\texttt{Age}_i + \gamma_4\texttt{Employed}_i + u_i\end{aligned}$$

- If we want to estimate average causal effect of family size on labor supply, we will need two things:
  1. a data set with all 6 variables;
  2. `numkids` must be randomly assigned conditional on the other 4 variables
- Now how do we interpret $\widehat{\beta}_1$? And can we visualize $\widehat{\beta}_1$ when there's multiple dimensions to the data? Yes, using the regression anatomy theorem, we can.

Theorem 3.1.7: Regression Anatomy Theorem

Assume your main multiple regression model of interest:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i$$

and an auxiliary regression in which the variable x_{1i} is regressed on all the remaining independent variables

$$x_{1i} = \gamma_0 + \gamma_{k-1} x_{k-1i} + \gamma_{k+1} x_{k+1i} + \cdots + \gamma_K x_{Ki} + f_i$$

and $\tilde{x}_{1i} = x_{1i} - \widehat{x}_{1i}$ being the residual from the auxiliary regression. The parameter β_1 can be rewritten as:

$$\beta_1 = \frac{Cov(y_i, \tilde{x}_{1i})}{Var(\tilde{x}_{1i})}$$

In words: The regression anatomy theorem is about interpretation. It says that $\widehat{\beta}_1$ is simply a scaled covariance with the $\tilde{x}_1$ residual used instead of the actual data x.

I think a more detailed proof could be helpful, so I'm leaving it in the slides for now.

Regression Anatomy Proof

To prove the theorem, note $E[\tilde{x}_{ki}] = E[x_{ki}] - E[\widehat{x}_{ki}] = E[f_i]$, and plug y_i and residual $\tilde{x}_{ki}$ from x_{ki} auxiliary regression into the covariance $cov(y_i, \tilde{x}_{ki})$

$$\begin{aligned}
\beta_k &= \frac{cov(y_i, \tilde{x}_{ki})}{var(\tilde{x}_{ki})} \\
&= \frac{cov(\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \dots + \beta_K x_{Ki} + e_i, \tilde{x}_{ki})}{var(\tilde{x}_{ki})} \\
&= \frac{cov(\beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \dots + \beta_K x_{Ki} + e_i, f_i)}{var(f_i)}
\end{aligned}$$

1. Since by construction $E[f_i] = 0$, it follows that the term $\beta_0 E[f_i] = 0$.
2. Since f_i is a linear combination of all the independent variables with the exception of x_{ki}, it must be that

$$\beta_1 E[f_i x_{1i}] = \dots = \beta_{k-1} E[f_i x_{k-1i}] = \beta_{k+1} E[f_i x_{k+1i}] = \dots = \beta_K E[f_i x_{KI}] = 0$$

Regression Anatomy Proof (cont.)

3. Consider now the term $E[e_i f_i]$. This can be written as:

$$\begin{aligned} E[e_i f_i] &= E[e_i f_i] \\ &= E[e_i \tilde{x}_{ki}] \\ &= E[e_i(x_{ki} - \widehat{x}_{ki})] \\ &= E[e_i x_{ki}] - E[e_i \tilde{x}_{ki}] \end{aligned}$$

Since e_i is uncorrelated with any independent variable, it is also uncorrelated with x_{ki}: accordingly, we have $E[e_i x_{ki}] = 0$. With regard to the second term of the subtraction, substituting the predicted value from the x_{ki} auxiliary regression, we get

$$E[e_i \tilde{x}_{ki}] = E[e_i(\widehat{\gamma}_0 + \widehat{\gamma}_1 x_{1i} + \cdots + \widehat{\gamma}_{k-1} x_{k-1}i + \widehat{\gamma}_{k+1} x_{k+1i} + \cdots + \widehat{\gamma}_K x_{Ki})]$$

Once again, since e_i is uncorrelated with any independent variable, the expected value of the terms is equal to zero. Then, it follows $E[e_i f_i] = 0$.

Regression Anatomy Proof (cont.)

4. The only remaining term is $E[\beta_k x_{ki} f_i]$ which equals $E[\beta_k x_{ki}\tilde{x}_{ki}]$ since $f_i = \tilde{x}_{ki}$. The term x_{ki} can be substituted using a rewriting of the auxiliary regression model, x_{ki}, such that

$$x_{ki} = E[x_{ki}|X_{-k}] + \tilde{x}_{ki}$$

This gives

$$\begin{aligned} E[\beta_k x_{ki}\tilde{x}_{ki}] &= E[\beta_k E[\tilde{x}_{ki}(E[x_{ki}|X_{-k}] + \tilde{x}_{ki})]] \\ &= \beta_k E[\tilde{x}_{ki}(E[x_{ki}|X_{-k}] + \tilde{x}_{ki})] \\ &= \beta_k\{E[\tilde{x}_{ki}^2] + E[(E[x_{ki}|X_{-k}]\tilde{x}_{ki})]\} \\ &= \beta_k var(\tilde{x}_{ki}) \end{aligned}$$

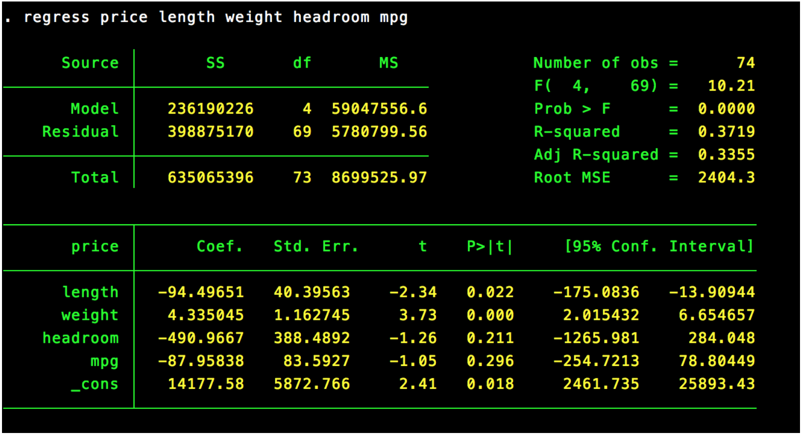
which follows directly from the orthogonoality between $E[x_{ki}|X_{-k}]$ and $\tilde{x}_{ki}$. From previous derivations we finally get

$$cov(y_i, \tilde{x}_{ki}) = \beta_k var(\tilde{x}_{ki})$$

which completes the proof. □

STATA command: `reganat` (i.e., regression anatomy)

```
. ssc install reganat, replace
. sysuse auto
. regress price length weight headroom mpg
. reganat price length weight headroom mpg, dis(weight length) biline
```



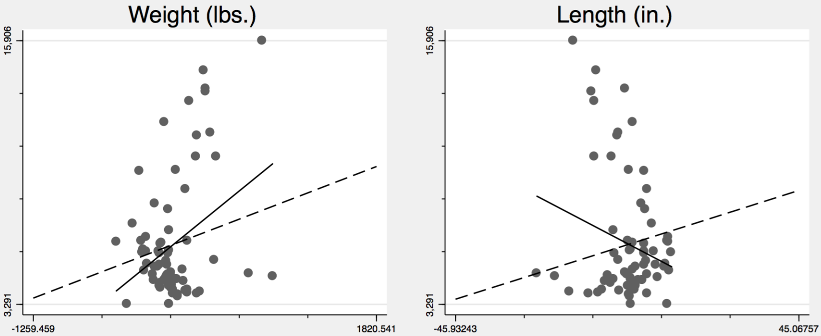
```
. regress price length weight headroom mpg

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  4,    69) =   10.21
       Model |  236190226     4  59047556.6           Prob > F      =  0.0000
    Residual |  398875170    69  5780799.56           R-squared     =  0.3719
-------------+------------------------------           Adj R-squared =  0.3355
       Total |  635065396    73  8699525.97           Root MSE      =  2404.3

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      length |  -94.49651   40.39563    -2.34   0.022    -175.0836   -13.90944
      weight |   4.335045   1.162745     3.73   0.000     2.015432    6.654657
    headroom |  -490.9667   388.4892    -1.26   0.211    -1265.981     284.048
         mpg |  -87.95838    83.5927    -1.05   0.296    -254.7213    78.80449
       _cons |   14177.58   5872.766     2.41   0.018     2461.735    25893.43
------------------------------------------------------------------------------
```

Regression Anatomy

Dependent variable: Price



Weight (lbs.)

Length (in.)

Multivariate slope: 4.335 (1.163)
Bivariate slope: 2.044 (0.377)

Multivariate slope: -94.497 (40.396)
Bivariate slope: 57.202 (14.080)

Covariates: Length (in.), Weight (lbs.), Headroom (in.), Mileage (mpg).

Regression lines: Solid = Multivariate, Dashed = Bivariate.

Big picture

1. Regression provides the best linear predictor for the dependent variable in the same way that the CEF is the best unrestricted predictor of the dependent variable
2. If we prefer to think of approximating $E(y_i|x_i)$ as opposed to predicting y_i, the regression CEF theorem tells us that even if the CEF is nonlinear, regression provides the best linear approximation to it.
3. Regression anatomy theorem helps us interpret a single slope coefficient in a multiple regression model by the aforementioned decomposition.

- Omitted variable bias formula

Omitted Variable Bias

- A typical problem is when a key variable is omitted. Assume schooling causes earnings to rise:

$$Y_i = \beta_0 + \beta_1 S_i + \beta_2 A_i + u_i$$

$$\begin{aligned} Y_i &= \text{log of earnings} \\ S_i &= \text{schooling measured in years} \\ A_i &= \text{individual ability} \end{aligned}$$

- Typically the econometrician cannot observe A_i; for instance, the Current Population Survey doesn't present adult respondents' family background, intelligence, or motivation.

- What are the consequences of leaving ability out of the regression? Suppose you estimated this short regression instead:

$$Y_i = \beta_0 + \beta_1 S_i + \eta_i$$

  where $\eta_i = \beta_2 A_i + u_i$; β_0, β_1, and β_2 are population regression coefficients; S_i is correlated with η_i through A_i only; and u_i is a regression residual uncorrelated with all regressors by definition.

Derivation of Ability Bias

- Suppressing the *i* subscripts, the OLS estimator for β_1 is:

$$\widehat{\beta_1} = \frac{Cov(Y,S)}{Var(S)} = \frac{E[YS] - E[Y]E[S]}{Var(S)}$$

- Plugging in the true model for Y, we get:

$$\begin{aligned}
\widehat{\beta_1} &= \frac{Cov[(\beta_0 + \beta_1 S + \beta_2 A + u), S]}{Var(S)} \\
&= \frac{E[(\beta_0 S + \beta_1 S^2 + \beta_2 SA + uS)] - E(S)E[\beta_0 + \beta_1 S + \beta_2 A + u]}{Var(S)} \\
&= \frac{\beta_1 E(S^2) - \beta_1 E(S)^2 + \beta_2 E(AS) - \beta_2 E(S)E(A) + E(uS) - E(S)E(u)}{Var(S)} \\
&= \beta_1 + \beta_2 \frac{Cov(A,S)}{Var(S)}
\end{aligned}$$

- If $\beta_2 > 0$ and Cov$(A,S)>$ 0 the coefficient on schooling in the shortened regression (without controlling for A) would be upward biased

Summary

- When $Cov(A, S) > 0$ then ability and schooling are correlated.
- When ability is unobserved, then not even multiple regression will identify the causal effect of schooling on wages.
- Here we see one of the main justifications for this class – what will we do when the treatment variable is endogenous? Because endogeneity means the causal effect has not been identified.

- Interpreting coefficient movements

When can we be sure there is no OVB left in what we have?

- Imagine you start off from a situation where there are confouncers, i.e. $Cov(A, S) > 0$
- You can now add more controls and your $\widehat{\beta_1}$ estimate changes.
- This is an improvement –can we learn about how much is left to improve?

Altonji, Elder, Taber: JPE 2005 idea

- Consider the following simple OLS model:

$$Y_i = \beta_0 + \beta_1 D_i + X_i \beta + \epsilon_i$$

- The OLS assumption is:

$$Cov(D, \epsilon) = 0$$

- This assumption implies that the error term ϵ_i is uncorrelated with the independent variable D_i.

- Assume equality of selection on unobservables and observables:

$$\frac{Cov(\epsilon, D)}{Var(\epsilon)} = \frac{Cov(X\beta, D)}{Var(X\beta)}$$

- When is this assumption reasonable?
  - A number of highly relevant observed control variables are available.
  - Data from high-quality surveys specifically targeted to answer the type of research question you are after.

- The assumption:

$$\frac{Cov(\epsilon, D)}{Var(\epsilon)} = \frac{Cov(X\beta, D)}{Var(X\beta)}$$

can be used directly to estimate the OLS bias:

$$\widehat{\beta_1} = \beta_1 + \beta_2 \frac{Cov(\epsilon, \tilde{D})}{Var(\tilde{D})}$$

by substituting:

$$\frac{Cov(\epsilon, D)}{Var(\epsilon)} = \frac{Cov(X\beta, D)}{Var(X\beta)} \cdot \frac{Var(\epsilon)}{Var(\tilde{D})}$$

Estimate the following term step by step:

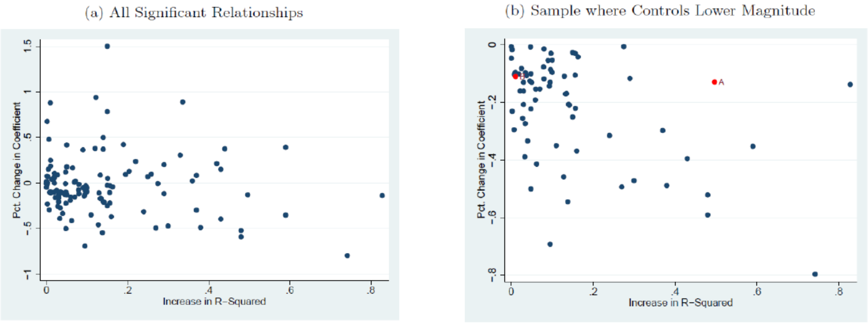
$$\frac{Cov(\epsilon, D)}{Var(\epsilon)} = \frac{Cov(X\beta, D)}{Var(X\beta)} \cdot \frac{Var(\epsilon)}{Var(\tilde{D})}$$

- Estimate an OLS of outcome on all X (excluding treatment) – get Var(ϵ) and $X\hat{\beta}$.
- Use the predicted index and regress it on treatment, D – get $\frac{\text{Cov}(X\beta,D)}{\text{Var}(X\beta)}$.
- Regress OLS of treatment on all X – get Var($\tilde{D}$).

What Emily Oster adds to AET

Critique:

- Coefficient movements alone are not sufficient to calculate bias.
- In particular, this can lead to assuming stability if the included control variable does not explain much of the outcome variance.
- Basic idea is to relate coefficient movements to changes in R^2.



(a) All Significant Relationships

(b) Sample where Controls Lower Magnitude

Notes: These figures show the relationship between the percent change in coeficient and the increase in R-squared in sample of highly cited papers from top journals in economics. The sample is discussed in Section 2 in Oster (2014).

Oster's idea:

- Exploit information on coefficients and R^2 values to compute bounds of treatment effect
- Calculate value of proportionality δ: captures the explanatory power of unobserved variables as a proportion of the explanatory power of observed variables.
- Quality of the control will be diagnosed by the movement in R^2 when the control is included.

- Calculate an identified set for the treatment effect together with the following bias-adjusted coefficient:

$$\beta' = \tilde{\delta}\frac{(\dot{\beta} - \tilde{\beta})(R_{\max} - \tilde{R})}{(\tilde{R} - \dot{R})}$$

- If $\tilde{\delta} = 1$

- Calculate model with few basic explanatory variables only!
- Calculate model that controls for a rich set of explanatory variables!
- Look at R-squared - is it considerably higher than in the "raw" model?
- Look at the coefficient of interest - did it stay fairly stable?

- Researchers can either estimate ... δ
  - Assume a value for R_{max} and calculate $\tilde{\delta}$ for which $\beta = 0$.
  - Oster (2013) argues that $\tilde{\delta} = [0, 1]$ is a useful bound.
  - It is possible to assume $R_{\text{max}} = 1$.
  - Oster (2013) proposes a rule of thumb to calculate $R_{\text{max}} = \min\{2.2\tilde{R}, 1\}$.
- ... or β
  - Assume a value for R_{max} and δ to obtain a set of bounds for $\tilde{\beta}$.
  - Consider $\beta^{*'}(R_{\text{max}}, \delta)$, where $R_{\text{max}} = \min\{\tilde{R}, 1\}$ and assume that proportional selection is positive: $\delta = [0, \tilde{\delta}]$.
  - Define bounds for β where one side of the bound is $\tilde{\beta}$ and the other side is $\beta^{*'}(1, \bar{\delta})$.

- A lot depends on the hypothetical R_{max}
- Original NBER discussion paper:

$$R_{max} = min 2.2 \tilde{R}, 1$$

- Updated versions and final publication:

$$R_{max} = min 1.3 \tilde{R}, 1$$

- Differnces in R_{max} value affects δ depending on increase between short and full model.

- A lot depends on the hypothetical R_{max}
- Original NBER discussion paper:

$$R_{max} = min 2.2\tilde{R}, 1$$

- Updated versions and final publication:

$$R_{max} = min 1.3\tilde{R}, 1$$

- Differnces in R_{max} value affects δ depending on increase between short and full model.

- AET and Oster-method try to formalize what many papers do informally: put a number of *stability of estimates*
- This requires remaining assumptions
- Current state of the art: if you need these methods to defend your estimates, you are in a weak situation.
- Referees might ask you to provide Oster-bounds, so you need to be able to do this.

- 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.

- 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.

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

- 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)	(2)	(3)	(4)
	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}^{AE}$	$\hat{\beta}^{P}$
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:

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

- 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?