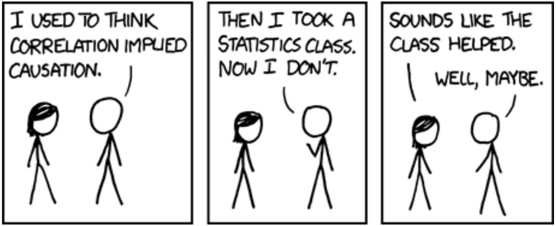
Advanced Applied Econometrics

Prof. Dr. Felix Weinhardt, BSE


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
- Coefficient movements and selection on unobservables (no slides)
- Fisher-inference

Part 1: 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:

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

- 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

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

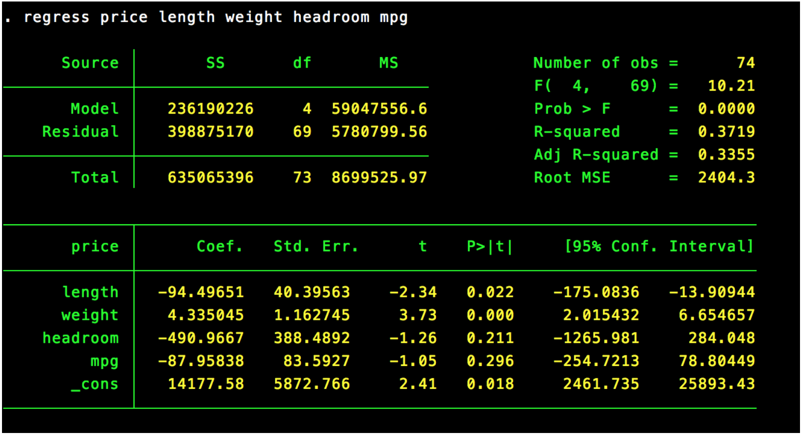
which follows directly from the orthogonoality between $E[x_{ki}|X_{-k}]$ and $\tilde{x}_{ki}$. From previous derivations we finally get

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

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
```



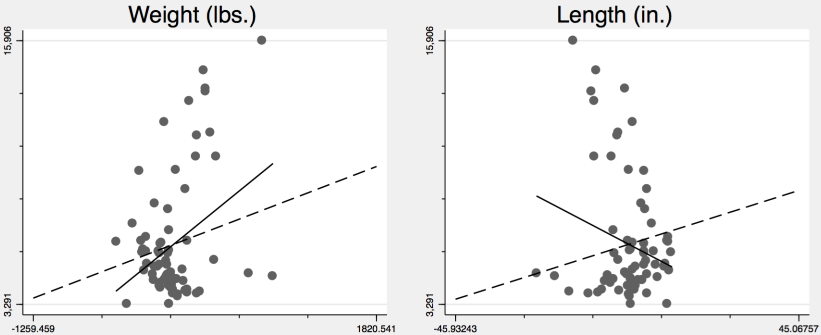
```
. regress price length weight headroom mpg

      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



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.

Part 2: Omitted Variable Bias

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.

Part 4: 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.

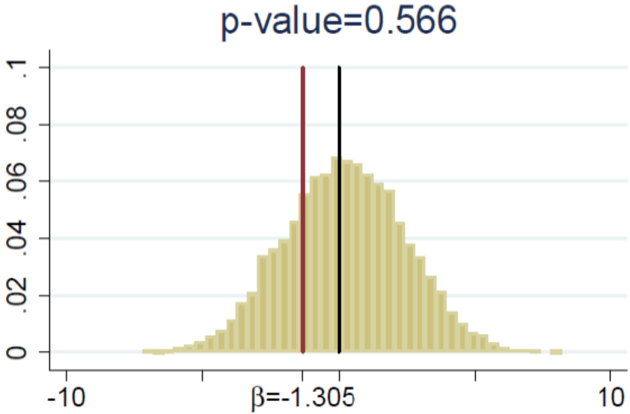
Fisher in today's work

- 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



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