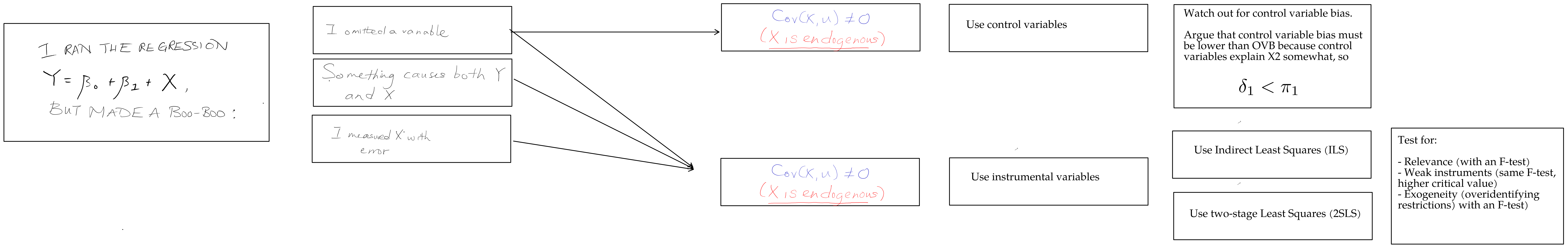


I RAN THE REGRESSION $Y = \beta_0 + \beta_1 + X$, BUT MADE A Boo-Boo:

I omitted a variable

Something causes both Y and X

I measured X with error

$Cov(X, u) \neq 0$ (X is endogenous)

$Cov(X, u) \neq 0$ (X is endogenous)

Use control variables

Use instrumental variables

Watch out for control variable bias.

Argue that control variable bias must be lower than OVB because control variables explain X2 somewhat, so

$$\delta_1 < \pi_1$$

Use Indirect Least Squares (ILS)

Use two-stage Least Squares (2SLS)

Test for:

- Relevance (with an F-test)
- Weak instruments (same F-test, higher critical value)
- Exogeneity (overidentifying restrictions) with an F-test)

Measurement error in X

Even if $cov(X, e_x) = cov(Y, e_x) = 0$:

$Cov(X, u) \neq 0$ (X is endogenous)

because:

$$Y = \beta_0 + \beta_1 (X^* + e_x) + u$$

$$= \beta_0 + \beta_1 X^* + (e_x + u)$$

Pop LR of Y on X^* will give

$$\beta_1^* = \frac{Cov(Y, X^*)}{var(X^*)}$$

$$= \frac{Cov(\beta_0 + \beta_1 X^* + (e_x + u), X^*)}{Var(X^*)}$$

$$= \beta_1 + \frac{Cov(e_x, X^*)}{Var(X^*)}$$

$$= \beta_1 + \frac{Cov(e_x, X - e_x)}{Var(X^*)}$$

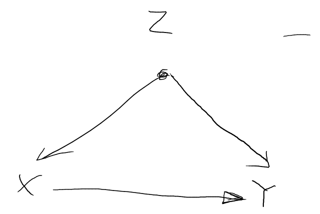
$$= \beta_1 - \frac{var(e_x)}{Var(X^*)}$$

$$= \beta_1 \left(1 - \frac{Var(e_x)}{Var(X^*)}\right)$$

$$= \beta_1 \left(\frac{var(X)}{var(X) + var(e_x)}\right)$$

(attenuation bias)

Simultaneous causation



$Cov(X, u) \neq 0$ (X is endogenous)

because:

$$Y = \beta_0 + \beta_1 X_1 + \overset{\overset{Z}{\downarrow}}{u}$$

A change in u causes an increase in both X and Y.

Omitted variable bias

$Cov(X, u) \neq 0$ (X is endogenous)

because:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \quad (long)$$

$$Y = \gamma_0 + \gamma_1 X_1 + v \quad (short)$$

Pop LR on Y on X_1 will give

$$\gamma_1 = \frac{Cov(Y, X_1)}{Var(X_1)}$$

$$= \frac{Cov(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + u, X_1)}{Var(X_1)}$$

$$= \beta_1 + \beta_2 \frac{Cov(X_2, X_1)}{Var(X_1)}$$

$$= \beta_1 + \beta_2 \pi_1$$

Using control variables (and bias)

Control variables must satisfy:

1. Relevance

$$Cov(Z, X) \neq 0$$

2. Exogeneity

$$Cov(Z, u) = 0$$

But no need to satisfy exclusion.

(But then shouldn't you have already put them in the causal model?)

Note that the regression of

$$Y = \gamma_0 + \gamma_1 X + \gamma_2 Z + \eta$$

will have control variable bias if, in a pop LR of X2 on X1,

$$X_2 = \delta_0 + \delta_1 X_1 + \delta_2 Z_1 + e$$

$$\delta_1 = \frac{Cov(X_2, X_1)}{Var(X_1)} \neq 0$$

d1 = 0 when ("enough of X2 is explained by Z1") such that the effect of Z1 on Z2 = 0.

Otherwise we have control variable bias:

$$\gamma_1 = \beta_1 + \beta_2 \delta_1$$

Using instrumental variables

Instrumental variables must satisfy:

1. Relevance

$$Cov(Z, X) \neq 0$$

2. Exogeneity

$$Cov(Z, u) = 0$$

3. Exclusion: in a population LR of

$$Y = \beta_0 + \beta_1 X + \delta Z + u$$

$$\delta = 0$$

The control variable bias formula is given by

$$\pi_1 = \delta_1 + \delta_2 \frac{Cov(W, X_1)}{Var(X_1)}$$

As long as d2 > 0 (the control variable is correlated with the unobserved var) and Cov(W, X1) > 0 (the control variable is somewhat corr. with the var that can be observed), adding the control var will decrease OVB.

ILS vs 2SLS

Step 1: Run a population regression of

$$X_1 = \pi_0 + \pi_1 Z + v$$

Step 2 (ILS): Substitute this pop LR into the causal model (structural equation), to get the "reduced form":

$$Y = \beta_0 + \beta_1 \pi_0 + \beta_1 \pi_1 Z + \beta_1 v + u$$

$$\rightarrow Y = \gamma_0 + \gamma_1 Z + \epsilon$$

Step 3 (ILS): Simply compare the coeffs:

$$\frac{\gamma_1}{\pi_1} = \frac{\beta_1 \pi_1}{\pi_1} = \beta_1$$

Step 2 (2SLS): Get the predicted values of X from the population regression:

$$X^* = \pi_0 + \pi_1 Z$$

Substitute the predicted values of X into the causal model to get the "reduced form":

$$Y = \beta_0 + \beta_1 (X^* + v) + u$$

where here we know that X* is exogenous for certain due to the fact that it is the part of X 'explained' by Z, which is exogenous.

$$\hat{\beta}_1 = \frac{\hat{\gamma}_1}{\hat{\pi}_1} = \frac{c\hat{o}v(Y, Z)}{c\hat{o}v(X, Z)} \quad (ILS)$$

$$\hat{\beta}_1 = \frac{c\hat{o}v(Y, \hat{X})}{v\hat{a}r(\hat{X})} \quad (2SLS)$$

Testing for relevance, weak instruments and exogeneity with F-tests

Instrumental variables: interpretation of results

6. We have the following data on $n = 973$ men in the US aged 30–34, in 1976 (from the NLSY):

Variable	Description
`lwage`	log of hourly wage (in cents)
`educ`	years of completed schooling
`age`	in years
`black`	$= 1$ if black, $= 0$ otherwise
`south`	$= 1$ if lives in the southern US
`iqscore`	score on an aptitude test
`libcrd14`	$= 1$ if library card in home when individual was aged 14, $= 0$ otherwise
`famed`	father's years of education
`mamed`	mother's years of education

We obtain the following estimates:

OLS and 2SLS regressions

Dependent variable: `lwage`

	(1)	(2)	(3)	(4)
`educ`	0.049	0.037	0.116	0.075
	(0.004)	(0.007)	(0.041)	(0.018)
`age`	0.028	0.037	0.045	0.040
	(0.008)	(0.010)	(0.012)	(0.010)
`black`	−0.281	−0.210	−0.256	−0.234
	(0.033)	(0.052)	(0.059)	(0.053)
`south`	−0.152	−0.091	−0.093	−0.091
	(0.027)	(0.033)	(0.035)	(0.033)
`iqscore`		0.003	−0.004	−0.000
		(0.001)	(0.004)	(0.002)

where (1) and (2) refer to OLS regressions; (3) to a 2SLS regression using `libcrd14` as an instrument for `educ`, and (4) to a 2SLS regression using `libcrd14`, `famed` and `mamed` as instruments for `educ`. (In all cases, an intercept term was also estimated.) We also have F statistics corresponding to the following tests:

H_0: **coefficients are zero**	(5)	(6)
all variables	91	108
`age`, `black`, `south`, `iqscore`	80	54
`libcrd14`	25	8
`famed`, `mamed`		42
`libcrd14`, `famed`, `mamed`		29

where (5) refers to an OLS regression of `educ` on `age`, `black`, `south`, `iqscore` and `libcrd14` (and a constant); and (6) to a regression that additionally includes `famed` and `mamed` as regressors. Finally, in the context of model (4), a (heteroskedasticity-robust) test of 'overidentifying restrictions' yielded a test statistic of 2.25.

(a) Interpret the estimated coefficient on educ in (1). Compute a 99% confidence interval for this coefficient and interpret it.

$$lwage = \beta_0 + \beta_{educ}\, educ + \beta_{age}\, age + \beta_{black}\, black + \beta_{south}\, south + u$$

A one-year increase in education is associated with a 4.9% percent increase in wage holding all other vars in the regression constant.

$$99\% \text{ confidence interval} : \left\{ \hat{\beta}_{educ} \mp 2.58\, SE(\beta_{educ}) \right\}$$

$$99\%\ CI_{actual} : \left\{ 0.049 \pm 2.58(0.004) \right\}$$

$$= \left\{ \qquad \qquad \right\}$$

The 99% confidence interval is the interval contain the true value of the β_{educ} coefficient in 99 % of random samples

(b) Compare the estimated coefficients on educ in models (1) and (2), and explain the differences using the omitted variables bias formula. Do you think regression (1) consistently estimates the causal effect of education on wages? How about regression (2)? Explain.

Suppose the true causal model includes ability as an unobservable variable in the determination of one's wage:

$$\log wage = \beta_0 + \beta_1\, educ + \beta_2\, age + \cdots + \underline{\beta_{ab} Ability} + u$$

However, we only have the short regression in (1) of

$$\log wage = b_0^s + b_1^s\, educ + b_2^s\, age + \cdots + e.$$

This sample linear regression will consistently estimate

(from sample to population → is this OK?)

$$\hat{b}_1^s = \frac{\hat{Cov}(\log wage, \widetilde{educ})}{\hat{Var}(\widetilde{educ})} \xrightarrow{d} \frac{Cov(lwage, educ)}{Var(educ)}$$

Substituting the causal model,

$$= \frac{\hat{Cov}(\beta_0 + \beta_1 educ + \beta_2 age + \cdots + \beta_n ability + u, \widetilde{educ})}{\hat{Var}(\widetilde{educ})}$$

$$\vdots$$

$$= \beta_1 + \beta_n \frac{\widehat{Cov}(ability, \widetilde{educ})}{\widehat{Var}(\widetilde{educ})}$$

FWL: $Y = b_0 + b_1 X_1 + b_2 X_2$

$$b_1 = \frac{Cov(\tilde{X}_1, Y)}{Var(\tilde{X}_1)}$$

$= \beta_1 + \beta_n \pi_1$, where π_1 is the coefficient of educ on a (purely hypothetical) sample linear regression of ability on educ and the other demographic variables.

Therefore, omitted variable bias will cause the estimated coefficient on $\widetilde{educ}$ to go up when ability is positively correlated with income and education.

When we include IQ in regression 2, it is natural to suppose that $Cov(IQ, educ) > 0$ and $Cov(IQ, ability) > 0$.

Consider then a purely hypothetical regression of ability on IQ and education with no causal interpretation,

$$Ability = \delta_0 + \delta_1 educ + \delta_2 IQ + \delta_3 age + \ldots + e$$

The sample LR in (2) thus consistently estimates

$$\hat{b}_1^L \xrightarrow{d} \frac{Cov(\gamma_0 + \gamma_1 educ + \beta_2 age + \ldots, \widetilde{educ})}{Var(\widetilde{educ})}$$

By construction, $Cov(\widetilde{educ}, age) = Cov(\widetilde{educ}, IQ) = 0$ and all the other vanables as well, so

$$b_1^L = \beta_1 + \beta_n \delta_1 \quad \text{where } \delta_1 = \frac{Cov(\widetilde{educ}, Ability)}{Var(\widetilde{educ})}$$

where δ_1 is the coeff on educ in the regession of ability on educ, demographic variables, AND IQ.

Given that IQ explains ability somewhat, we would expect that $\delta_1 < \pi_1$. Hence,

$$b_1^L > b_1^S$$

which is why we see the coefficient on educ fall in (2) compared to in (1).

Additionally, neither regression coeff. on educ can be interpreted causally interpreted, unless we believe that IQ is a valid control : enough of one's ability is explained by IQ such that the covariance between education and ability, controlling for IQ, falls to zero. And of course seeing as we have OVB in (1), b_1^S can't be causally interpreted.

(c) Why do you think **educ** is instrumented for in (3) and (4)?

We instrument a variable because we worry that $Cov(educ, u) \neq 0$; if this is the case, OR does not hold. We have seen in part (b) that some notion of "ability" not specified in the short regression is responsible for driving wages. Thus $Cov(educ, u) \neq 0$ and we have to instrument for it, to recover the true causal effect of education on wages.

(d) Compare the estimated coefficients on `educ` in (2) and (3). How would you account for the difference in these estimates?

educ in (2) has coeff 0.037 SE = 0.007

in (3) has coeff 0.116 . with SE = 0.041

? Why ?

→ Measurement error in educ → Least squares attenuation bias?

→ Something to do with weak instrument?

(e) State the conditions required of a valid instrumental variable. Do you think that `libcrd14`, `famed`, `mamed` satisfy these conditions? Discuss the extent to which these conditions can be verified empirically, and how any of the results reported above might be used for this purpose.

(f) Why might the standard error for the estimated coefficient on educ be so much smaller in (4) than in (3)? Does this have any relevance for the validity of famed and mamed as instruments for educ? Explain.

(4) uses famed and mamed and libcard as instruments

(3) uses libcard only

$\hat{SE}(\beta_{educ})$ goes down

We are aware that the 2SLS estimator satisfies

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} N(0, w^2_{\beta_1, IV})$$

$$w^2_{\beta_1, IV} = \frac{\mathbb{E}(X^* - \mu_{X^*})^2 u^2}{[\mathbb{E}(X^* - \mu_{X^*})^2]^2}$$

where $\mu_{X^*} = \mathbb{E}X^* = \mathbb{E}X$ because $\mathbb{E}(v) = 0$

Numerator

$$w^2_{\beta, IV} = \frac{\sigma_u^2 \, var(X^*)}{[var(X^*)]^2} = \frac{var(u)}{var(X^*)}$$

where $X^* = \pi_0 + \pi_1 Z_1 + \dots \pi_m Z_m$

predicted values obtained from the first stage reg.

By additive variance,

$$var(X^*) = var(\pi_0) + var(\pi_1 Z_1) + \dots + 2\overset{=0}{Cov(Z_1, Z_2)} \dots$$

and so if the variance of the part of X explained by Z_1 is high,

$w^2_{\beta, IV}$ becomes small and so $SE(\beta_{educ})$ goes down.

3. The following table gives output from an OLS regression run on data from a survey of 1000 adults in the United Kingdom.

Dependent variable: log(wage); wage in pounds per hour		
	Estimate	**Std. Error**
Experience (total years in employment)	0.05	0.02
Gender (=1 if male, =0 otherwise)	0.08	0.03
Region dummies (=1 if lives in indicated region, 0 otherwise)		
Wales	−0.15	0.07
Scotland	−0.06	0.08
Northern Ireland	−0.21	0.13
Constant	2.23	0.42

(a) [20%] Interpret the coefficient estimate on the ‘Wales’ dummy.

(b) [10%] Provide an estimate of the effect of two additional years of experience on log(wage).

(c) [20%] Construct a 90 per cent confidence interval for the effect estimated in part (b).

(d) [20%] Compute the *p*-value for the null hypothesis that men earn the same wages as workers of any other gender. What is its interpretation?

(e) [30%] If experience and the region dummies were excluded from the regression, how would you expect the estimated coefficient on gender, and its standard error, to change?

$$\begin{aligned} Y = \beta_0 + \beta_1 Exp + \beta_2 Gender &+ \beta_3 Wales \\ &+ \beta_4 Scot \\ &+ \beta_5 NI \\ &+ u \end{aligned}$$

(Long)

Short: $Y = \gamma_0 + \gamma_1 Gender$

$$\gamma_1 = \frac{Cov(\beta_1 Exp + \beta_2 Gender + Dummies + u, Gender)}{Var(Gender)}$$

$$= \beta_1 \frac{Cov(Exp, Gender)}{Var(Gender)} + \beta_2 + \beta_3 \frac{Cov(Dummies, Gender)}{Var(Gender)}$$

Gender and region are probably uncorrelated,

so $\text{Cov}(\text{Dummies}, \text{Gender}) \longrightarrow 0$

$$\Rightarrow \quad \beta_2 + \beta_1 \frac{\text{Cov}(\text{Exp}, \text{Gender})}{\text{Var}(\text{Gender})}$$

<u>Asymptotic variance:</u>

$$Y = \beta_0 + \beta_1 X + u$$

$$se(\hat{\beta}_1) \doteq \frac{1}{\sqrt{n}} \quad \frac{se(\hat{u})}{se(\hat{X})}$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 + u$$

$$se(\hat{\beta}_1) = \frac{1}{\sqrt{n}} \cdot \frac{se(\hat{u})}{se(\hat{\tilde{X}}_1)}$$

Remember this!

Here in our sample,

$$se(\hat{\beta}_2^{s}) = \frac{1}{\sqrt{n}} \quad \frac{se(\hat{u}_s)}{se(\hat{gender}_s)}$$

$$se(\hat{\beta}_2^{L}) = \frac{1}{\sqrt{n}} \quad \frac{se(\hat{u}_L)}{se(\hat{\widetilde{gender}}_L)} \downarrow$$

The more you add, the ~~b~~more this goes down

where

$$gender = \gamma_0 + \gamma_1 D \ldots + \ldots + \widetilde{gender}$$

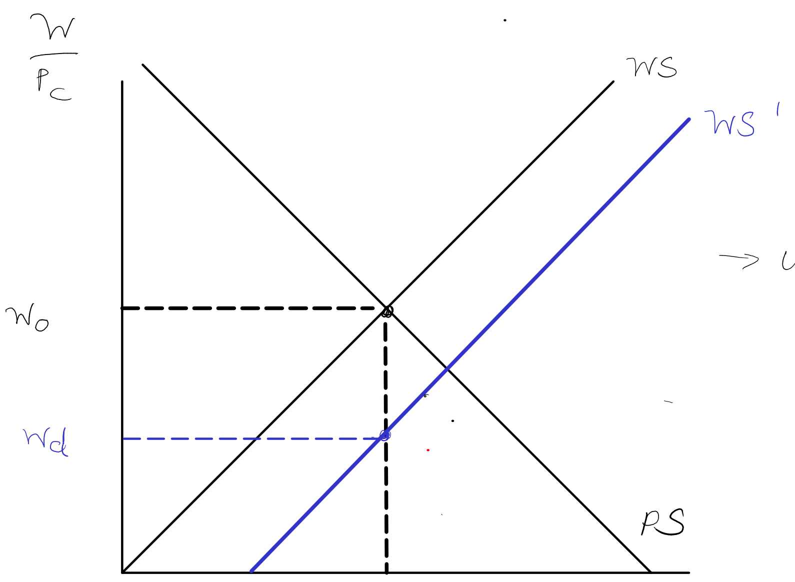
Ambiguous effect - if

$\text{Cov}(\text{Exp}, \text{Gender}) > \text{Cov}(\text{Exp}, \text{Wage})$

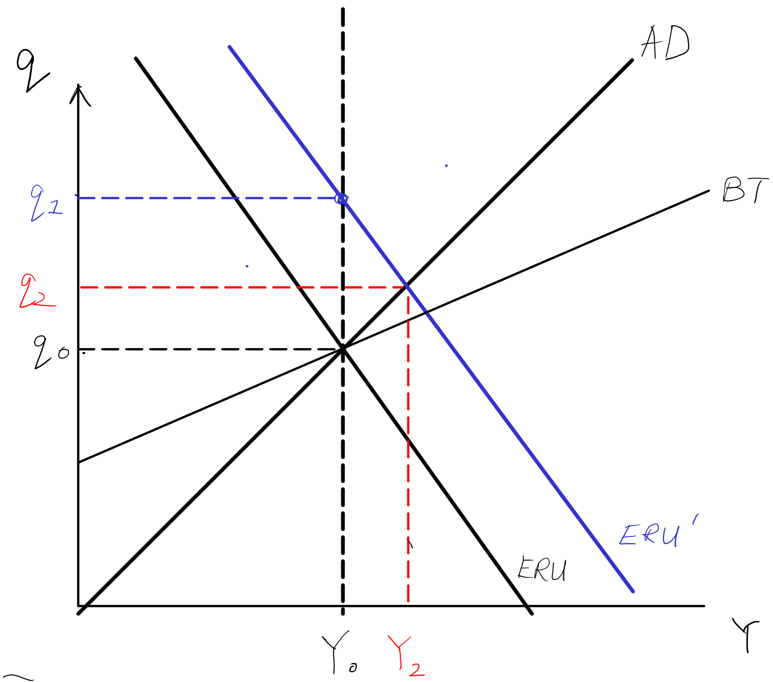
then $se(\hat{\beta}_2^{L}) < se(\hat{\beta}_2^{S})$

□

If experience explains more of wage than it does gender, than se($\hat{\beta}_2$) will come down when adding more regressors.



→ Workers demand lower wages and firms give Wd, but lower prices so inflation falls.



→ In the next period, inflation falls below target and CB lowers rates

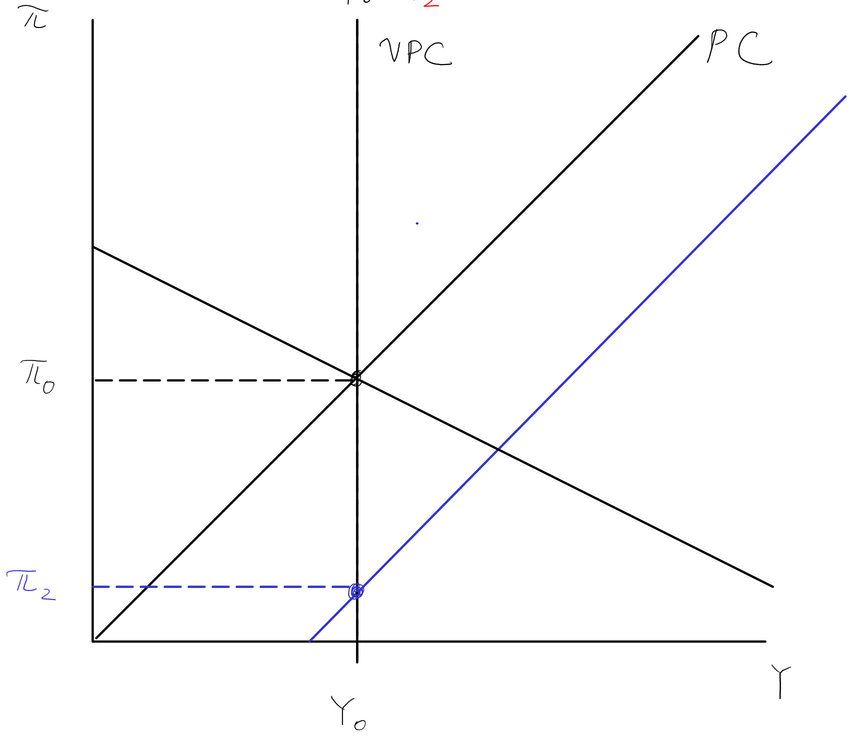
t = 0 after shock:

·workers demand less firms don't raise prices

inf falls.

$q\ \frac{P^{\cancel{*}}e}{P^{\cancel{*}}}$, $P\downarrow$ $q\uparrow$

real depreciation.



Mov