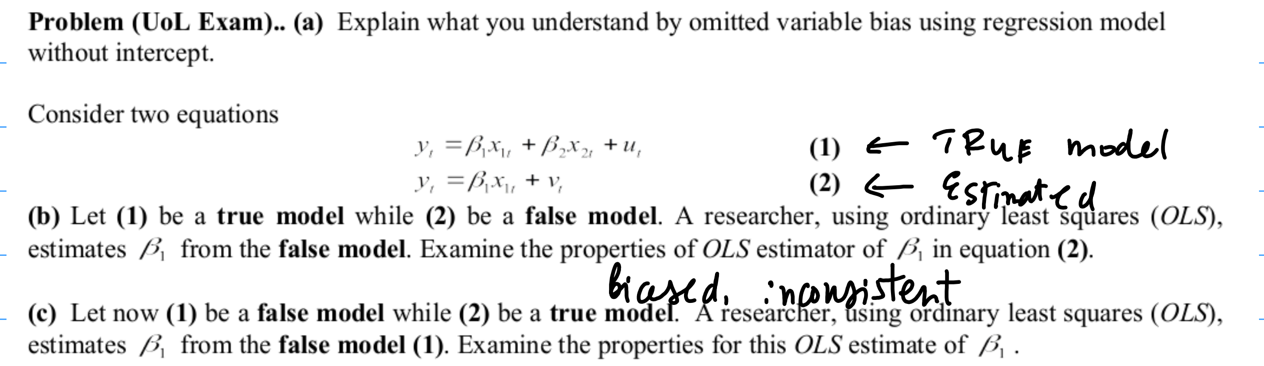
Misspecification.

Omitted variable bias



Problem (UoL Exam).. (a) Explain what you understand by omitted variable bias using regression model without intercept.

Consider two equations

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + u_t \qquad \textbf{(1)} \leftarrow \text{TRUE model}$$

$$y_t = \beta_1 x_{1t} + v_t \qquad \textbf{(2)} \leftarrow \text{Estimated}$$

(b) Let **(1)** be a **true model** while **(2)** be a **false model**. A researcher, using ordinary least squares (*OLS*), estimates β_1 from the **false model**. Examine the properties of *OLS* estimator of β_1 in equation **(2)**.

biased, inconsistent

(c) Let now **(1)** be a **false model** while **(2)** be a **true model**. A researcher, using ordinary least squares (*OLS*), estimates β_1 from the **false model (1)**. Examine the properties for this *OLS* estimate of β_1 .

b) for (2) : $\hat{\beta}_1 = \dfrac{\sum X_{1i} \cdot Y_i}{\sum X_{1i}^2} = \dfrac{\sum X_{1i} \cdot (\beta_1 X_{1i} + \beta_2 \cdot X_{2i} + u_i)}{\sum X_{1i}^2} =$

{Deterministic regressors}

$$= \beta_1 \frac{\cancel{\sum X_{1i}^2}}{\cancel{\sum X_{1i}^2}} + \beta_2 \frac{\sum X_{1i} X_{2i}}{\sum X_{1i}^2} + \frac{\sum X_{1i} \cdot u_i}{\sum X_{1i}^2}$$

$E(u_i) = 0$

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \cdot \underbrace{\frac{\sum X_{1i} \cdot X_{2i}}{\sum X_{1i}^2}}_{\overset{\shortparallel}{\text{bias}}}$$

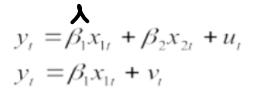
sign bias $\quad \beta_2 > 0 \,, \quad \sum X_{1i} X_{2i} > 0$

$\Rightarrow$ bias > 0

no bias $\quad \beta_2 = 0 \,, \quad X_1^T X_2 = \sum X_{1i} X_{2i} = 0$

orthogonal

c)



$y_t = \hat{\beta}_1 x_{1t} + \beta_2 x_{2t} + u_t$ (1) False

$y_t = \beta_1 x_{1t} + v_t$ (2) True

$\hat{\beta}_1$ unbiased, consistent, inefficient

Problem (ICEF Exam). □ Explain why omitting explanatory variable from the regression equation can lead to the violation of Gauss-Markov conditions so standard errors and all tests become invalid.

□ What are consecuences of omitting explanatory variable from the regression equation for the estimatuion regression equations.

$Y_i = \beta_1 + \beta_2 \cdot X_{2i} + \beta_2 \cdot X_{3i} + u_i$ (TRUE)

X_3 – relevant variable

$cov(X_2, X_3) \neq 0$

$Y_i = \beta_1 + \beta_2 \cdot X_{2i} + u_i^*$ $\qquad u_i^* = \beta_2 \cdot X_{3i} + u_i$

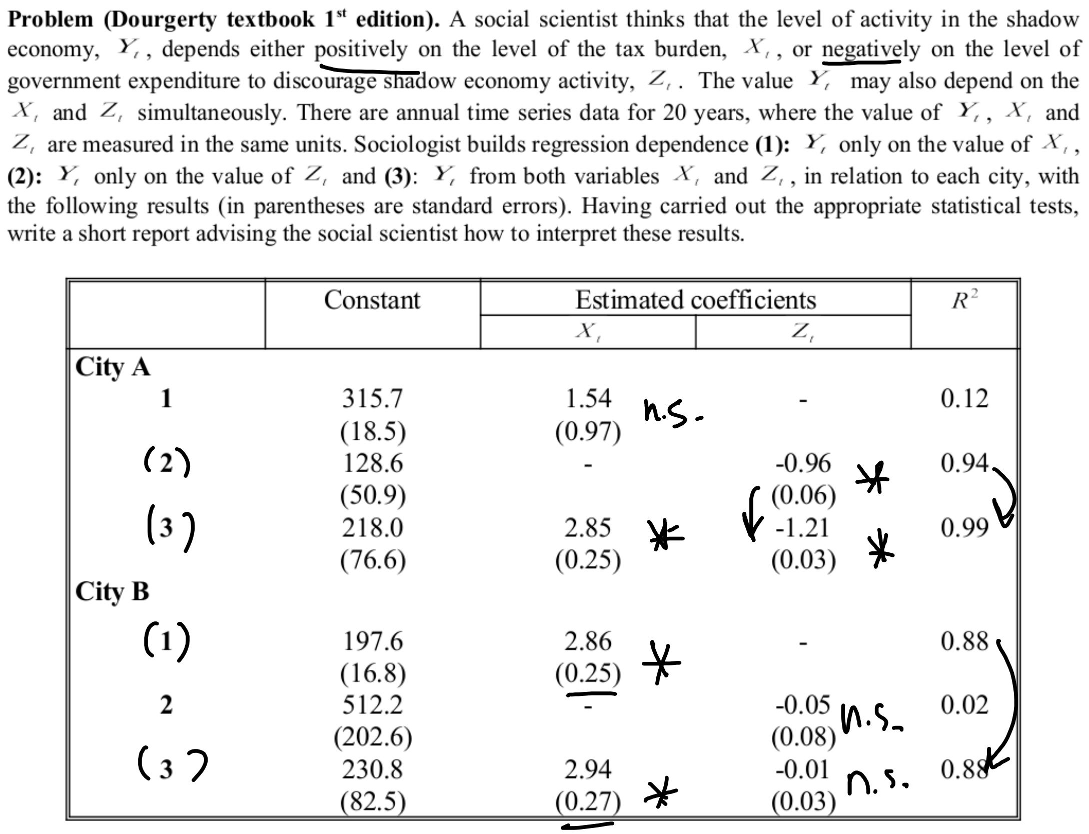
$cov(X_{2i}, u_i^*) \neq 0$

Stochastic Regressors GMT Assumptions:

$\boxed{7}$ $E(X_i u_i) = 0$

X_i – exogenuous

otherwise X_i – endogenuous



Problem (Dourgerty textbook 1st edition). A social scientist thinks that the level of activity in the shadow economy, Y_t, depends either positively on the level of the tax burden, X_t, or negatively on the level of government expenditure to discourage shadow economy activity, Z_t. The value Y_t may also depend on the X_t and Z_t simultaneously. There are annual time series data for 20 years, where the value of Y_t, X_t and Z_t are measured in the same units. Sociologist builds regression dependence **(1):** Y_t only on the value of X_t, **(2):** Y_t only on the value of Z_t and **(3)**: Y_t from both variables X_t and Z_t, in relation to each city, with the following results (in parentheses are standard errors). Having carried out the appropriate statistical tests, write a short report advising the social scientist how to interpret these results.

	Constant	Estimated coefficients		R^2
		X_t	Z_t	
City A				
1	315.7 (18.5)	1.54 n.s. (0.97)	-	0.12
(2)	128.6 (50.9)	-	-0.96 * (0.06)	0.94
(3)	218.0 (76.6)	2.85 * (0.25)	-1.21 (0.03) *	0.99
City B				
(1)	197.6 (16.8)	2.86 (0.25) *	-	0.88
2	512.2 (202.6)	-	-0.05 n.s. (0.08)	0.02
(3)	230.8 (82.5)	2.94 (0.27) *	-0.01 n.s. (0.03)	0.88

City A: model (3) z_i - relevant

City B: model (1) z_i - irrelevant

Question 4 (UoL Exam). Explain the RESET test as a general test for functional form misspecification and discuss the drawbacks and advantages of this test. In your answer consider the following multiple linear regression model:

$$y_i = \gamma_1 + \gamma_2 x_{2i} + \gamma_3 x_{3i} + u_i , \quad i = 1, \ldots, n ,$$

where x_{2i} and x_{3i} are exogenous variables known to affect $E(y_i)$.

1) $\hat{y}_i \Leftarrow y \mid x_2 , X_3$ $\qquad x_{4i} \quad x_4 = x_{2i}^2$

↑ ↑

2) $y_i = \gamma_1 + \gamma_2 \cdot x_{2i} + \gamma_3 x_{3i} + \gamma_4 \hat{y}_4^2 + e_i$

$H_0 : \gamma_4 = 0$

Disadvantages : test doesn't indicate actual form

Advantage : no loss in d.o.f.

address multicollinearity

Question 5 (ICEF Exam).

A researcher has data on output per worker, *y*, and capital per worker, *k*, both measured in thousands of pounds, for 50 firms in the manufacturing sector of the U.K. for 2016. She hypothesizes that output per worker depends on capital per worker and perhaps also the technological sophistication of the firm, *tech*:

$$y = \beta_1 + \beta_2 k + \beta_3 tech + u$$

where *u* is a disturbance term. She is unable to measure *tech* and decides to use expenditure per worker on research and development in 1998, *exp*, as a proxy for it.

Proxy: $\cdot\ corr(exp, tech) \neq 0$

$\cdot\ corr(exp, u) = 0$

(a) □ What do you mean by good or poor proxy?

□ Explain the consequences of using *exp* as a proxy for *tech* if it is a good proxy.

□ Explain the consequences of using *exp* as a proxy for *tech* if it is a poor proxy.

$$\cdot\ tech = \lambda + \mu \cdot exp + \nu$$

$$\cdot\ y = \beta_1 + \beta_2 \cdot k + \beta_3 (\lambda + \mu \cdot exp + \nu) + u$$

$$= (\beta_1 + \beta_3 \cdot \lambda) + \beta_2 k + \underline{\beta_3 \cdot \mu} \cdot exp + u^*$$

The researcher fits the following regressions (standard errors in parentheses):

$$(1)\quad \hat{y} = 1.02 + 0.32k \qquad R^2 = 0.749$$
$$\qquad (0.45)\ \ (0.04)$$

bias > 0

$$(2)\quad \hat{y} = 0.34 + 0.29k + 0.05\exp \qquad R^2 = 0.750$$
$$\qquad (0.61)\ \ (0.22)\ \ (0.15)$$

The correlation coefficient for *k* and *exp* was 0.92.

⇒ exp is poor proxy for tech

(b) □ Discuss these regression results assuming that *y* does depend on both *k* and *tech*.

c) Discuss these regression results assuming that *y* depends only on *k*.

b) $y = \alpha + \beta_1 \cdot k + \beta_2 \cdot tech + u$

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \cdot \frac{Cov(k, tech)}{Var(k)} =$$

$\beta_2 > 0 \qquad Cov(k, tech) > 0$

$\Rightarrow$ bias > 0