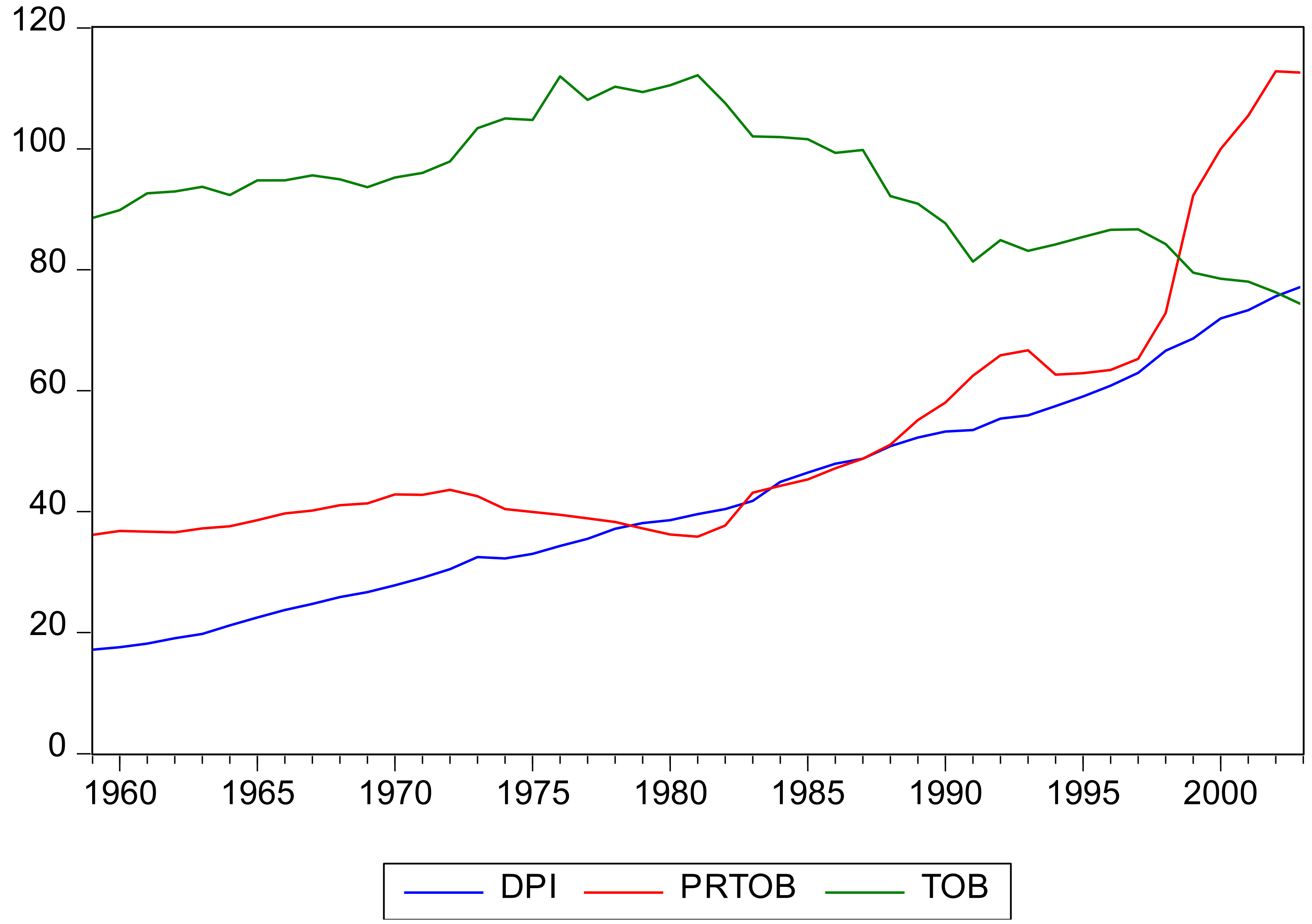
Elements of Econometrics. Lecture 9. Model Misspecification.

FCS, 2022-2023

Data set: Demand; Disposable Personal Income (DPI), Relative Price for Tobacco (PRTOB) and Demand for Tobacco (TOB), 45 annual observations, USA:



Dependent variable: log(TOB), 45 observations:

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	5.483083	0.282674	19.39721	0.0000
LOG(DPI)	-0.114311	0.034158	-3.346537	0.0017
R-squared	0.206632	Mean dependent var		4.538406

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.828287	0.123689	39.03575	0.0000
LOG(DPI)	0.193227	0.025143	7.685108	0.0000
LOG(PRTOB)	-0.483213	0.032874	-14.69888	0.0000
R-squared	0.870876	Mean dependent var		4.538406

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.216060	1.019210	0.211988	0.8332
LOG(DPI)	0.764471	0.127301	6.005233	0.0000
LOG(PRTOB)	-0.385691	0.034577	-11.15448	0.0000
TIME	-0.021276	0.004678	-4.548159	0.0000
R-squared	0.914176	Mean dependent var		4.538406

VARIABLE MISSPECIFICATION

Consequences of variable misspecification			
		True model	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Fitted model	$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2$	**Correct specification, no problems**	*First we consider the case of Omission of a relevant variable.*
	$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$		**Correct specification, no problems**

There are two types of Variable Misspecification: Omission of a relevant variable and Including an irrelevant one.

VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$ – true model $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2$ – fitted model (X_3 omitted)

$$\hat{\beta}_2 = \frac{\sum(X_{2i} - \bar{X}_2)(Y_i - \bar{Y})}{\sum(X_{2i} - \bar{X}_2)^2}$$

$$= \frac{\sum(X_{2i} - \bar{X}_2)\big((\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i) - (\beta_1 + \beta_2 \bar{X}_2 + \beta_3 \bar{X}_3 + \bar{u})\big)}{\sum(X_{2i} - \bar{X}_2)^2}$$

$$= \frac{\sum\big(\beta_2 (X_{2i} - \bar{X}_2)^2 + \beta_3 (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3) + (X_{2i} - \bar{X}_2)(u_i - \bar{u})\big)}{\sum(X_{2i} - \bar{X}_2)^2}$$

$$= \beta_2 + \beta_3 \frac{\sum(X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum(X_{2i} - \bar{X}_2)^2} + \frac{\sum(X_{2i} - \bar{X}_2)(u_i - \bar{u})}{\sum(X_{2i} - \bar{X}_2)^2}$$

We simplify and demonstrate that $\hat{\beta}_2$ has three components: true value $\boldsymbol{\beta_2}$, bias and random component (error term).

VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

$$\mathrm{E}(\hat{\beta}_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2} + \mathrm{E}\left(\frac{\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u})}{\sum (X_{2i} - \bar{X}_2)^2} \right)$$

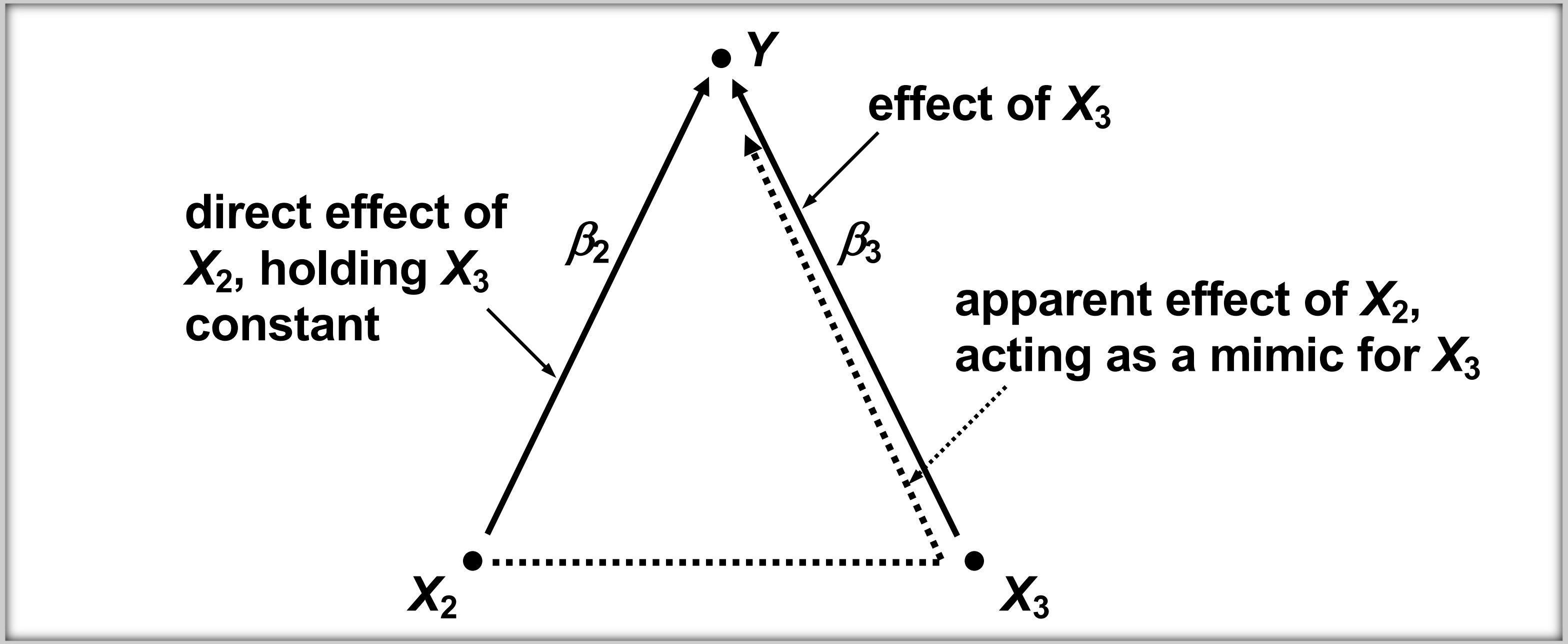
$$\begin{aligned}
\mathrm{E}\left(\frac{\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u})}{\sum (X_{2i} - \bar{X}_2)^2} \right) &= \frac{1}{\sum (X_{2i} - \bar{X}_2)^2} \mathrm{E}\left(\sum (X_{2i} - \bar{X}_2)(u_i - \bar{u}) \right) \\
&= \frac{1}{\sum (X_{2i} - \bar{X}_2)^2} \sum \mathrm{E}\{(X_{2i} - \bar{X}_2)(u_i - \bar{u})\} \\
&= \frac{1}{\sum (X_{2i} - \bar{X}_2)^2} \sum (X_{2i} - \bar{X}_2)\mathrm{E}(u_i - \bar{u}) \\
&= 0
\end{aligned}$$

By Assumption A.3, $\mathrm{E}(u)=0$. It follows that $\mathrm{E}(\bar{u}) = 0$. Hence the expected value of the error term is 0. Thus we have shown that the expected value of $\hat{\beta}_2$ is equal to the true value β_2 plus a bias term. As a consequence of the misspecification, the standard errors, *t* tests and *F* test are invalid.

VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad\qquad \hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2$$

$$\mathrm{E}(\hat{\beta}_2) = \beta_2 + \beta_3 \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sum (X_{2i} - \bar{X}_2)^2}$$



The reason is that, in addition to its direct effect β_2, X_2 has an apparent indirect effect as a consequence of acting as a proxy for the missing X_3.

VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u \qquad \hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2$$

Is the estimator of β_1 *biased?*

If yes, what is the value of bias?

$\mathrm{E}(\hat{\beta}_1) = \beta_1 + \mathrm{bias}(\hat{\beta}_1)$:
do at home before the class!

The intercept may be considered as one more explanatory variable.

It also shows some indirect effect of the omitted variable.

VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE

$$Y = \beta_1 + \beta_2 X_2 + u \qquad \hat{Y} = \hat{\beta}_2 X_2$$

Is the estimator of β_2 *biased?*

If yes, what is the value of bias?

$\mathrm{E}(\hat{\beta}_2) = \beta_2 + \mathrm{bias}(\hat{\beta}_2)$:

do at home before the class!

The intercept may be considered as one more explanatory variable.

If it is missing, then its indirect effect is reflected by another explanatory variable.

VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE, EXAMPLE

$$LGEARN = \beta_1 + \beta_2 S + \beta_3 ASVABC + u$$

Dependent Variable: LGEARN
Method: Least Squares
Included observations: 570

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.777056	0.132301	5.873384	0.0000
S	0.077404	0.010007	7.734779	0.0000
ASVABC	0.012379	0.002662	4.650030	0.0000

R-squared	0.227830	Mean dependent var	2.456463
S.D. dependent var	0.541347	S.E. of regression	0.476537
Sum squared resid	128.7586	F-statistic	83.64712
Durbin-Watson stat	1.728273		

We will illustrate the bias using an Earnings Function (EAEF 40). Assume that in the true model *LGEARN* depends only on *S* and *ASVABC*. Both are highly significant.

VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE, EXAMPLE

$$LGEARN = \beta_1 + \beta_2 S + \beta_3 ASVABC + u \qquad \widehat{LGEARN} = \hat{\beta}_1 + \hat{\beta}_2 S$$

Dependent Variable: LGEARN   Method: Least Squares   Included observations: 570

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.062208	0.119340	8.900668	0.0000
S	0.102202	0.008620	11.85613	0.0000

R-squared	0.198383	Mean dependent var	2.456463
S.D. dependent var	0.541347	S.E. of regression	0.48511
Sum squared resid	133.6689	F-statistic	140.5678
Durbin-Watson stat	1.746617		

$$\mathrm{E}(\hat{\beta}_2) = \beta_2 + \beta_3 \frac{\sum\left(ASVABC_i - \overline{ASVABC}\right)(S_i - \bar{S})}{\sum\left(S_i - \overline{S}\right)^2} =$$

$$= \beta_2 + \beta_3 \frac{\widehat{\mathrm{Cov}}\,(ASVABC, S)}{\widehat{\mathrm{Var}}\,(S)} = \beta_2 + \beta_3 \frac{11.13}{5.556} \approx 0.0774 + 0.0124 * 2 = 0.102$$

Covariance Matrix:

	S	ASVABC
S	5.556122	11.13060
ASVABC	11.13060	78.51596

VARIABLE MISSPECIFICATION I: OMISSION OF A RELEVANT VARIABLE, EXAMPLE

$$LGEARN = \beta_1 + \beta_2 S + \beta_3 ASVABC + u \qquad \widehat{LGEARN} = \hat{\beta}_1 + \hat{\beta}_3 ASVABC$$

Dependent Variable: LGEARN   Method: Least Squares   Included observations: 570

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.280356	0.121012	10.58043	0.0000
ASVABC	0.023352	0.002366	9.868220	0.0000

R-squared	0.146355	Mean dependent var	2.456463
S.D. dependent var	0.541347	S.E. of regression	0.5006
Sum squared resid	142.3446	F-statistic	97.38176
Durbin-Watson stat	1.761189		

$$\mathrm{E}(\hat{\beta}_3) = \beta_3 + \beta_3 \frac{\sum\left(ASVABC_i - \overline{ASVABC}\right)\left(S_i - \overline{S}\right)}{\sum\left(ASVABC_i - \overline{ASVABC}\right)^2} =$$

$$= \beta_3 + \beta_2 \frac{\widehat{\mathrm{Cov}}\,(ASVABC, S)}{\widehat{\mathrm{Var}}\,(ASVABC)} = \beta_3 + \beta_2 \frac{11.13}{78.516} \approx 0.0124 + 0.0774 * 0.142 = 0.0234$$

Covariance Matrix:

	S	ASVABC
S	5.556122	11.13060
ASVABC	11.13060	78.51596

VARIABLE MISSPECIFICATION: OMITTED VARIABLE BIAS

Omitted variable bias conclusion: all estimated coefficients will be biased

$$x_2 = \delta_0 + \delta_1 x_1 + v$$

If x_1 and x_2 are correlated, assume a linear regression relationship between them

$$\Rightarrow \quad y = \beta_0 + \beta_1 x_1 + \beta_2(\delta_0 + \delta_1 x_1 + v) + u$$

$$= \boxed{(\beta_0 + \beta_2\delta_0)} + \boxed{(\beta_1 + \beta_2\delta_1)} x_1 + \boxed{(\beta_2 v + u)}$$

If y is only regressed on x_1 this will be the estimated intercept

If y is only regressed on x_1, this will be the estimated slope on x_1

error term

More general case:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

True model (contains x_1, x_2, and x_3)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + w$$

Estimated model (x_3 is omitted)

- No general statements possible about direction of bias
- Analysis as in simple case if one regressor uncoreelated with others

VARIABLE MISSPECIFICATION II: INCLUSION OF AN IRRELEVANT VARIABLE

Consequences of variable misspecification			
		True model	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Fitted model	$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2$	Correct specification, no problems	Coefficients are biased (in general). Standard errors are invalid.
	$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$		Correct specification, no problems

Now we will investigate the consequences of including an irrelevant variable in a regression model.

VARIABLE MISSPECIFICATION II: INCLUSION OF AN IRRELEVANT VARIABLE

$$Y = \beta_1 + \beta_2 X_2 + u$$

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$$

$$Y = \beta_1 + \beta_2 X_2 + 0 X_3 + u$$

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma_u^2}{\sum (X_{2i} - \bar{X}_2)^2} \times \frac{1}{1 - r^2_{X_2, X_3}}$$

The estimator of β_2 in the multiple regression model is less efficient than the alternative one in the simple regression model. The standard errors remain valid, because the model is formally correctly specified, but they will tend to be larger than those obtained in a simple regression, reflecting the loss of efficiency.

VARIABLE MISSPECIFICATION II: INCLUSION OF AN IRRELEVANT VARIABLE, EXAMPLE

$$LGEARN = \beta_1 + \beta_2 S + \beta_3 ASVABC + u$$

$$\widehat{LGEARN} = \hat{\beta}_1 + \hat{\beta}_2 S + \hat{\beta}_3 ASVABC + \hat{\beta}_4 SF$$

Dependent Variable: LGEARN   Method: Least Squares   Included observations: 570

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.774212	0.133080	5.817655	0.0000
S	0.076872	0.010321	7.448213	0.0000
ASVABC	0.012257	0.002725	4.497968	0.0000
SF	0.001367	0.006397	0.213677	0.8309

R-squared	0.227892	Mean dependent var	2.456463
S.D. dependent var	0.541347	S.E. of regression	0.476939
Sum squared resid	128.7483	F-statistic	55.68611
Durbin-Watson stat	1.730361		

$$\sigma^2_{\hat{\beta}_2} = \frac{\sigma_u^2}{\sum\left(S_i - \overline{S}\right)^2} \times \frac{1}{1 - R_2^2};$$

$$For\ regression\ of\ S\ on\ ASVABC\ and\ SF\ {R_2}^2 = 0.33; \quad \sqrt{\frac{1}{1 - R_2^2}} = 1.22;$$

$$For\ regression\ of\ ASVABCon\ S\ and\ SF\ {R_3}^2 = 0.32; \quad \sqrt{\frac{1}{1 - R_3^2}} = 1.21;$$

VARIABLE MISSPECIFICATION II: INCLUSION OF AN IRRELEVANT VARIABLE

Consequences of variable misspecification

		True model	
		$Y = \beta_1 + \beta_2 X_2 + u$	$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$
Fitted model	$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2$	Correct specification, no problems	Coefficients are biased (in general). Standard errors are invalid.
	$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$	Coefficients are unbiased (in general), but inefficient. Standard errors are valid (in general)	Correct specification, no problems

The coefficients in general remain unbiased, but they are inefficient.

The standard errors remain valid, but are larger than could be.

PROXY VARIABLES

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + u$$

$$X_2 = \lambda + \mu Z$$

$$\begin{aligned} Y &= \beta_1 + \beta_2(\lambda + \mu Z) + \beta_3 X_3 + \ldots + \beta_k X_k + u \\ &= (\beta_1 + \beta_2 \lambda) + \beta_2 \mu Z + \beta_3 X_3 + \ldots + \beta_k X_k \end{aligned}$$

Comparison of regression with *Z* instead of X_2

1. The estimates for β_3, ..., β_κ are the same
2. S.e. and *t* for *the* estimates *of* β_3, ..., β_κ are the same
3. R^2 is the same
4. Impossible to obtain an estimate of β_2, unless μ is known
5. *t* statistic for *Z* is the same as that for X_2
6. Impossible to obtain an estimate of β_1

Suppose that a variable *Y* depends on a set of explanatory variables X_2, ..., X_k, and there are no data on X_2. Regression of *Y* on X_3, ..., X_k would yield biased estimates and invalid standard errors and tests. Suppose that Z is linearly related with X_2 and there is data for Z.

VARIABLE MISSPECIFICATION: UNINTENDED PROXIES

$$\widehat{LGEARN} = \hat{\beta}_1 + \hat{\beta}_2 S + \hat{\beta}_3 ASVABC + \hat{\beta}_4 WEIGHT$$

Dependent Variable: LGEARN Method: Least Squares Included observations: 560

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.395942	0.161431	2.452696	0.0145
S	0.079152	0.010000	7.915092	0.0000
ASVABC	0.011799	0.002649	4.455022	0.0000
WEIGHT	0.002243	0.000505	4.438712	0.0000

R-squared	0.246187	Mean dependent var	2.455518
S.D. dependent var	0.539762	S.E. of regression	0.469897
Sum squared resid	122.7668	F-statistic	60.52769
Durbin-Watson stat	1.804027		

Why WEIGHT variable is significant there? It either actually influences earnings, or acts as a proxy for some omitted variable correlated with it.

VARIABLE MISSPECIFICATION: UNINTENDED PROXIES

$$\widehat{LGEARN} = \hat{\beta}_1 + \hat{\beta}_2 S + \hat{\beta}_3 ASVABC + \hat{\beta}_4 WEIGHT + \hat{\beta}_5 MALE$$

Dependent Variable: LGEARN Method: Least Squares Included observations: 560

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.585710	0.160493	3.649445	0.0003
S	0.075840	0.009744	7.783129	0.0000
ASVABC	0.012058	0.002577	4.679895	0.0000
WEIGHT	0.000442	0.000584	0.757383	0.4491
MALE	0.265646	0.046476	5.715789	0.0000

R-squared	0.288093	Mean dependent var	2.455518
S.D. dependent var	0.539762	S.E. of regression	0.457060
Sum squared resid	115.9419	F-statistic	56.14911
Durbin-Watson stat	1.857755		

The variable WEIGHT acted as a proxy for MALE. Corr(WEIGHT, MALE)=0.54.