Chapter 8 Model Specification

남준우·허인 (2018), 제7장

Gujarati/Porter (2018), Chapter 13

(1) Omission of Relevant Variables

(2) Inclusion of Irrelevant Variables

(3) Decision

(4) Information Criterion

Model and Assumptions(Revisited)

① Model:

$$y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

► X_{ji}: jth variable, ith observation.

► k= # of independent variables including constant term.

② Assumptions

(a) $E(\varepsilon_i) = 0$ for all i.

(b) $V(\varepsilon_i) = \sigma^2$ for all i.

(c) $Cov(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$.

(d) No exact linear relationship among X variables.

(e) Variation in each column of X.

(f) X's are non-random.

◎ Issues of Model Specification:

(a) Choice of Variables: Include or Omit?

(b) Functional Form

(c) Measurement Error

(d) Error Structure

(e) Normality Assumption

(f) Endogeneity of Independent Variables

◎ Inclusuon or Omission of Variables?

► We do not know whether the true model is

(a) $y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$ or (b) $y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + \varepsilon_i$

► Parameter of interest: β_2

- Omit Z or not(insert)? $\Rightarrow$

(a) $y_i = b_1 * + b_2 * X_i + e_i *$ vs. (b) $y_i = b_1 + b_2 X_i + b_3 Z_i + e_i$

► For the effect of variable X, which one will you report, $b_2 *$ or b_2?

- The <u>misspecification error</u> occurs:

① Omission of Relevant Variables

② Inclusion of Irrelevant Variables

► It is known that $V(b_2) = \dfrac{V(b_2{}^*)}{1-\gamma_{XZ}{}^2}$.

(1) Omission of Relevant Variables($\beta_3 \neq 0$)

True model: $y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + \varepsilon_i$ $\Rightarrow$ SRF: $y_i = b_1 + b_2 X_i + b_3 Z_i + e_i$

Assumed model: $y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$, $\Rightarrow$ Estimated: $y_i = b_1{*} + b_2{*} X_i + e_i{*}$

① Unbiased?

$$E(b_2{*}) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})E(y_i)}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \beta_2 + \beta_3 \frac{\sum_{i=1}^{n}(X_i - \bar{X})Z_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \equiv \beta_2 + \beta_3\gamma \quad \neq \beta_2 \qquad \text{if } S_{xz} \neq 0.$$

► Omitted Variable Bias: $E(b_2{*}) - \beta_2 = \beta_3 \dfrac{S_{XZ}}{S_X{}^2}$.

Furthermore,

② <u>Variance?</u>

Since $V(b_2*) = \dfrac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$, $V(b_2) = \dfrac{V(b_2*)}{1-\gamma_{XZ}{}^2}$,

$$E\left(\hat{V}(b_2*)\right) = E\left(\sigma_{b_2}*^2\right) \neq V(b_2), \quad \text{where} \quad \hat{V}(b_2*) = \frac{s^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

► Inferences are invalid.

(<u>Special case</u>) $\gamma_{XZ} = 0$.

(Example) $Fin_i = \beta_1 + \beta_2 Hedu_i + \beta Wedu_i + \beta_4 KL6_i + \varepsilon_i$

- Parameter of interest: β_2.

► Omit Wedu?

► Omit KL6?

(2) Inclusion of Irrelevant Variables$(\beta_3 = 0)$

True model: $y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$ $\Rightarrow$ SRF: $y_i = b_1 * + b_2 * X_i + e_i *$

Assumed model: $y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + \varepsilon_i$, $\Rightarrow$ Estimated: $y_i = b_1 + b_2 X_i + b_3 Z_i + e_i$

► $b_2 = \dfrac{S_{Xy} S_Z{}^2 - S_{Zy} S_{XZ}}{S_X{}^2 S_Z{}^2 - S_{XZ}{}^2}$

① Unbiased?

$E(b_2) = \beta_2$.

② <u>Efficient?</u>

$$V(b_2) = \frac{V(b_2^*)}{1-\gamma_{XZ}^{\;2}} \text{ ,}$$

$$V(b_2) \;\; \geq \;\; V(b_2^*) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2},$$

► If $\gamma_{XZ} \neq 0$, $V(b_2) \geq V(b_2^*)$.

- Inclusion of irrelevant variables $\Rightarrow$ unbiased but inefficient.

► As $|\gamma_{XZ}|\uparrow$, $V(b_2)\uparrow$, t-ratio of $b_2\downarrow$.

► Reminds for the multicollinearity.

► Sometimes, infact, b_2 is significant, but high $|\gamma_{XZ}|$ makes it insignificant.

(3) Decision: Mean Squared Error(MSE)

- Two misspecified cases.

► Consider MSE and prefer smaller MSE.

① Omitted Relevant Variables(b_2*)

$$MSE(b_2*) = V(b_2*) + Bias(b_2*)$$

$$= \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} + \left(\gamma\beta_3\right)^2$$

$$= \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} + \left(\beta_3 \frac{S_{XZ}}{S_X{}^2}\right)^2$$

② <u>Inclusion of Irrelevant Variables</u>(b_2)

$$MSE(b_2) = V(b_2) + Bias(b_2)$$

$$= \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2(1-\gamma_{XZ}{}^2)}.$$

③ <u>Decision</u>

- $MSE(b_2{}^*) >< MSE(b_2)$ depends on size of β_3 and γ_{XZ}.

► If $|\beta_3|$ is high, $MSE(b_2{}^*) > MSE(b_2)$.

$\Rightarrow$ Prefer b_2.

$\Rightarrow$ Include X_3.

► If $|\gamma_{XZ}|$ is high, $MSE(b_2{}^*) < MSE(b_2)$.

$\Rightarrow$ Prefer $b_2{}^*$.

$\Rightarrow$ Omit X_3.

(4) Information Criterion

► Omit or not?

► Use X or Z for certain variables?

- Consider both explanatory power and size (k).

① $\bar{R}^2$

► Prefer model of higher $\bar{R}^2$.

② Akaike Information Criterion(AIC):

$$AIC = \left(\frac{\sum_{i=1}^{n} e_i^{\,2}}{n} \right) \cdot e^{2k/n} \qquad \text{or} \qquad \log\left(\frac{\sum_{i=1}^{n} e_i^{\,2}}{n} \right) + \frac{2k}{n}.$$

► Prefer model of smaller AIC.

③ Schwarz Information Criterion(SC):

$$SC = \left(\frac{\sum_{i=1}^{n} e_i^{\,2}}{n} \right) \cdot n^{k/n} \qquad \text{or} \qquad \log\left(\frac{\sum_{i=1}^{n} e_i^{\,2}}{n} \right) + \frac{k}{n}\log(n).$$

► Prefer model of smaller SC.

- Conflict across criteria is possible.

(Examples) Artprice file

Model 1: OLS, using observations 1-250
Dependent variable: logprice

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
const	−%s	2.02697	−%#.4g	0.1092	
AGE	0.172248	0.0615860	2.797	0.0056	***
ARD	0.0885329	0.0385425	2.297	0.0225	**
EXB	−%s	0.00256821	−%#.4g	0.7290	
LIFE	0.363957	0.170114	2.139	0.0334	**
SIZE	0.0288943	0.00481476	6.001	<0.0001	***
sq_AGE	−%s	0.000458325	−%#.4g	0.0153	**
sq_SIZE	−%s	1.43586e-05	−%#.4g	<0.0001	***

Mean dependent var	3.909161	S.D. dependent var	1.304825
Sum squared resid	328.2114	S.E. of regression	1.164580
R-squared	0.225806	Adjusted R-squared	0.203412
F(7, 242)	10.08328	P-value(F)	4.64e-11
Log-likelihood	−388.7593	Akaike criterion	793.5185
Schwarz criterion	821.6902	Hannan-Quinn	804.8568

(Example) 공사유형 재분류에 따른 회귀분석 추정 결과(n=274)

log(낙찰률)	(1)	(2)	(3)	(4)
log(설계금액)	-0.0056	-0.00481	-0.00131	-0.00165
종합심사낙찰제	0.0372	0.0364	0.0401*	0.0401*
일괄입찰	0.1091***	0.109***	0.125***	0.126***
대안입찰	0.0986**	0.102**	0.120***	0.120***
철도시설공단	0.0268	0.0252	0.0294	0.0292
공사유형분류	분류 (1)	분류 (2)	분류 (3)	분류 (4)
log(CBSI)	0.0692***	0.0674***	0.0613***	0.0617***
실업률	-0.0139	-0.0129	-0.0130	-0.0135
분기=2	-0.0253*	-0.0249*	-0.0240	-0.0242*
분기=3	-0.0570***	-0.0566***	-0.0575***	-0.0579***
분기=4	-0.0032	-0.00322	-0.00333	-0.00337
log(입찰참가자수)	-0.0535***	-0.0536***	-0.0460***	-0.0458***
절편	-0.2448	-0.258	-0.312	-0.302
관찰치 수	274	274	274	274
R-squared	0.6623	0.6619	0.6538	0.6538
Adj R-Squared	0.6267	0.6351	0.6365	0.6365
AIC	-563.27	-574.94	-582.48	-582.48
BIC	-465.72	-499.06	-531.90	-531.89

***, **, *는 각각 1%, 5%, 10% 수준에서 통계적으로 유의함을 나타낸다.