Spesifikasi Model: Pemilihan Variabel-variabel Bebas

Imam Awaluddin

Jurusan Ekonomi Pembangunan
Fakultas Ekonomi dan Bisnis
Universitas Lampung
2024

Spesifikasi Model: Pemilihan Variabel-variabel Bebas

Spesifikasi Model

- **Salah satu asumsi klasik adalah model regresi yang digunakan dalam analisis telah dispesifikasikan dengan "benar".**
- Jika tidak dispesifikasikan dengan "benar" maka terjadi bias spesifikasi model.
- Namun pencarian spesifikasi model yang "benar" seperti mencari Cawan Suci (the Holy Grail).

Menentukan Persamaan Ekonometrik dan Kesalahan Spesifikasi

- Sebelum persamaan apa pun dapat diestimasi, persamaan harus dispesifikasikan secara lengkap
- **Spesifikasi** suatu persamaan ekonometrika terdiri dari **tiga bagian**, dikatakan pemilihan secara benar dalam:
  - Variabel-variabel bebas
  - Bentuk fungsional
  - Bentuk dari *stochastic error term*
- Suatu **kesalahan spesifikasi** disebabkan karena satu atau beberapa pilihan ini dilakukan **tidak tepat**

Kriteria Pemilihan Model

- Data harus tersedia à prediksi dari model harus secara logis dimungkinkan
- Konsisten dengan teori.
- Memiliki regresor-regresor eksogen yang lemah à regresor harus tidak berkorelasi dengan faktor kesalahan.
- Memiliki parameter yang konstan à parameter harus stabil.
- Memiliki data yang koheren à residual yang diestimasi harus murni acak (*white noise*)
- Menyeluruh à model harus menyertakan semua model tandingan, model lain bukan sebagai perbaikan model yang dipilih.

Bias-bias Kesalahan Spesifikasi

- Menghilangkan variabel yang relevan (*omitted variable(s)*) à ***underfitting a model***.
- Menyertakan variabel yang tidak relevan (*irrelevant variabel(s)*) à ***overfitting a model***.
- Bentuk fungsi yang salah
- Kesalahan pengukuran

$Q = AK^{\alpha}L^{\beta}$

$\ln Q = \ln A + \alpha \ln K + \beta \ln L$

Kesalahan-kesalahan Spesifikasi model

- Spesifikasi yang salah dari faktor kesalahan stokastik
- Asumsi bahwa faktor kesalahan terdistribusi normal

$Q = AK^{\alpha}L^{\beta} \rightarrow \alpha + \beta = 1$ CRS

$\ln Q = \ln A + \alpha \ln K + \beta \ln L$

$\ln Q = \beta_0 + \beta_1 \ln K + \beta_2 \ln L + \varepsilon$

Underfitting a Model (Ommited Variabel(s))

Model yang benar:

$$Y_i = b_1 + b_2 X_{2i} + b_3 X_{3i} + u_i$$

Model yang digunakan menghilangkan X_3:

$$Y_i = a_1 + a_2 X_{2i} + v_i$$

Apa konsekuensi menghilangkan X_3?

$v = \beta_3 X_3 + u$

Underfitting a Model (Ommited Variabel(s))

Konsekuansi menghilangkan X_3:

1. Jika variabel X_3 berkorelasi dg X_2 mk koefisien korelasi kedua variabel r_{23} tidak nol $\hat{a}_1$ dan $\hat{a}_2$ menjadi bias dan tidak konsisten.
2. Jika variabel X_3 tidak berkorelasi dg X_2 $\hat{a}_1$ bias, meskipun $\hat{a}_2$ tidak lagi bias.
3. Varians dari faktor gangguan s^2 tidak terestimasi secara benar.
4. Varians $\hat{a}_2$ merupakanestimator yang bias dari varians estimator sebenarnya $\hat{b}_2$.
5. Kesimpulan signifikansi statistik parameter tidak valid.
6. Forecast tidak dapat diandalkan.

Contoh

Dengan menggunakan data Table 6.1 Buku Gujarati dengan variabel-variabel:

- CM = angka kematian balita
- PNPB = PNB per kapita
- FLR = tingkat kepandaian membaca Wanita

Proxy

Misal model yang "benar" adalah:

$$\mathrm{CM}_i = \beta_1 + \beta_2\mathrm{PNBP}_i + \beta_3\mathrm{FLR}_i + u_i$$

Hasil estimasinya adalah sebagai berikut:

Hasil Estimasi Model yang "benar"

Dependent Variable: CM
Method: Least Squares
Date: 06/08/12 Time: 07:38
Sample: 1 64
Included observations: 64

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	263.6416	11.59318	22.74109	0.0000
FLR	-2.231586	0.209947	-10.62927	0.0000
PGNP	-0.005647	0.002003	-2.818703	0.0065

R-squared	0.707665	Mean dependent var	141.5000
Adjusted R-squared	0.698081	S.D. dependent var	75.97807
S.E. of regression	41.74780	Akaike info criterion	10.34691
Sum squared resid	106315.6	Schwarz criterion	10.44811
Log likelihood	-328.1012	F-statistic	73.83254
Durbin-Watson stat	2.186159	Prob(F-statistic)	0.000000

Kita lihat koefisien PGNP adalah sebesar -0.005647

Contoh Ommited Variable Bias

Jika kita menghilangkan variabel FLR, sehingga modelnya menjadi:

$$\mathrm{CM}_i = \beta_1 + \beta_2\mathrm{PNBP}_i + u_i$$

Hasil estimasinya adalah sebagai berikut:

Hasil Estimasi Model yang "salah"

Dependent Variable: CM
Method: Least Squares
Date: 06/08/12 Time: 07:46
Sample: 1 64
Included observations: 64

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	157.4244	9.845583	15.98935	0.0000
PGNP	-0.011364	0.003233	-3.515661	0.0008

R-squared	0.166217	Mean dependent var	141.5000
Adjusted R-squared	0.152769	S.D. dependent var	75.97807
S.E. of regression	69.93413	Akaike info criterion	11.36374
Sum squared resid	303228.5	Schwarz criterion	11.43120
Log likelihood	-361.6396	F-statistic	12.35987
Durbin-Watson stat	1.931458	Prob(F-statistic)	0.000826

Kita lihat koefisien PGNP menjadi sebesar -0.011364

-0,005 V

Contoh: Tabel 6.4

Dependent Variable: FLR
Method: Least Squares
Date: 06/08/12 Time: 07:52
Sample: 1 64
Included observations: 64

1,96.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	47.59716	3.555330	13.38755	0.0000
PGNP	0.002562	0.001167	2.195027	0.0319

R-squared	0.072108	Mean dependent var	51.18750
Adjusted R-squared	0.057142	S.D. dependent var	26.00786
S.E. of regression	25.25385	Akaike info criterion	9.326586
Sum squared resid	39540.94	Schwarz criterion	9.394051
Log likelihood	-296.4508	F-statistic	4.818143
Durbin-Watson stat	2.081666	Prob(F-statistic)	0.031915

Parsimony.

Overfitting a Model (Irrelevant Variabel (S))

à Memasukkan variabel yang tidak relevan.

Misalkan model yang benar adalah:

(1) $$Y_i = b_1 + b_2 X_{2i} + u_i$$

Namun kita gunakan model berikut:

(2) $$Y_i = a_1 + a_2 X_{2i} + a_3 X_{3i} + v_i$$

à X_3 merupakan variabel yang tidak perlu.

Overfitting a Model (Irrelevant Variabel (S))

Konsekuensi:

- Estimator-estimator OLS dari parameter2 dari model yang "tidak benar" semuanya tidak bias dan konsisten
- Error variance diestimasi dengan benar ✓
- Interval kepercayaan dan prosedur pengujian hipotesis tetap valid $SE(\hat{\beta})\uparrow \longrightarrow t_{\hat{\beta}_i} = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$
- Namun parameter yang diestimasi tidak efisien (varians dari α yang diestimasi lebih besar dari varians β model sebenarnya). → $\cancel{B}$LUE

Contoh

Dengan menggunakan data "Moody" buku Studenmund:

Model yang "benar" adalah:

$$Y_i = \beta_0 + \beta_N N_i + \beta_P P_i + \beta_I I_i + \epsilon_i$$

- Di mana: Y_i adalah pendapatan masing-masing restoran Woody ke-i, N adalah jumlah pesaing dalam radius 2 mil dari restoran Woody, P adalah jumlah penduduk yang tinggal dalam radius 3 mil dari restoran Woody, dan I adalah pendapatan rata-rata rumah tangga yang diukur dalam variabel P.
- Hasil estimasinya adalah sebagai berikut:

Contoh

$$\hat{Y}_i = 102,192 - 9075N_i + 0.355P_i + 1.288I_i$$
$$(2053) \quad (0.073) \quad (0.543)$$
$$t = -4.42 \quad 4.88 \quad 2.37$$
$$N = 33 \quad \bar{R}^2 = .579$$

- Misalkan kita menambahkan suatu variabel yang disebut A_i yaitu angka tiga digit terakhir alamat restoran Woody ke-i, dan hasilnya adalah sebagai berikut:

Contoh irrelevant variable

$$\hat{Y}_i = 98,125 - 8975N_i + 0.360P_i + 1.301I_i + 58.07A_i$$
$$(2082) \quad (0.074) \quad (0.550) \quad (95.21)$$
$$t = -4.31 \quad +4.86 \quad +2.37 \quad +0.61$$
$$N = 33 \quad \bar{R}^2 = .569$$

1,96

- Dari hasil tersebut apakah variabel A_i adalah variabel yang penting?
- Hubungkan relevansinya dengan teori
- Lihat nilai adjusted R^2, dan terutama adalah nilai t koefisien variabel A.

Omitted Variables

- Two reasons why an important explanatory variable might have been left out:
  - we forgot…
  - it is not available in the dataset, we are examining
- Either way, this may lead to **omitted variable bias** (or, more generally, **specification bias**)
- The reason for this is that when a variable is not included, it **cannot** be **held constant**
- Omitting a relevant variable usually is evidence that the entire equation is a suspect, because of the likely bias of the coefficients.

The Consequences of an Omitted Variable

- Suppose the true regression model is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \qquad (6.1)$$

  Where ϵ_i is a classical error term
- If X_2 is omitted, the equation becomes instead:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i^* \qquad (6.2)$$

  Where:

$$\epsilon_i^* = \epsilon_i + \beta_2 X_{2i} \qquad (6.3)$$

- Hence, the explanatory variables in the **estimated** regression (6.2) are **not independent** of the error term (unless the omitted variable is **uncorrelated** with **all** the **included** variables—something which is **very unlikely**)
- But this **violates Classical Assumption III**!

The Consequences of an Omitted Variable (cont.)

- What happens if we estimate Equation 6.2 when Equation 6.1 is the truth?
- We get **bias**!
- What this means is that:

$$E(\hat{\beta}_1) \neq \beta_1 \qquad (6.4)$$

- The **amount** of bias is a function of the impact of the **omitted** variable on the dependent variable times a function of the **correlation** between the **included** and the **omitted** variable
- Or, more formally:

$$\text{Bias} = \beta_{om} f(r_{in,om}) \qquad (6.7)$$

- So, the **bias exists unless**:
  1. the true coefficient equals zero, or
  2. the included and omitted variables are uncorrelated

Correcting for an Omitted Variable

- In theory, the solution to a problem of specification bias seems easy: add the omitted variable to the equation!
- Unfortunately, that's easier said than done, for a couple of reasons
  1. Omitted variable bias is **hard to detect**: the **amount** of bias introduced can be **small** and not immediately detectable
  2. Even if it has been decided that a given equation is suffering from omitted variable bias, how to decide exactly **which** variable to include?
- Note here that **dropping** a variable is **not** a viable strategy to help cure omitted variable bias:
  – If anything you'll just generate even **more** omitted variable bias on the remaining coefficients!

Correcting for an Omitted Variable (cont.)

- What if:
  – You have an **unexpected result**, which leads you to believe that you have an **omitted variable**
  – You have two or more **theoretically sound** explanatory variables as potential "candidates" for inclusion as the omitted variable to the equation is to use
- How do you **choose** between these variables?
- One possibility is **expected bias analysis**
  – **Expected bias:** the likely bias that omitting a particular variable would have caused in the estimated coefficient of one of the included variables

Correcting for an Omitted Variable (cont.)

- Expected bias can be estimated with Equation 6.7:

$$\text{Expected bias} = \beta_{om} \cdot f(r_{in,om}) \qquad (6.7)$$

- When do we have a viable candidate?
  - When the **sign** of the **expected bias** is the **same** as the **sign** of the **unexpected result**
- Similarly, when these signs **differ**, the variable is **extremely unlikely** to have caused the unexpected result

Irrelevant Variables

- This refers to the case of including a variable in an equation when it does **not** belong there
- This is the **opposite** of the **omitted** variables case—and so the impact can be illustrated using the same model
- Assume that the true regression specification is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i \qquad (6.10)$$

- But the researcher for some reason includes an extra variable:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i^{**} \qquad (6.11)$$

- The misspecified equation's error term then becomes:

$$\epsilon_i^{**} = \epsilon_i - \beta_2 X_{2i} \qquad (6.12)$$

Irrelevant Variables (cont.)

- So, the inclusion of an irrelevant variable will **not cause bias** (since the true coefficient of the irrelevant variable is zero, and so the second term will drop out of Equation 6.12)
- However, the inclusion of an irrelevant variable will:
  - **Increase** the **variance** of the estimated coefficients, and this increased variance will tend to **decrease** the absolute magnitude of their **t-scores**
  - **Decrease** the $\bar{R}^2$ (but not the R^2)
- Table 6.1 summarizes the consequences of the omitted variable and the included irrelevant variable cases (unless $r_{12} = 0$)

Table 6.1 Effect of Omitted Variables and Irrelevant Variables on the Coefficient Estimates

Table 6.1 Effect of Omitted Variables and Irrelevant Variables on the Coefficient Estimates

Effect on Coefficient Estimates	Omitted Variable	Irrelevant Variable
Bias	Yes	No
Variance	Decreases	Increases

Four Important Specification Criteria

- We can summarize the previous discussion into four criteria to help decide whether a given variable belongs in the equation:
  1. **Theory:** Is the variable's place in the equation unambiguous and theoretically sound?
  2. **t-Test:** Is the variable's estimated coefficient significant in the expected direction?
  3. $\bar{R}^2$: Does the overall fit of the equation (adjusted for degrees of freedom) improve when the variable is added to the equation?
  4. **Bias:** Do other variables' coefficients change significantly when the variable is added to the equation?
- If **all** these conditions hold, the variable **belongs** in the equation
- If **none** of them hold, it does **not belong**
- The tricky part is the **intermediate cases**: use sound judgment!

Specification Searches

- **Almost any result** can be obtained from a given dataset, by simply specifying different regressions until estimates with the desired properties are obtained
- Hence, the **integrity** of all empirical work is open to question
- To counter this, the following three points of **Best Practices in Specification Searches** are suggested:
  1. Rely on **theory** rather than statistical fit as much as possible when choosing variables, functional forms, and the like
  2. **Minimize the number of equations** estimated (except for **sensitivity analysis**, to be discussed later in this section)
  3. **Reveal**, in a footnote or appendix, **all alternative specifications** estimated

Sequential Specification Searches

- The sequential specification search technique allows a researcher to:
  - Estimate an undisclosed number of regressions
  - Subsequently present a final choice (which is based upon an unspecified set of expectations about the signs and significance of the coefficients) as if it were only a specification
- Such a method **misstates** the **statistical validity** of the regression results for two reasons:
  1. The **statistical significance** of the results is **overestimated** because the estimations of the previous regressions are ignored
  2. The **expectations** used by the researcher to choose between various regression results **rarely**, if ever, are **disclosed**

Bias Caused by Relying on the *t*-Test to Choose Variables

- Dropping variables **solely** based on **low t-statistics** may lead to two different types of errors:

  1. An **irrelevant** explanatory variable may sometimes be included in the equation (i.e., when it does **not belong** there)

  2. A relevant explanatory variables may sometimes be dropped from the equation (i.e., when it **does** belong)
- In the first case, there is **no bias** but in the second case there **is bias**
- Hence, the estimated coefficients will be **biased** every time an **excluded** variable belongs in the equation, and that **excluded** variable will be **left out** every time its estimated coefficient is not statistically significantly different from zero
- So, we will have **systematic bias** in our equation!