MODEL SPECIFICATIONS

DR. SUBRATA ROY

Associate Professor

Department of Commerce, MGCUB

Before estimating a regression model, model specification should be checked. Specification errors in the regression model estimation may result series problem with robustness and reliability of the estimation. So specification checking is constructed under three subtitles which are choosing **independent variable, functional form and stochastic term**.

Choosing Independent Variables

The researchers define the independent/exogenous variables based on the economic theory, experience, common sense, critical thinking and research objectives. The independent variables which are chosen should satisfy the estimation assumptions. So any violation from those assumptions may cause unreliable and incorrect findings/estimation.

If the economic theory requires including one or more independent variables in the regression equation, then they should be definitely included in the model. So leaving a related variable outside of the model can cause estimation bias and including irrelevant variables in the regression equation leads to higher variances of the estimation. It is very difficult to decide whether to include or exclude an independent variable in the regression equation, if it is not included in the economic theory.

Omitting a relevant variable:

Generally, data unavailability and researchers' ignorance may cause a regression equation without relevant independent variable. Omitted variables bias is the bias resulting from leaving a relevant independent variable out of the model. In a multiple regression equation the β_k (beta) coefficient represents the amount of changes in the dependent variable when independent variable supposes x_k change one unit holding constant the other independent variables in the equation. If a relevant variable is excluded, it is not held constant for the calculation and interpretation of coefficient β_k. Because omission of a relevant independent variable may cause bias pushing the expected value of the estimated coefficient out from the true value of the population coefficient.

Assume that the independent variable z_i is omitted from the regression equation.

$$y_i = \alpha_0 + \beta_1 x_i + e_i \qquad 1$$

The error term includes the omitted independent variable as follows:

$$e_i = \beta_2 z_i + \varepsilon_i \qquad 2$$

then assumption $\boldsymbol{E(e_i) = \beta_2 z_i \neq 0}$ does not hold and the estimation will be biased.

Also, if xi and zi are correlated then the efficiency assumption $\boldsymbol{cov(e_i,x_i) \neq 0}$ does not hold.

Thus, OLS overestimate the coefficient of independent variable $\boldsymbol{\beta_1}$.

$$E(\hat{\beta}_1^*) \neq \beta_1$$

So omission of a closely related independent variable implies that Ordinary Least Square (OLS) is no longer BLUE (Best Linear Unbiased Estimator).

If a relevant variable is omitted from the regression equation:

i. Estimate of the coefficient of that variable is absent

ii. The coefficients of the included independent variables are likely to be biased

If the omitted variable is not correlated with the included independent variable and the coefficient of omitted independent variable is ***zero***, then ***OLS is BLUE***.

How will you detect?

i. Check the economic theory to identify relevant independent variable

ii. Check the compatibility of the sign of estimated coefficients with the related economic theory

WALD TEST (Coefficient restrictions)

The Wald test (Wald, 1943) statistics is the estimation of unrestricted regression without imposing the coefficient restrictions specified by the null hypothesis. It measures how close the unrestricted estimation satisfies the restrictions under the null hypothesis. The unrestricted estimation satisfies the restrictions if the imposed restrictions are true.

The Wald test statistic is computed as under:

$y = X\beta + e$, is a linear model 3

Linear restrictions are as follows:

H_0: $R\beta - r = 0$

R: q * k matrix

R: a q vector

Wald statistics:

$W = (Rb - r)^1(s^2R(X^1X)^{-1}R^1(Rb - r), \chi^2(q)$ 4

If it is assumed that errors are independent and identically normally distributed, ***F-statistic*** can be computed as under:

$$F' = \frac{(\bar{u}^I\bar{u} - u^Iu)/q}{u^Iu/(N-K)} = \frac{W}{q} \quad 5$$

$\bar{u}$ **:** Residual vector of the restricted regression. ***F-statistic*** does compare the residual sum of squares computed with and without the imposed restrictions. ***F-statistic*** is small if the restrictions are true and there is little difference in two residual sums of squares.

Omitted Variables Test

This test gives to add one or more variables to an existing equation and to test whether the set makes a significant contribution in explaining the variation in the dependent variable. The null hypothesis in this case is:

H_0: the additional variable or variables are not significant

The F-statistic is based on the difference between the residual sums of squares of the restricted and unrestricted regressions.

Here, one thing is kept into mind that number of observation of original and test equation should be equal in

order to employ the omitted variables test.

What are the solutions then?

i. Expected bias analysis may be applied. Omitting a variable may have caused it in the estimated coefficient of one of the independent variable in the model. Estimated bias can be estimated as under:

***Expected bias* = β_{om}. *f*($r_{in.om}$)**

ii. $r_{in.om}$ denotes the ***correlation coefficient*** between included (***in***) and omitted (***om***) variable in the regression. Sign of expected bias should be checked with the sign of the unexpected result. ***If they are same***, then the variable could be the reason of ***bias***. This analysis should be used if there is obviously a bias.

iii. If the omitted variable is obvious/evident and available, it should be included in the model.

iv. If the omitted variable is not available, a proxy variable should be found which is closely related to the omitted variable and included in the model.

What happens if omitted variable is included in the regression model?

Including an ***irrelevant variable*** in the regression equation may cause an increase in the variables of the estimated coefficients of included independent variables. It may ***reduce the precision of the estimation*** and does not cause bias.

Suppose that an independent variable $\mathbf{z_i}$ is not related or it is obvious that it has no relationship with the dependent variable such as:

$$y_i = \alpha_0 + \beta_1 x_i + e_i \quad (6)$$

$$y_i = \alpha_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i \quad (7)$$

If z_i is not related, $\beta_2 = 0$. Hence all the assumption would be satisfied then ***OLS*** estimators are not biased and remain ***BLUE***.

Standard errors on the estimated coefficients are larger in the model which includes irrelevant ***dependent variable*** than the optimum model. It must be omitted from the regression model.

How to choose correct variables?

i. Following the Economic theory

ii. Check whether it is significant with correct sign

iii. Has $\bar{R}^2$ improved

iv. Check the other coefficients sign (+ or -), after the variable is included.

If all the above ***four conditions are satisfied***, then ***variable belongs to the regression*** equation. Following techniques are strongly recommended to be employed for ***specification bias*** problem:

a. ***Scanning to develop a tesTable theory***: It is about analyzing a data set for the purpose of developing a ***tesTable*** theory or hypothesis an economic theory or hypothesis should have been tested on a different data set before giving reference to theory or hypothesis.

b. ***Sensitivity analysis:*** It refers to employing different ***alternative specifications*** to determine whether the estimation result is robust. How sensitive an estimation result is to a change in different specifications should

be examined.

Redundant variable test

This test allows testing for statistical significance of a subset of the included variables. It tests whether the coefficients of the variables in a regression equation ***are zero***. If they are equal to zero, they should be omitted from the equation.

CHOOSING A FUNCTIONAL FORM

To specify the classical linear regression model, a specific functional form should be chosen. Any functional form that is linear in parameters can be chosen. If the incorrect functional form is chosen, then the model is mis-specified. If the model is mis-specified then it may not be a reasonable approximation of the true data generation process. We make a functional form specification error when we choose the wrong functional form.

Constant term should be included in the regression model unless there is some strong reason for opposite such as the data is in the close neighborhood. Not including a constant term causes ***inflated t-ratio***.

Functional Forms:

The Log-Log Regression model (Double log):

$\ln y_i = \alpha_0 + \beta_1 \ln x_{1i} + \beta_2 \ln x_{2i} + \ldots\ldots\ldots\ldots + \beta_k \ln x_{ki} + e_i$ 8

when x_i changes 1%, y changes β_k%, holding the other regressors constant.

Exponential regression model:

$$y_i = \alpha x_i^{\beta 1} \varepsilon^{ei} \quad 9$$

It can also be expressed in logs form as under:

$\ln y_i = \alpha_0 + \beta_1 \ln x_{1i} + e_i$ 10

It is called linear in logs and can be estimated by ***OLS*** on the condition that classical assumptions are satisfied.

Cobb-Douglas Production function:

$$Q = \alpha_0 L^{\beta_1} K^{\beta_2} \Rightarrow \log Q \Rightarrow \log \alpha_0 + \beta_1 \log L + \beta_2 \log K \quad 11$$

When L changes 1%, Q changes by β_1%

Lin-Log Model (Semi log)

$y_i = \alpha_0 + \beta_1 \ln x_{1i} + \ldots.. + e_i$ 12

When x_1 changes 1%, y changes 0.01*β_1, holding other variables constant.

The impact of a variation in x_i on y decreases as x_i gets larger.

Log-In Model:

$\ln y_i = \alpha_0 + \beta_1 x_{1i} + \ldots\ldots + e_i$ 13

When x_1 changes one unit, y_i changes 100 * β_1% holding the other regressors constant. The impact of a variation in xi on y_i increases with y_i.

Quadratic forms:

$$y_i = \alpha_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{2i} + e_i \qquad i = 1,\ldots\ldots,n \qquad 14$$

We know that cost curve is U-shaped and cost is quadratic in output with $\beta_1<0$ and $\beta_2>0$. Yi increases with x_{1i} but decrease with x_{1i}^2

Inverse form:

$$y_i = \alpha_0 + \beta_1 \frac{1}{x_{1i}} + \beta_2 x_{2i} + e_i \qquad 15$$

The slope moves to zero when x_{1i} is large.

Intercept dummy independent variable:

The intercept dummy independent variable changes the intercept but the slope remains constant. One dummy variable is used for two categories. If there are more than two categories more than one dummy independent variable can be set in the regression equation. To include two same dummy variables cause perfect multi-collinearity and violate assumption.

Slope dummy independent variable:

$$y_i = \alpha_0 + \beta_1 x_i + \beta_2 D_i + \beta_3 x_i D_i + e_i \qquad 16$$

There are two equations as under:

$$y_i = \alpha_0 + \beta_2 + (\beta_1 + \beta_3) x_i + e_i, \quad \text{if } D_t = 1 \qquad 17$$

$$\alpha_0 + \beta_1 x_i + e_i, \quad \text{if } D_i = 0 \qquad 18$$

Here each equation can be estimated separately.

In some cases according to the need, interaction term should be included in equation 16 if there is a reason to believe that the slopes are different across categories.

Lags form:

$$y_t = \alpha_0 + \beta_1 y_{t-1} + \beta_2 x_t + e_i \qquad 19$$

Here the length of time between cause and effect is called lag.

y_{t-1} is lagged independent variable. β_1 measures the impact of previous observation on the current observation on the current observation. If lag structure take place over more than one time period, it is called distributed lags.

Mixed functional form:

In this case, y_i are a semi-log function of x_i, a quadratic function of φ_i, and a linear function of z_i. The marginal effect and elasticity for each of these variables is given by the formulas above.

Note: Formation of mixed functional form requires expert knowledge of econometrics to handle such models.

Now the question is how to evaluate functional form?

In the residual plot, whether there is a systematic pattern between e_i and x_i, should be checked. Different functional forms should be used.

Consequences of choosing the wrong functional form

The OLS estimator will be biased and not valid.

Detection and correction of functional form specification errors

Following two alternative approaches are used to choose a specific functional form for a model:

i. Maintained hypothesis methodology

ii. Theory/testing methodology

Maintained hypothesis methodology: This approach uses theory and/ or tractability to choose a specific functional form. Once a specific functional is chosen, it is treated as a maintained hypothesis and not tested using the sample data. Choosing functional form based on tractability is not good practice always. Relying on theory alone may not be a good practice. This is because there are many situations when theory has nothing to say about the appropriate functional form. If the wrong functional form is chosen, then the parameter estimates will be biased and all tests of hypothesis, strictly speaking, will be incorrect.

Theory/Testing Methodology: This approach involves the following steps as under

i. A set of specific functional forms that are consistent with theory is identified. If a specific functional form is inconsistent with the theory being used to guide the specification of the statistical model, then it should not be considered.

ii. Statistical tests should be conducted to determine which specific functional form should be chosen

iii. One of the following two approaches may be used in the third step;

1) Testing-down approach; 2) Testing-up approach

Testing-down approach: When using the testing-down approach, we begin with a general model and test-down to a more specific model. To test for nonlinear terms and interaction terms, begin with a general model that includes one or more of these terms. For example, the general model might include nonlinear terms such as X^2 and/or lnX, or interaction terms such as X*A. A t-test and/or an F-test can be employed to test whether these terms belong in the model.

Testing-up approach: When using this approach, begin with a specific model and test-up to a more general model. To test for nonlinear terms and interaction terms, a specific model that does not include one or more of these terms may be appropriate. For example, the specific model might not include nonlinear terms such as X^2 and/or lnX or interaction terms such as X*A. We then use a Lagrange multiplier test to test whether these terms should be added to the model.

Other tests for functional form

A number of other criteria and statistical tests are also used to test for functional form. Some of these are below:

i. Adjusted R^2

ii. Ramsey's Reset Test

iii. Recursive Residual Test

SPECIFICATION TESTS

The specification test is conducted by considering the ***residuals*** of the regression estimates (Residual tests).

1. Correlograms and Q-Statistics: This test displays the autocorrelation and partial autocorrelation of the squared residuals up to specified number of lags. It is available for LS, TSLS, nonlinear LS, binary, ordered, censored and count methods.

2. ***Correlograms of squared residuals***: This test displays the autocorrelation and partial autocorrelations of the squared residuals up to specified number of lags defined. It can be used to check ARCH (Autoregressive Conditional Heteroskedasticity) in the residuals. If autocorrelation and partial autocorrelation is equal to zero, there is no ARCH effect in the residuals.

3. Histogram and Normality test: This test provides you a histogram and descriptive statistics of the residuals, including Jarque-Bera statistics (JB, 1981) for testing normality. If the residuals are normally distributed, the histogram should be bell-shaped and J-B test statistics should not be significant.

4. Serial correlation LM test: This test is an alternative to the Q-statistics for testing serial correlation. Unlike Durbin Watson test (AR(1)), LM test can be used to test higher order ARMA (Auto regressive moving average) errors.

The ***null hypothesis of the LM test is there is no serial correlation*** up to chosen lag order. The F-statistics is an omitted variable test for the joint significance of all lagged residuals. Omitted variables are residuals not independent variables. R^2 statistic is the ***Breusch-Godfrey*** LM test statistics.

5. ARCH-LM test: This is LM test for ARCH (Engle 1982) in the residuals. Ignoring ARCH effect can cause inefficiency in estimation. The ***null hypothesis is that there is no ARCH effect in the residuals***. It is computed from the residual test regression as follows:

$$e_t^2 = \alpha_0 + \beta_1 e_{t-1}^2 + \beta_2 e_{t-2}^2 + ... + \beta_q e_{t-q}^2 + v_t \qquad 20$$

Where 'e' is the residual and v is the residual of residual.

This is a regression of the squared residuals (***auxiliary residuals regression***) on constant and lagged squared residuals up to order q. ***The F-statistic*** is an omitted variable test for the joint significance of all lagged squared residuals. ***Engles LM*** test statistic is equal to number of observations item R^2 statistic.

6. Whites Heteroskedasticity Test: It is a test for heteroskedasticity in the residuals from a LS regression (White 1980). In the presence of heteroskedasticity standard errors are no longer valid but ***OLS*** estimates are still consistent. In order to correct heteroskedasticity, weighted least squares estimation method can be employed or chosen the robust standard error option to correct the standard errors.

White test is test of ***null hypothesis of that there is no Heteroskedasticity***.

$$y_t = \alpha_0 + \beta_1 x_t + \beta_2 z_t + e_t \quad (21)$$

The test statistics is based on auxiliary regression as under:

$$e_t^2 = \alpha_0 + \alpha_1 x_t + \alpha_2 z_t + \alpha_3 x_t^2 + \alpha_4 z_t^2 + \alpha_5 x_t z_t + v_t \quad (22)$$

The White test statistic is computed the number of observations times $\mathbf{R^2}$ from the regression.