Intermediate Econometrics

Professor: Xiaojun Song

Timekeeper: Rui Zhou

Spring 2023

目录

8 More on Specification and Data Issues

8.1 Test for Functional Form Misspecification

First revisit MLR.4 of Zero Conditional Mean:

$$E[u|\vec{x}] = 0$$

The most usual case that violates zero-conditional-mean assumption is, **misspecification of functional form**. (Leave out some important explanatory variables)

If MLR.4 doesn't hold, then the slope coefficients may **not** be unbiased and consistent. Even though there are only a few variables correlating with u, since other variables would inevitably correlate with them, the slope coefficients would be biased and inconsistent.

8.1.1 RESET (Regression Specification Error Test)

The idea of RESET is to include squares and possibly higher-order fitted values in the regression. (which is similar to the reduced White test.) Note that squared terms and interaction terms are included since $\hat{y}^2, \hat{y}^3$ have been induced.

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + error$$

We test for the exclusion of the $\hat{y}$ terms (both $\hat{y}^2, \hat{y}^3$). If they cannot be excluded, this is evidence for omitted higher order terms and interactions, which indicates the misspecification of functional form. Note that it's accepted and enough to test only $\hat{y}^2$ and $\hat{y}^2$.

We conduct a F test for $\hat{y}$ terms.

$$H_0 : \delta_1 = \delta_2 = 0$$

The constructed F statistic will asymptotically follows

$$F \sim F_{2,n-k-1-2}$$

The number of restricted coefficients is 2, and the degree of freedom for model's residuals is $n-1-(k+2) = n-k-3$.

Remarks

- If $\hat{y}$ itself is included, then multicollinearity problem arises.
- RESET provides little guidance as to **where** misspecification comes from. Even if significant, some higher order terms are omitted, but no further specific information can be implied.

8.1.2 Testing against Nontested Alternatives

Suppose we postulate two candidate model for y:

$$\text{Model 1: } y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$
$$\text{Model 2: } y = \beta_0 + \beta_1 \log x_1 + \beta_2 \log x_2 + u$$

Neither Model 1 nor Model 2 is a special case under another. They are **nonnested** alternatives of each other. We can specify **either** Model 1 or Model 2 to be the null model. For example, if take Model 1 as H_0^1:

$$H_0^1 : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$
$$H_1^1 : y = \beta_0 + \beta_1 \log x_1 + \beta_2 \log x_2 + u$$

If take model 2 as H_0^2

$$H_0^2 : y = \beta_0 + \beta_1 \log x_1 + \beta_2 \log x_2 + u$$
$$H_1^2 : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

In either cases, define a general and comprehensive model that contains both models as subcases, and test:

$$y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 \log x_1 + \gamma \log x_2 + u$$

Then the corresponding null hypothesis to H_0^1 is:

$$H_0^1 : \gamma_3 = \gamma_4 = 0$$

If H_0^1 cannot be rejected, we say the model under H_0^1 is **preferred**, but **not** necessarily *true* model. It's possible that neither Model 1 nor Model 2 is true model. True model stands, if and only if zero conditional mean holds. Similarly, we can treat model 2 as the null model and test against $H_2^0 : \gamma_1 = \gamma_2 = 0$. But note that which model is specified as null model really matters!

However, if $E(u|x_1, x_2) = 0$, which means the model under H_0^1 is true model, this will imply that $E(u|f(x_1), g(x_2)) = 0$ whenever $f(x_1), g(x_2)$ is different from x_1, x_2. That is true because starting from the true model, no more term should correlate with the error. Therefore, for any alternative hypothesis, theoretically it proves to be not preferred, since error in the model under alternative hypothesis is not considered practically. Jointly speaking, the nonnested model test has no real effect. An alternative way that enforces the consideration of the error in alternative hypothesis will make sense regardless of H_0's model's correctness:

1. Regress y on $\log x_1, \log x_2$, and obtain fitted value $\tilde{y}$.
2. Regress y on x_1, x_2 and $\tilde{y}$, and then check t statistic of coefficient of $\tilde{y}$.

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 \tilde{y} + u \Longrightarrow H_0 : \alpha_3 = 0$$

Remarks

- A **clear winner** need not emerge. In one pair of hypothesis, one model may be preferred; while in another pair of hypothesis, its counterpart may be preferred.
- The preferred model may still differ from the true one.
- This cannot be used if the models differ in **dependent** variable, i.e., the same y.

8.2 Proxy Variables

Some variables, though indispensable in the model, hard to interpret numerically. For those unobservable explanatory variables, the interpretations of them are rather opaque, so are their coefficients. And such tricky issue will further influence our estimation for rest of the slope coefficients. If this is the case, a **proxy** variable should be introduced for the unbiasedness of other variables of our interest.

8.2.1 General Approach

Assume that the population regression function is

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

where x_3^* is unobservable. However, we have found a proxy x_3 for x_3^*.

$$x_3^* = \delta_0 + \delta_3 x_3 + v_3$$

Plug-in solution to omitted variable problem:

$$\text{Regress } y \text{ on } x_1, x_2, x_3$$

However, a "good" proxy variable must satisfy the following assumptions.

8.2.2 Assumptions

1. u is uncorrelated with x_1, x_2, x_3^*.
   - $E(u|x_1, x_2, x_3^*) = 0 \Longrightarrow Corr(x_3, u) = 0$. Note that x_3 differs from x_3^*!
   - The proxy is "just a proxy" for the omitted variable, and does **not** belong into the population regression. The proxy won't enter the true model, contrary to an explanatory variable.
2. v_3 is uncorrelated with x_1, x_2, x_3^*.
   - $E(x_3^*|x_1, x_2, x_3) = E(x_3^*|x_3) \Longrightarrow Corr(x_1, v_3) = Corr(x_2, v_3) = 0$.
   - The proxy variable is a good proxy for the omitted variable.

- i.e., using other variable in addition will not help predict the omitted variable.

Note that proxy variable is **not equivalent** to its corresponding unobserved variable.

8.2.3 Overall Model

Under these assumption,

$$y = (\beta_0 + \beta_3\delta_0) + \beta_1 x_1 + \beta_2 x_2 + (\beta_3\delta_3)x_3 + (u + \beta_3 v_3)$$

In this regression model, the error term in uncorrelated with all the explanatory variables. Consequently, all slope coefficients will be correctly estimated using OLS, unbiased and consistent (i.e., correctly identified). Meanwhile, the coefficient for the proxy variable may also be of interest, and it is a multiple of the coefficient of the omitted variable.

8.2.4 Lagged DV as Proxy

Omitted unobserved factors may be proxied by the value of the dependent variable from an earlier time period.

For example, city crime rate has an inertia that when considering this year's crime rate, crime rate in the past year matters. Moreover, including the past crime rate will at least partly **control for** the many *omitted* factors that also determine the crime rate in a given year.

$$crime = \beta_0 + \beta_1 crime_{-1} + \beta_2 unem + u$$

Another way to interpret this equation is that, one compares cities which had the "same" crime rate last year. This avoids comparing cities that differ very much in unobserved crime factors.

8.3 Random Coefficient Model

Models with random slopes, are also called random coefficient models. Back to the simple linear regression model, however, in this time parameters for intercept and slope coefficients vary among individuals.

$$y_i = a_i + b_i x_i$$

It is pitiful that, with only dataset $\{y_i, x_i\}_{i=1}^n$, a_i, b_i cannot be estimated. Luckily, we can take apart each random coefficient into two part, the fixed (average) and the zero-mean random part.

$$y_i = (\alpha + c_i) + (\beta + d_i)x_i$$

where, α as the average intercept, and c_i as random component for intercept; β as the average slope, and d_i as random component for slope.

If we group all the deterministic parts together and attribute the random parts to the error term, we can obtain the following equation:

$$\Leftrightarrow y_i = (\alpha + \beta x_i) + (c_i + d_i x_i)$$

Assumptions are:

$$E(c_i|x_i) = E(d_i|x_i) = 0 \Leftrightarrow E(a_i|x_i) = E(b_i|x_i) = 0$$

which means the individual random components are independent of the explanatory variable.

Under the assumptions above, we can get

$$E(c_i + d_i x_i|x_i) = 0$$

then, for the random coefficient model, the slope coefficients are both **unbiased** and **consistent**.

However, the error term is inate with heteroskedasticity, since that will depend on x_i obviously.

$$\Longrightarrow Var(c_i + d_i x_i|x_i) = \sigma_c^2 + \sigma_d^2 x_i^2$$

Fortunately, though with heteroskedasticity, the specific functional form is known. WLS or OLS with standard errors will consistently estimate the **average** intercept and average slope in the population.

Finally note that, relatively precise as random coefficient model is, it is **incapable** of making predictions, since individuals have random parts and differ from each other.

8.4 Measurement Error

Consider two kinds of measurement error:

- on dependent variable
- on explanatory variable(s)

8.4.1 Dependent Variable

Suppose the regression function for the population is as follows:

$$y^* = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$$

where the starred y^* cannot be observed directly. In fact, you can only observed y with error e_0:

$$e_0 = y - y^*$$

Take the measurement error on dependent variable into account:

$$\begin{aligned} & y^* = y - e_0 = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u \\ \Longleftrightarrow \quad & y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + (u + e_0) \end{aligned}$$

If e_0 is supposed to be uncorrelated with $x_1, ..., x_k$, i.e., $E(u + e_0|x_1, ..., x_k) = 0$, then all slope coefficients will be unbiased and consistent. Pratically, regress y on $x_1, ..., x_k$. This kind of measurement error imposes no influence on the estimation of slope coefficients.

However, the estimated slope coefficients will be less precise, since their variance will be larger because of measurement error e_0.

$$Var(u + e_0) = \sigma_u^2 + \sigma_e^2 > \sigma_u^2$$

8.4.2 Explanatory Variable

Suppose the regression function for the population is as follows:

$$y = \beta_0 + \beta_1 x^* + u$$

where the starred x^* cannot be observed directly. The observed x deviates from true x^* systematically at e.

$$x = x^* + e$$

Taking measurement error into account, then the estimated regression is

$$y = \beta_0 + \beta_1 x + (u - \beta_1 e_1)$$

Typically, there are two ways to treat the measurement error on explanatory variable.

If $Cov(x, e) = 0$, then $E(x^*|x) = x$. x is regarded as a proxy variable of x^*, and β_1 is unbiased and consistent.

If $x = x^* + e$, the observed x is pratically decomposed into two parts, unobserved part x^* and measurement error e. Instead of assuming linear independence of observed value x and measurement error e, we assume unobserved part x^* to be uncorrelated with e:

$$Cov(x^*, e) = 0$$

This assumption is so important that it is called classical errors-in-variables assumption, **CEV** in short.

To consider potential issues of endogeneity, compute $Cov(x, u - \beta_1 e)$:

$$Cov(x, u - \beta_1 e) = Cov(x^* + e, u - \beta_1 e) = -\beta_1 \cdot Cov(x, e) + Cov(u, e) + \beta_1 \cdot Cov(x^*, e)$$

where $Cov(u, e)$ is safely supposed to be 0. Thus under CEV,

$$Cov(x, u - \beta_1 e) = -\beta_1 \cdot Cov(x, e) = -\beta_1 \cdot Cov(x^* + e, e) = -\beta_1 \sigma_e^2$$

Suppose $y = \beta_0 + \beta_1 x + v$, and regress y on x,

$$\text{plim } \widehat{\beta}_1 = \beta_1 + \frac{Cov(v, x)}{Var(x)} = \beta_1 (1 - \frac{\sigma_e^2}{\sigma_x^2}) = \beta_1 \cdot \frac{\sigma_{x^*}^2}{\sigma_x^2 + \sigma_e^2} = \beta_1 \cdot \frac{Var(x^*)}{Var(x)}$$

Notice that $\frac{\sigma_{x^*}^2}{\sigma_x^2+\sigma_e^2} \in (0,1)$ holds forever, then $\hat{\beta}_1$ has a **attenuation bias** (bias towards 0) compared with β_1. The magnitude of the effect will be attenuated towards zero.

In MLR, the problem will be more tricky.

$$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + ... + \beta_k x_k + u, \text{ with } x_1 = x_1^* + e$$
$$\Longrightarrow y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + (u - \beta_1 e)$$

Possible that $(u - \beta_1 e)$ correlates with x_1. In this case, even if $(u - \beta_1 e)$ is uncorrelated with all other variables, as long as any of the remaining variables correlates with x_1, it's likely that all those slope coefficients will be biased and not consistent. However, the deterministic equation of plim $\hat{\beta}_j$ is rather complicated and hence not covered here.

8.5 Missing Data

Missing data is viewed as a special case of sample selection (= nonrandom sampling), as the observations with missing information cannot be used.

If the sample selection is based on **independent** variables, there's no problem as a regression conditions on the independent variables (i.e., exogenous sample selection). In general, sample selection is no problem if it is uncorrelated with the error term of a regression.

If the sample selection is based on the **dependent** variable or the error term, this entails a problem (i.e., endogenous sample selection).

Missing Completely at Random (MCAR) means that missingness is unrelated to both u and $x_1, ..., x_k$.

$$\Pr(m_k = 1 | u, x_1, ..., x_k) = \Pr(m_k = 1)$$

Compared to MCAR, Missing at Random means that

$$\Pr(m_k = 1 | x_1, ..., x_k) = \Pr(m_k = 1)$$

where the missing mechanism is allowed to correlate with $x_1, ..., x_k$, but not allowed with u.

8.5.1 Missing Indicator Method (MIM)

Suppose we are missing some information on explanatory variable x_k, and with full information on $y, x_1, ..., x_{k-1}$.

One conservative but direct way is to work only with complete cases and get the complete-case estimator. However, the cost is hefty, though robust. Missing Indicator Method (MIM), will make better use of current data.

MIM creators two new variables:

$$z_{ik} = \begin{cases} x_{ik}, & \text{if } x_{ik} \text{ is observed} \\ 0, & \text{if } x_{ik} \text{ is missing} \end{cases}$$

$$m_{ik} = \begin{cases} 1, & \text{if } x_{ik} \text{ is observed} \\ 0, & \text{if } x_{ik} \text{ is missing} \end{cases}$$

where m_{ik} is a missing data indicator.

Then, regress y_i on $x_{i1}, ..., x_{i,k-1}, z_{ik}, m_{ik}$, using **all** observations.

Unfortunately, MIM is valid only under a strong assumption:

$$Cov(x_k, x_j) = 0, \forall j \neq k$$

The assumption is too strong to hold in practice.

Note that omitting m_{ik} is the same as assuming $x_{ik} = 0$ whenever it is missing.

8.6 Outliers & Influential Observations

If outliers come from mistakes that occurred when keying in the data, one should just discard the affected observations. But if outliers are the result of the data generating process, the decision on whether or not to discard the outliers is not so easy.

8.6.1 Least Absolute Deviations Estimation (LAD)

Algorithm in OLS will exemplify the outliers' influence, and causing pratical shock to the estimated result. Leat Absolute Deviations Estimation is a competitive alternative.

The least absolute deviations estimator minimizes the sum of absolute deviations, which may be more robust to outliers as deviations are not squared.

$$\min_{b_0, b_1, ..., b_k} \sum_{i=1}^{m} |y_i - b_0 - b_1 x_{i1} - ... - b_k x_{ik}|$$

The least absolute deviation estimator estimates the parameters of the **conditional median**, instead of the conditional mean with OLS. The conditional median and conditional mean coincide only when u in $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$ is **symmetric**. Furthermore, the least absolute deviations estimators is a special case of quantile regression.