会计与财务研究方法论

Lecture 1: Basic Assumptions of OLS Regression

上海财经大学
会计学院、会计与财务研究院
靳庆鲁

Internal and External Validity

- An empirical analysis is internally valid if the statistical inferences about the causal effects are valid.

- The analysis is externally valid if inferences can be generalized from the setting of study to other settings.

Threats to Internal Validity

- *1. Heteroskedasticity*
- *2. Time-series Dependence*
- *3. Cross-sectional Dependence*
- *4. Both C-S and T-S Dependences*
- *5. Multi-Collinearity*
- *6. Omitted Variable Bias*
- *7. Functional Form Misspecification*
- *8. Measurement Error*
- *9. Endogeneity*
- ……

- **Threats to External Validity**

ex/ Difference in institutional settings, laws, culture, etc.

- **Basic Assumptions of OLS Regression (Related to Internal Validity)**

- （1）$E(u_i)=0$ （随机项均值为零）
- （2）随机项服从正态分布 (No skewness and kurtosis)
- （3）$Cov(X, u_i)=0$（随机项与解释变量X不相关）
- （4）解释变量之间无高度相关 (No multi-collinearity)
- （5）$Var(u_i)=\sigma^2$ (No heteroscedasticity)

- （6）$Cov(u_i, u_j)=0$ (No cross-sectional dependence)
- （7）$Cov(u_t, u_{t+1})=0$ (No time-series dependence)
- （8）$Cov(u_{it}, u_{jt+1})=0$ (No cross-sectional & time-series dependence)

不满足基本假定的情形

- 1、随机项均值不等于0的情况不会影响解释变量的系数，只会影响截距项。
- 我们不讨论上述假定是否违背。

- 2、随机项正态性假设一般在大样本下会近似成立。

- 3、模型设计不合理，遗漏相关变量，会导致随机项与解释变量X相关（这是因为遗漏的解释变量同模型中保留的解释变量往往是相关的，因此就造成随机项与模型中的解释变量相关）。应该重新考察模型的合理性，通过文献回顾，补充控制变量。

不满足基本假定的情形（2）

- 4、解释变量之间高度相关=>**多重共线性**
- 5、随机项方差不等于常数=>**异方差**
- 6、 $Cov(u_i, u_j)$ ^=0 => **Cross-sectional dependence**
- 7、 $Cov(u_t, u_{t+1})$^=0 => **Time-series dependence**
- 8、**Cross-sectional and Time-series dependence**

- **Disturbance term ~ i.i.d**（随机项服从正态分布）

Normality

- Kurtosis
  - Kurtosis=0　与正态分布的陡缓程度相同。
    Kurtosis>0　比正态分布的高峰更加陡峭——尖顶峰
    Kurtosis<0　比正态分布的高峰来得平台——平顶峰

Kurtosis= $\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^4 / SD^4 - 3$

- Skewness
  - Skewness=0　分布形态与正态分布偏度相同
    Skewness>0　正偏差数值较大，为正偏或右偏。长尾巴拖在右边。
    Skewness<0　负偏差数值较大，为负偏或左偏。长尾巴拖在左边。

- Skewness= $\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^3/SD^3$
- | Skewness| 越大，分布形态偏移程度越大。
- Solutions:
  - Non-Parameter tests
  - Rank Regression
  - Data Transform (Box-Cox Transform)

Box-Cox Transform

Methods		Box-Cox Transform
new=raw^3	cube	reduce severe negative skew
new=raw^2	square	reduce mild negative skew
new=raw		no change
new=raw^0.5	square root	reduce mild positive skew
new=ln(raw) or new=log10(raw)	$\log_e$ or $\log_{10}$	reduce positive skew
new=-(raw^-0.5)	negative reciprocal root	reduce severe positive skew
new=-(raw^-1)	negative reciprocal	reduce very severe positive skew
new=-(raw^-2)	negative reciprocal square	reduce very severe positive skew
new=-(raw^-3)	negative reciprocal cube	reduce very severe positive skew

Skewness / Kurtosis tests for Normality

- A Skewness-Kurtosis test, can more formally evaluate the null hypothesis that the sample at hand came from a normally-distributed population.

- STATA:
- sktest varlist

- **Cov(X, u) ^=0（随机项与解释变量X相关）**

Omitted variable bias

- If we miss out an important variable it not only means our model is poorly specified it also means that any estimated parameters are likely to be biased. If the true model were

$$Y_t = \alpha_0 + \alpha_1 * X_t + \alpha_2 * Z_t + \mu$$

and we estimate

$$Y_t = \beta_0 + \beta_1 * X_t + \mu$$

- Then the omitted variable can be considered as a function of X in a conditional or auxiliary regression

$$Z_t = \gamma_0 + \gamma_1 * X_t + \omega_t$$

- So we have estimated

$$Y_t = \beta_0 + \beta_1 * X_t + \beta_2 * (\gamma_0 + \gamma_1 X_t + \omega_t) + \mu_t$$

Or

$$Y_t = (\beta_0 + \beta_2 * \gamma_0) + (\beta_1 + \gamma_1 * \beta_2) * X_t + \beta_2 * \omega_t + \mu_t$$

$$Y_t = \partial_0 + \partial_1 * X_t + \varepsilon_t$$

- So

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 * [\frac{\sum x_t z_t}{x_t^2}]$$

- So there will be a bias as the coefficient of X picks up the part of the influence of Z that was correlated with X.
  - The coefficient estimate can have positive or negative bias
  - Its standard error will also be biased positively
  - The bias on the coefficient estimate can either cancel or reinforce the bias in the standard error when we do a t test.

    So it is not clear what the effect will be.
  - Note that while incorrect omission of variables leads to biased estimates of the parameters that are included, so better to include the wrong variables rather than exclude the right ones.

- **Multi-Collinearity（多重共线性）**

- 多重共线性的后果
- 多重共线性的检验
- 克服多重共线性的方法

多重共线性的后果

- 1、完全共线性下参数估计量不存在
- 2、近似共线性下普通最小二乘法参数估计量很难通过显著性检验

3、参数估计量经济含义不合理

如果模型中两个解释变量具有线性相关性，例如X_1和X_2，那么它们中的一个变量可以由另一个变量表征。

这时，X_1和X_2前的参数并不反映各自与被解释变量之间的结构关系，而是反映它们对被解释变量的共同影响。

所以各自的参数已经失去了应有的经济含义，于是经常表现出似乎反常的现象，例如本来应该是正的，结果恰是负的。

多重共线性的检验

- 1、使用相关系数法检验自变量之间的相关性(>0.6)；
- 2、使用方差膨胀因子法检验（VIF>5).
- 3、经验准则：Adjusted R-square大，而t值不显著；那么很可能存在多重共线性。

Multi-Collinearity Modifications

- 1. Demeaning the regressors (Aiken and West, 1991)
- This rule only applies when there are interactions.
- When estimating Y = a + b1*X1 + b2*X2 + b12*X1*X2, and correlation between X1 and X1*X2 and/or between X2 and X1*X2 are large, based on demeaned regressors may significantly reduce multicollinearity.
  - Y = a + b1*(X1-X1*) + b2*(X2-x2*) + b12*(X1-X1*)*(X2-X2*), where X1* and X2* are within sample mean of X1 and X2, which are both continuous variables.

- What if X1 is continuous but X2 is a dummy (Chen, 2009)
  - Demeaning the continuous regressors

- 2. Orthogonalize high correlated variables (Burrill, 1997)
  - STATA: orthog x1 x2, gen(ox1 ox2)

- 3. 差分法
- 4. 改用相对变量的形式
- 5. PCA (主成分分析法)

- **Heteroscedasticity （异方差）**

- 异方差的后果
- 异方差的检验
- 克服异方差的方法

异方差的后果

1、估计参数的无偏性仍然成立

异方差的存在对参数无偏性的成立没有影响。

2、参数估计的方差不再最小

同方差假定是OLS估计方差最小的前提条件，所以随机项存在异方差时，将不能再保证最小二乘法估计的方差最小。

3、对参数显著性检验的影响

由于异方差的影响，使得无法正确估计参数的标准误，导致参数估计的t统计量不能正确确定。

4、对预测的影响

尽管基于OLS估计的参数仍然无偏，但是由于估计参数的标准误不正确，从而对Y的预测也将不再有效。

异方差检验
White (1980) 检验

基本思想：

不需要关于异方差的任何先验信息，只需要在大样本的情况下，将OLS估计后的残差平方对解释变量、解释变量的平方及其交叉项进行回归，利用相应的检验统计量来判断异方差性。

检验的基本步骤：

以一个二元线性回归模型为例，设模型为：

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \qquad (1)$$

并且，设异方差与 X_{2t}, X_{3t} 的一般关系为

$$\sigma_t^2 = \alpha_1 + \alpha_2 X_{2t} + \alpha_3 X_{3t} + \alpha_4 X_{2t}^2 + \alpha_5 X_{3t}^2 + \alpha_6 X_{2t} X_{3t} + v_t$$

其中 v_t 为随机误差项。

1、求回归估计式并计算e_t^2

用OLS估计方程（1），计算残差 $e_t = Y_t - \hat{Y}_t$， 并求残差 的平方e_t^2。

2、求辅助函数

用残差平方e_t^2作为异方差σ_t^2的估计，并建立 $X_{2t}, X_{3t}, X_{2t}^2, X_{3t}^2, X_{2t}X_{3t}$ 的辅助回归，即

$$\hat{e}_t^2 = \hat{\alpha}_1 + \hat{\alpha}_2 X_{2t} + \hat{\alpha}_3 X_{3t} + \hat{\alpha}_4 X_{2t}^2 + \hat{\alpha}_5 X_{3t}^2 + \hat{\alpha}_6 X_{2t} X_{3t} \quad (2)$$

3、计算nR^2

基于OLS估计方程（2）得到辅助回归函数的 R^2，n为样本量。

4、提出假设

$H_0:\alpha_2=\cdots=\alpha_6=0,\qquad H_1:\alpha_j\ (j=2,3,\cdots,6)$ 至少有一个不为零。

5、检验

在原假设成立下，有 nR^2 渐进服从自由度为5的 χ^2 分布。给定显著性水平 α ,查 χ^2 分布表得临值 $\chi^2_\alpha(5)$ ，如果 nR^2 > $\chi^2_\alpha(5)$,则拒绝原假设，表明模型中随机误差存在异方差 。

注意： White(1980)检验不仅能够检验异方差的存在性，同时在多变量的 情况下，还能判断出是哪个变量引起的异方差。

Heteroscedasticity Modifications

- 1) White-adjusted Statistics
- In PROC REG, you can use the SPEC option to test for homoskedastic errors. If the SPEC test is significant, then you can use the ACOV option along with a series of TEST statements to get statistical tests on the parameter estimates that are adjusted for heteroskedasticity.

White-adjusted Statistics SAS Program

- /* run the spec test on the model to see if the model is homoskedastic. For details on this test, see White's 1980 paper*/

- ```
  proc reg data=test;
        model y=x1 x2 / spec;
  run;
  ```
  ```

- /*If the chi-square test is significant, you need to use the ACOV option. The ACOV option calculates symptotic standard errors for your parameter estimates. The normal output from PROC REG does not change when you use the ACOV option. To get the updated tests of the paramter estimates, you have to use TEST statements*/

- ```
  proc reg data=test;
        model y=x1 x2 / acov;
        x1:test x1;
        x2:test x2;
  run;

  quit;
  ```
  ```

- 2) STATA
- reg depvar indepvar, robust

- **Autocorrelation (Time-series dependence)**

- 自相关的后果
- 自相关的检验
- 克服自相关的方法

自相关的后果

- 原因：自相关经常出现在以时间序列为样本的模型中，原因在于大多数经济时间数据都有一个明显的特点:**惯性**。

- 后果：
- **1) OLS估计的参数仍然具有一致性（无偏性）。**
- **2) 变量的显著性检验失去意义**
- **3) 模型的预测失效**

自相关检验

自相关检验方法有多种，但基本思路相同：

首先，采用OLS 法估计模型，以求得随机项的“近似估计量

$$\tilde{e}_i = Y_i - (\hat{Y}_i)_{0ls}$$

然后，通过分析这些“**近似估计量**”之间的相关性，以判断随机项是否具有自相关。

1）回归检验法

以$\tilde{e}_t$为被解释变量，以各种可能的相关量，诸如以$\tilde{e}_{t-1}$、$\tilde{e}_{t-2}$、$\tilde{e}_t^2$等为解释变量，建立各种方程：

$$\tilde{e}_t = \rho\tilde{e}_{t-1} + \varepsilon_t$$

$$\tilde{e}_t = \rho_1\tilde{e}_{t-1} + \rho_2\tilde{e}_{t-2} + \varepsilon_t$$

……

如果存在某一种函数形式，使得方程显著成立，则说明原模型存在自相关。

回归检验法的优点是：（1）能够确定序列相关的形式，（2）适用于任何类型自相关问题的检验。

2）Durbin-Watson 检验法

D-W检验是杜宾（J.Durbin）和瓦尔森(G.S.Watson)于1951年提出的一种检验序列自相关的方法。**假定：**

随机误差项μ_i为一阶自回归形式：

$$\mu_i=\rho\mu_{i-1}+\varepsilon_i$$

D.W. 统计量:

针对原假设：H_0: ρ=0， 构如下造统计量：

$$D.W. = \frac{\sum_{t=2}^{n} (\tilde{e}_t - \tilde{e}_{t-1})^2}{\sum_{t=1}^{n} \tilde{e}_t^{\,2}}$$

D.W检验步骤:

（1）计算DW值

（2）给定α，由n和k的大小查DW分布表，得临界值d_L和d_U

（3）比较、判断

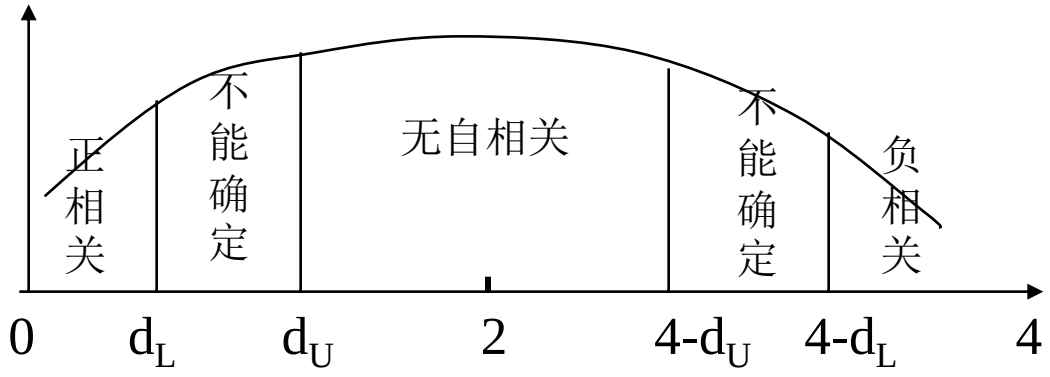
若　$0<D.W.<d_L$　　存在正自相关

$d_L<D.W.<d_U$　　不能确定

$d_U<D.W.<4-d_U$　　无自相关

$4-d_U < D.W. < 4-d_L$　　不能确定

$4-d_L < D.W. < 4$　　存在负自相关



当D.W.值在2左右时，模型不存在一阶自相关。

Autocorrelation Modifications

1). Newey-West Standard Error Adjust (N-W)

2). FM-i, Z2-i, CL-i

1. Newey-West (1987) Standard Error Adjust

newey depvar indepvar, lag(1)

2. FM-i, Z2-i, CL-i

FM-i: Fama-MacBeth t-statistic based on mean and standard error of cross-section of coefficients from time-series (firm-specific) regressions.

Z2-i: Z2 statistic based on mean and standard error of cross-section of t-statistics from time-series (firm-specific) regressions.

CL-i: robust standard errors clustered by firm

reg depvar indepvar, cluster(stkcd)

logit depvar indepvar, cluster(stkcd)

- **Cross-sectional Dependence**

- **The consequences of Cross-sectional Dependence**
- **Tests of Cross-sectional Dependence**
- **The solution of Cross-sectional Dependence**

The consequences of Cross-sectional Dependence

- A growing literature comes to the conclusion that panel data sets are likely to exhibit substantial c-s dependence, which may arise due to the presence of common shocks.

An-example: Asset Pricing

- An best example for the effect of c-s dependence when t-s dependence is not present is asset pricing, which is because market efficiency implies that excess stock returns are serially uncorrelated, allowing us to focus exclusively on the effects of c-s dependence.

- Assuming that c-s dependence is caused by the presence of common factors, which are unobserved (and as a result, the effect of these components is felt through the disturbance term) but they are uncorrelated with the included regressors, the standard fixed-effects (FE) and random effects (RE) estimators are consistent, although the estimated standard errors are biased.

- On the other hand, if the unobserved components that create interdependencies across cross-sections are correlated with the included regressors, the FE and RE estimators will be biased and inconsistent.

- In this case, one may follow an alternative method proposed by Pesaran (2006). This alternative apply an instrumental variables (IV) type approach using standard FE IV, or RE IV estimators. However, in practise, it would be difficult to find instruments that are correlated with the regressors and not correlated with the unobserved factors.

Tests of Cross-sectional Dependence

- When the time dimension (T) of the panel is larger than the cross-sectional dimension (N), one may use the LM test, developed by Breusch and Pagan (1980), which is readily available in Stata using the command xttest2.

- When N is large and T is small, *the most commonly encountered situation in panels,* one may use an alternative method proposed by Hoyos and Sarafidis (2007), which is now available in Stata using the command xtcsd.

- Use http://www.econ.cam.ac.uk/phd/red29/xtcsd_baltagi.dta
- tsset id t
- Fixed effect: xtreg lngsp lnpcap lnpc lnemp unemp, fe
- Pesaran's (2004) CD test: xtcsd, pesaran abs
- (abs option get the average *absolute corr. Between* c-s units)
- Frees (1995) test: xtcsd, frees
- Friedman (1937) test: xtcsd, friedman
- Random effect: xtreg lngsp lnpcap lnpc lnemp unemp, re
- xtcsd, pesaran
- xtcsd, frees
- xtcsd, friedman

The Modifications of Cross-sectional Dependence

1). FM-t, FM-NW

2). Z2-t

3). CL-t

1. FM-t, FM-NW

FM-t: Fama-MacBeth t-statistic based on mean and standard error of time-series of coefficients from cross-sectional regressions.

FN-NW: FM-t statistic with Newey-West (1987) correction.

2. Z2-t

Z2-t: Z2 statistic based on mean and standard error of time-series of t-statistics from cross-section regressions.

3. CL-t

CL-t: robust standard errors clustered by time (year or quarter)

reg depvar indepvar,cluster(year)

logit depvar indepvar,cluster(year)

- **Cross-sectional and Time-series Dependence**

- Much of the empirical accounting literature uses panel data sets, In these data sets often the variables of interest are both cross-sectionally and serially correlated.

- Several examples for both c-s and t-s dependence: Cost of Capital, Conservatism, Audit Fees, Accounting Items, Executive Compensation, and Corporate Governance.

The Modifications of Cross-sectional Dependence

- **CL-2: robust standard errors clustered by firm and time.**

- **cluster2 depvar indepvar, fcluster(stkcd) tcluster(year)**
- **logit2 depvar indepvar, fcluster(stkcd) tcluster(year)**

Summary of methods to cross-sectional and time-series dependence

- This table summarizes the robustness to c-s and/or t-s dependence of methods commonly used in accounting research to calculate standard errors. Yes(No) indicates the method is (is not) robust to the indicated form of dependence.
- **Key to methods:** ***OLS:*** OLS standard errors; ***White:*** White (1980) standard errors; ***NW:*** Newey-West (1987) standard errors; ***FM-i:*** Fama-MacBeth t-statistic based on mean and standard error of c-s of coefficients from t-s regressions; ***FM-t:*** Fama-MacBeth t-statistic based on mean and standard error of t-s of coefficients from c-s regressions; ***FM-NW:*** FM-t statistic with Newey-West (1987) correction; ***Z2-i:*** Z2 statistic based on mean and standard error of c-s of t-statistics from t-s regressions; ***Z2-t:*** Z2 statistic based on mean and standard error of t-s of t-statistics from c-s regressions; ***CL-i:*** robust standard errors clustered by firm (or industry); ***CL-t:*** *robust standard errors* clustered by time; ***CL-2:*** robust standard errors clustered by firm and time.

Summary of methods to cross-sectional and time-series dependence

	Form of Dependence		
Method	Cross-sectional	Time-series	Cross-sectional & Time-series
OLS	No	No	No
White	No	No	No
N-W	No	**Yes**	No
FM-i	No	**Yes**	No
Z2-i	No	**Yes**	No
CL-i	No	**Yes**	No
FM-t	**Yes**	No	No
FM-NW	**Yes**	No	No
Z2-t	**Yes**	No	No
CL-t	**Yes**	No	No
CL-2	**Yes**	**Yes**	**Yes**