Econometrics Note 2

2024年7月11日 15:31

Note 2

2022年2月7日 10:13

L_1

1) Assumptions

高斯-马尔可夫 (brace over 1–5 labelled 高斯马)

1. $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$

   (Specified & Linear)
2. Random sampling $\{y_i, x_{1i}, \cdots, x_{ki}\}$, n observations
3. Explanatory variables (X满秩时)无完全共线性
4. ZCM $E[u_i|X_i] = E[u_i|x_{1i}, x_{2i}, \ldots, x_{ki}] = 0$
5. Homoskedasticity $Var[u_i|X_i] = \sigma^2$

1~5 则 OLS估计量为 BLUE

6. $u \sim N(0, \sigma^2)$ 正态分布误差

1~6 可基于任何大小的样本t和F检测推演β

$$\frac{\hat{\beta}_i - \beta_i}{s.e.(\hat{\beta}_i)} \sim t_{n-k-1}$$ 自由度(k为能影响斜率的变量个数)

$$F \sim F_{q,n-k-1}$$ (检验中受限变量自由度)

若无A6 仅用于样本量大的 (Asymptotic Theory)

$$\frac{\hat{\beta}_i - \beta_i}{s.e.(\hat{\beta}_i)} \overset{a}{\sim} t_{n-k-1} \overset{a}{=} N(0,1)$$

$$F \overset{a}{\sim} F_{q,n-k-1}$$

观测量式

$y_i = \beta_0 + \beta_1 x_i + u_i$

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \beta_1 = \frac{Cov(y_i, x_i)}{Var(x_i)} \end{cases}$$

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1-R_j^2)}$$

矩阵式

$Y = X\beta + u$

$$\underset{(k\times1)}{\hat{\beta}} = \underset{(k\times n)(n\times k)}{(X'X)^{-1}} \underset{(k\times n)(n\times1)}{X'Y}$$

$$\underset{(k\times k)}{Var(\hat{\beta})} = \underset{(1\times1)}{\sigma^2} \underset{(k\times n)(n\times k)}{(X'X)^{-1}}$$

$$\underset{(k\times1)}{\hat{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

$$\underset{(k\times k)}{Var(\hat{\beta})} = \begin{pmatrix} Var(\hat{\beta}_1) & & & \\ cov(\hat{\beta}_1, \hat{\beta}_2) & Var(\beta_2) & & \\ \vdots & & \ddots & \\ cov(\hat{\beta}_1, \hat{\beta}_k) & cov(\hat{\beta}_2, \hat{\beta}_k) & \cdots & Var(\hat{\beta}_k) \end{pmatrix}$$

$\therefore Var(\hat{\beta})_{ij} = Var(\hat{\beta})_{ji}$

2) Random Regressors

$$\hat{\beta} = \beta + (X'X)^{-1}X'u$$

(随机变量) 且 $E[\beta|X] = \beta$ (无偏)

如 X fixed: $E[\hat{\beta}] = \beta$, 假设 $E[u] = 0$

如 X random: $E[\hat{\beta}] = \beta$, 假设 $E[u|X] = 0$

(有条件X)

Law of Iterated Expectation (迭代期望) (将有条件期望转为无条件期望)

$$E[\hat{\beta}] = E_X(E[\hat{\beta}|X]) = E_X(\beta) = \beta$$

E[.] 是X分布下的期望

3) p-Values

p-value 下的 decision rule

Reject H_0 if p-value $< \alpha$

$\alpha = P$ (I型误差 — H_0为真) 依研究设置

P值是当H_0为真的极端统计检测的概率

数值越小, 犯I类误差或伪正的概率越小

4) Asymptotic preliminaries

Relax some assumptions, 无法获取 exact finite sample distribution

但可以近似分布

· $n \to \infty$ 样本量趋于无限 (概念框架, 不用真无限)

· LLN (Law of Large Numbers) 大数定律: 真实均值下样本平均覆盖度

· CLT (Central Limit Theorem) 正态分布下平均覆盖度

在考虑 $n \to \infty$ 时的估计量的反应有用

OLS估计量: $\hat{\beta} = (X'X)^{-1}X'Y$

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & \vdots & \\ 1 & x_{1n} & x_{2n} \end{pmatrix}$$

$$= \begin{pmatrix} n & \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{2i} \\ \sum_{i=1}^n x_{1i} & \sum_{i=1}^n x_{1i}^2 & \sum_{i=1}^n x_{1i}x_{2i} \\ \sum_{i=1}^n x_{2i} & \sum_{i=1}^n x_{1i}x_{2i} & \sum_{i=1}^n x_{2i}^2 \end{pmatrix}$$

Weak

The WLLN and CLT consider the behaviour of averages:

$$\frac{1}{n}\mathbf{X}'\mathbf{X} = \begin{pmatrix} \frac{1}{n}n & \frac{1}{n}\sum x_i \\ \frac{1}{n}\sum x_i & \frac{1}{n}\sum x_i^2 \end{pmatrix}$$

i.e. each element of $\mathbf{X}'\mathbf{X}$ can be represented in terms of an average!

Similarly for $\mathbf{X}'\mathbf{y}$:

$$\frac{1}{n}\mathbf{X}'\mathbf{y} = \begin{pmatrix} \frac{1}{n}\sum y_i \\ \frac{1}{n}\sum (x_i y_i) \end{pmatrix}$$

Asymptotic Properties of OLS (A1 ~ A4)

$$\hat{\beta} = (X'X)^{-1}X'Y$$

代入 $Y = X\beta + u$ ∴ $= (X'X)^{-1}X'(X\beta + u)$

$$= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u$$

$$= \beta + (X'X)^{-1}X'u$$

β到左，$\frac{1}{n}$到右 $\hat{\beta} - \beta = \frac{1}{n}\left((X'X)^{-1}X'u\right)$

$$\underbrace{\hat{\beta} - \beta}_{\text{'bias'}} = \underbrace{\left(\tfrac{1}{n}X'X\right)^{-1}}_{A} \underbrace{\tfrac{1}{n}X'u}_{B} = A^{-1}B$$

random variable

(Consistency)

如 $\hat{\beta} - \beta \xrightarrow{p} 0$，则 $\hat{\beta}$ 恒定

如 $E[x_i u_i] = 0$ (A4)；$\{x_i, u_i\}$ 为 iid (A2) 则 LLN 给出的 $B \xrightarrow{p} 0$

仍可在 A^{-1} 上用LLN且假设覆盖度的矩阵为恒定值 M.

$$\hat{\beta} - \beta \xrightarrow{p} \underbrace{M}_{A^{-1}} \cdot \underbrace{0}_{B} = 0$$

(Normality)

$$t = \frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)}$$

运用 $\hat{\beta} - \beta$：$\sqrt{n}(\hat{\beta} - \beta) = \underbrace{\left(\tfrac{1}{n}X'X\right)^{-1}}_{M} \underbrace{\left(\tfrac{1}{\sqrt{n}}X'u\right)}_{\overset{a}{\sim} N(0,P)}$

运用CLT：

$$\sqrt{n}(\hat{\beta} - \beta) \overset{a}{\sim} N(0, MPM')$$

In **Semester 1**:

GM assumptions (A1 to A5) $\Longrightarrow$ OLS is BLUE

CLM assumptions $\Longrightarrow \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim N(\mathbf{0}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$

$$t = \frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} \sim t_{df}$$

In **Semester 2**:

(A1 to A4) are sufficient to derive $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{a}{\sim} N(\mathbf{0}, \mathbf{MPM}')$

With A5 (Homoskedasticity) $\Longrightarrow \widehat{Avar}(\boldsymbol{\beta}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$

$$t = \frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} \overset{a}{\sim} N(0, 1)$$

Importantly: A6 is no longer necessary

5) Auxiliary Regressions (辅助回归)

检验多种限制下：

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + u_i$$

$$H_0: \beta_2 = \beta_3 = \beta_4 = 0$$

1、Estimate restricted model：

$$y_i = \beta_0 + \beta_1 x_{1i} + u_i$$

& obtain 估测残差

$$\tilde{u}_i = y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_{1i}$$

(β_0 & β_1 是从 restricted model 中的 OLS 估计量)

2、$\tilde{u}$ 不变，x_{1i}，x_{2i}，x_{3i}，x_{4i} 的回归；观测其中的 R^2：

$$\tilde{u}_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \alpha_4 x_{4i} + e_i$$

3、计算统计检验，$LM = nR^2$

(Lagrane Multiplier)

零假设下该统计检验为 asymptotically χ^2 在 $q=3$ 自由度下的分布

只计算统计检验，$LM = nR^2$

(Lagrane Multiplier)

零假设下该统计检验为asymptotically χ^2 在 $q=3$ 自由度下的分布

Idea: 只有零假设 not valid 下 null model 中的回归残差才有关联

· 若 null is valid, 辅助回归的回归值和null残差应接近于0和R^2

L2 Heteroskedasticity

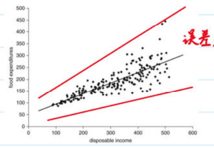
1) 同方差性假设

$$Y = X\beta + u$$

高—马假设 5：Homosekedasticity

$Var[u_i|x_i] = \sigma^2$ or $Var[u|X] = \sigma^2 I_n$

如果不是，则为异方差性



误差越来越大

2) 异方差性的后果

因为A5建立BLUE的OLS估计量

如A5不再成立

BLUE中Best (Efficient)，不再成立

$\therefore \hat{\beta} = (X'X)^{-1}X'Y$

仍是 valid linear unbiased，但并非最佳的那个

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

$A_1 - A_4$： $Var(u_i|x_i) = \sigma_i^2$

OLS 估计量：$\hat{\beta}_1 = \beta_1 + \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

$$\therefore \quad Var(\hat{\beta}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma_i^2}{SST_x^2}$$

$$Y = X\beta + u$$

假设异方差误差：$\boxed{Var(u|X) = \Omega} \neq \sigma^2 I_n$

$Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$

则：$Var(\beta) = (X'X)^{-1}X' \cdot Var(u|X)X(X'X)^{-1}$

$= (X'X)^{-1}X'\Omega X(X'X)^{-1}$

3) 加权最小平方 (Weighted Least Squares)

(更改误差权重，使之变成一个新的无异方差的模型 (小的拉大，大的加小))

$Y = X\beta + u$, $Var(u|X) = \Omega$

则矩阵P：

$$\underset{(N\times N)}{\Omega} = \underset{(N\times N)}{P} \underset{(N\times N)}{P'}$$

$$P^{-1}\Omega P'^{-1} = I_n$$

假设误差项服从正态分布：$u \sim N(0, \Omega)$

则转变后的误差项：

$v = P^{-1}u \sim N(0, P^{-1}\Omega P'^{-1}) = N(0, I_n)$

乘以P^{-1}　　Mean 0仍为0 (线性转换)　　同方差误差项

($Y = X\beta + u$ 全部重复乘P^{-1})

$\therefore$ 模型为：$P^{-1}Y = P^{-1}X\beta + P^{-1}u$

$$\tilde{Y} = \tilde{X}\beta + v$$

$v \sim N(0, I_n)$；进行OLS估计，β不变

WLS 估计β：

$$\hat{\beta}_{WLS} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}/\tilde{Y}$$

WLS 估计 β:

$$\hat{\beta}_{WLS} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y}$$
$$= (X'(P^{-1})'P^{-1}X)^{-1}X'(P^{-1})'P^{-1}Y$$
$$= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$$

方差: $Var(\hat{\beta}_{WLS}) = (X'\Omega^{-1}X)^{-1}$

(当 A1-A4 成立时, $\hat{\beta}_{WLS}$ 是 BLUE ($\hat{\beta}_{WLS} \neq \hat{\beta}_{OLS}$))

4) 因 Ω 未观测且未知, P 如何得知?

假设 σ_i^2 是倍数关系, z_i

$$P = \begin{pmatrix} \sqrt{z_1} & 0 & \cdots & 0 \\ 0 & \sqrt{z_2} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sqrt{z_N} \end{pmatrix}$$

如: $expend_i = \gamma_0 + \gamma_1 inc_i + u_i$, Suspect $Var(u_i|inc_i) = \sigma^2 inc_i$

$$\tilde{\mathbf{Y}} = \mathbf{P}^{-1}\mathbf{Y} = \begin{pmatrix} \sqrt{inc_1} & 0 & \cdots & 0 \\ 0 & \sqrt{inc_2} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sqrt{inc_N} \end{pmatrix}^{-1} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$
$$= \begin{pmatrix} y_1/\sqrt{inc_1} \\ y_2/\sqrt{inc_2} \\ \vdots \\ y_N/\sqrt{inc_N} \end{pmatrix}$$

($\tilde{y}_i$ 为 y_i 重新加权后的, $\widetilde{inc}_i$ 同理)

并未将 P 的方差矩阵减到 I_n

如: $P^{-1}u = v$

v 的方差矩阵 $P^{-1}\Omega P^{-1}$, Ω 是动态矩阵, 其每项动态元素均等同于 $\sigma^2 inc_i$

∴ v 的方差矩阵:

$$\frac{\sigma^2 inc_i}{\sqrt{inc_i}\sqrt{inc_i}} = \sigma^2$$ (恒定值)

∴ v 误差项为同方差, 成变量估计量为 BLUE

5) FGLS (Fesible)

如果不止 Suspect 1 个方差影响残差方差, $Var(u_i)$

Argue: $Var(u_i)$ 有某种线性关系 z_i

需确保替换进 P 的对角 (Diagonal of P) 的元素的有效性

GLS 让 Ω 成为非对角 (non-diagonal)

估计线性组合 → 可行的 GLS

6) HS 检测

如: $Sprice_i = \alpha_0 + \alpha_1 livarea_i + \alpha_2 age_i + u_i$

一些回归方差的测量与一些变量系统性地相关

(Auxiliary Regression)

(并非好的回归残差) $\hat{u}_i = \gamma_0 + \gamma_1 livarea_i + \gamma_2 age_i + \epsilon_i$

($\hat{\gamma}_1$ 与 $cov(\hat{u}_i, livarea_i)$ 成比例关系, 故 $\hat{\gamma}_1 = 0$)

$|\hat{u}_i|$ 或 $\hat{u}_i^2$ 才是好的回归残差

(辅助回归) $\hat{u}_i^2 = \gamma_0 + \gamma_1 livarea_i + \gamma_2 age_i + \epsilon_i$

假设检验:

H_0: 没有异方差性, $\gamma_1 = \gamma_2 = 0$ 或 $R^2 = 0$

H_A: 异方差性, $\gamma_1, \gamma_2 \neq 0$, 或 $R^2 > 0$

$$LM = nR^2 \overset{a}{\sim} \chi^2_k$$

Lagrange Multiplier (拉格朗日乘子法)

Asymptotic (渐近)

大部分情况采取:

1、估测回归模型, 并记录估计的残差 ($\hat{u}_i$, $i = 1, 2, \ldots, n$)

$$Sprice_i = \alpha_0 + \alpha_1 livarea_i + \alpha_2 age_i + u_i$$

2、用辅助回归 (Auxiliary Regression), 并记录辅助回归的 R^2

$$\hat{u}_i^2 = \gamma_0 + \gamma_1 livarea_i + \gamma_2 age_i + \epsilon_i$$

3、计算统计检验

2、用辅助回归（Auxiliary Regression），并记录辅助回归的 R^2

$$\hat{u}_i^2 = \gamma_0 + \gamma_1 livarea_i + \gamma_2 age_i + \epsilon_i$$

3. 计算统计检验

$$LM = nR^2$$

4. 测试假设

$H_0: \gamma_1 = \gamma_2 = 0 \longrightarrow nR^2 = 0$

$\longrightarrow$ 同方差残差

H_A: 任意 $\gamma_j \neq 0$，在 $j = 1, 2, \ldots, k \longrightarrow nR^2 > 0$

$\longrightarrow$ 异方差残差

$$LM = nR^2 \overset{a}{\sim} \chi_k^2$$

如果 $nR^2 > \chi^2_{k,\alpha,crit}$：拒绝 H_0

批评：辅助回归能包括任何所怀疑是有异方差因素的变量

7）怀特检验（White's test）

（可用于辅助回归的额外项）

$$\hat{u}_i^2 = \gamma_0 + \gamma_1 livarea_i + \gamma_2 age_i + \gamma_3 livarea_i^2 + \gamma_4 age_i^2 + \gamma_6 livarea_i \times age_i + \epsilon_i$$

（加入平方与相交项）（是否能找到至少一个与误差项平方有关的变量）

一样的

替代公式：

$$\hat{u}_i^2 = \gamma_0 + \gamma_1 \widehat{sprice_i} + \gamma_2 \widehat{sprice_i}^2 + \epsilon_i$$

（仅用第一个模型中的一个变量的拟合值及拟合值的平方）

另伯特（怀特）标准差

如果误差项中有异方差：

$$Var(\hat{\beta}) = (X'X)^{-1} X' \Omega X (X'X)^{-1}$$

为了估测 $Var(\hat{\beta})$，要

$$\Omega = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_N^2 \end{pmatrix}$$

再将 Ω 的估计量 $\hat{\Omega}$ 代入 $\widehat{Var}(\hat{\beta})$ 一并估计

$$\hat{\Omega} = \begin{pmatrix} \hat{u}_1^2 & 0 & \cdots & 0 \\ 0 & \hat{u}_2^2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \hat{u}_N^2 \end{pmatrix}$$ （直接将 $\hat{u}_i^2$ 用于 $\hat{\Omega}$ 不好）

但可以：$\widehat{Var}(\hat{\beta}) = (X'X)^{-1} X' \hat{\Omega} X (X'X)^{-1}$

$\because$ 接近于正态分布

$$\hat{\beta} \overset{a}{\sim} N(\beta, (X'X)^{-1} X' \hat{\Omega} X (X'X)^{-1})$$

$\therefore \widehat{Var}(\hat{\beta})$ 可用于 t 检测

此为异方差稳健标准差（Heteroskedastic robust interference / Standard errors）

L_3 Time Series Properties

1）时间序列

$\{\ldots, y_{t-s}, y_{t-s+1}, \ldots, y_{t-2}, y_{t-1}, y_t, \ldots y_{t+j} \ldots\}$

过去的事情可影响将来

则 Y 为时间序列模型（或随机过程）

$$y_t = x_t \beta + u_t$$

$x_t = \{z_{1t}, z_{2t} \ldots\}$

$x_t = \{z_t, z_{t-1}, \ldots\}$（历史事件）

$x_t = \{z_t, z_{t-1}, y_{t-1}, \ldots\}$

2）稳定性（Stationarity）& 弱相关性（Weak Dependence）

1. 协方差平稳

· $\{y_t : t = 1, 2, \ldots\}$，当 1）$E[y_t]$ 为恒定的 2）$Var(y_t)$ 不变，3）$h \geq 1$ 在任何 t 下，$Cov(y_t, y_{t+h})$ 仅和 h 非 t 有关时，有着有限二阶矩（finite secound moment）（$E[y_t^2] < \infty$）的时间序列模型为协方差平稳

2、弱相关

· 不准确定义：当 y_t 和 y_{t+h} 随 h 的无弹跳（bound）增大"几乎不相关"时，稳定的时间序列模型 $\{y_t : t = 1, 2, \ldots\}$ 为弱相关（现在的某事与过去的随着过去的时间越来越久相关性降低）

2、弱相关

· 不准确定义：当 y_t 和 y_{t+h} 随 h 的无弹跳（bound）增大"几乎不相关"时，稳定的时间序列模型 $\{y_t : t=1,2,\dots\}$ 为弱相关（现在的某事与过去的随着过去的时间越来越久相关性降低）

稳定协方差的时间序列条件：

1. 时间序列方差不随时间变化
2. 时间序列的均值不随时间变化
3. 序列的协方差取决于滞后值而非观测时间

3) OLS 假设

TS1：真实模型 $y_t = X_t\beta + u_t,\ t=\{1,2,\dots,T\}$ 为稳定协方差且有弱相关性

TS2：与 X_t 无完全相关性（Perfect Correlation）

TS3：零均值（ZCM）：$E[u_t|X_t]=0$

TS4：同方差：$Var(u_t|X_t)=\sigma^2$

TS5：无误差序列相关：$Cov(u_t, u_s|X_t, X_s)=0$ 对于所有 $s\neq t$

· TS1 ~ TS3，OLS 为一致（Consistent）

· TS1 ~ TS5，OLS 估计量近乎于正态分布：t.F.LM 检测皆有效

伪回归（Spurious regression）（违背 OLS 假设）

（不稳定 且/或 无弱相关会造成问题）

（对 LLNs 和 CLTs 有效）

4) 单变（univariate）TS 模型

变量 × 过去时间段的误差

$$y_t = \underbrace{\beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2}}_{\text{Auto regression}} + \cdots + \underbrace{u_t + \phi_1 u_{t-1} + \phi_2 u_{t-2}}_{\text{Moving regression}} + \cdots$$

向量自回归模型（Autoregression）

如：与 k 有关的，AR(k)，

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \cdots + y_{t-k} + u_t$$

（包括了大多时间序列常见的特征）

主要特征：$\rho_h = Corr(y_t, y_{t-h})$ 对于滞后 $h=1,2,3,\dots$

5) 随机游走（Random walk）

如：AR：

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t$$

（i.i.d（独立同分布）；均值为0；$Var(u_t)=\sigma_u^2$）

（$|\beta_1|<1$）

且已知：$y_t : t=0,1,\dots$ 为随机过程

则：unconditional moments：

$$E[y_t] = \mu = \frac{\beta_0}{1-\beta_1}$$

$$Var[y_t] = \sigma^2 = \frac{\sigma_u^2}{1-\beta_1^2}$$

（独立时间，故有稳定协方差）

6) AR 的自相关

$$\cdot\ Cov(y_t, y_{t-1}) = \sigma_u^2 \frac{\beta_1}{1-\beta_1^2} = \sigma^2\beta_1 \implies \rho_1 = corr(y_t, y_{t-1}) = \beta_1$$

$$\cdot\ Cov(y_t, y_{t-2}) = \sigma_u^2 \frac{\beta_1^2}{1-\beta_1^2} = \sigma^2\beta_1^2 \implies \rho_1 = corr(y_t, y_{t-2}) = \beta_1^2$$

$$\vdots$$

$$\cdot\ Cov(y_t, y_{t-k}) = \sigma_u^2 \frac{\beta_1^k}{1-\beta_1^2} = \sigma^2\beta_1^k \implies \rho_1 = corr(y_t, y_{t-k}) = \beta_1^k$$

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t$$

$$y_t = \frac{\beta_0}{1-\beta_1} + \sum_{i=0}^{\infty} \beta_1^i u_{t-1}$$

$$y_{t-s} = \frac{\beta_0}{1-\beta_1} + \sum_{i=0}^{\infty} \beta_1^i u_{t-s-i}$$

$$E[y_t] = E[y_{t-s}] = \frac{\beta_0}{1-\beta_1}$$

代入：

$$y_t - E[y_t] = \sum_{i=0}^{\infty} \beta_1^i u_{t-i}$$

$$y_t - E[y_{t-s}] = \sum_{j=0}^{\infty} \beta_1^j u_{t-s-j}$$

$$Cov[y_t, y_{t-s}] = E[(y_t - E[y_t])(y_{t-s} - E[y_{t-s}])]$$

$$= E\left[\sum_{i=0}^{\infty} \beta_1^i u_{t-i} \sum_{i=0}^{\infty} \beta_1^i u_{t-s-i}\right]$$

则：

$$Cov[y_t, y_{t-s}] = \beta_1^s(1+\beta_1^2+\beta_1^4+\cdots)\sigma_u^2$$

又：$(1+\beta_1^2+\beta_1^4+\cdots) = \frac{1}{1-\beta_1^2}$；当 $(1-\beta_1^2)(1+\beta_1^2+\cdots)=1$

故：$Cov(y_t, y_{t-s}) = \frac{\beta_1^s}{1-\beta_1^2}\sigma_u^2$

L_4 I. 检测并纠正非平稳性

1) 平稳性检测

一、画图

二、用时间序列的自相关模型

（如在许多滞后中有大且持续的相关性，则有可能为非平稳时间序列）

三、测试单位根（迪基-福勒检验（Dickey-Fuller Unit Root test））

（如有单位根，则有非稳定序列）

三、测试单位根（迪基-福勒检验（Dickey-Fuller Unit Root test））

（如有单位根，则有非稳定序列）

DF Test： $\Delta y_t = \beta_0 + \beta_1 t + \beta_3 y_{t-1} + u_t$

$H_0: \beta_3 = 0$

$H_A: \beta_3 < 0$

$\Delta y_t = y_t - y_{t-1}$

（拒绝则确认为稳定性）

或用扩张迪基-福勒检验（Augmented Dickey - Fuller test）

ADF test： $\Delta y_t = \beta_0 + \beta_1 t + \beta_3 y_{t-1} + \beta_4 \Delta y_{t-1} + \cdots + u_t$

$H_0: \beta_3 = 0$

$H_A: \beta_3 < 0$

2）纠正非平稳性

（使用 1.Detrending 和 2.Differencing 法）

1. Detrending（除趋势）

（在具有上升趋势的模型中，可以添加趋势项）

$$LEX_t = \alpha_0 + \alpha_1 t + u_t$$

（t 为时间线变量，随时间增加）

除趋势数据则为：

$$\hat{u}_t = LEX_t - \hat{\alpha}_0 - \hat{\alpha}_1 t$$

2. Differencing a Series（差分时间序列）

$$\Delta y_t = y_t - y_{t-1}$$

$$\Delta^2 y_t = \Delta(\Delta y_t) = y_t - 2y_{t-1} + y_{t-2}$$

3）单变量时间序列

·AR(1) 模型

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t$$

（误差项均值为0，i.i.d.（独立同分布）与 $Var(u_t) = \sigma_u^2$）

·非条件期望

$$E[y_t] = \mu = \frac{\beta_0}{1-\beta_1}$$

（是 y_t 在除过程变量下无其它任何知识的期望）

条件瞬间及预报

·类似于估计

·预测未来值 $y_{t+1}, y_{t+2}, \ldots, y_{t+h}$

·步骤一、二、…，h 在预报之前

·y_{t+1} 的非条件均值仍是 μ

信息集合：$I_t = y_t, y_{t-1}, y_{t-2}, \ldots$

4）条件期望

AR(1)： $E[y_{t+1}|I_t] = E[\beta_0 + \beta_1 y_t + u_{t+1}|I_t]$

$= \beta_0 + \beta_1 E[y_t|I_t] + E[u_{t+1}|I_t]$

$= \beta_0 + \beta_1 y_t + 0$

（大部分情况：$E[y_{t+1}|I_t] \neq E[y_t]$）

如：AR(1)： $y_t = 0.2 + 0.5y_{t-1} + y_t$

非条件期望：$E[y_{t+1}] = \frac{\beta_0}{1-\beta_1} = \frac{0.2}{1-0.5} = 0.4$

信息集合：$y_{t-3} = 0.02$，$y_{t-2} = 0.55$，$y_{t-1} = 0.71$，$y_t = 0.64$

条件期望：$E[y_{t+1}|I_t] = 0.2 + 0.5y_t = 0.52$

$E[y_{t+2}|I_t] = E[0.2 + 0.5y_{t+1} + u_{t+2}|I_t]$

·条件预报与当前信息预报值的非条件期望指示相交

·稳定性和弱相关性在当 $|\beta_1| < 1$ 时的 AR(1) 中

·稳定过程反应出相同的均值

AR(1) —> 单调收敛（Monotonic Convergence）

AR(k)，k > 1 —> 更复杂的

L5 自相关的问题

1) 误差相关

误差项中包含解释变量以外的其它所有

· 横截面/纵向数据聚类 (Clustering)

· Class room effect

· Spatial relationship

· TS数据中的时间关系

· 时间彼此靠近的观测值可能有与彼此相关的误差项

· 称为自相关/序列相关

2) 自相关

假设：(TS数据中的常见问题是误差无自相关)

TS： $y_t = x_t\beta + u_t$

假设 TS5：无自相关

$$Corr(u_t, u_{t-s} \mid x_t, x_{t-s}) = 0 \quad \forall s \neq 0$$

故：自相关违反TS5

自相关形成原因：

1、经济事件有持久性影响，会造成一系列正或负的残差

2、特定错误模型(misspecified) —— 功能形式

3、特定错误模型 —— 省略变量

例：一个一元线性回归但误差项非 i.i.d (独立同分布)但相关

AR(1)：

$$y_t = x_t\beta + u_t$$

$$u_t = \rho u_{t-1} + v_t$$

$$v_t \sim N(0, \sigma_v^2)$$

· 当 $\rho \neq 0 \Longrightarrow corr(u_t, u_{t-1}) \neq 0$. 变量 ρ 决定关联有多强

· v_t 可能与零平均值和方差 σ_v^2 有正态分布

如：$u_t = \rho u_{t-1} + v_t$ 对 u_t 为真，则 u_{t-1} 同样有效

$$u_{t-1} = \rho u_{t-2} + v_{t-1}$$

代入 $u_{t-1} = \rho u_{t-2} + v_{t-1}$

$$u_t = \rho(\rho u_{t-2} + v_{t-1}) + v_t$$

$$= v_t + \rho v_{t-1} + \rho^2 v_{t-2} + \rho^3 v_{t-3} + \cdots$$

(决定当今扰乱) (扰乱系统的随机) (之前的扰乱)

(|ρ| > 1，过去对如今有较大影响，反之易然)

u_t 回归量的性质：

$$E(u_t) = 0, \quad Var(u_t) = \sigma^2 = \frac{\sigma_v^2}{1-\rho^2}$$

$$corr(u_t, u_{t-1}) = \rho, \quad cov(u_t, u_{t-1}) = \frac{\sigma^2\rho}{1-\rho^2}$$

$$Corr(u_t, u_{t-k}) = \rho^k, \quad cov(u_t, u_{t-k}) = \frac{\sigma^2\rho^k}{1-\rho^2}$$

(违反TS5，故 $\rho = 0$)

3) 自相关的后果

$$\begin{cases} y_t = x_t\beta + u_t \\ u_t = \rho u_{t-1} + v_t \end{cases}$$

可写为 $Y = X\beta + u$

$$Var(u) = \Omega \neq \sigma^2 I_T$$

$$= \frac{\sigma^2}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \rho \\ \rho^{T-1} & \rho^{T-2} & \cdots & \rho & 1 \end{pmatrix}$$

(T为样本量)

· 非对角协方差：$cov(u_t, u_{t-k}) = \frac{\sigma^2\rho^k}{1-\rho^2}$ $(\neq \sigma^2 I_T)$

(AR(1)中不同的残差过程会有不同的矩阵 Ω)

· 异方差下 $\hat\beta$ 不变，$Var(\hat\beta)$ 也相同

$$Var(\hat\beta) = (X'X)^{-1}X'\Omega X(X'X)^{-1}$$

$$= \sigma^2(X'X)^{-1} \quad \text{(只有当 } \Omega = \sigma^2 I_T \text{ 时)}$$

假设 TS1−5，$\hat\beta \stackrel{a}{\sim} N(\beta, \sigma^2(X'X)^{-1})$

$$Var(\hat{\beta}) = (X'X)^{-1}X'\Omega X(X'X)^{-1}$$

$$= \sigma^2(X'X)^{-1}$$ （只有当 $\Omega=\sigma^2 I_T$ 时）

假设TS1-5：$\hat{\beta} \overset{a}{\sim} N(\beta, \sigma^2(X'X)^{-1})$

假设TS1-3：$\hat{\beta} \overset{a}{\sim} N(\beta, (X'X)^{-1}X'\Omega X(X'X)^{-1})$

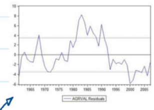
检验 β 中 i^{th} 元素里的零假设，β_i

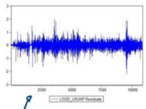
如：$H_0: \beta_i = a$（t-检验）

$$t = \frac{(\hat{\beta}_j - a)}{s.e.(\hat{\beta}_j)} \overset{a}{\sim} N(0,1)$$

（如 $\hat{\beta} \overset{a}{\sim} N(\beta, \sigma^2(X'X)^{-1})$ 中 $s.e.(\hat{\beta}_j)$ 计算错，统计检验不近似于正态分布）

4）检测自相关





· 均值为零上下的观测值

· 更大的随机

· u_t 与前观测值 u_{t-1} 相关

（用 Breusch-Godfrey 而非 Durbin-Watson 检验）

如：
$$\begin{cases} y_t = X_t\beta + u_t \\ u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_k u_{t-k} + v_t \end{cases}$$

· X_t 可能有常数、滞后关联变量、滞后解释变量

· $\hat{u}_t$，预估 u_t 和审查 $\hat{u}_t$ 与滞后变体 $\hat{u}_{t-1}, \hat{u}_{t-2}\ldots$ 等的关系

辅助回归（有自相关下）步骤：

1、OLS估计回归模型

2、保留 $\hat{u}_t$

3. $\hat{u}_t = \gamma + X_t\delta + \rho_1\hat{u}_{t-1} + \rho_2\hat{u}_{t-2} + \cdots + \rho_k\hat{u}_{t-k} + v_t$

4、检验 k 的残差自相关

$H_0: \rho_1 = \rho_2 = \cdots = \rho_k = 0$

H_A: any $\rho_i \neq 0$，对于 $i=1,\ldots,k$

用 $LM(nR^2 \overset{a}{\sim} \chi^2_k)$ 当 n # 辅助回归中的观测值（要求条件同方差 v_t）（Odd phrase）

· P 是最长的滞后，故用 z_{t-4} 为 X_t 的要素时失去 P

· 对于自相关的辅助回归需包含全部解释变量

· LM检验的灵活性：辅助回归会包括滞后回归（$i=1,3,12$）无需全部的（$k=12$）

（最大的是否要包括，取决于数据频率）

5）解决自相关

出现原因（自相关）

1、有误的模型

· 用了非稳定性数据

· 遗失变量（滞后相关/解释 变量）

2、真实自相关

行为：1、修补模型

1、用正确的功能式（稳定性转换）

2、将相关滞后变量包括进来

2、运用稳健干扰程序 —— 即：Newey West

例如：

$$\Delta UR_t = \beta_0 + \beta_1\Delta CPI_t + \beta_2\Delta FFIR_t + u_t$$

· 这种广义最小二乘法（General Least Square）重制模型使完美的误差项

· 当为复杂自相关且/或 x_t 中有滞后相关变量且/或残差为条件异方差的

则：用 Variance - Covariance Matrix 去获得有效的大样本推断

$$Var(\hat{\beta}) = (X'X)^{-1}X'\Omega X(X'X)^{-1}$$

（自相关残差）

参照异方差残差：

$$\Omega = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_T^2 \end{pmatrix}$$

参照异方差残差：

$$\Omega=\begin{pmatrix}\sigma_1^2 & 0 & \cdots & 0\\ 0 & \sigma_2^2 & & \vdots\\ \vdots & & \ddots & 0\\ 0 & \cdots & 0 & \sigma_T^2\end{pmatrix}$$

和：

$$\hat\Omega_{White}=\begin{pmatrix}\hat u_1^2 & 0 & \cdots & 0\\ 0 & \hat u_2^2 & & \vdots\\ \vdots & & \ddots & 0\\ 0 & \cdots & 0 & \hat u_T^2\end{pmatrix}$$

自相关残差（Ω为非零非对角元素）

$$\Omega=\begin{pmatrix}\sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1T}\\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \cdots & \sigma_{2T}\\ \sigma_{31} & \sigma_{23} & \sigma_3^2 & \cdots & \sigma_{3T}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \sigma_{T1} & \sigma_{T2} & \sigma_{T3} & \cdots & \sigma_T^2\end{pmatrix}$$

估测的σ_{ij}即$\hat\sigma_{ij}=\hat u_i\hat u_j$

$$\hat\Omega=\begin{pmatrix}\hat u_1^2 & \hat u_1\hat u_2 & \hat u_1\hat u_3 & \cdots & \hat u_1\hat u_T\\ \hat u_2\hat u_1 & \hat u_2^2 & \hat u_2\hat u_3 & \cdots & \hat u_2\hat u_T\\ \hat u_3\hat u_1 & \hat u_3\hat u_2 & \hat u_3^2 & \cdots & \hat u_3\hat u_T\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \hat u_T\hat u_1 & \hat u_T\hat u_2 & \hat u_T\hat u_3 & \cdots & \hat u_T^2\end{pmatrix}$$

但，$\hat\Omega$对于 $Var(\hat\beta)=(X'X)^{-1}X'\Omega X(X'X)^{-1}$ 无用！

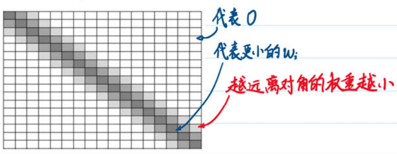
$\because \hat\Omega=\hat u\hat u'$

$$\therefore X'\hat\Omega X = X'\hat u\hat u' X$$
$$=(X'\hat u)(\hat u' X)$$
$$=0$$

$\therefore$使用Newey and West（1987）的方法

$$\hat\Omega_{NW}=$$
$$\begin{pmatrix}\hat u_1^2 & w_1\hat u_1\hat u_2 & w_2\hat u_1\hat u_3 & 0 & \cdots & 0\\ w_1\hat u_1\hat u_2 & \hat u_2^2 & w_2\hat u_2\hat u_3 & w_2\hat u_2\hat u_4 & \cdots & 0\\ w_2\hat u_1\hat u_3 & w_1\hat u_2\hat u_3 & \hat u_3^2 & w_1\hat u_3\hat u_4 & \cdots & 0\\ 0 & w_2\hat u_2\hat u_4 & w_1\hat u_3\hat u_4 & \hat u_4^2 & \cdots & 0\\ \vdots & \vdots & \ddots & & \ddots & \vdots\\ \vdots & \vdots & & \ddots & & w_2\hat u_{T-2}\hat u_T\\ \vdots & \vdots & & & \hat u_{T-1}^2 & w_1\hat u_{T-1}\hat u_T\\ 0 & \cdots & 0 & w_2\hat u_{T-2}\hat u_T & w_1\hat u_{T-1}\hat u_T & \hat u_T^2\end{pmatrix}$$

方法图解



（示例中0权重仅对前两个非对角时）

$$\widehat{Var_{nw}}(\hat\beta)=(X'X)^{-1}X'\hat\Omega_{nw}X(X'X)^{-1}$$

（以上操作同样适合异方差残差；且变量估计不变，仍具有TS1－TS3）

再进行检测（用$\widehat{Var_{nw}}(\hat\beta)$可达到近似）

$$t-test_{NW}=\frac{\hat\beta_i-0.5}{s_{\hat\beta_{i,nw}}}\overset{a}{\sim}N(0,1)$$

因为原先的OLS t检测不知其近似分布

$$t-test_{OLS}=\frac{\hat\beta_i-0.5}{s_{\hat\beta_{i,OLS}}}\overset{a}{\sim}\ ?$$

L6 季节性与因果关系

1）季节性影响

来源：
- 天气（冬季能源消耗上涨）
- 行政（Administrative）（学年初和末）
- 社会、文化、宗教（圣诞购物）
- 历法变量（闰年）

影响：

如：时间序列 $Y_t=X_t\beta+u_t$

若误差项中包含季节性，一些假设（ZCM，异方差，自相关…）失效

解决：纳入虚拟变量

$$Y_t=X_t\beta+\alpha_2 feb_t+\alpha_3 mar_t+\cdots+u_t$$

解决：纳入虚拟变量

$$Y_t = X_t\beta + \alpha_2 feb_t + \alpha_3 mar_t + \cdots + u_t$$

（移除了一个变量，无多重共线性问题）

2）时间序列因果关系

加和乘法模型（多变量时间序列）

趋势（T_t），季节性（S_t），周期性（C_t），不寻常（I_t）

$$Y_t = T_t + C_t + S_t + I_t$$

$$Y_t = T_t \times C_t \times S_t \times I_t$$

思路（总）：· 描绘时间序列图

· 单位根检验（ADF）

· 平稳性检验

不平稳的话：· 协整检验（恩格尔-格兰杰检验）

思路：

（如果存在协整关系，便可用传统检验方法分析两个变量间的长期均衡关系）

（各个经济变量间同时也有短期的失衡（除了长期均衡外）故要修正误差）

· 协整关系（Cointegrating Relationships）

（非平稳序列的经济变量的线性组合有可能是平稳的）

如：两个非稳定性序列（Y_t, X_t）

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 t + \epsilon_t$$

· 求出 $\hat{\epsilon}_t$，并用前几周的测试（平稳性检验）确认两个非平稳序列的线性组合是平稳的（趋势互相抵消）

· 故 Y_t 和 X_t 为 cointegrated（协整），即两个经济变量为长期均衡关系

一、协整关系检验（Engle - Granger / 恩格尔-格兰杰检验）

（可以用ADF及DF检验测试是否有基于 $\hat{\epsilon}_t$ 的协整关系）

恩格尔-格兰杰

假设存在协整关系

步骤一：基于OLS回归模型得到系数（β_1, β_2）

通过系数得到其估计量 $\hat{\beta}_1, \hat{\beta}_2$，在得到残差估计量 $\hat{\epsilon}_t$

步骤二：基于ADF检测残差序列的估计量 $\hat{\epsilon}_t$

若平稳（$\hat{\epsilon}_t$），则接受原假设；非平稳 $\hat{\epsilon}_t$，则拒绝

二、修正误差

Granger Representation Theorem 指出，对任何有协整关系的 Y_t, X_t，可以用修正机制表示之间的关系

如：长期均衡关系为：

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 t + \epsilon_t$$

短期关系为：

$$\Delta Y_t = \alpha_0 + \alpha_1 \Delta X_t + \alpha_2 \epsilon_{t-1} + V_t$$ （纠正模型 / Error Correlation Model）

（差分运算符）　（应是负的）

（Y的变化取决于X变化和滞后均衡误差项）

（α_2 的绝对值描述模型返回均衡的速率）

L8 设定检验与周氏检验

1）Specification Testing

如：MLR1：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

潜在问题：

· 异方差

· 自相关

· 省略变量

检验过程：

· RESET 检测（正确的函数方程）（REgression Specification Error Test）

· Chow 检测（结构差分）

一、遗失变量（Missing Variable）

如：$y_t = \alpha + \beta x_t + u_t$

1、添加新变量 z_t

$$y_t = \alpha + \beta x_t + \gamma z_t + u_t$$

2、T检验 $H_0: \gamma = 0$ 足以

若 x_t 和 y_t 为二次函数

（添加变量） $y_t = \alpha + \beta x_t + \gamma x_t^2 + u_t$

T检验 $H_0: \gamma = 0$

若无法明确替代法，则需指数是有‘些许错误’的检验

（无需知晓之前的‘东西’） 用拉姆齐 RESET（Ramsey RESET）

辅助回归

1、估测处于 H_0 下的式子（$y_t = \alpha + \beta x_t + u_t$）的模型为正确形式

2、用估计量 $\hat{\alpha}$ 和 $\hat{\beta}$ 以获得拟合值 $\hat{y}_t$

$$\hat{y}_t = \hat{\alpha} + \hat{\beta} x_t$$

3、估测辅助回归

$$y_t = \alpha + \beta x_t + \gamma_1 \hat{y}_t^2 + \gamma_2 \hat{y}_t^3 + u_t$$

标准F检验 $H_0: \gamma_1 = \gamma_2 = 0$

（用 $\hat{u}_t$ 为相关变量结果相同）

· 若拒绝 H_0，则 $y_t = \alpha + \beta x_t + u_t$ 设定中有些问题

· 设定偏误（Mis-specification）可导致其它问题

2）结构断裂

策略一（邹氏检验）：

Pool： $\widehat{sprice_i} = \hat{\beta}_0 + \hat{\beta}_1 livarea_i$

$R_p^2 = 0.685$；$RSS = 4.51 * 10^{11}$；$n = 130$

No Pool： $\widehat{sprice_i} = \hat{\gamma}_0 + \hat{\gamma}_1 livarea_i$

$R_{np}^2 = 0.647$；$RSS = 3.67 * 10^{12}$；$n = 2480$

若 price 和 living area 的关系并非取决于若 house 是 pool，则两模型是相同的

（受限模型 restricted model：用所有数据估计）

（非限制模型 unrestricted model：以数据的两个子集估计 i.e. $pool_i = 0 / 1$）

（零假设 H_0：用所有数据和以数据半样本估计的拟合值无差异）

（备择假设 H_A：以上二者有差异）

策略二：

$$\widehat{Sprice_i} = \hat{\delta}_0 + \hat{\delta}_1 livarea_i + \hat{\delta}_3 (livarea_i * pool_i)$$

$R_{dum}^2 = 0.665$；$RSS = 4.12 * 10^{12}$；$n = 2610$

虚拟变量：

$$\hat{\delta}_0 = \hat{\gamma}_0$$

$$\hat{\delta}_0 + \hat{\delta}_2 = \hat{\beta}_0$$

$$\hat{\delta}_1 = \hat{\gamma}_1$$

$$\hat{\delta}_1 + \hat{\delta}_3 = \hat{\beta}_1$$

∴ $\hat{\delta}_2 =$ 有 house 且无 pool 下稳定的差

$\hat{\delta}_3 =$ 有 house 且无 pool 下（living area 在 house price 的影响）斜率之差

$$\widehat{sprice_i} = \hat{\delta}_0 + \hat{\delta}_1 livarea_i + \hat{\delta}_2 pool_i + \hat{\delta}_3 (livarea_i * pool_i)$$

（F检验，$\hat{\delta}_2 = \hat{\delta}_3 = 0$，查是否 $\frac{sprice_i}{livarea_i}$ 与 house 有和无 pool 的关系一致）

$$F = \frac{\frac{RSS_r - RSS_u}{k}}{\frac{RSS_u}{dof_u}}$$

k ← 限制个数

dof_u ← 非限制模型中的 dof（自由度）

Dummy　　　　　　Chow

Dummy

$RSS_r = 4.14 * 10^{12}$ from __

$RSS_u = 4.12 * 10^{12}$ from __

$k = 2$ 为 $H_0: \hat{\delta}_2 = \hat{\delta}_3 = 0$

$dof_u = 2610 - 4 = 2606$

n

Chow

$RSS_r = 4.14 * 10^{12}$ from __

$RSS_u = 4.12 * 10^{12}$ from sum of __ and __

$k = 2$ 为每个模型有2个 coefficients 需估计

$dof_u = (2480 + 130) - (2 * 2) = 2606$

(no pool + pool 的观测值个数)

$$\therefore F = \frac{\frac{(4.14 - 4.12)}{2}}{\frac{4.12}{2610 - 4}} = 6.946$$

$F \sim F_{2,inf}$. $F_{cv,0.01} = 4.61$

拒绝 H_0

Dummy

- 两(或多)组进行求导
- 注意因引入过多虚拟变量致完全共线性 (perfect colineary problem)

Chow

- 只有当两子样本中的误差方差(error variance)是相同的时 $F_{k,T-2k}$ 分散(distributed)
- 若拒绝 H_0, 不知道是否是因交点还是斜率(或二者均有)导致的变动 ⟶ 用 dummy variables

3) 结构断裂的影响

对估测与预报

Inflation 的 AR(4)估计:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3} + \beta_4 y_{t-4} + u_t$$

Inflation 以 t 估测

非条件均值: 3.38% (过高)

Chow 检验:

(检验以全样本和子样本估计)

若无断裂 ⟶ 总拟合无变化

限制模型样本: 1989-Q2 ~ 2011-Q4

$n = 87$, $RSS_r = 351.8058$

非限制 1: 1989-Q2 ~ 1991-Q4

$n = 7$, $RSS_{u1} = 20.3287$

非限制 2: 1992-Q1 ~ 2011-Q4

$n = 7$, $RSS_{u2} = 197.0599$

$RSS_u = RSS_{u1} + RSS_{u2} = 20.3287 + 197.0599 = 217.388$

问: 拟合显著性是否不同?

Chow test:

$$F = \frac{\frac{RSS_r - RSS_u}{k}}{\frac{RSS_u}{df_u}}$$

$F \sim F_{k, T-2k}$

(参数个数)

H_0: 以所有数和子样本数据的拟合无差异

H_A: 二者不同

$F_{crit,0.05} = 2.32$, $F_{crit,0.01} = 3.23$

拒绝 H_0

L9 二元响应模型

1)

微观计量中有些问题为受限因变量,该情况为二元响应模型

(模型中事件发生的可能性)

如: 是否被雇佣

是否贷款成功

是否赚钱

如：是否被雇佣

是否贷款成功

是否赚钱

以 单位观测单位的特征看事件发生的条件概率

用虚拟变量模拟因变量)

$$y=\begin{cases}1 & \text{发生}\\ 0 & \text{未发生}\end{cases}$$

X 为解释变量向量

则（探究）的发生概率：$P(y=1|X)$

2) 线性概率模型（LPM）

如：$arrest_i = \beta_0 + \beta_1 income_i + \beta_2 race_i + \epsilon_i$

（=1事件中止；=0继续） （周收入$100） （=1有色种族；=0白人）

OLS 拟合模型：

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 income_i + \hat{\beta}_2 race_i$$

（推演各项数值）

假设遵守ZCM假设

$$E[y|income, race] = \beta_0 + \beta_1 income + \beta_2 race$$

当因变量为二元变量时

$$E[y|X] = P(y=1|X)$$

X中给出y的期望 = y为1的概率

故：

$\hat{y}$ — 预计中止的概率

$\hat{\beta}_0$ — 白人无收入时中止概率

$\hat{\beta}_1$ — 收入升至$100所改变的概率

$\hat{\beta}_2$ — 白人和有色人种间中止概率之差

但：·会估计"有问题"的概率

·异方差

·误差分布并非正态

·且因线性型式，简单的推演恐有局限

（假设最初很穷，额外$100有很大影响；

若初始富有，$100无较大影响，但

模型中影响均设为一致）

3) Logit 模型

·限制Y在 0-1 之间

·允许变量与x变量间有非线性关系（即允许当x减小时，y_i以较慢速度接近0；当x增大时接近1的速度更慢）

指数函数

定义出变量 I_i^* 当：

$$I_i^* = X_i\beta + \epsilon_i$$

$Y_i = 1$ 当 $I_i^* \geq 0$ 时 （当解释变量向量和误差项有正影响时，Y_i发生）

$Y_i = 0$ 当 $I_i^* \leq 0$ 时 （当～负影响时，Y_i不发生）

如此，便能告诉我们（如： 抽烟与否时）概率会如何变化

（定义指数函数与概率的联系很重要，并将定义使用哪些有限的因变量模型）

定义概率

故：$Pr[Y_i=1] = Pr[I_i^* \geq 0]$

$= Pr[X_i\beta + \epsilon_i \geq 0]$

$= Pr[\epsilon_i \geq -X_i\beta_i]$ （Y_i发生的概率 = 误差项的影响≥解释变量$X_i\beta_i$的反影响的概率）

可进一步假设误差项的分布为对称分布（Symmetric）：

$= P_r[\epsilon_i \geq -X_i\beta_i]$ （Y_i发生的概率 = 误差项的影响≥解释变量$X_i\beta_i$的反影响的概率）

可进一步假设误差项的分布为对称分布（Symmetric）：

$P_r[\epsilon_i \geq -X_i\beta] = P_r[\epsilon_i \leq X_i\beta]$

（所以分布的特定样式对于确定指数函数如何与我们成不成为吸烟者的概率非常重要，因为仍是 X 的函数）

4) Logistic累积分布函数（Logistic Cumulative Distribution Function）

（告诉给定分布 ≤x 时观测值的概率）

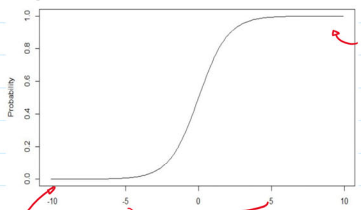
（Logistic 概率函数） $P_i = \frac{1}{1+e^{-z_i}}$

$z_i = X_i\beta$

$1-P_i = \frac{1}{1+e^{z_i}}$

（当 $z_i \in (-\infty, \infty)$ 时，$P_i \in (0,1)$ 且为非线性关系

Logistic CDF



$X_i\beta$

当x非常大时趋近1

x非常小时趋近0

中间时变化取决于增加的速率

5) 估测

（从比值比（odds ratio）入手探究）

$$\frac{P_i}{1-P_i} = \frac{1+e^{z_i}}{1+e^{-z_i}} = e^{z_i}$$

（两侧为自然对数）

$$\ln\left(\frac{P_i}{1-P_i}\right) = z_i = X_i\beta + \epsilon_i$$

（有log，所以叫 logit）

仍假设 odds ratio的对数为线性关系

P_i仍有界限 ，但变化的logit 变量无界限，能确保概率在范围中

（即：因为对以上所有都以 logic的结构设置了，概率都会是看起来真实的'，且参数现在表示赞成吸烟比不赞成者的几率的 log 如何增加）

仍要估计问题

（估计中 对于个人而言 $\ln(\frac{1}{0})$；$\ln(\frac{0}{1})$；$\ln(0)$ 为无限）

用最大可能性法估计，无需用 这些值（下周）

$$\hat{logit}_i = 2.74 - 0.02\,age_i - 0.09\,educ_i + 4.72e-06\,income_i - 0.002\,P_{cigs}$$

通过$\hat{z}_i$估计概率：

$$\hat{P}_i = \frac{1}{1+e^{-\hat{z}_i}}$$

$$\hat{z}_i = X_i\hat{\beta}$$

（找到每个因素在抽烟的概率上的边际效应是有难度的，因其取决于初始时 变量的值，通常在平均值下完成）

Logitic 模型：

$$P_r(y_i = 1 | X_i) = \frac{exp(X_i\beta)}{1+exp(X_i\beta)} = \frac{1}{1+exp(-X_i\beta)}$$

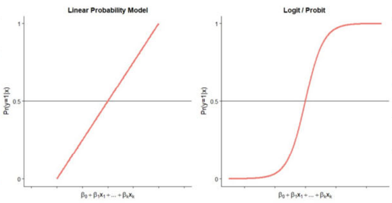
$$X\beta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Probit（多元概率比回归）模型：

$$Pr(y=1|X_i)=\Phi(X_i\beta)=\int_{\infty}^{X_i\beta}\phi(v)\,dv$$

（标准正态随机变量的概率）

区别：



（Logit / Probit 中的）

Pros：

预测概率是处在 0和1之间

非线性 $\Longrightarrow$ 非固定边际效应

Cons：

要更复杂的估计技巧

非线性 $\Longrightarrow$ 不能直接推出预测的 coefficients

6）例题

变量：
- age_i $=28$
- $educ_i$ $=15$ years
- $income_i$ $=\$12{,}500$
- $pcigs79_i$ 烟费 $=60.6$
- $smoker_i$ 是(1)否(0)抽烟

1、OLS 回归

$$smoker_i=\beta_0+\beta_1 educ_i+\beta_2 age_i+\beta_3 income_i+\beta_4 pcigs79_i+u_i$$

2、估测

$$Pr(y_i=1|X_i)=\frac{exp(X_i\beta)}{1+exp(X_i\beta)}=\frac{1}{1+exp(-X_i\beta)}$$

$$X_i\beta=\beta_0+\beta_1 educ_i+\beta_2 age_i+\beta_3 income_i+\beta_4 pcigs79_i+u_i$$

代码：
```
logit <- glm(smoker ~ educ + age + income + pcigs79, data = smoking, family = "binomial")
```

3. 找拟合值

（估计抽烟概率的拟合值需 $\widehat{logit_i}$）

$$\widehat{logit}_i=\hat{\beta}_0+\hat{\beta}_1 educ_i+\hat{\beta}_2 age_i+\hat{\beta}_3 income_i+\hat{\beta}_4 pcigs79_i$$

（再代入求拟合概率）

$$\hat{P}_i=\frac{1}{1+e^{-\widehat{logit}_i}}$$

数值：

$$\widehat{logit}_i=\hat{\beta}_0+\hat{\beta}_1*15+\hat{\beta}_2*28+\hat{\beta}_3*12500+\hat{\beta}_4 60.6=-0.4968$$

转换后：

$$\hat{P}_i=\frac{1}{1+e^{-1\times-0.4968}}=0.378$$

∴ 37.8% 的概率抽烟

边际效应

（以上的 LPM/正态回归中每个点对概率影响的权重相同
对于非线性转换需从对概率影响的边际效应着手）

（先只看单一变量 educ，在均值周围的所有 *来估计额外一年 educ 的边际效应）

$$\widehat{logit}_{mean} = \hat{\beta}_0 + \hat{\beta}_1 \overline{educ} + \hat{\beta}_2 \overline{age} + \hat{\beta}_3 \overline{income} + \hat{\beta}_4 \overline{pcigs79}$$

概率估计：

$$\hat{P}_{mean} = \frac{1}{1+e^{-\widehat{logit}_{mean}}}$$

额外一年：

$$\widehat{logit}_{mean^*} = \hat{\beta}_0 + \hat{\beta}_1 (\overline{educ}+1) + \hat{\beta}_2 \overline{age} + \hat{\beta}_3 \overline{income} + \hat{\beta}_4 \overline{pcigs79}$$

$$\hat{P}_{mean^*} = \frac{1}{1+e^{-\widehat{logit}_{mean^*}}}$$

边际效应带来的概率差：

$$\hat{P}_{mean^*} - \hat{P}_{mean} = -0.02108$$

抽概率每年↓ 2.1%

边际效应的变动：

$$\widehat{logit}_{mean^{**}} = \hat{\beta}_0 + \hat{\beta}_1 (\overline{educ}+2) + \hat{\beta}_2 \overline{age} + \hat{\beta}_3 \overline{income} + \hat{\beta}_4 \overline{pcigs79}$$

额外两年

$\hat{P}_{mean^{**}} - \hat{P}_{mean^*} = -0.02053$ （接近 −0.02108 但所带影响更小）（每年权重下降）

第一年的教育对抽烟人数下降的影响大于第二年和之后（速率递减）
教育对抽烟人数下降的总影响增大

L10 极大似然估计

1)

随机变量：一个将未预料事件的结果绘制为一个数的公式

概率分布：正态：$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

伯努利：$f(x;p) = \begin{cases} p & \text{if } x=1, \\ q=1-p & \text{if } x=0. \end{cases}$

泊松：$f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$

似然：已发生事件会产生观察到的结果的假设概率
（有已知结果的过去事件）

$$L(\theta) = f(x_1;\theta)\ldots f(x_n;\theta)$$

（样本的概率密度）

例： g_i（每局进球数）为随机变量，服从正态分布（数据从 68 场比赛）

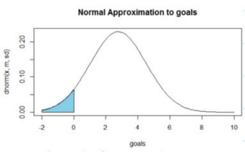
$g_i = \alpha + u_i \qquad u_i \sim N(0, \sigma^2)$

$E(g_i) = \alpha$

$P(g_i \geq 1) = 0.846$

$P(g_i \geq 3) = 0.450$

$P(g_i < 0) = 0.056$



连续性 r.v. 正态分布有误出现

最大似然估计:

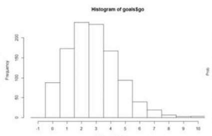
已知: $g_i \in \mathbb{N}^+$

设: $g_i \sim Poisson(\lambda)$

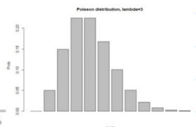
$$f(g_i) = \frac{\lambda^{g_i} e^{-\lambda}}{g_i!}$$

(需估计参数为λ)

$E(g_i) = \lambda$, $Var(g_i) = \lambda$



正态



泊松

Logarithms 有如下三种性质

1、 $\log(ab) = \log(a) + \log(b)$

2、 $\log\left(\frac{a}{b}\right) = \log(a) - \log(b)$

3. $\log(a^n) = n\log(a)$

2) 参数估计

(求最优下 ML 参数的估计)

Poisson 分布密度, r.v 为:

$$f(g_i;\lambda) = \frac{\lambda^{g_i} e^{-\lambda}}{g_i!}$$

(要预测的参数为λ)

求给定λ下, 前两个结果(得分为0和5时)发生的似然

i.i.d: 独立同分布

$L(\lambda; g_1, g_2) \overset{iid}{=} f(g_1;\lambda)\, f(g_2;\lambda)$

(似然等于该两值, 但当数值数目变多, 产生问题)

运用 log 来解决:

$\ln L(\lambda; g_1, g_2) \overset{iid}{=} \ln f(g_1;\lambda) + \ln f(g_2;\lambda)$

最大似然估计:

(选样本中前两个观测值: 0, 5, 3...)

设参数可取 λ=2, 5 or 8

λ=2: $\ln L(\lambda=2; g_1, g_2) = -2 + (-3.322) = -5.322$

λ=5: $\ln L(\lambda=5; g_1, g_2) = -6.740$ 最高的, 最有可能的

λ=8: $\ln L(\lambda=8; g_1, g_2) = -10.390$

（$L(\lambda; g_1, g_2)$ 的值越大，数据源于 Respective Distribution 的可能性越大）

☆ 注意：仅用前两个数据（first two data point），且无需找到所有可能的参数

先前问题的形式为：

$$\hat{\lambda}_{ML} = \underset{\lambda}{argmax} \ln L(\lambda; g_1, g_2, \dots g_{681})$$

（用 glm 代码）

$\hat{\lambda}_{ML} = 2.779441$

$P(g_i \geq 1) = 0.9379$

$P(g_i \geq 3) = 0.5256$

$P(g_i < 0) = 0$

例：若探究'是否 top 6 的队伍为主场'

引入虚拟变量 top_i

$$top_i \begin{cases} 1 & \text{若 top6 队伍为主场} \\ 0 & \text{若无一个 top6 队伍为主场} \end{cases}$$

OLS 调整为：

$$g_i = \alpha + \beta_1 top_i + u_i$$

$$E(g_i | top_i) = \alpha + \beta_1 top_i$$

相应的，分布密度调整：

$$f(g_i | top_i; \lambda) = \frac{\lambda_i^{g_i} e^{-\lambda_i}}{g_i!}$$

（λ 改变为 ⟹ 多种条件下的期望 $E(g_i | top_i) = \lambda_i$）

$$\lambda_i = \exp(\gamma_0 + \gamma_1 top_i)$$

（确保 λ_i 为正）

则：

$$(\hat{\gamma}_{0,ML}, \hat{\gamma}_{1,ML}) = argmax \ln L(\gamma_0, \gamma_1; g_1, g_2, \dots g_{681}, top_1, top_2, \dots, top_{681})$$

3）极大似然与 OLS

考虑以下模型：

$$y = X\beta + u \qquad u_i \sim N(0, \sigma^2)$$

求 OLS 估计量的方式：

1、最小平方：

$$\min (y - X\beta)'(y - X\beta) \Longrightarrow \min \sum_{i=1}^{n} \hat{u}_i^2$$

$$FOC: X'(y - X\hat{\beta}) = 0$$

（一阶导为 0 法）

2、Method of Moments

$$X'\hat{u} = X'(y - X\hat{\beta}) = 0$$

解 FOC 得 $\hat{\beta} = (X'X)^{-1}X'y$

同一模型但分布为 r.v. u_i

正态分布 pdf（概率密度函数）：

$$f(z;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right)$$

似然函数：

$$L(u_1,\dots,u_n;\beta,\sigma^2)=\frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n}\exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i-x_i\beta)^2\right)$$

该情况下 MLE 和 OLS 一样