endog & exog

- endog --> y --> Endogenous(内生变量) **模型内需解释的变量**
- exog ---> x ---> Exogenous(外生变量) **模型无需解释的变量**

Gaussian-Markov Theory

$$if\ E(\varepsilon)=0, Var(\varepsilon)=\sigma^2 I_n,\ \mathrm{E}(X,\varepsilon)=0 \implies OLS=BLUE$$

计量经济学假设

- 1. 线性假设：

$$Y=X\beta+\varepsilon$$

  使用$F-test$检验

- 2. 零均值假设/严格外生假设：

$$\mathrm{E}(\varepsilon_i|X) \quad =\mathrm{E}(\varepsilon_i|X_1,\cdots,X_i,\cdots,X_n)=0$$

- 3. 同方差假设/球形干扰：

$$Var(\varepsilon)=\sigma^2 I_n=\begin{cases} Var(\varepsilon_i|X_i)=E(\varepsilon_i^2)=\sigma^2 \text{【同方差性】} \\ Cov(\varepsilon_i,\varepsilon_j)=0\text{【序列相关性】} \end{cases}$$

  - 使用$White-test,\ GQ-test,\ BP-test$检验异方差性，使用$WLS$补救
  - 使用$DW-test,\ BG-test,\ LM-test$检验序列相关性，使用$GLS$补救

- 4. 无共线性假设/满秩假设：

$$rank(X) = p$$

- 使用 $VIF, Cov\ Matrix,\ Cond.No$检验多重共线性

- 5. 正态性假设：

$$\varepsilon|X \backsim N(0, \sigma^2 I_n)$$

- 使用$JB - test$检验正态性

检验回归系数显著性

- 小样本+单约束 $\implies\ T - test$
- 渐进相等+大样本+多约束 $\implies \begin{cases} LR - test \\ Wald - test \\ LM - test \end{cases}$

$OLS\ [\varepsilon \backsim\ N(0, \sigma^2 I)]$

$$Y = X\beta + \varepsilon \implies \min_{\hat{\beta}} \sum_{i=1}^{n} (y_i - X_i\hat{\beta})^2$$

$$\implies \hat{\beta} = (X^T \cdot X)^{-1} \cdot X^T \cdot Y$$

- **特殊的**：$y = \beta_0 + \beta_1 x$

$OLS\ non\ -\ linearcurve$

- 是一种特殊形式的***OLS***

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \sin(x_1) + \beta_4 x_2^2 + beta_5 x_1 x_2 + \varepsilon$$

- 可将$x_1 x_2$, $\sin(x_1)$ or x_2^2 看作新的x

$OLS\ with\ dummy\ variables$

- 是一种特殊形式的***OLS***

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- 其中x_2为分类变量

衡量指标

1. R^2: 可决系数

因变量y的波动有多少比例可以由自变量来解释

$$R^2 = 1 - \frac{SSE}{SST}$$

- $SSE = \sum(\hat{y}_i - \bar{y}_i)^2$ 是残差平方和
- $SST = \sum(y_i - \bar{y}_i)^2$ 是总体标准差
- $Std\ err = \sqrt{SST/n}$ 是标准误
- $SSR = \sum(\hat{y}_i - y_i)^2$ 是拟合残差。
- $SST = SSR + SSE$

2. $Adjust\ R^2$: R^2的无偏估计

$$R^2(adj) = 1 - (1 - R^2)\frac{n-1}{n-p-1} = 1 - \frac{\frac{1}{n-p-1}SSR}{\frac{1}{n-1}SST}$$

3. $F - statistic$： F检验，判断线性关系

- $H_0 : \beta_1 = \cdots = \beta_p = 0$

$$F = \frac{SSR/p}{SSE/(n-p-1)} \backsim F(p, n-p-1)$$

4. $Log - Likelihood$： 对数似然值，判断拟合优度

- 高斯分布模型:

$$\ln(L(\theta|x)) = \ln(\prod_{i=1}^{n} p(y|x;\beta)) = \ln(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y_i - X_i\beta)^2}{2\sigma^2}})$$

5. $Akaike\ Information\ Criterion - AIC$： 判断拟合优度

$$AIC = -2 \times \ln(L(\theta|x)) + 2p$$

6. $Bayesian\ Information\ Criterion - BIC$： 判断拟合优度

- 相比于AIC, BIC 可以有效地减少 **过拟合** 和**维度灾难** （<u>维度越大，效果越差</u>)

$$BIC = -2 \times \ln(L(\theta|x)) + p \times \ln(n)$$

7. $T - test$： 回归系数检验

- $H_0 : \beta_i = 0$
- $H_1 : \beta_i \neq 0$

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

8. $Durbin - Watson\ test$: 检验样本残差间是否$AR(1)$

- $\varepsilon_i = \rho\varepsilon_{i-1} + v_t$
- $H_0 : \rho = 0$
- $H_1 : \rho \neq 0$

$$d = \frac{\sum_{i=2}^{n}(\varepsilon_i - \varepsilon_i)^2}{\sum_{i=2}^{n}\varepsilon_i^2}$$

- d 越接近2越好，$d = 1 \backsim 3$ 没问题，$d < 1$ 否定H_0

9. $Jarque - Bera\ test$: 检验数据是否具有正态性

- 使用时：$N > 30$
- $H_0 : X \backsim N$

$$J - B = \frac{n}{6}(Skew^2 + \frac{(Kurtosis - 3)^2}{4})$$

10.$Likelihood\ Ratio - test$ ： 只适用线性

- 作用同 $Wald - test$, 使用时需分别计算约束与非约束模型的 $\log L(\beta)$
- $H_0 : \hat{\beta}_{1\times q} = \{\hat{\beta}_i, \cdots, \hat{\beta}_j\}_{1\times q} = 0, q \leq p$
- $\beta_{constent} = \complement_{\beta_0}(\beta_0 \cap \beta_{constent})$, $\beta_{non-constent} = \beta_0$

$$LR = 2(\log L(\beta_{constent}) - \log L(\beta_{non-constent})) \backsim \chi^2(q)$$

- q 为约束变量个数（下同）

11. $Wald-test$：(适用线性and非线性方程)

检验使用规模更大的模型是否比规模更小的模型有**更好的拟合度**，若没有，则没有必要选择规模更大的模型。(规模：回归系数个数)

- $H_0: f(\hat{\beta}) = f(\beta_1, \cdots, \hat{\beta}_p) = 0$

$$W = (f(\hat{\beta}))^T[Var(f(\hat{\beta}))]^{-1}(f(\hat{\beta})) \backsim \chi^2(q)$$

- $eg: f(\hat{\beta}) = \beta_1\beta_2 - \beta_3 = 0$
- $Var(f(\hat{\beta})) = (\frac{\partial f(\hat{\beta})}{\partial \hat{\beta}})(Var(\hat{\beta}))(\frac{\partial f(\hat{\beta})}{\partial \hat{\beta}})^T$

12. $LagrangeMultiplier-test$:(适用线性and非线性方程)

- 思路：若x_i, x_j 可去，则x_i, x_j与y无关，更不会进入ε中，由$Gaussian-Markov\ Theory$可知，ε与其他 $x_k, k \neq i, j$也无关。
  Step1: 对 $y = \sum_{k \neq i,j} \beta_k x_k + \varepsilon$ 进行回归，得到残差序列 ε。
  Step2: 对辅助回归方程$\varepsilon = \sum_{k=0}^{p} \alpha_k x_k + \varepsilon_u$ 进行回归，并计算可决系数R_u^2。
  Step3: 构建LM统计量：$LM = nR_u^2 \backsim \chi^2(q)$。
- LM统计量还可以检测残差的高阶延后 $AR(p)$
  辅助统计量为：$(BG-test)$

$$\varepsilon = \sum_{k=1}^{m} \beta_k x_k + \sum_{i=1}^{p} \rho_i u_{t-i} + v_t$$

$$LM = (n-p)R_u^2 \backsim \chi^2(p)$$

13. $White - test$：(检验异方差)

- $H_0: \forall \varepsilon_i \backsim N(0, \sigma^2) \implies$ 同方差
- 辅助方程：$\varepsilon_i^2 = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{1i}^2 + \alpha_4 x_{2i}^2 + \alpha_5 x_{1i} x_{2i} + v_i$ **(包含：一次项、二次项&交叉项)**

$$W = nR_W^2 \backsim \chi^2(m)$$

- m 为辅助回归模型回归系数个数

14. $BP - test$：(检验异方差，$White - test$的特例)

- $H_0: \forall \varepsilon_i \backsim N(0, \sigma^2) \implies$ 同方差
- 辅助方程：$\varepsilon_i^2 = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + v_i$ **(包含：一次项)**

$$W = nR_B^2 \backsim \chi^2(m)$$

- m 为辅助回归模型回归系数个数

15. $GQ - test$：(检验异方差, 使用$F - test$)

Step1：将模型观测值按照解释变量大小升序排列

Step2:将排序后样本中间删掉 c 个样本 **(一般取样本数的四分之一)**，将余下的样本平均分为两部分，每部分有$\frac{n-c}{2}$个观测值。

Step3:对两个样本进行回归，得到残差平方和SSE_1, SSE_2 构建统计量

$$F = \frac{\frac{SSE_1}{\frac{n-c}{2} - 1}}{\frac{SSE_2}{\frac{n-c}{2} - 1}} \backsim F(\frac{n-c}{2} - k - 1, \frac{n-c}{2} - k - 1)$$

15. $Gleiser-test\ and\ Park-test$：

(检验异方差, 并寻找出异方差形式 $\sigma_i = f(X_i)$)

Step1：对模型进行 OLS 回归，得到残差序列 $\{\hat{\varepsilon}_i\}_{i=1}^n$

Step2：对于辅助回归模型进行回归

$$Gleiser: \ln|\hat{\varepsilon}_i| = \ln\sigma^2 + \alpha\ln X_i + v_i$$
$$Park: \ln(\hat{\varepsilon}_i{}^2) = \ln\sigma^2 + \alpha\ln X_i + v_i$$

其中，$\ln\sigma^2$为辅助回归模型的常数(截距)项

Step3：对于辅助回归模型的回归系数 α 进行 $T-test$, 检验回归系数显著性。若显著，则存在异方差性并确定异方差形式 $f(X_i)$

- 检验的辅助回归模型与异方差形式 $f(X_i)$的关系为:

$$\sigma_i = f(X_i) = \sigma^2 X_i^\alpha e^{v_i} \xRightarrow{ln} \ln\sigma^2 + \alpha\ln X_i + v_i$$

PS:

- $Df\ Residuals$: 残差自由度 $(n-p-1)$
- $Df\ Model$: 模型参数个数 (p)
- $Skewness->Skew$: 偏度

$$Skew = E[(\frac{(Mean-Median)}{Std})^3] = \frac{1}{n}\sum_{i=1}^n[(\frac{X_i-\mu}{\sigma})^3]$$

- $Kurtosis$: 峰度

$$Kurtosis = E[(\frac{(Mean - Median)}{Std})^4] = \frac{1}{n}\sum_{i=1}^{n}[(\frac{X_i - \mu}{\sigma})^4]$$

- $P > |t|$: $P - value$
- $[0.025\ 0.975]$: 系数的95%置信区间
- $Cond.No.$: $if\ Cond.No. < 100$，共线性程度小, $if\ 100 < Cond.No. < 1000$，共线性程度较大, $if\ Cond.No. > 1000$，共线性程度严重

GLS 广义最小二乘法 $[\varepsilon \backsim N(0, \sigma^2\Sigma),\ \Sigma$可逆$]$

$$\tilde{Y} = \tilde{X}\beta + \varepsilon \implies \min_{\hat{\beta}} \frac{1}{2p}(Y - X\hat{\beta})^T\Sigma^{-1}(Y - X\hat{\beta})$$

$$\implies \hat{\beta} = (X^T\Sigma^{-1}X)^{-1} \cdot X^T\Sigma^{-1}Y = (\tilde{X}^T \cdot \tilde{X})^{-1} \cdot \tilde{X}^T \cdot \tilde{Y}$$

- $\Sigma = C^TC$
- $\tilde{X} = CX, \tilde{Y} = CY$

误差项遵循 $AR(1)$

$$\varepsilon_i = \beta_0 + \rho\varepsilon_{i-1} + \eta_i,\ \eta_i \backsim N(0, \Sigma)$$

- 可使用GLSAR等带有AR的 **滞后回归模型**
- $FGLS$ (可行广义最小二乘法)：因一般残差的协方差阵不一致，需使用样本残差协方差阵进行一致估计后得出残差的协方差, 再使用GLS

Quantile Regression: 分位数回归

$$Y=X\beta+\varepsilon \implies \min_{\hat{\beta}}\frac{1}{n}(\sum_{i:Y_i<X\hat{\beta}}(1-\tau)|y_i-X\hat{\beta}|\sum_{i:Y_i\geq X\hat{\beta}}\tau|y_i-X\hat{\beta}|)$$

- 这里的分位数是 **下分位数**，即回归线下包含$\tau\times 100\%$的数据

分位数

$$\tau=\mathrm{P}(y\leq y_\tau)=\mathrm{F}_\tau(y)$$

LAD Estimator: 最小一乘回归(0.5分位数回归)

$$\min_{\hat{\beta}}\frac{1}{n}\sum_{i=1}^{n}|y_i-X\hat{\beta}|$$

分位数相关检验

- 拟合优度检验:
  需将解释变量矩阵与系数向量分为两部分

$$X=(1_{n\times 1},Z_{n\times p})_{n\times p},\quad \hat{\beta}_{(\tau)}=(\hat{\beta}_{0(\tau)},\hat{\beta}_{2\backsim p(\tau)})$$

$\hat{\beta}_{2\backsim p(\tau)}\neq 0$ 时的残差

$$\hat{Q}=\min\sum_{i:y_i<X\hat{\beta}}(1-\tau)|y_i-\hat{\beta}_{0(\tau)}-Z\hat{\beta}_{2\backsim p(\tau)}|+\sum_{i:y_i\geq X\hat{\beta}}(1-\tau)|y_i-\hat{\beta}_{0(\tau)}-Z\hat{\beta}_{2\backsim p(\tau)}|$$

$\hat{\beta}_{2\smile p(\tau)}=0$ 时的残差

$$\tilde{Q}=\min\sum_{i:y_i<X\hat{\beta}}(1-\tau)|y_i-\hat{\beta}_{0(\tau)}|+\sum_{i:y_i\geq X\hat{\beta}}(1-\tau)|y_i-\hat{\beta}_{0(\tau)}|$$

拟合优度：
使用分位数回归和使用水平直线回归的拟合残差之差

$$R^*_{(\tau)}=1-\frac{\hat{Q}}{\tilde{Q}}$$

- 系列分位数检验
  - 斜率相等检验：检验不同分位数回归下相同特征的回归系数是否相同 $H_0:\beta_{i(\tau_1)}=\cdots=\beta_{i(\tau_m)},\ i=1,\cdots,k$, 使用$Wald-test$进行检验，其服从 $\chi^2((k-1)(m-1))$
  - 对称性检验: 检验y的分布是否对称 $H_0:\frac{\beta_{0,(\tau_j)}+\beta_{0,(\tau_{m-j+1})}}{2}=\beta_{0,(0.5)}$, 使用$Wald-test$进行检验，其服从 $\chi^2(k(m-1)/2)$

$RecursiveLS$: 递归最小二乘回归

- 该方法与数据不同时间批次出现有关
- 为减少多次回归的计算量，**引入参数更新方法**
  已知:

$$\hat{\beta}_0=(X_0^TX_0)^{-1}X_0^TY=\Sigma_0^{-1}X_0^TY$$

$$\implies \Sigma_0\hat{\beta}_0 = X_0^TY \tag{1}$$

在引入新数据X_1 后，再次求解就会变为：

$$\hat{\beta}_1 = \left(\begin{bmatrix} X_0 \\ X_1 \end{bmatrix}^T \begin{bmatrix} X_0 \\ X_1 \end{bmatrix} \right)^{-1} \begin{bmatrix} X_0 \\ X_1 \end{bmatrix}^T \begin{bmatrix} Y_0 \\ Y_1 \end{bmatrix} = \Sigma_1^{-1} \begin{bmatrix} X_0 \\ X_1 \end{bmatrix}^T \begin{bmatrix} Y_0 \\ Y_1 \end{bmatrix}$$

$$\implies \Sigma_1\hat{\beta}_1 = \begin{bmatrix} X_0 \\ X_1 \end{bmatrix}^T \begin{bmatrix} Y_0 \\ Y_1 \end{bmatrix}$$

经过递推可得：

$$\Sigma_k = \Sigma_{k-1} + X_k^TX_k \tag{2}$$

由(1)和(2)可推出：

$$\begin{aligned} \begin{bmatrix} X_0 \\ X_1 \end{bmatrix}^T \begin{bmatrix} Y_0 \\ Y_1 \end{bmatrix} &= X_0^TY_0 + X_1^TY_1 \\ &= \Sigma_1\hat{\beta}_0 + X_1^T(Y_1 - X_1\hat{\beta}_0) \end{aligned}$$

可递推出：

$$\hat{\beta}_k = \hat{\beta}_{k-1} + \Sigma_k^{-1}X_k^T(Y_k - X_k\hat{\beta}_{k-1}) \tag{3}$$

我们可以根据两个递推式来更新回归系数，但对矩阵求逆无疑是困难且复杂的。因此我们使用Sherman-Morrison-Woodbury 引理来简化运算：

$$\begin{gathered} Lemma : Sherman - Morrison - Woodbury \\ (A+UV^T)^{-1} = A^{-1} - A^{-1}U(I+V^TA^{-1}U)^{-1}V^TA^{-1} \end{gathered}$$

将(2)代入(3)中，并应用引理将两个递推式转化为：

$$P_k = \frac{1}{\lambda}(P_{k-1} - P_{k-1}X_k^T(\lambda I + X_kP_{k-1}X_k^T)^{-1}X_kP_{k-1})$$
$$\hat{\beta}_k = \hat{\beta}_{k-1} + P_kX_k^T(Y_k - X_k\hat{\beta}_{k-1})$$

其中，$P_k = \Sigma_k^{-1}$, λ为遗忘因子。且 $\lambda I + X_kP_{k-1}X_k^T$ 为 **一个实数**，取逆=取倒数

- 遗忘因子的加入，解决了旧数据因数量过大而淹没新数据的问题。使用遗忘因子的算法称为 $FFRLS$, 当 $lambda = 1$ 时，$FFRLS \implies RecursiveLS$
- $RecursiveLS$ 常用作自适应滤波器
- $CUSUM - test$ 是残差图评价方法的拓展，他引入了递归残差的累计残差和。

可使用 $CUSUM$检验 和 $CUSUM$平方检验

H_0 : 参数稳定

H_1 : 参数非稳定

判断标准： 若图中曲线超出设定的置信区间则否定 H_0

参数稳定性假设:

有两个使用不同数据子集进行回归的回归模型：

$$Y = X\beta + \varepsilon$$
$$Y = X\alpha + v$$

若：$\alpha = \beta$, 则称模型参数稳定。

$Rolling\ Regression$: 滚动窗口回归

- 滚动窗口有两种形式:

$$\begin{cases} \text{固定起始值:} & \text{给定起始窗口长度, 给定最大窗口长度,} \\ & \text{在回归过程中窗口长度逐渐递增} \\ \text{固定窗口值:} & \text{给定窗口固定长度} \\ & \text{在回归过程中窗口长度始终不变} \end{cases}$$

- 显然这与滑动窗口的两种形式类似，本质上:

$$Rolling\ Regression = Sliding\ Window + OLS$$

WLS 加权最小二乘法 $[\varepsilon_i \backsim N(0, \sigma_i^2)$, 异方差性$]$

$$Var(\varepsilon) = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} \implies w_i = \frac{1}{\sigma_i^2}$$

$$\implies W = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix}$$

$$Y = X\beta + \varepsilon \implies \min_{\hat{\beta}} \frac{1}{2p}(Y - X\hat{\beta})^T \Sigma^{-1}(Y - X\hat{\beta})$$

$$\implies \hat{\beta} = (X^T W^{-1} X)^{-1} \cdot X^T W^{-1} Y$$

- $FWLS$ (可行加权最小二乘法)：因一般残差的协方差阵不一致，需使用样本残差协方差阵进行一致估计后得出残差的协方差, 再使用WLS
- 显然WLS是GLS的一种特殊形式。因此在异方差和序列相关同时存在时，使用GLS进行回归。
