Intermediate Econometrics

Professor: Xiaojun Song

Timekeeper: Rui Zhou

Spring 2023

目录

1 简单回归模型 **2**
1.1 OLS . . . 2
1.2 拟合优度 . . . 3
1.3 非线性回归模型 . . . 3
1.4 OLS 估计量的期望和方差 . . . 3

2 多元线性回归：估计 **6**
2.1 OLS . . . 6
2.2 OLS 估计量的期望和方差 . . . 6
2.3 拟合优度 . . . 7
2.4 误设模型问题 . . . 7
2.5 OLS 估计量的有效性 . . . 9

3 多元回归分析：推断 **10**
3.1 正态抽样分布 . . . 10
3.2 单个参数的 t 检验 . . . 10
3.3 参数线性组合的检验 . . . 11
3.4 多个线性约束的检验 . . . 11
3.5 经济显著性 . . . 12

4 多元回归分析：OLS 的渐进性 **13**
4.1 一致性 . . . 13
4.2 渐进正态和大样本推断 . . . 14

1 简单回归模型

简单线性回归模型 (SLR Model)：$y = \beta_0 + \beta_1 x + u$

- y：因变量、被解释变量、响应变量、回归子
- x：自变量、解释变量、控制变量、回归元
- u：误差项、干扰项

定义截距：$\mathrm{E}(u) = 0$

零条件均值 (zero conditional mean) 假定：$\mathrm{E}(u|x) = 0$

总体回归函数 (PRF)：$\mathrm{E}(y|x) = \beta_0 + \beta_1 x$

1.1 OLS

1.1.1 OLS 的必要条件

1. $\mathrm{E}(u) = 0$
2. $\mathrm{E}(xu) = 0 \Longleftarrow \mathrm{Cov}(x, u) = 0$

1.1.2 OLS 估计量

- $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = frac\sum_{i=1}^{n}(x_i - \bar{x})y_i\sum_{i=1}^{n}(x_i - \bar{x})^2$
- $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$

样本回归方程 (SRF)：$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \Longrightarrow \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (\hat{u}_i := y_i - \hat{y}_i)$

1.1.3 OLS 估计量的代数性质

- $\sum_{i=1}^{n} \hat{u}_i = 0$（残差和为 0，残差均值为 0）
- $\sum_{i=1}^{n} x_i \hat{u}_i = 0$ （回归元和残差的样本协方差为 0）
- $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$（均值点在 OLS 回归线上）

注：

- $\beta_1 = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)} = \hat{\rho}_{xy} \cdot \frac{\hat{\sigma}_y}{\hat{\sigma}_x}$
- 妙用 $\sum_{i=1}^{n} c(x_i - \bar{x}) = 0$，其中 c 是任意常数

1.2 拟合优度

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$
$$SSE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$
$$SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$
$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

拟合优度 R^2 衡量的是模型可以解释的数据变异性占总变异性的比例。

注：

- 特别地，R^2 实际上是因变量的预测值和真实值的相关系数的平方。
- R^2 的进一步含义为，回归模型与虚无模型对比能够多解释的变异性。

1.3 非线性回归模型

- $y \sim x \to \Delta y = \beta_1 \Delta x$
- $y \sim \log x \to \Delta y = \frac{\beta_1}{100}\%\Delta x$
- $\log y \sim x \to \%\Delta y = (100\beta_1)\Delta x$
- $y \sim x \to \%\Delta y = \beta_1 \%\Delta x$

1.4 OLS 估计量的期望和方差

1.4.1 基本假设

- SLR.1-线性于参数：总体回归函数为 $y = \beta_0 + \beta_1 x + u$
- SLR.2-随机抽样：$\{(x_i, y_i) : i = 1, 2, ..., n\} \to y_i = \beta_0 + \beta_1 x_i + u_i$
- SLR.3-自变量样本有波动：$SST_x > 0$
- SLR.4-零条件均值：$\mathrm{E}(u|x) = 0$
- SLR.5-同方差性：$\mathrm{Var}(u|x) = \sigma^2$

1.4.2 期望和无偏性

在 SLR.1~SLR.4 下，OLS 估计量是无偏的，$\mathbf{E}(\hat{\beta}_1)=\beta_1, \mathbf{E}(\hat{\beta}_0)=\beta_0$。

$$
\begin{aligned}
\because \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i-\bar{x})y_i}{SST_x} = \frac{\sum_{i=1}^n (x_i-\bar{x})(\beta_0+\beta_1 x_i+u_i)}{SST_x} \\
&= \frac{\sum_{i=1}^n (x_i-\bar{x})\beta_0 + \sum_{i=1}^n (x_i-\bar{x})\beta_1 x_i + \sum_{i=1}^n (x_i-\bar{x})u_i}{SST_x} \\
&= \frac{0+\beta_1\sum_{i=1}^n (x_i-\bar{x})x_i + \sum_{i=1}^n (x_i-\bar{x})u_i}{SST_x} \\
&= \frac{0+\beta_1\sum_{i=1}^n (x_i-\bar{x})(x_i-\bar{x}) + \sum_{i=1}^n (x_i-\bar{x})u_i}{SST_x} \\
&= \beta_1 + \frac{\sum_{i=1}^n (x_i-\bar{x})u_i}{SST_x} \\
\therefore \mathrm{E}(\hat{\beta}_1) &= \beta_1 + \frac{1}{SST_x}\cdot\sum_{i=1}^n (x_i-\bar{x})\mathrm{E}(u_i) = \beta_1 \\
\because \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} \\
&= \frac{1}{n}\cdot\sum_{i=1}^n y_i - \hat{\beta}_1\bar{x} = \frac{1}{n}\cdot\sum_{i=1}^n (\beta_0+\beta_1 x_i+u_i) - \hat{\beta}_1\bar{x} \\
&= \beta_0 + \beta_1\bar{x} + \frac{1}{n}\cdot\sum_{i=1}^n u_i - \hat{\beta}_1\bar{x} = \beta_0 + (\beta_1-\hat{\beta}_1)\bar{x} + \frac{1}{n}\cdot\sum_{i=1}^n u_i \\
\therefore \mathrm{E}(\hat{\beta}_0) &= \beta_0 + \frac{1}{n}\cdot\sum_{i=1}^n \mathrm{E}(u_i) = \beta_0
\end{aligned}
$$

若 SLR.1~SLR.4 中有一个不成立，那么无偏性一般是不成立的；尤其是 SLR.4，误差项中可能存在与自变量相关的因素。

1.4.3 方差

在 SLR.1~SLR.5 下，$\mathrm{Var}(\hat{\beta}_1)=\frac{\sigma^2}{SST_x}, \mathrm{Var}(\hat{\beta}_0)=\frac{n^{-1}\sigma^2\sum_{i=1}^n x_i^2}{SST_x}$

$$
\begin{aligned}
\because \hat{\beta}_1 &= \beta_1 + \frac{\sum_{i=1}^n (x_i-\bar{x})u_i}{SST_x} \\
\therefore \mathrm{Var}(\hat{\beta}_1) &= \mathrm{Var}\left(\frac{\sum_{i=1}^n (x_i-\bar{x})u_i}{SST_x}\right) = \frac{1}{(SST_x)^2}\cdot\sum_{i=1}^n (x_i-\bar{x})^2\mathrm{Var}(u_i) = \frac{\sigma^2}{SST_x}
\end{aligned}
$$

误差方差通常未知，误差方差的无偏估计为

$$\hat{\sigma}^2 = \frac{1}{n-2} \cdot \sum_{i=1}^{n} \hat{u}_i^2$$

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$$

$$\mathrm{se}(\hat{\beta}_1) = \frac{\hat{\sigma}}{(SST_x)^{\frac{1}{2}}}$$

2 多元线性回归：估计

多元线性回归模型 (MLR Model)：$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$

2.1 OLS

2.1.1 OLS 的必要条件

- $\mathrm{E}(u) = 0$
- $\mathrm{E}(x_j u) = 0, \forall j$

2.1.2 OLS 的估计量

OLS 斜率系数为（以 $\hat{\beta}_1$ 为例，其余相同）

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \hat{r}_{i1} y_i}{\sum_{i=1}^{n} \hat{r}_{i1}^2}$$

其中 $\hat{r}_{i1}$ 是基于现有样本，将 x_1 对其他所有回归元（$x_2, ..., x_k$）回归得到的残差。

$\hat{\beta}_1$ **度量的是，排除其他所有回归元的影响之后，y 和 x_1 的样本关系**。

偏效应指的是，**其他条件不变时**，x_j 对 y 的影响体现为 β_j。（注意，截距项可能不具有实际意义）

$\hat{\beta}_0$ 可以通过回归线通过样本中心点这一特性直接得出。

2.1.3 OLS 估计量的代数性质

- 残差的样本均值为 0 $\Leftrightarrow \sum_{i=1}^{n} \hat{u}_i = 0, \bar{y} = \bar{\hat{y}}$
- 每个自变量和残差之间的样本协方差为 0 $\Leftrightarrow \sum_{i=1}^{n} x_{ij} \hat{u}_i = 0, \forall j = 1, 2, ..., k \Leftrightarrow \sum_{i=1}^{n} \hat{y}_i \hat{u}_i = 0$，也即 OLS 拟合值和残差之间的样本协方差为 0
- 点 $(\bar{x}_i, \bar{x}_2, ..., \bar{x}_k, \bar{y})$ 总在回归线上 $\Leftrightarrow \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + ... + \hat{\beta}_k \bar{x}_k$

2.2 OLS 估计量的期望和方差

2.2.1 基本假设

- MLR.1-线性与参数：总体回归函数为 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$
- MLR.2-随机抽样：$\{(x_{i1}, x_{i2}, ..., x_{ik}, y_i) : i = 1, 2, ..., n\} \rightarrow y_i = \beta_0 + \beta_1 x_i + u_i$

- MLR.3-不存在完全共线性：样本中没有一个自变量是常数（共线于斜率），自变量之间没有严格的线性关系需要注意，可能存在等价转换之后的共线性，如 $\log w$ 和 $\log w^2$
- MLR.4-零条件均值：$\mathrm{E}(u|x_1, x_2, ..., x_k) = 0$
- MLR.5-同方差性：$\mathrm{Var}(u|x_1, x_2, ..., x_k) = \sigma^2$

2.2.2 OLS 估计量的期望和无偏性

在 MLR.1~MLR.4 下，$\mathrm{E}(\hat{\beta}_j) = \beta_j, \forall j = 0, 1, 2, ..., k$

2.2.3 OLS 估计量的方差

在 MLR.1~MLR.5 下，$\mathrm{Var}(\hat{\beta}_j) = \frac{\sigma^2}{SST_j \cdot (1-R_j^2)}, \forall j = 1, 2, ..., k$

其中，定义方差膨胀因子 $VIF_j = \frac{1}{1-R_j^2}$，$VIF < 10$ 被认为是底线。

2.3 拟合优度

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = \frac{(\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{y}))^2}{(\sum_{i=1}^n (y_i - \bar{y})^2) \cdot (\sum_{i=1}^n (\hat{y}_i - \bar{y})^2)}$$

注：

- 在回归中添加自变量时，R^2 至少不会减少
- 判断模型中是否应该加入变量，判定标准应该是**偏效应是否非零**，而不是 R^2 是否增大
- 不同类型、不同自变量的模型之间的比较应该看调整后的 R^2

2.4 误设模型问题

2.4.1 引入无关变量

引入**无关变量**不影响 OLS 估计量的无偏性，但会**影响方差**。

2.4.2 遗漏变量

假设真实的回归方程应为 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$，但遗漏 x_2 后的误设回归模型为 $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$。而 x_2 对 x_1 的回归有 $\tilde{x}_2 = \tilde{\delta}_0 + \tilde{\delta}_1 x_1$，综合来有

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1$$
$$\text{Bias}(\tilde{\beta}_1) = \text{E}(\tilde{\beta}_1) - \beta_1 = \beta_2 \tilde{\delta}_1$$

若 $\text{Bias}(\tilde{\beta}_1) = \text{E}(\tilde{\beta}_1) - \beta_1 = \beta_2 \tilde{\delta}_1 > 0$，则说明有向上的偏误；若 $\text{Bias}(\tilde{\beta}_1) = \text{E}(\tilde{\beta}_1) - \beta_1 = \beta_2 \tilde{\delta}_1 < 0$，则说明有向下的偏误；若 $\text{E}(\tilde{\beta}_1)$ 比 β_1 更接近 0，则说明有向零的偏误。

在更一般的情形下，**如果一个解释变量与误差相关，那么这通常会导致所有的 OLS 估计量都产生偏误**。只要被遗漏的变量与其他解释变量都有关，但因误设而被放入误差中。

如果 $\beta_2 = 0$，则 $\tilde{\beta}_1$ 和 $\hat{\beta}_1$ 都是无偏的。并且有 $\text{Var}(\tilde{\beta}_1) = \frac{\sigma^2}{SST_1}, \text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{SST_1 \cdot (1-R_1^2)}$，因此 $\text{Var}(\tilde{\beta}_1) < \text{Var}(\hat{\beta}_1)$。实际上，这说明了往模型中加入无关变量，会加剧多重共线性的问题，使得 $\hat{\beta}_1$ 的估计效率下降。

如果 $\beta_2 \neq 0$，则 $\tilde{\beta}_1$ 将会是有偏的，不过方差上的劣势可以通过样本量弥补，因此偏好将 x_2 纳入回归。

如果 $\delta_1 = 0$，由于**偏误是以样本为条件的**，所以特定样本中 $\tilde{\delta}_1$ 并不一定为零，可能仍然会有偏误。

2.4.3 多重共线性

$$VIF_j = \frac{1}{1 - R_j^2}$$

如果 $VIF_j > 10$，或者 $R_j^2 \to 1$，那么意味着存在多重共线性。

形式上看，多重共线性会让系数的误差变大，不仅更难显著，也意味着估计难以精确。同时，存在多重共线性的回归元之间由于有高度的线性关系，可以相互表出，计算系数可能都不准确。此外，**多重共线性的影响可能不是通过扩大样本量能够直接消除的**。

不过，即便**多重共线性对单一系数的 t 检验能够产生影响，F 检验对多个变量之间的多重共线性免疫**。

2.4.4 OLS 估计量的标准误

方差是很难得到的，只能通过特定的样本估计标准误，在 MLR.1~MLR.5 下，无偏估计为

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \cdot \sum_{i=1}^{n} \hat{u}_i^2 \qquad \text{E}(\hat{\sigma}^2) = \sigma^2$$

$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ 为回归标准误 (SER, standard error of regression)，且有

$$\text{se}(\hat{\beta}_j) = \frac{\hat{\sigma}}{\sqrt{SST_j \cdot (1 - R_j^2)}} = \frac{\hat{\sigma}}{\sqrt{n \cdot \text{sd}(x_j) \cdot (1 - R_j^2)}}$$

2.5 OLS 估计量的有效性

高斯-马尔科夫定理：在 MLR.1~MLR.5 下，$\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ 分别是 $\beta_0, \beta_1, ..., \beta_k$ 的最优线性无偏估计量。(BLUE, best linear unBiased estimator)

- 最优：具有最小的方差
- 线性：β_j 的一个估计量 $\hat{\beta}_j$ 是线性的充分必要条件是，$\hat{\beta}_j$ 能表示成因变量数据的一个线性函数，也即 $\hat{\beta}_j = \sum_{i=1}^{n} w_I y_i$，其中 w_i 可以是关于自变量样本值的一个函数
- 无偏：$\mathrm{E}(\hat{\beta}_j) = \beta_j$

注：称 MLR.1~MLR.5 为高斯-马尔科夫假定。

3 多元回归分析：推断

3.1 正态抽样分布

当以样本中自变量的值为条件时，OLS 估计量的抽样分布取决于其背后的误差分布。为使 $\hat{\beta}_j$ 的抽样分布易于处理，假定总体中不可观测的误差是正态分布的。

MLR.6-正态性：总体误差独立于解释变量 $x_1, x_2, ..., x_k$，并且 $u \sim N(0, \sigma^2)$。

MLR.6 是一个强假定，相当于在 MLR.4 & MLR.5 的基础上假设了 u 的分布；因此，MLR.6 也可以推出 MLR.4 & MLR.5。

CLM 假定下，总体可以表示为 $y|\vec{x} \sim N(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k, \sigma^2)$。

在 CLM 假定下，OLS 估计量是最小方差无偏估计量。这比高斯-马尔科夫假定下具有更强的效率，不需要把比较局限于对 y_i 为线性的估计量内。

正态抽样分布定理：在 CLM 假定下，以自变量的样本值为条件，有 $\hat{\beta}_j \sim N(\beta_j, \mathbf{Var}(\hat{\beta}_j))$

证明过程只要理解 OLS 估计量是样本误差的一个线性组合、并且样本误差是独立同分布的随机变量即可。在此基础上，该定理也可以加强，OLS 估计量的任何线性组合也都服从人态分布。

注：

- 称 MLR.1~MLR.6 为经典线性模型假定 (CLM assumptions, classical linear model)
- 误差服从正态分布：误差是影响因变量但又无法观测的诸多因素之和，可以借助中心极限定理断定误差具有近似正态分布。但近似的效果取决于误差中的因素、各自的分布差异，以及彼此之间的独立性。
- 大样本下 OLS 估计量的正态性近似成立

3.2 单个参数的 t 检验

标准化估计量的 t 分布：在 CLM 假定下，$\frac{\hat{\beta}_j - \beta_j}{\mathrm{se}(\hat{\beta}_j)} \sim t_{n-k-1}$，其中 $k+1$ 是总体模型中未知参数的个数。

假设检验和构造置信区间是等价的，后者可以直接判断是否拒绝特定的原假设。

注：

- 原假设中对 β_j 的判断，不一定是 0
- 原假设和备择假设的方向，注意检验是单尾还是双尾，注意对应的 p 值和临界值
- 大样本的 t 检验可以用标准正态分布近似
- “不能拒绝原假设” 不等同于 “接受原假设”

3.3 参数线性组合的检验

假设有回归方程 $\hat{y}=\hat{\beta}_o+\hat{\beta}_1x_1+\hat{\beta}_2x_2$，如果要检验参数的线性组合，如 $\beta_1=p+q\beta_2$，那么可以**最一般性地表出** β_1 **为** $\beta_1=p+q\beta_2+\theta$，如果原假设成立，那么 $\theta=0$。将 $\beta_1=p+q\beta_2+\theta$ 代入回归方程，可以得到

$$\begin{aligned} y&=\beta_0+\beta_1x_1+\beta_2x_2=\beta_0+(p+q\beta_2+\theta)x_1+\beta_2x_2\\ \Longrightarrow y-px_1&=\beta_0+\theta x_1+\beta_2(qx_1+x_2)\end{aligned}$$

这时只要构造新的因变量为 $y-px_1$，新的自变量为 qx_1+x_2，并检验 x_1 的系数是否显著地异于 0 即可。同时，这也可以最简单地得到 $\mathrm{se}(\hat{\theta})=\mathrm{se}(\hat{\beta}_1-q\hat{\beta}_2)$，而后者是很难计算得出的。

要计算 $\mathrm{se}(\hat{\beta}_1-q\hat{\beta}_2)$，那么必须能够计算 $\mathrm{Cov}(\hat{\beta}_1,\hat{\beta}_2)$。记住，OLS 估计量都是 y_i 的线性函数，而 y_i 之间是独立的。

$$\begin{aligned}\mathrm{Cov}(\hat{\beta}_1,\hat{\beta}_2)&=\mathrm{Cov}(\frac{\sum_{i=1}^n\hat{r}_{i1}y_i}{\sum_{i=1}^n\hat{r}_{i1}^2},\frac{\sum_{i=1}^n\hat{r}_{i2}y_i}{\sum_{i=1}^n\hat{r}_{i2}^2})\\&=\frac{1}{\sum_{i=1}^n\hat{r}_{i1}^2\cdot\sum_{i=1}^n\hat{r}_{i2}^2}\cdot\mathrm{Cov}(\sum_{i=1}^n\hat{r}_{i1}y_i,\sum_{i=1}^n\hat{r}_{i2}y_i)\\&=\frac{1}{\sum_{i=1}^n\hat{r}_{i1}^2\cdot\sum_{i=1}^n\hat{r}_{i2}^2}\cdot\sum_{i=1}^n\hat{r}_{i1}\hat{r}_{i2}\cdot\sigma^2\\ \Longrightarrow\mathrm{Var}(w_1\hat{\beta}_1+w_2\hat{\beta}_2)&=\sigma^2\cdot\frac{\sum_{i=1}^n(w_2\hat{r}_{i1}+w_1\hat{r}_{i2})^2}{\sum_{i=1}^n\hat{r}_{i1}^2\cdot\sum_{i=1}^n\hat{r}_{i2}^2}\\ \Longrightarrow\mathrm{Var}(\sum_{i=1}^k w_i\hat{\beta}_i)&=\frac{1}{2}\sum_{m\neq n}\mathrm{Var}(w_n\hat{\beta}_m+w_m\hat{\beta}_n)+\sum_{m\neq n}w_mw_n\cdot\mathrm{Cov}(\hat{\beta}_m,\hat{\beta}_n)\end{aligned}$$

3.4 多个线性约束的检验

假设总体参数为 0 的原假设为排除性约束，同时包含多个参数的检验的原假设为多重约束，对多重约束进行的检验被称为多重假设检验或联合假设检验。

在排除性约束下的模型为受约束模型，它嵌套于不受约束模型之中，两个模型的自由度之差即为受约束的（即待检验的）自变量个数。

联合假设检验看的是不受约束模型和受约束模型之间的 SSR 相对变化，定义 F 统计量

$$F=\frac{(SSR_r-SSR_ur)/q}{SSR_ur/(n-k-1)}\sim F_{q,n-k-1}$$

其中 q 即为两个模型的自由度之差，是受约束的（即待检验的）自变量个数；F 统计量的分母是不受约束模型的误差的方差无偏估计。F 统计量的 SSR 表出形式是普适的，无论因变量在约束和不受约束模型是否一致。

如果拒绝原假设，即说受约束的自变量在对应显著性水平上是**联合统计显著**的，否则是联合不显著的。

联合假设检验可以穿透多重共线性，而多重共线性往往会干扰 t 检验的结果。实际上，单变量的双边 t 检验等价于对应的 F 检验，但 t 检验可以应对**单边**检验，且 t 统计量更**容易获得**。

当受约束模型和不受约束模型的因变量相同时，F 统计量还可以表示成 R^2 型。

$$F = \frac{(R_{ur}^2 - R_r^2)/q}{(1 - R_{ur})^2/(n-k-1)} \sim F_{q,n-k-1}$$

特别地，检验回归的整体显著性时，相当于对所有自变量施加约束，F 统计量的 R^2 型可以表示为

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)} \sim F_{k,n-k-1}$$

对于一般的线性约束，可以拆解为排除性约束和非排除性约束。对于后者，如 $\beta_1 = f(\vec{\beta})$，其中 $f(\vec{\beta})$ 是 $\vec{\beta}$ 的线性函数，仍然用 $\beta_1 = f(\vec{\beta}) + \theta$ 的方式最一般化地表出。一般来说，有常数会对应创建新的因变量，有其他回归系数的线性关系会对应创建新的自变量。

3.5 经济显著性

变量的经济显著性（或实际显著性）完全由系数的大小和符号决定；而统计显著性完全由 t 统计量的大小决定。

- 在样本容量扩大时有理由使用更小的显著性水平，经济上和统计上的显著性更可能达成一致，抵偿逐渐减小的标准误。
- 优先检查统计显著性，对于经济显著性应当注意变量在方程中出现的方式（比如单位、对数等）
- 不符合预期但不显著的变量可以忽略，但可能折射了模型变量的问题

4 多元回归分析：OLS 的渐进性

4.1 一致性

OLS 的一致性：在假定 MLR.1~MLR.4 下，对所有的 $j = 0, 1, ..., k$，OLS 估计量 $\hat{\beta}_j$ 都是 β_j 的一致估计。

实际上，一致性只需要假定误差项和单一自变量的零相关即可（相当于放款了 MLR.4），因此可以写出假定

MLR.4'-零均值和零相关：对所有的 $j = 1, 2, ..., k$，都有 $\mathrm{E}(u) = 0$ 和 $\mathrm{Cov}(x_{ji}, u) = 0$。

关于 MLR.4' 和 MLR.4 的讨论：

- MLR.4' 显得直接，因为这即为 OLS 推导过程的条件；
- 推导一致性时，满足 MLR.4' 即可，但这样的 OLS 估计值可能会有偏（但一致的）；
- MLR.4 更强，因为在有限样本下更希望考虑 OLS 估计量的精确性质；
- MLR.4 意味着模型正确地设定了总体回归函数，即 $\mathrm{E}(y|\vec{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$，于是可以得到解释变量对 y 的期望值的偏效应。如果只假设 MLR.4'，$\beta_0 + \beta_1 x_1 + ... + \beta_k x_k$ 就不一定代表了总体回归函数，有可能面临自变量的非线性函数与误差相关的可能性。

如果 u 和 $x_1, x_2, ...x_k$ 中的任何一个相关，那么通常也会导致 OLS 估计量失去一致性；同时，估计量还失去了无偏性。

对于估计量 $\hat{\theta}$，不一致性可以表述为 $\mathrm{plim}\ \hat{\theta} - \theta$，这也可以称为渐进偏误。

在简单回归模型中，$\hat{\beta}_1$ 的不一致性为

$$\mathrm{plim}\ \hat{\beta}_1 - \beta_1 = \frac{\mathrm{Cov}(x_1, u)}{\mathrm{Var}(x_1)}$$

可以进一步将其拓展，推导遗漏变量偏误的渐进类似情况。那么有 $\mathrm{plim}\ \hat{\beta}_1 = \beta_1 + \beta_2 \delta_1$，其中 $\delta_1 = \frac{\mathrm{Cov}(x_1, x_2)}{\mathrm{Var}(x_1)}$。因此，实际上可以把这种不一致性看作偏误，但区别在于，不一致性是用 x_1 的总体方差与 x_1 和 x_2 之间的总体协方差表示的，而偏误则基于其对应样本量（因为推导偏误时以特定样本为条件）。

不一致性和遗漏变量偏误既有联系也有区别，比如思考如下例子。给定总体回归方程为 $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$，其中 x_1 和 x_2 是独立的。对于一个给定的样本，这时如果分别进行如下三个线性回归

$$(1) \quad \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$
$$(2) \quad \hat{y} = \tilde{\alpha}_0 + \tilde{\alpha}_1 x_1$$
$$(3) \quad \hat{y} = \tilde{\gamma}_0 + \tilde{\gamma}_2 x_2$$

假设 x_2 对 x_1 的回归斜率系数为 $\tilde{\delta}_1$，若 x_1 和 x_2 在总体中不相关，那么 $\delta_1 = 0$，$\tilde{\alpha}_1$ 仍然是 β_1 的一致估计量，$\tilde{\gamma}_2$ 仍然是 β_2 的一致估计量。但遗漏变量偏误是基于特定样本的，在特定样本中通常 $\tilde{\delta}_1 \neq 0$，因此 $\tilde{\alpha}_1$ 不一定是 β_1 的无偏估计量，$\tilde{\gamma}_2$ 也不一定是 β_2 的无偏估计量。

4.2 渐进正态和大样本推断

注意，**正态性对于 OLS 的无偏性不起任何作用，也不影响 OLS 在 MLR.1~MLR.5 之下成为最优线性无偏估计的结论**。

尽管 y_i 不是来自一个正态分布，但仍然可以用到中心极限定理证明 OLS 的估计量满足渐进正态性——在大样本容量的情况下，OLS 估计量是近似正态分布的。

OLS 的渐进正态性：在 MLR.1~MLR.5 下，

- $\sqrt{n}(\hat{\beta}_j - \beta_j) \overset{a}{\sim} N(0, \sigma^2/a_j^2)$，其中 σ^2/σ_j^2 是渐进方差，斜率系数 $a_j^2 = \text{plim}(\frac{\sum_{i=1}^n \hat{r}_{ij}^2}{n})$，$\hat{r}_{ij}$ 是 x_j 对其余自变量进行回归所得到的残差。称 $\hat{\beta}_j$ 是渐近正态分布的
- $\hat{\sigma}^2$ 是 $\sigma^2 = \text{Var}(u)$ 的一个一致估计量
- $(\hat{\beta}_j - \beta_j)/\text{sd}(\hat{\beta}_j) \overset{a}{\sim} N(0,1), (\hat{\beta}_j - \beta_j)/\text{se}(\hat{\beta}_j) \overset{a}{\sim} N(0,1)$

注：从渐近的观点来看，$\hat{\sigma}$ 和 σ 是 “等价” 的，因此无论是考虑 $\text{sd}(\hat{\beta}_j)$ 还是 $\text{se}(\hat{\beta}_j)$，标准化的 $\hat{\beta}_j$ 都服从于渐近标准正态分布

在大样本之下，$\hat{\beta}_j$ 的估计方差的形式也和此前的相似。

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

随着样本容量的扩大，$\hat{\sigma}^2$ 概率收敛为 σ^2，$SST_j \to n\sigma_j^2$，因此 $\widehat{\text{Var}(\hat{\beta}_j)}$ 以 $\frac{1}{n}$ 的速度收敛至零。