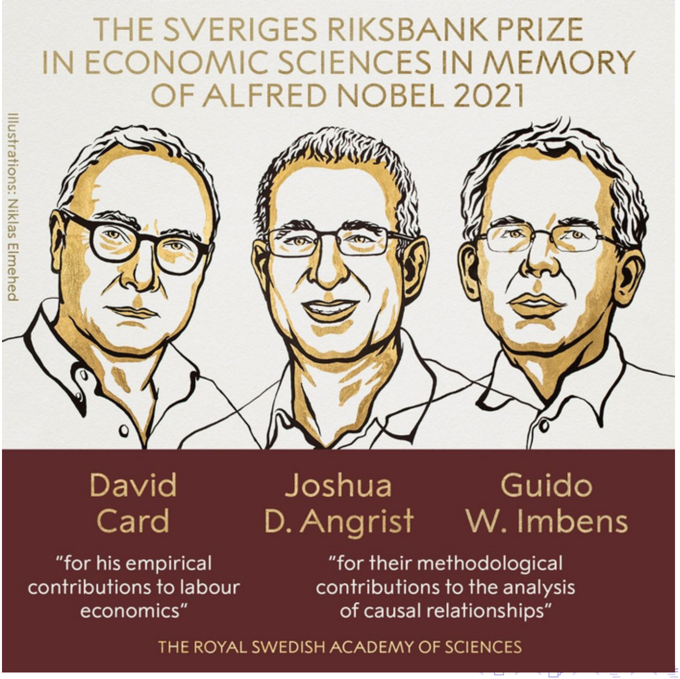
计量经济学STAT30021

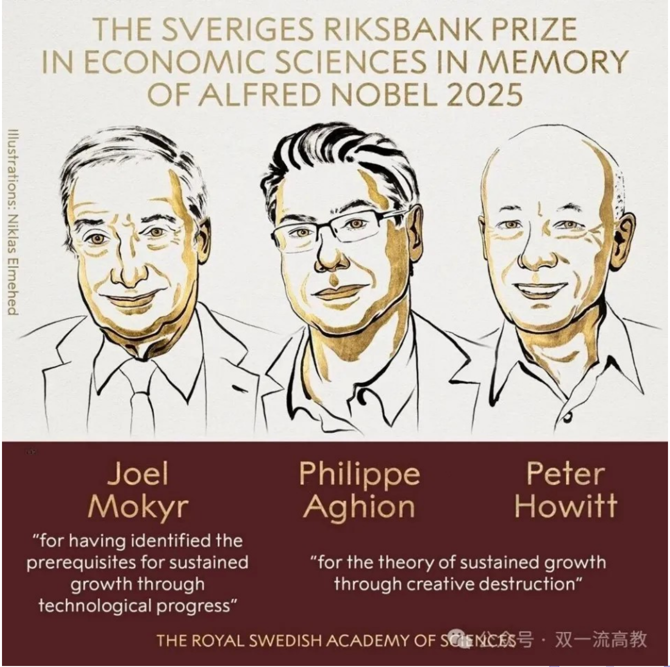
第六讲：多元线性回归分析的统计推断

肖志国

复旦大学管理学院

2025年10月


THE SVERIGES RIKSBANK PRIZE IN ECONOMIC SCIENCES IN MEMORY OF ALFRED NOBEL 2021
Illustrations: Niklas Elmehed
David Card
Joshua D. Angrist
Guido W. Imbens
"for his empirical contributions to labour economics"
"for their methodological contributions to the analysis of causal relationships"
THE ROYAL SWEDISH ACADEMY OF SCIENCES


THE SVERIGES RIKSBANK PRIZE IN ECONOMIC SCIENCES IN MEMORY OF ALFRED NOBEL 2025
Illustrations: Niklas Elmehed
Joel Mokyr
Philippe Aghion
Peter Howitt
"for having identified the prerequisites for sustained growth through technological progress"
"for the theory of sustained growth through creative destruction"
THE ROYAL SWEDISH ACADEMY OF SCIENCES
公众号 · 双一流高教

AEA Papers and Proceedings 2025, 115: 62–67
https://doi.org/10.1257/pandp.20251047

How Different Uses of AI Shape Labor Demand: Evidence from France†

By Philippe Aghion, Simon Bunel, Xavier Jaravel, Thomas Mikaelsen, Alexandra Roulet, and Jakob Søgaard*

Little is known about the firm-level effects of AI adoption on labor demand. Recent work has focused on estimating causal effects of AI adoption at the individual employee level *within* the firm (e.g., Brynjolfsson, Li, and Raymond 2023; Noy and Zhang 2023; Toner-Rodgers 2024), documenting positive effects of AI on worker-level productivity. Instead, in this paper we use comprehensive measures of AI adoption to document the relationship between AI and employment at the firm level.

Using French firm-level data on AI adoption between 2018 and 2020,[1] we establish four results. First, we document that AI-adopting firms are larger, more productive, more skill intensive, and primarily concentrated in IT and scientific activities.

Second, using difference-in-differences, we show that AI adoption is positively associated with an increase in total firm-level employment and sales. This finding is consistent with the idea that AI adoption induces productivity gains allowing the firm to expand its scope and raise its labor demand. The productivity effect appears to be stronger than potential displacement effects, whereby AI takes over the tasks of certain workers, reducing labor demand (Acemoglu and Restrepo 2018).[2]

多元线性回归模型

假设随机变量Y与k维随机向量$X=(X_1,\ldots,X_k)$满足如下关系：

$$Y=\beta_0+\beta_1X_1+\cdots+\beta_kX_k+u, \tag{1}$$

其中u满足

$$E[u|X]=0, Var(u|X)=\sigma^2. \tag{2}$$

设$(Y_i,X_i), i=1,\ldots,n$为来自总体(Y,X)的一个简单随机样本。则有：

$$Y_i=\beta_0+\beta_1X_{i1}+\cdots+\beta_kX_{ik}+u_i, \tag{3}$$

其中$\mathbf{u}=(u_1,\ldots,u_n)'$满足

$$E[u_i|X_i]=0, Var(\mathbf{u}|X_1,\ldots,X_n)=\sigma^2I. \tag{4}$$

最小二乘估计量的矩阵表示

令

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & X_{11} & \dots & X_{1k} \\ \vdots & \vdots & \vdots & \\ 1 & X_{n1} & \dots & X_{nk} \end{bmatrix}, \mathbf{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix},$$

则方程(3)可以写成如下形式：

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{u} \tag{5}$$

且β的最小二乘估计量$\hat{\beta}$满足的方程为：

$$\mathbf{X}'(\mathbf{Y} - \mathbf{X}\hat{\beta}) = 0. \tag{6}$$

或者等价的：

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{7}$$

我们定义$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$, $\hat{\mathbf{u}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\beta}$。

最小二乘估计量的性质

Theorem (无偏性)

在*Gauss-Markov*假设*MLR.1-MLR.4*条件下，β的*OLS*估计量$\hat{\beta}$是β的无偏估计量：

$$E[\hat{\beta}|\mathbf{X}] = \beta.$$

Theorem (方差公式)

在*Gauss-Markov*假设*MLR.1-MLR.5*条件下，$\hat{\beta}$的条件方差为：

$$Var[\hat{\beta}|\mathbf{X}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

具体地，对于任意$j = 1, \ldots, k$，有

$$Var(\hat{\beta}_j|\mathbf{X}) = \frac{\sigma^2}{SST_j(1 - R_j^2)}, \tag{8}$$

方差σ^2的估计

令

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{1}{n-k-1}\hat{\mathbf{u}}'\hat{\mathbf{u}} \tag{9}$$

则有：

Theorem (定理3.3)

在*Gauss-Markov*假设*MLR.1-MLR.5*都满足的条件下，$\hat{\sigma}^2$是σ^2的无偏估计量：

$$E[\hat{\sigma}^2|X] = \sigma^2.$$

Theorem (Gauss-Markov 定理)

在*Gauss-Markov*假设*MLR.1-MLR.5*都满足的条件下，β的*OLS*估计量$\hat{\beta}$是β的所有线性无偏估计量中方差最小的一个。

单个系数估计的标准差 vs 标准误

我们称σ为回归模型的标准差(standard deviation)，$\hat{\sigma}$为回归模型的标准误(standard error)。标准差是理论值(因而是未知的)，标准误是通过数据计算出来的标准差的估计值(因而是可以计算的)。

- $\hat{\beta}_j$的标准差为

$$sd(\hat{\beta}_j|X) = \frac{\sigma}{\sqrt{SST_j(1-R_j^2)}},$$

- $sd(\hat{\beta}_j|X)$的估计量，也就是$\hat{\beta}_j$的标准误为

$$se(\hat{\beta}_j|X) = \frac{\hat{\sigma}}{\sqrt{SST_j(1-R_j^2)}}, \tag{10}$$

多元回归模型的统计推断

通过多元回归模型，我们可以回答下面的问题：

- X_j对Y的（边际）影响是多少：$\hat{\beta}_j$
- 对上述的边际影响的估计的可靠性有多大？换句话说，估计量（随不同数据样本而变化）的波动幅度是多少：$se(\hat{\beta}_j)$

上述问题涉及的是变量之间的定量关系是多少以及可靠性有多大的问题。科学研究中，数据分析通常需还要回答另外一类更本质的问题：

- 在一组选定的解释变量中，哪些X_j是对Y有显著影响的，哪些是没有真正影响的？

这类问题叫做**假设检验**。由于假设检验涉及到概率计算，所以我们需要知道$\hat{\beta}_j$的分布。

正态性假设与抽样分布

- 估计量$\hat{\beta}_j$是一个随机变量
- 期望值和方差是这个随机变量的一阶矩和二阶矩
- 一阶矩和二阶矩对一个随机变量的描述是很粗略的
- 为了进行统计推断，也就是说，**判断统计决策正确或者错误的概率**，我们需要知道$\hat{\beta}_j$的概率分布(在这里称为抽样分布)
- 给定解释变量的时候，$\hat{\beta}_j$的条件概率分布显然依赖于误差项$u_1, \ldots, u_n$的概率分布
- 关于误差项$u_1, \ldots, u_n$的概率分布，一个最常用的假设是：

  正态性假设MLR.6 u和$X_1, \ldots, X_k$独立，且服从期望为零，方差为σ^2的正态分布：

$$u \sim N(0, \sigma^2). \tag{11}$$

满足假设MLR.1-MLR.6的线性回归模型称为**经典线性回归模型**。

正态抽样分布

Theorem (OLS估计量的条件正态分布)

在*MLR.1-MLR.6*条件下，

- $\hat{\beta}$基于所有解释变量的条件分布为均值为β，协方差为$Var(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$的多元正态分布。也就是

$$\hat{\beta}|x \sim N\left(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)$$

- $\hat{\beta}_j$基于所有解释变量的条件分布为均值为β_j，方差为$Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1-R_j^2)}$的正态分布。也就是

$$\hat{\beta}_j|x \sim N\left(\beta_j, \frac{\sigma^2}{SST_j(1-R_j^2)}\right),$$

从而$\frac{\hat{\beta}_j-\beta_j}{sd(\hat{\beta}_j)}$服从*(*无条件的*)*标准正态分布。

标准化估计量的t分布

Theorem (标准化估计量的t分布)

在*MLR.1-MLR.6*条件下，$\hat{\beta}_j$的标准化形式

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)}$$

服从自由度为$n-k-1$的t分布，也就是

$$\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k-1} \tag{12}$$

注：此分布为基于解释变量**X**的条件分布。但是由于条件分布t_{n-k-1}与**X**无关，所以此分布也是$\hat{\beta}_j$的无条件分布。

以上结果是我们在回归分析中进行假设检验的基础。

什么是假设检验？

考虑下面的例子：

- 在某次(大型)竞选中有两个候选人C,T
- 公布的投票结果显示，C的支持率为43%，T的支持率为57%
- 候选人C坚信他自己的支持率高于43%
- C雇请了一个咨询机构来调查投票结果是否公正
- 该机构随机调查了100名选民，询问他们是否在投票中选了C
- 假定这次调查结果显示，100个人中有53个在竞选中投票给了C
- 问：C能够下结论说竞选结果有问题吗？

假设检验基础

- 调查数字看起来似乎说明C在竞选的投票率被少算了
- 一定会是这样吗?
- 即便确实在全体选民中C的支持率为43%，仍然有可能出现100次调查中有53个是投票给C的(这个概率为?)
- 问题：咨询机构调查的样本数据有多大程度是与竞选结果43%对立的?

科学地解决这个问题的方法是建立一个假设检验

- 令θ表示全体选民中C的真正的支持率
- “公布的结果是准确的”这个假设可以表示成：

$$H_0: \ \theta = 0.43$$

这个假设被称作原假设

- 另外的一个假设就是“A的实际支持率高于43%”，也就是

$$H_1: \ \theta > 0.43$$

这个假设被称作备择假设。

假设检验基础

假设检验的基本任务就是要决定在什么情况下拒绝H_0

- 在什么情况下样本数据提供的信息更有利于H_1?
- 直观上来看，100个被调查者中投票给C的人越多，说明H_0越可能不成立，H_1越可能成立
- 如果100个被调查者只有44个表示投票给了C，这似乎并不能提供足够的证据来否决H_0
- 但是我们也不需要100个被调查者有100个表示投票给了C来否决H_0
- 那么到底当100个中有多少个投票给了A时，我们能够不令人怀疑地拒绝H_0呢?

检验统计量与拒绝域

具体而言，一个假设检验方法(又称决策规则)由两部分组成：检验统计量T，以及拒绝域$\mathcal{R}$

检验统计量T是我们用来决定是否拒绝H_0的一个量。

- T是样本数据的函数，通过实际数据计算出来的T的值我们记作t。
- 选择什么样的量作为检验统计量取决于原假设和备择假设的形式，以及出于方便的考虑。
- 一个通常的要求是，在H_0成立的情况下，我们选取的检验统计量T具有一个已知的概率分布(此分布不依赖于样本数据，或者未知参数)

拒绝域$\mathcal{R}$表示的是使得我们拒绝H_0的T的取值范围。

- 当$T \in \mathcal{R}$ 时，我们拒绝H_0，当$T \notin \mathcal{R}$时，我们不能拒绝H_0。
- 拒绝域$\mathcal{R}$的形式通常应该和H_1吻合。

单样本t-检验

设$Z_1, \ldots, Z_n$为来自正态分布$N(\mu, \sigma^2)$的简单随机样本，其中$\mu \in R, \sigma > 0$为未知参数。取显著性水平为α，检验假设：

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu > \mu_0$$

- 检验统计量$T = \frac{\sqrt{n}(\bar{Z} - \mu_0)}{S_Z}$，其中

$$S_Z = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Z_i - \bar{Z})^2}$$

- 拒绝域$\mathcal{R}$: 当$T > t_\alpha(n-1)$时，拒绝原假设
- 实际操作：给定数据之后，我们计算T的实现值t。如果$t > t_\alpha(n-1)$，我们拒绝原假设，否则我们不能拒绝原假设。

第一类与第二类错误

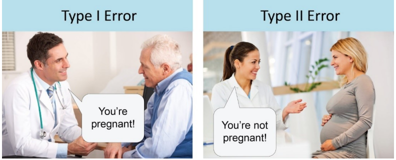
- 假设检验是对未知参数取值范围的判断
- 理论上，任何假设检验中，我们都有可能犯错误(为什么?)
- 如果我们在假设检验中犯了错误，那么它必属于以下两种错误中的一种
- 第一类错误(弃真错误)：当H_0为真时，我们拒绝了H_0
- 第二类错误(取伪错误)：当H_1为真时，我们没有拒绝H_0

Figure 1

		Reality	
		H_0 Is True	H_1 Is True
Conclusion	Do Not Reject H_0	Correct Conclusion	Type II Error
	Reject H_0	Type I Error	Correct Conclusion

第一类与第二类错误

$$H_0: \quad \text{Not pregnant} \quad v.s. \quad H_1: \quad \text{Pregnant}$$



	Test result: Pregnant	Test result: Not Pregnant
Reality: Not Pregnant	Type I error (probability = α) *False Positive*	Probability = 1-α *Correct Negative*
Reality: Pregnant	Power (1-β) *Correct Positive*	Type II error (probability = β) *False Negative*

第一类与第二类错误

一般的，记$\Theta = \Theta_0 \cup \Theta_1$以及

$$H_0 : \theta \in \Theta_0 \quad v.s. \quad H_1 : \theta \in \Theta_1.$$

一个假设检验规则 $\{$拒绝H_0若$T \in \mathcal{R}\}$ 的功效函数是：

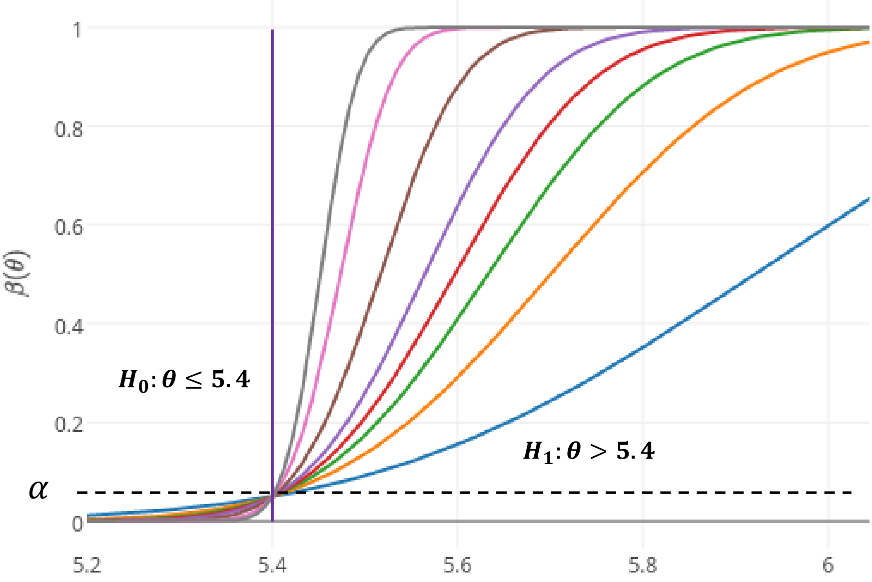
$$power(\theta) = P(\text{拒绝}H_0|\theta \in \Theta).$$

- 犯第一类错误的概率通常记作α，犯第二类错误的概率通常记作β，也就是

$$\alpha(\theta) = P(\text{拒绝}H_0|\theta \in \Theta_0), \ \beta = P(\text{不拒绝}H_0|\theta \in \Theta_1)$$

- 一个理想的假设检验方法应该让α,β都为零
- 但是α,β不可能同时为零。可以证明，当α大时，β小；反之亦然
- 通用的寻找最优检验方法是先为α设置一个上限（显著性水平），然后在所有满足条件的方法中寻找β最小的那个。

功效函数Power Function



功效

我们定义功效（power）为$1-\beta$。也就是说：功效越小，犯第二类错误的概率越大，也就是越容易错误的接受了原假设；功效越大，犯第二类错误的概率越小，也就是越不容易错误的接受了原假设。

注意我们常说的功效和功效函数不完全一样。功效是功效函数在Θ_1上的那部分的取值。

后面我们要学到一个常用的双重差分法（DiD），这个方法涉及到一个叫做“平行趋势假设”的检验。最新的研究表明，常用的平行趋势检验的方法功效都很低。也就是说，我们很容易就错误地（也就是，在不该接受的时候）接受了“平行趋势假设合理”这样一种判断。

假设检验的类型

双侧检验:

$$H_0 : \theta = \theta_0 \text{ vs } H_1 : \theta \neq \theta_0$$

单侧检验:

$$\begin{aligned}
H_0 : \theta = \theta_0 &\quad \text{vs} \quad H_1 : \theta > \theta_0 \\
H_0 : \theta = \theta_0 &\quad \text{vs} \quad H_1 : \theta < \theta_0 \\
H_0 : \theta \leq \theta_0 &\quad \text{vs} \quad H_1 : \theta > \theta_0 \\
H_0 : \theta \geq \theta_0 &\quad \text{vs} \quad H_1 : \theta < \theta_0
\end{aligned}$$

假设检验的简易办法：p-值

- 假设检验的任务就是要找到一个拒绝H_0的准则。p-值规则是一种等价于拒绝域规则的准则。
- p-值规则的背后逻辑是：应该发生的都是大概率事件，小概率事件不应该发生。如果现实中小概率事件出现了，那么必然某个地方出问题了。
- p-值指的是“在原假设成立的情况下，出现现有观测值（以及比现有观测值更极端情形）的概率”。
- 显然，根据定义，这个概率越小，表示现有观测值在原假设下是小概率事件，是不应该发生的。
- 由于现实数据是不会出错的，那么**p-值越小只能说明原假设越有问题**。也就是说，p-值越小，表示样本数据提供的信息越不利于H_0。
- p-值多小才算足够小？这取决于研究者所在研究领域的共识。一般大家用0.05作为衡量p-值大小的标准。

多元线性回归中的默认t-检验

在多元线性回归模型中，其中$\beta_j, j = 1, \ldots, k$为我们通常感兴趣的未知参数。取显著性水平为$\alpha$，检验假设：

$$H_0 : \beta_j = 0 \text{ vs } H_1 : \beta_j \neq 0$$

- 检验统计量$T_{\hat{\beta}_j} = \hat{\beta}_j / se(\hat{\beta}_j)$称为t-统计量
- 拒绝域$\mathcal{R}$: 当$|T_{\hat{\beta}_j}| > t_{\frac{\alpha}{2}}(n-k-1)$时，拒绝原假设
- 假设给定数据之后，我们计算$T_{\hat{\beta}_j}$的实现值为t_j。则如果$|t_j| > t_{\frac{\alpha}{2}}(n-k-1)$，我们拒绝原假设，否则我们不能拒绝原假设。
- p-值等于$P(|t(n-k-1)| > t_j)$，其中$t(n-k-1)$表示自由度$n-k-1$为的t-分布
- 当p-值小于等于0.05时，我们通常说X_j对Y的影响是显著的

多元线性回归中的t-检验

类似地，我们也可以进行如下的t-检验，其中b_j为已知常数：

1. 情形1：
$$H_0: \beta_j = b_j \text{ vs } H_1: \beta_j > b_j$$
决策规则： 当$T_j > t_\alpha(n-k-1)$时，拒绝原假设
2. 情形2：
$$H_0: \beta_j = b_j \text{ vs } H_1: \beta_j < b_j$$
决策规则： 当$T_j < -t_\alpha(n-k-1)$时，拒绝原假设
3. 情形3：
$$H_0: \beta_j = b_j \text{ vs } H_1: \beta_j \neq b_j$$
决策规则： 当$|T_j| > t_{\frac{\alpha}{2}}(n-k-1)$时，拒绝原假设

注：在以上三种情况中，

$$T_j = \frac{\hat{\beta}_j - b_j}{se(\hat{\beta}_j)}$$

总体显著性的F-检验

考虑检验下面的假设：

$$H_0 : X_1, \ldots, X_k \text{ 不能解释Y的变化}$$

写成参数β_j的形式为：

$$H_0 : \beta_1 = \cdots = \beta_k = 0 \text{ vs } H_1 : \beta_1, \ldots, \beta_k \text{中至少一个不为0}$$

- 检验统计量$F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$称为F-统计量
- 拒绝域$\mathcal{R}$: 当$F > F_\alpha(k, n-k-1)$时，拒绝原假设，其中$F_\alpha(k, n-k-1)$为$F(k, n-k-1)$分布的上-α分位点

其他线性约束条件的检验

考虑检验下面的假设：

$$H_0:\beta_1=\beta_2 \ vs \ H_1:\beta_1\neq\beta_2$$

检验办法：

- 令$\theta_1=\beta_1-\beta_2$，也就是$\beta_1=\theta_1+\beta_2$
- 在回归方程中用θ_1替换掉β_1
- 新的回归方程将是以$\theta_1,\beta_2,\ldots,\beta_k$为参数的
- 对新的回归方程引用如下的t-检验：

$$H_0:\theta_1=0 \ vs \ H_1:\theta_1\neq 0$$

其他线性约束条件的检验

考虑检验下面的假设：

$$H_0: \beta_3 = \beta_4 = \beta_5 = 0 \ \ vs \ \ H_1: H_0\text{不成立}$$

检验办法：

- 定义原线性模型为无约束模型，满足$\beta_3 = \beta_4 = \beta_5 = 0$的模型为有约束模型
- 对无约束模型和有约束模型分别作OLS估计，计算其残差平方和SSR_{ur}, SSR_r
- 定义检验统计量为

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} \tag{13}$$

  其中q为H_0中约束等式个数
- 则拒绝域为：当$F > F_\alpha(q, n-k-1)$时，拒绝原假设，其中$F_\alpha(q, n-k-1)$为$F(q, n-k-1)$分布的上-α分位点

F检验的R^2表示以及p-值

F检验的公式(13)可以等价地表示为：

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

其中R_{ur}^2, R_r^2分别为无约束模型以及有约束模型的R^2.

假设我们通过实际数据计算出来的F检验统计量的值为f，那么F假设检验的p-值为：

$$p - \text{值} = P\left(F(q, n-k-1) > f\right).$$

不能拒绝 v.s. 接受

- 在假设检验中，会有两种情况：(1) 数据提供的信息拒绝H_0,或者(2)数据提供的信息不能拒绝H_0
- 对于第二种情况，有些书上写成“接受H_0”：这是不对的
- 不能拒绝 $\neq$ 接受!
- 例子. 假设在某个线性回归中我们计算得$\hat{\beta}_j = 1$, $se(\hat{\beta}_j) = 2$。考虑假设检验

$$H_{0a} : \beta_1 = 0 \ vs \ H_{1a} : \beta_1 \neq 0$$

由此得$t_{\hat{\beta}_j} = 1/2 = 0.5$。从而我们不能拒绝$H_{0a} : \beta_1 = 0$。但是如果我们考虑假设检验

$$H_{0b} : \beta_1 = 0.5 \ vs \ H_{1b} : \beta_1 \neq 0.5$$

我们可以计算得$t_{\hat{\beta}_j} = (1 - 0.5)/2 = 0.25$。从而我们也不能拒绝$H_{0b} : \beta_1 = 0.5$。

统计显著性vs 实际显著性

- 一个变量的统计显著性是指其p值很小，也就是说该变量的系数能够被很准确的估计（标准误相对于估计值来说很小）。
- 一个变量的实际显著性是指其估计量的绝对大小，也就是其系数估计的绝对值。
- 在实际的回归分析中，我们不仅要看变量的统计显著性，更要看其实际显著性。
- 比如，假设在一个对工资进行的回归分析中，工作经验（experience）和受教育年限是两个测量单位相同的变量。假设前者的估计值为0.99，p值为0.09，后者的估计值为0.02，p值为0.01。那么按照统计显著性，受教育年限比工作经验重要。但是从实际的影响强度来看，工作经验要远比受教育年限重要。
- 一个相关问题：$H_0 : \beta = 1$ vs $H_1 : \beta \neq 1$。假设$\hat{\beta} = 0.99, se(\beta) = 0.001$。请问$\beta$是否等于1？