计量经济学原理笔记

Kevin

目录

1 引论

1.1 什么是计量经济学

对于多数经济决策或选择问题，仅知道经济变量之间相关及其方向还不够，我们还要知道这种相关性的大小，也就是说，某个变量的变化对另一个变量影响有多少（how much）。计量经济学就是研究如何利用经济学、商业和社会科学的理论及数据，结合统计学工具，来回答上面的“多少”的学科。

举个 FRB（美联储）进行宏观调控的例子。当发现价格上升时（表明通胀率上升），FRB 必须作出决策：是否需要减缓经济增长速度？比如，提高向成员银行收取的利率（贴现率），或提高银行间隔夜利率（联邦基金利率）。提高这些利率将在经济中产生连锁反应，引发其它利率的上升，投资成本上升，消费者减少耐用品购买。最终，总需求下降，从而降低通胀水平。那么，到底需要加息多少，既能降低通胀，也能维持经济稳定增长？这取决于企业和个人对加息以及削减投资对 GNP 影响的反应。其中，主要的弹性（elasticities）和乘子（multipliers）称为参数（parameters）。参数的值是未知的，需要通过经济数据样本来估计。

计量经济学就是给定数据来估计参数。决策者每天都在面临类似的问题：

- 增加警力与降低暴力犯罪的问题
- 广告支出与销售额
- 大学学费问题
- 产品未来十年的需求，是否投资新厂和设备
- 房产商预测未来几年的人口和收入，开设赌场是否能盈利
- 投资组合决策问题，多少投资于股票基金，多少投资于货币市场
- 公共交通费增加，如何影响人们对交通工具的选择

1.2 计量模型

某汽车销售的计量模型可以表示为:

$$Q^d = f(P, P^s, P^c, INC) + e$$

其中，Q^d 表示某汽车的需求量，P 是该汽车的价格，P^s 是替代品的价格，P^c 是补充品（比如汽油）的价格，INC 是收入水平。随机误差 e 表示其他影响销售的因素，它反映了经济活动内在的不确定性。接下来，需要对经济变量加以代数关系限定。比如，将需求关系的系统部分设定为线性函数:

$$f(P, P^s, P^c, INC) = \beta_1 + \beta_2 P + \beta_3 P^s + \beta_4 P^c + \beta_5 INC$$

对应的计量模型是:

$$Q^d = \beta_1 + \beta_2 P + \beta_3 P^s + \beta_4 P^c + \beta_5 INC + e$$

上面的系数是未知参数，需要基于数据来估计。函数形式代表对变量关系的假设，如何选择与理论和数据都兼容的函数形式是一个挑战。每个计量模型都包含系统部分和不可观测的随机部分，前者来自经济理论，包含对函数形式的假设；后者表示 “噪声”。

基于计量模型和数据样本，我们对真实世界进行统计推断并学习到东西。统计推断过程包括估计（Estimating）、预测（Predicting）和检验（Testing）等。

1.3 数据类型

经济学和社会学中的数据往往是观测数据，而不是控制试验得到的，这使得学习经济参数更加困难。控制试验的例子有产品销售、Tennessee's Project Star 项目，但这种例子在社会科学中非常少。非试验数据的典型例子是调查数据（survey data），比如著名的 CPS 项目。

经济数据有多种类型: 有微观（micro）或宏观（macro）之分，有流量或存量之分，也有定量或定性之分。计量经济学一般研究的数据类型有: 时间序列（time series）、横截面（cross-sectional）和面板（panel）数据。其中，面板数据也叫经线数据（longitudinal），关键特征是每一个微观单位都带有一段时期的观测。如果每个单位的观测期数相同，称之为平衡面板（balanced panel）。观测期数通常小于单位数量。

1.4 研究过程

首先，选择一个感兴趣的 topic。通常在研究这个 topic 过程中，其他问题也会随之而来，这些新问题会提供关于更多原始 topic 的线索（灵感），或者使你更感兴趣的研究路径。一旦选定主题，接下来的研究过程一般包括：

- 经济理论是思考问题的起点。有哪些变量，关系方向如何？每个研究项目都是从建立经济模型开始，列出一系列感兴趣的问题（假设）。当然，研究过程中可能会出现更多问题。
- 经济模型建立后，要转换为计量模型，我们需要选择一个函数形式，并对误差项做一些假定。
- 搜集样本数据，基于初始假设和数据特点，选择某个统计分析方法
- 估计未知参数（借助统计软件包），作出预测，检验假设。
- 对模型进行诊断，即检查假设是否正确。比如，解释变量是否显著？函数形式合适吗？
- 分析评估经济结果和实证意义，哪些问题需要进一步研究？

2 简单线性回归模型

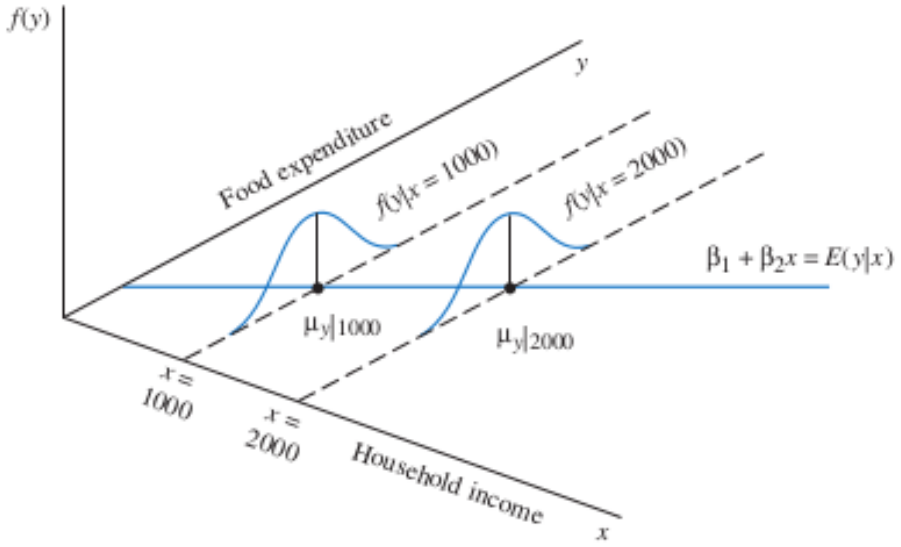
看一个简单但重要的经济例子。假定我们要考察家庭收入与食物支出之间的关系。从一个特定总体中随机选择一些家庭，该特定总体可以是某个省市或全国。假设我们目前只关心周收入为 1000 元的家庭，从中选取一些家庭，进行询问：上周你家人均食物支出多少？这里的周食物支出（记为 y）是随机变量。记收入为 x。

为了考察支出与收入之间的关系，我们需要建立一个经济模型及相应的计量模型。由经济理论知道，平均周人均支出 $E(y|x) = \mu_{y|x}$ 是收入 x 的函数，对于不同的收入水平（比如 2000 元），相应的平均周人均支出也不一样。换言之，周人均支出的概率分布取决于收入水平（条件分布）。假定这种关系是线性的：

$$E(y|x) = \mu_{y|x} = \beta_1 + \beta_2 x$$

上式的条件均值称为简单回归函数（simple regression function）。其中的 β_2 实际上就是边际消费倾向。这个经济模型是实际经济行为的一种抽象。给定

$x = 1000$，y 是一个均值为 $\beta_1 + \beta_2(1000)$ 的随机变量，对于每一个 x 都如此：



线性回归函数是计量模型的基础，再对数据做一些假定，就得到计量模型：

- 对于每个 x，y 的分布具有同方差：

$$\text{var}(y|x) = \sigma^2$$

- y（的样本点）之间不相关（协方差为 0），即它们之间不具有线性关系：

$$\text{cov}(y_i, y_j) = 0$$

（更强的假设是，y 之间统计独立）

- x 不是随机的，且必须取至少两个值
- （可选）对每个 x，y 都服从正态分布：

$$y \sim N\left[(\beta_1 + \beta_2 x), \sigma^2\right]$$

上面的模型直接对被解释变量作出假定，出于统计上方便的目的，我们可以用另一种方式对假定进行描述。回归分析的本质是，被解释变量的每个观测值都可以分解成两个部分：一个是系统性分量，一个是随机分量。系统性分量就是 y 的条件均值（回归函数），它本身并不是随机的。随机分量是

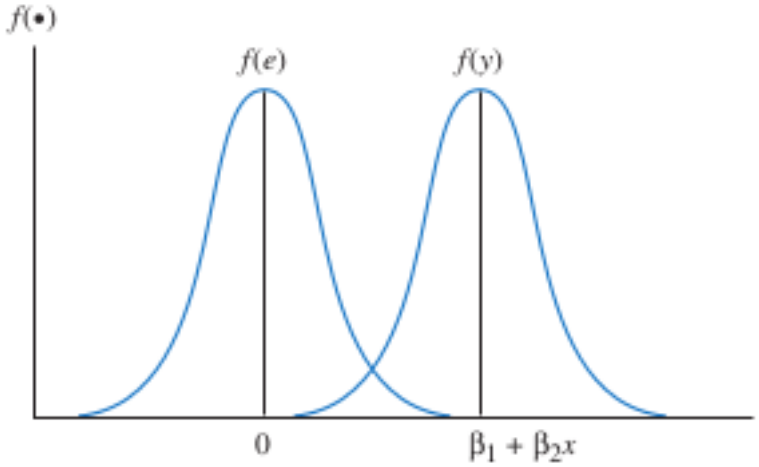
y 与其条件均值之间的差，称为随机误差项（random error term），定义为：

$$e = y - E(y|x) = y - \beta_1 - \beta_2 x$$

从而，简单线性回归模型可重写为：

$$y = \beta_1 + \beta_2 x + e$$

由于 y 随机，因此 e 也随机，基于上述有关假定，我们可以得到 e 的性质：均值为 0，具有同方差 σ^2。也就是说，y 和 e 的概率密度函数只是位置不一样：



现在我们可以进一步讨论上述关于 x 假定。x 非随机意味着它的值是已知的，在统计上，这样的值被称为 “随机抽样中固定”。实际上这种情况很少见，但作出这样的假定不影响后面的结果，而且在记号处理上比较方便。既然 x 非随机，我们就不需要再使用条件符号 | 了（也有一些情况下不能假定 x 固定，后面详述）。

ASSUMPTIONS OF THE SIMPLE LINEAR REGRESSION MODEL-II

SR1. The value of y, for each value of x, is

$$y = \beta_1 + \beta_2 x + e$$

SR2. The expected value of the random error e is

$$E(e) = 0$$

which is equivalent to assuming that

$$E(y) = \beta_1 + \beta_2 x$$

SR3. The variance of the random error e is

$$\text{var}(e) = \sigma^2 = \text{var}(y)$$

The random variables y and e have the same variance because they differ only by a constant.

SR4. The covariance between any pair of random errors e_i and e_j is

$$\text{cov}(e_i, e_j) = \text{cov}(y_i, y_j) = 0$$

The stronger version of this assumption is that the random errors e are statistically independent, in which case the values of the dependent variable y are also statistically independent.

SR5. The variable x is not random and must take at least two different values.

SR6. (*optional*) The values of e are normally distributed about their mean

$$e \sim N(0, \sigma^2)$$

if the values of y are normally distributed, and vice versa.

为了估计参数，我们用最小二乘法则（least square principle）：每个点到拟合线的垂直距离的平方和最小。拟合线为：

$$\hat{y}_i = b_1 + b_2 x_i$$

样本点到拟合线的垂直距离，称为最小二乘残差（least square residuals）：

$$\hat{e}_i = y_i - \hat{y}_i = y_i - b_1 - b_2 x_i$$

LS 法则说的是，最佳的拟合系数 b_1, b_2，将使下面的残差平方和最小：

$$SSE = \sum_{i=1}^{N} \hat{e}_i^2$$

问题是如何方便地求得这样的系数？只要求解一个最优化问题：

$$\min_{\beta_1,\beta_2} S(\beta_1,\beta_2)=\sum_{i=1}^{N}(y_i-\beta_1-\beta_2 x_i)^2$$

该问题的解称为最小二乘估计（least square estimators）：

$$b_2=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$$

$$b_1=\bar{y}-b_2\bar{x}$$

其中，$\bar{y}$ 和 $\bar{x}$ 分别为 y 和 x 的样本均值。将数据点代入上面两个式子，就得到截距和斜率的估计值（estimates）。事实上，b_1 和 b_2 是随机变量，对于每一个特定样本，得到 b_1 和 b_2 的一个观测。所以，作如下区分：上面的一般公式成为最小二乘估计（estimators），应用于某个特定样本得到的是最小二乘估计值（estimates）。

对于食物支出的例子，得到最小二乘估计值后，可以给出拟合回归线：

$$\hat{y}_i=83.42+10.21x_i$$

R 代码如下：

```
food <- read.table('data/food.dat')
colnames(food) <- c('food_exp','income')
food_lm <- lm(food_exp ~ income, data=food)
summary(food_lm)
```

得到最小二乘估计值后，就可以用它们来解释所考虑的经济模型，也可以作预测。（注意，解释截距时需要非常小心，因为通常我们并没有 $x=0$ 的点。）

从上面的模型，我们可以得到**收入弹性**（income elasticity）：

$$\varepsilon=\frac{\Delta E(y)/E(y)}{\Delta x/x}=\beta_2\cdot\frac{x}{E(y)}$$

回归线上每一点的弹性并不相同，通常我们计算均值点上的弹性：

$$\hat{\varepsilon}=b_2\frac{\bar{x}}{\bar{y}}=10.21\times\frac{19.60}{283.57}=0.71$$

这个收入弹性估计值意味着，当 x 和 y 取其样本均值时，平均意义上，收入每增加 1%，食物支出增加 0.71%。可以发现，由于收入弹性小于 1，这里的食物属于必需品（necessity）而不是奢侈品（luxury）。

我们得到了最小二乘估计，那么它们好不好？这个问题本质上是无法回答的。因为我们永远无法知道真实值，到底这些估计与真实值多接近，我们无法得知。与其讨论估计的质量，我们退一步，考察估计过程的质量。如果选择另一个样本，即使数量和收入水平都相同，也将得到不同的估计。这种抽样差异（sampling variation）是不可避免的。因此，作为估计过程，b_1 和 b_2 都是随机变量，它们的性质称为抽样性质（sampling properties）。

考察估计 b_2。上面的公式称为均值离差形式（deviation from mean form）。利用假设 SR1，可将 b_2 写成线性形式：

$$b_2 = \sum_{i=1}^{N} w_i y_i, \quad w_i = \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}$$

由于 w_i 取决于非随机的 x，因此 w_i 也是非随机的。上式意味着，每个估计量都是 y_i 的加权平均，这样的估计量称为线性估计量（linear estimators）。b_2 还可以进一步写成理论上更方便的形式：

$$b_2 = \beta_2 + \sum w_i e_i$$

先看均值。对上式取期望，容易得到 $E(b_2) = \beta_2$，说明在前面的模型假定下，b_2 是 β_2 的无偏估计。其中用到“误差均值为 0”（SR2）和“解释变量非随机”（SR5）两个假设。容易证明，b_1 也是 β_1 的无偏估计。无偏性并不是说每一个样本作回归得到的估计值（estimate）都接近真实参数，而是说在平均意义下接近。假如我们从同一个总体中重复抽样，分别作回归，得到的估计值的均值将接近真实参数。换句话说，我们不能说最小二乘估计值是无偏的，只能说最小二乘估计过程是无偏的。

再看方差和协方差。方差衡量估计量的精度（precision），它告诉我们不同样本得到的估计值之间变化有多大。因此，常常讨论估计量的抽样方差（sampling variance）或抽样精度（sampling precision）。方差越小，抽样精度越大。当 SR1-SR5 成立时，我们有：

$$\mathrm{var}(b_1) = \sigma^2 \left[\frac{\sum x_i^2}{N \sum (x_i - \bar{x})^2} \right]$$

$$\mathrm{var}(b_2) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

$$\mathrm{cov}(b_1, b_2) = \sigma^2 \left[\frac{-\bar{x}}{\sum (x_i - \bar{x})^2} \right]$$

影响方差和协方差的因素有：

- 随机误差项方差 σ^2
- 解释变量离差平方和
- 样本容量
- 解释变量观测值的平方和（影响 b_1 的方差）
- 解释变量的样本均值

GAUSS − MARKOV 定理：对于满足假设 SR1-SR5 的线性回归模型，最小二乘估计量 b_1 和 b_2 是具有最小方差的线性无偏估计（BLUE）。

下面考察 LS 估计量的概率分布。如果假设 SR6 成立，则有：

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2 \sum x_i^2}{N\sum(x_i-\bar{x})^2}\right)$$

$$b_2 \sim N\left(\beta_2, \frac{\sigma^2}{\sum(x_i-\bar{x})^2}\right)$$

如果误差项不服从正态分布，只要样本容量足够大，由中心极限定理（CLT）可知，LS 估计量近似服从正态分布。

最后只剩一个未知参数 σ^2，我们有：

$$\mathrm{var}(e_i) = \sigma^2 = E(e_i^2)$$

但 e_i 不可观测，不能对其取均值作为方差的估计。一个合理的替代品是残差，可以证明：

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N-2}$$

是误差方差的无偏估计，其中分母中的 2 就是估计参数的个数。

有了误差方差的无偏估计，我们就可以估计最小二乘估计量的方差和协方差：

$$\widehat{\mathrm{var}(b_1)} = \hat{\sigma}^2\left[\frac{\sum x_i^2}{N\sum(x_i-\bar{x})^2}\right]$$

$$\widehat{\mathrm{var}(b_2)} = \frac{\hat{\sigma}^2}{\sum(x_i-\bar{x})^2}$$

$$\widehat{\mathrm{cov}(b_1,b_2)} = \hat{\sigma}^2\left[\frac{-\bar{x}}{\sum(x_i-\bar{x})^2}\right]$$

这些估计方差的平方根称为“标准误”（standard errors）：

$$\mathrm{se}(b_1) = \sqrt{\widehat{\mathrm{var}(b_1)}}$$

$$\mathrm{se}(b_2) = \sqrt{\widehat{\mathrm{var}(b_2)}}$$

标准误度量了最小二乘估计量在重复抽样中的抽样变异性（sampling variability），通过 Monte Carlo 模拟，可以看到标准误确实接近真实抽样变异。

世界并非线性。事实上，很多经济关系都是曲线（curvilinear）关系。通过变量转换，简单线性模型也可以刻画非线性关系。考虑房地产价格的例子。价格（PRICE）是房子尺寸（SQFT）的函数，一般认为，大房子尺寸增加的价值比小房子大，并不是同步增加的。可以用平方项或对数项替换线性项。加入平方项，得到二次模型：

$$PRICE = \alpha_1 + \alpha_2 SQFT^2 + e$$

这也是一个简单线性回归模型，此时 $y = PRICE$，$x = SQFT^2$。当然，模型参数 α_2 不再表示斜率。用前面的方法得到参数估计 $\hat{\alpha}_1$ 和 $\hat{\alpha}_2$，则拟合方程为：$\widehat{PRICE} = \hat{\alpha}_1 + \hat{\alpha}_2 SQFT^2$，斜率为：

$$\frac{d\left(\widehat{PRICE}\right)}{dSQFT} = 2\hat{\alpha}_2 SQFT$$

除了加入平方项，也可引入对数项，得到对数线性模型（log-linear model）。对数线性函数形式为：$\ln(y) = a + bx$。对数线性函数相当于指数函数，其中，$y > 0$，任意一点的斜率为 by，意味着当 $b > 0$ 时，y 越大其边际效应也越大。有时也称这种情况为 “按递增速度增加”。容易计算，对数线性模型的弹性为 bx。系数 b 实际上是所谓的半弹性（semi-elasticity）：

$$\eta = \frac{100(dy/y)}{dx} = 100b$$

表示 x 变化一个单位时 y 变化的百分比。对前面的例子应用对数线性模型得到：

$$\ln(PRICE) = \gamma_1 + \gamma_2 SQFT + e$$

对应变量采用对数变换，可以将右偏的数据正则化，很多经济变量，比如价格、收入、工资等都具有偏性分布，对其取对数很常见。可以用类似方法求得估计，从而得到斜率为 $0.0004113\widehat{PRICE}$，弹性为 $0.0004113SQFT$，半弹性为 0.04%（意味着每增加 1 平方英尺，房价上升 0.04%；或者说，增加 100 平方英尺，房价上升 4%）。

指示变量（indicator variable）是只取 0 或 1 的二元变量，用来表示性别、种族、位置等非定量变量。以 `utown.dat` 数据集为例，其中 `utown` 是

指示变量，当房子靠近大学时为 1，否则为 0。如果将 UTOWN 作为解释变量，回归模型为：

$$PRICE = \beta_1 + \beta_2 UTOWN + e$$

回归函数可以写成：

$$E(PRICE) = \beta_1 + \beta_2 UTOWN = \begin{cases} \beta_1 + \beta_2 & \text{if } UTOWN = 1 \\ \beta_1 & \text{if } UTOWN = 0 \end{cases}$$

这意味着两类区域的平均房价不同。β_2 并不是斜率，而是表示两类区域的均价之差。由于模型中除了区域位置外没有其他变量，可以认为指示变量将观测划分为两个总体。利用 LS 估计，可得到回归方程为：

$$\widehat{PRICE} = b_1 + b_2 UTOWN = 215.7325 + 61.5091 UTOWN$$

因此，University Town（$UTOWN = 1$）的均价为 277241.60，Golden Oaks（$UTOWN = 0$）的均价为 215732.50。回归模型中的参数 β_1 表示 Golden Oaks 的样本均值，β_2 表示两类区域的样本均值之差。

附录：Monte Carlo 模拟所谓 MC 模拟，先设定一个数据生成过程（DGP），再通过计算机随机数生成器创建人工数据样本。基于这些样本，研究估计量的重复抽样性质。简单线性回归模型的 DGP 为：

$$y_i = E(y_i|x_i) + e_i = \beta_1 + \beta_2 x_i + e_i, \quad i = 1, \ldots, N$$

为了得到 y_i，我们先创建回归关系的系统部分 $E(y_i|x_i)$，再加上一个随机误差 e_i。为了创建回归函数，需要三步：

- 选择样本容量 N。对于前面的例子来说，我们选 $N = 40$。
- 选择 x_i。设定 $x_1, x_2, \ldots, x_{20} = 10, x_{21}, x_{22}, \ldots, x_{40} = 20$。
- 选择系数 $\beta_1, \beta 2$。有趣的是，对于满足 SR1-SR5 的最小二乘估计，这些参数的实际大小并无太大影响。估计方差和协方差并不依赖它们。为了与回归结果大致一致，设定 $\beta_1 = 100, \beta_2 = 10$。

在这样的设定下，我们可以创建 40 个数据点：

$$E(y_i|x_i = 10) = 100 + 10 \times 10 = 200, \quad i = 1, \ldots, 20$$

$$E(y_i|x_i = 20) = 100 + 10 \times 20 = 300, \quad i = 21, \ldots, 40$$

为了创建随机误差，需要用随机数生成器。计算机生成的随机数是伪随机数，大约生成 10^{13} 个值后就会重复（recycling），不过这些数应该够了。假定 SR6 成立，且选择 $\sigma^2 = 2500$。有了随机误差，就可以得到一个样本（mc1.dat）。对其进行估计，得到：

$$\hat{y} = 75.7679 + 11.9683x_i$$

且 $SRR = \hat{\sigma} = 51.5857$。从单个样本得到的估计说明不了问题（即使比较接近真实值），需要通过生成多个样本才能得到重复抽样性质。MC 的目标就是重复抽样。假设我们得到 $M = 1000$ 个样本，就可以得到 1000 组系数和误差的估计值。对系数取均值，来验证是否无偏；同样，可以对系数的样本方差、误差方差的估计值进行验证。此外，还可以验证估计量是否服从正态分布。

3 区间估计和假设检验

最小二乘估计给出的是回归模型参数的点估计（point estimates），表示对回归函数 $E(y) = \beta_1 + \beta_2 x$ 的推断（inference）。这里" 推断 "的意思是，从已知或假定出发，经过推理得出结论。区间估计（interval estimation）和假设检验（hypothesis testing）是统计推断的两个重要工具。IE 和 HT 依赖假设 SR6。如果 SR6 不成立，则当样本容量足够大，可以得到近似结果。

3.1 区间估计

当 SR1-SR6 成立时，b_2 服从正态分布，对其进行标准化：

$$Z = \frac{b_2 - \beta_2}{\sqrt{\sigma^2 / \sum (x_i - \bar{x})^2}} \sim N(0, 1)$$

由正态分布表可知，$P(-1.96 \leq Z \leq 1.96) = 0.95$，从而有：

$$P\left(b_2 - 1.96\sqrt{\sigma^2 \Big/ \sum (x_i - \bar{x})^2} \leq \beta_2 \leq b_2 + 1.96\sqrt{\sigma^2 \Big/ \sum (x_i - \bar{x})^2}\right) = 0.95$$

上式定义了一个在 0.95 概率下包含参数 β_2 的区间。其中的两个端点给出了区间估计量（interval estimator）。在重复抽样中，这样构造的区间中有 95% 将会包含真实参数 β_2。这个估计量依赖于未知的误差方差，需要用其

估计量代替，但将 σ^2 换成 $\hat{\sigma}^2$ 后，上面的正态分布就变成了具有 2 个自由度的 t 分布：

$$t = \frac{b_2 - \beta_2}{\sqrt{\hat{\sigma}^2 / \sum (x_i - \bar{x})^2}} = \frac{b_2 - \beta_2}{\sqrt{\widehat{\mathrm{var}(b_2)}}} = \frac{b_2 - \beta_2}{\mathrm{se}(b_2)} \sim t_{(N-2)}$$

b_1 也是如此。一般意义下，对于满足假定 SR1-SR6 的简单线性回归模型，我们有：

$$t = \frac{b_k - \beta_k}{\mathrm{se}(b_k)} \sim t_{(N-2)}, \quad k = 1, 2$$

上式是简单线性回归模型进行区间估计和假设检验的基础。

t 分布的临界值 t_c 满足：$P(t \geq t_c) = P(t \leq -t_c) = \alpha/2$，其中 α 是概率，通常取 0.01 或 0.05。自由度为 m 的临界值 t_c 即分位数 $t_{(1-\alpha/2,m)}$。容易发现，$P(-t_c \leq t \leq t_c) = 1 - \alpha$。因此，区间估计可以如下构造：

$$p\left[-t_c \leq \frac{b_k - \beta_k}{\mathrm{se}(b_k)} \leq t_c\right] = 1 - \alpha$$

从而，

$$P[b_k - t_c \mathrm{se}(b_k) \leq \beta_k \leq b_k + t_c \mathrm{se}(b_k)] = 1 - \alpha$$

区间端点 $b_k \pm t_c \mathrm{se}(b_k)$ 是随机变量，它们定义了 β_k 的一个区间估计量。当基于某个样本得到 b_k 和 $\mathrm{se}(b_k)$，则得到 β_k 的一个区间估计值，称为 $100(1-\alpha)\%$ 区间估计值（interval estimate），也叫 $100(1-\alpha)\%$ 置信区间（confidence interval）。

3.2 假设检验

所谓假设检验，就是将我们对总体特征的猜测与数据样本信息作比较。给定一个经济和统计模型，我们设定关于经济行为的假设，再将这些假设用模型参数来表达。假设检验包括五个要素：

(1) 原假设（null hypothesis）H_0
(2) 备择假设（alternative hypothesis）H_1
(3) 检验统计量（test statistic）
(4) 拒绝域（rejection region）
(5) 结论（conclusion）

对于原假设 $\beta_k = c$，可能有三种备择假设：

- $H_1 : \beta_k > c$。这种形式在经济学中比较常用，因为经济理论往往给出变量之间的关系（符号），比如，食物支出例子中，经济理论认为食物是正常品（normal goods），即 $\beta_2 > 0$，因此，备择假设常设为 $H_1 : \beta_2 > 0$。
- $H_1 : \beta_k < c$。
- $H_1 : \beta_k \neq c$。

与原假设有关的样本信息包含在检验统计量的样本值中。基于统计量的值，我们决定是否拒绝原假设。检验统计量的一个特性是，当原假设成立时，检验统计量的分布是完全清楚的，而当原假设不成立时，它可能具有其他分布。对于我们现在的模型，检验统计量服从 t 分布。当 $H_0 : \beta_k = c$ 成立，我们可以用 c 替代 β_k，从而：

$$t = \frac{b_k - c}{\text{se}(b_k)} \sim t_{(N-2)}$$

拒绝域与备择假设的形式有关。构造拒绝域，除了检验统计量，还必须知道备择假设和显著性水平（level of significance）。拒绝域包含那些不太可能的值，也就是当原假设成立时，这些值出现的概率很小。其中的逻辑链可以表述为：”**如果检验统计量的值落入拒绝域，则该统计量不太可能服从假定的分布，从而原假设也不太可能为真**“。当备择假设成立时，检验统计量的值会不同寻常的大或不同寻常的小（这里的大小取决于显著性水平，一般取 0.01，0.05 或 0.10）。

如果原假设确实为真，而我们拒绝了它，称为第一类错误（Type I error）。犯第一类错误的概率就等于显著性水平，即 $P(\text{Type I error}) = \alpha$。只要拒绝原假设，就会不可避免犯第一类错误，当然如果这种错误成本比较高，可以通过减小 α 来控制。另一方面，如果原假设为假，而我们并未拒绝它，则称为第二类错误（Type II error）。在实际运用中，我们无法控制或计算犯第二类错误的概率，因为它依赖于未知参数 β_k。

假设检验完成后，我们可以给出结论。注意，不要用 “接受某假设 “的说法，只能说拒绝或者不拒绝原假设。无法拒绝原假设并不意味着原假设为真!

具体地，三种备择假设对应的拒绝域构造如下。

- 单侧检验，备择假设为 “>”：拒绝域为 $t \geq t_{(1-\alpha,N-2)}$
- 单侧检验，备择假设为 “<“：拒绝域为 $t \leq t_{(\alpha,N-2)}$
- 双侧检验，备择假设为 “$\neq$”：拒绝域为 $t \leq t_{(\alpha/2,N-2)}$ 或 $t \geq t_{(1-\alpha/2,N-2)}$

总结一下，假设检验的标准步骤为：

1. 设定原假设和备择假设
2. 设定统计量及其分布（当原假设成立时）
3. 选择一个显著性水平，决定拒绝域
4. 计算统计量的样本值
5. 给出结论

以食物支出为例。我们想知道到底 β_2 是不是与 0 有显著差异，即显著性检验。步骤如下：

1. $H_0 : \beta_2 = 0, H_1 : \beta > 0$
2. 统计量为前面的 t 统计量，本例中 $c = 0$，因此，$t = \frac{b}{\text{se}(b_2)} \sim t_{(N-2)}$
3. 取 $\alpha = 0.05$。右侧拒绝域的临界值为 95% 分位数：$t_{(0.95,38)} = 1.686$，因此拒绝域为 $t \geq 1.686$。
4. 计算统计量的样本值，由 $b_2 = 10.21, \text{se} = 2.09$，得到 $t = 4.88$
5. 由于 $t = 4.88 > 1.686$，所以我们拒绝原假设，也就是认为收入和食物支出之间存在统计上显著的正相关关系。

实践中，常常用 ***p*-value** 来表达检验结果。规则是，如果 p 值小于或等于显著性水平，则拒绝原假设。

4 预测、拟合优度和建模问题

预测（prediction）就是给定解释变量 x 的某个值，对未知的被解释变量 y 的值的预报。y 的未知值可能处于的范围称为预测区间（prediction interval）。考虑 y 的样本值与其预测值之间相关性，给出一个拟合优度的度量，称为 R^2。对样本中的每一个观测，预测值与实际值之间的差称为残差（residual）。基于残差构建的诊断测度用来检查回归分析中所用函数形式的正确性，并给出模型假设正确性的线索。

4.1 最小二乘预测

给定 SR-SR6，对某个 x_0，我们要预测相应的 y_0。假定 x_0 和 y_0 具有与模型相同的关系，即

$$y_0 = \beta_1 + \beta_2 x_0 + e_0$$

其中，e_0 为随机误差。假定 $E(y_0) = \beta_1 + beta_2 x_0$，$E(e_0) = 0$，$\text{var}(e_0) = \sigma^2$，$\text{cov}(e_0, e_i) = 0, i = 1, 2, \ldots, N$。由拟合回归线可得到 y_0 的最小二乘预测：

$$\hat{y}_0 = b_1 + b_2 x_0$$

因为 b_i 随机，因此 $\hat{y}_0$ 也是随机变量，定义预测误差为

$$f = y_0 - \hat{y}_0 = (\beta_1 + \beta_2 x_0 + e_0) - (b_1 + b_2 x_0)$$

我们期望预测误差比较小。容易发现，$E(f) = 0$，意味着 $\hat{y}_0$ 是 y_0 的无偏估计。而且，可以证明，当 SR1-SR5 成立时，$\hat{y}_0$ 是 y_0 的最佳线性无偏预测（BLUP）。预测误差的方差为：

$$\text{var}(f) = \sigma^2 \left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]$$

由上式可以考察影响预测误差方差的因素，有些因素与前面的估计模型类似，一个新的因子是 $(x_0 - \bar{x})^2$，度量 x_0 的离差，这意味着我们在具有较多样本信息的区域能作出更好的预测。实践中，我们用 $\hat{\sigma}^2$ 代替 σ^2，得到估计方差：

$$\widehat{\text{var}(f)} = \hat{\sigma}^2 \left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]$$

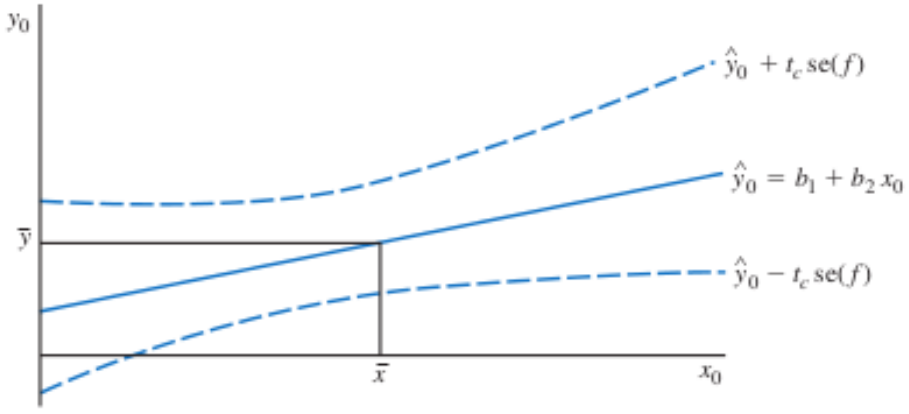
取平方根，得到预测标准误（standard error of the forecast）：

$$\text{se}(f) = \sqrt{\widehat{\text{var}(f)}}$$

从而，预测区间（prediction interval）为：

$$\hat{y}_0 \pm t_c \text{se}(f)$$

从 $\text{var}(f)$ 表达式可知，x_0 离均值 $\bar{x}$ 越远，预测可靠性越差，亦即预测区间越大。下图说明了不同 x_0 的点预测和区间预测之间的关系。



看食物支出的例子。当收入 $x_0 = 20$，可以预测食物支出为：

$$\hat{y}_0 = 83.4160 + 10.2096(20) = 287.6089$$

容易得到，预测区间为 $[104.1323, 471.0854]$。可见，预测区间很宽，表明此时的点预测并不可靠，即使 x_0 离均值还不算远。如果增加样本，可能会稍微改进估计和预测效果。但是，注意到这个例子中，误差方差和预测方差的估计很接近，表明预测的不确定性主要来自模型的不确定性。这并不奇怪，因为我们现在仅仅基于一个变量来预测本来非常复杂的家庭行为。## 拟合优度我们将 y_i 分解为可解释部分和不可解释部分：

$$y_i = E(y_i) + e_i$$

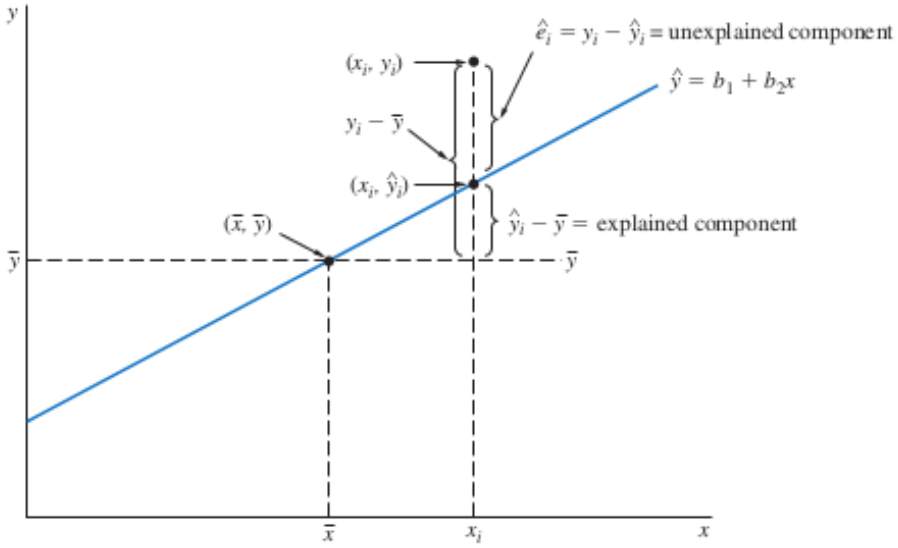
尽管这两个部分都是不可观测的，但因为系数可以估计，上式可以写成：

$$y_i = \hat{y}_i + \hat{e}_i$$

其中，$\hat{y}_i = b_1 + b_2 x$，$\hat{e}_i = y_i - \hat{y}_i$。注意到，最小二乘拟合回归线经过样本均值点。将上式两边减去 $\bar{y}$，可得：

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + \hat{e}_i$$

这表明，y_i 的离差由两部分组成，一个是由回归模型解释的部分 $\hat{y}_i - \bar{y}$，另一个是未解释部分 $\hat{e}_i$。



样本值的总体变异性由 y_i 的离差平方和给出，类似的，它可以分解为两个部分：

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum \hat{e}_i^2$$

上式用到了交叉项为 0 的事实（$\sum(\hat{y}_i - \bar{y})\hat{e}_i = 0$）。也就是说，离差平方和（总体变异）等于回归平方和（由模型解释）加上残差平方和（未由模型解释，由误差导致），一般写成：

$$SST = SSR + SSE$$

由此，我们定义判别系数，用来衡量 y 由 x 解释部分的比例：

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

简单线性回归模型中，R^2 和样本相关系数 r_{xy} 之间有两个有趣的关系：

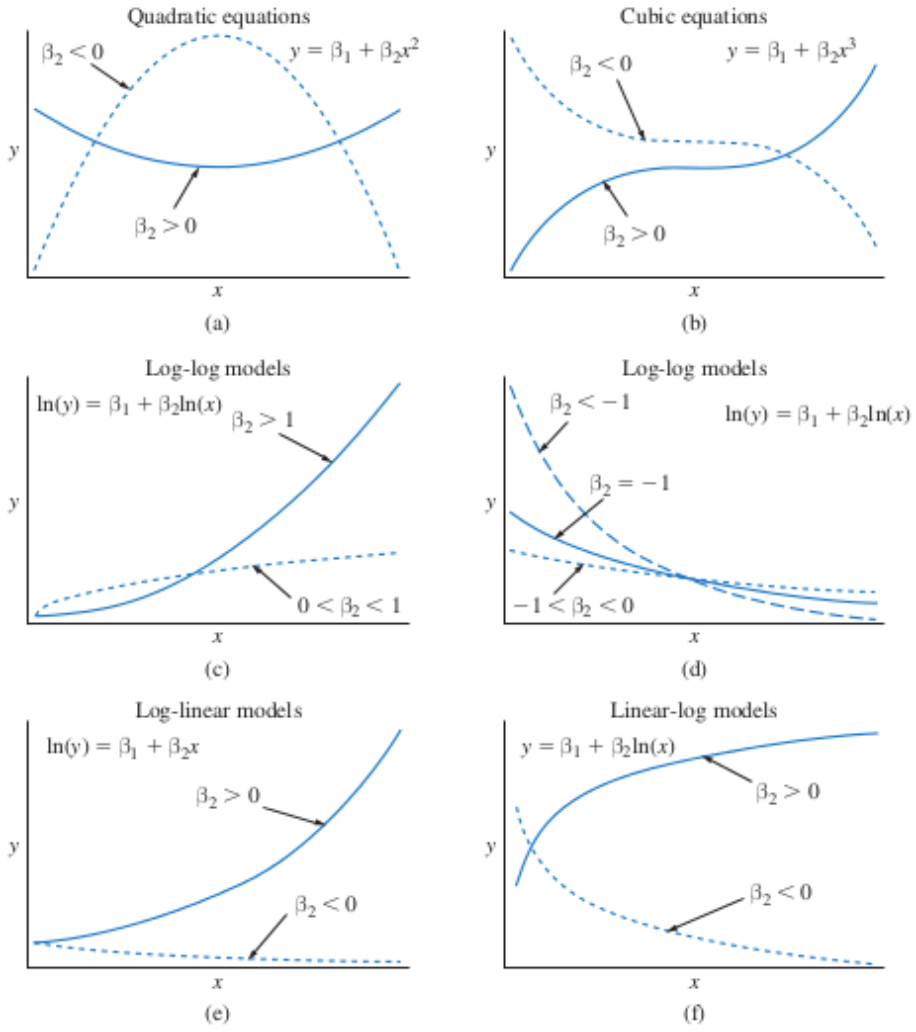
1. $r_{xy}^2 = R^2$
2. $R^2 = r_{y\hat{y}}^2$，因此，R^2 度量样本数据与其预测值之间的线性关系，这就是为什么 R^2 称为拟合优度（goodness-of-fit）。这个结果在多元线性回归模型中也成立。

4.2 Modelling Issues

数据缩放（scaling）不会改变变量之间的关系，但会影响对系数估计值的解释和某些 summary measures：

1. 缩放 x：
2. 缩放 y：
3. 同比例缩放：

建模中需要考虑函数形式选择。经济理论告诉我们，食物支出呈现边际效应递减的特征，即收入较大时，斜率较小。简单线性模型其实是很灵活的，可以通过变量变换描述非线性关系。前面已经提过二次模型和对数-线性模型。下面看其他包含幂、对数的形式。基于 quadratic、cubic 和对数三类变换，就可以刻画很多形式。



变量变换后，对模型结果的解释就需要改变。以包含对数的模型为例：

1. 双对数模型（log-log）：y 和 x 都必须大于 0，β_2 表示弹性。在上图 (c) 中，若 $\beta_2 > 1$，表示供给曲线；$0 < \beta_2 < 1$ 表示生产曲线。(d) 中，若 $\beta_2 < 0$ 表示需求曲线。log-log 模型中，弹性是常数，比较易于解释。
2. 对数-线性模型（log-linear）：若 $beta_2 > 0$，表示函数按递增速率递增；若 $\beta_2 < 0$，表示函数按递减速率递减。
3. 线性-对数模型（linear-log）：x 增加 1%，y 变化 $\beta_2/100$ 单位。

选择函数形式，一般遵循以下指导原则：

1. 符合经济理论对变量关系的描述；
2. 足够灵活，对数据拟合比较好；
3. 满足 SR1-SR6，确保最小二乘估计具有希望的性质。

在设定模型时，我们可能选择了不合适的函数形式，即使不是如此，也有可能模型假设不满足。有两个方法可用来探测这类错误。第一种方法是检查回归结果，比如错误的符号，某个理论上重要的变量不显著，等等。第二种方法是分析最小二乘残差，来发现同方差（SR3）、序列不相关（SR4）和正态性（SR6）假设不满足。异方差通常出现在横截面分析中，而序列相关出现在时间序列分析中。后面章节会给出正规的检验方法，现在先看如何通过残差图来诊断这些错误。

如果模型假设都满足，残差图应该呈现一种随机模式，并无明显的趋势或形态。若存在某种特定的模式，很可能某些假设未满足或出现了其他问题。如果残差出现明显的二次函数形式，表明线性模型设定有误，忽略了曲线关系。当残差出现正负值交替的模式时，表明误差项存在某种相关性（SR4 不满足）。由于残差图有时具有多种特征，往往不能据此得出确切的结论，难以分辨到底是模型设定问题，还是假设违背问题。尽管如此，残差图分析还是很常用。

如何检验模型的正态性？我们不能直接观测误差，只能从残差中分析。多数统计软件会给出残差的直方图，有的还给出关于残差的一些检验。当样本较小时，直方图并不是好的统计检验手段，因此，需要专门的统计量。常用的方法有 Jarque-Bera 检验，主要思想是看偏度是否显著不等于 0、峰度是否显著不等于 3，相应统计量为：

$$JB = \frac{N}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$$

当残差服从正态分布时，JB 统计量服从两个自由度的卡方分布。

4.3 多项式模型

看一个例子。数据集为 `wa_wheat.dat`。

4.4 Log-Linear 模型

4.5 Log-Log 模型

5 多元回归模型

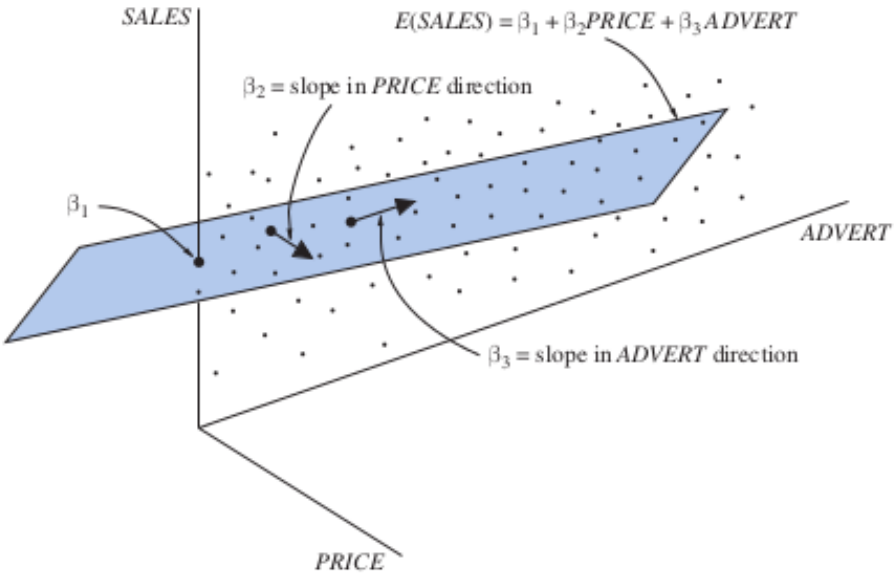
Big Andy 汉堡店需要评估价格策略和广告支出对销售的影响（数据集：andy.dat）。经济模型为：

$$SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT$$

由于城市大小对销售额影响很大，我们只选择小城市作为样本。$PRICE$ 可采用某个价格指数。相应的计量模型为：

$$SALES = E(SALES) + e = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + e$$

其中，系统部分是一个平面，称为回归平面。



一般的多元线性回归模型为：

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_K x_K + e$$

相应的，本例模型可以改写为：

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + e$$

我们作如下假设：

1. $E(e) = 0$
2. $\text{var}(e) = \sigma^2$
3. $\text{cov}(e_i, e_j) = 0$
4. $e \sim N(0, \sigma^2)$

从而，被解释变量具有性质：

1. $E(y) = \beta_1 + \beta_2 x_2 + \beta_3 x_3$
2. $\text{var}(y) = \text{var}(e) = \sigma^2$
3. $\text{cov}(y_i, y_j) = \text{cov}(e_i, e_j) = 0$
4. $y \sim N[(\beta_1 + \beta_2 x_2 + \beta_3 x_3), \sigma^2]$

对解释变量施加两个额外假设后，得到多元回归模型的 6 个假设：

ASSUMPTIONS OF THE MULTIPLE REGRESSION MODEL

MR1. $y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + e_i,\ i = 1, \ldots, N$

MR2. $E(y_i) = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} \Leftrightarrow E(e_i) = 0$

MR3. $\text{var}(y_i) = \text{var}(e_i) = \sigma^2$

MR4. $\text{cov}(y_i, y_j) = \text{cov}(e_i, e_j) = 0 \quad (i \neq j)$

MR5. The values of each x_{ik} are not random and are not exact linear functions of the other explanatory variables

MR6. $y_i \sim N[(\beta_1 + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}), \sigma^2] \Leftrightarrow e_i \sim N(0, \sigma^2)$

模型参数估计与简单线性回归类似，采用最小二乘法。对于汉堡店例子，可以得到：

$$b_1 = 118.91, \quad b_2 = -7.908, \quad b_3 = 1.863, \qquad R^2 = 0.448$$

误差方差估计为：

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{e}_i^2}{N - K}$$

其中 K 为系数个数。本例中，$\hat{\sigma}^2 = 23.874$，因此，回归标准误 $\hat{\sigma} = 4.8861$，也叫均方差（mean squared error，MSE）。有了标准误，就可以得到系数估计的抽样性质，比如方差和协方差等。类似简单线性回归，我们有 GAUSS-MARKOV 定理：对于满足 MR1-MR5 的多元回归模型，最小二乘估计是 BLUE。此外，正态性假设也与前面类似，此处大样本的情况，一般认为

$N-K=50$ 即可。容易计算系数估计的方差，比如对于 b_2，可以证明：

$$\text{var}(b_2)=\frac{\sigma^2}{(1-r_{23}^2)\sum_{i=1}^{N}(x_{i2}-\bar{x}_2)^2}$$

其中，r_{23} 为 x_2 和 x_3 的样本相关系数。可以发现，除了误差方差、样本大小、解释变量变异性之外，另一个影响估计方差的因素是 x_2 和 x_3 的样本相关系数。其中的机理是，当 x_2 的变动与其他解释变量无关联时，其变异性会增加估计的准确性；但当其变动与其他变量有关系，就很难区分这两种效应。后面关于" 共线性 "的讨论将表明，共线性将增大估计方差。

通常将系数估计的方差和协方差组合在一个协方差矩阵中：

$$\text{cov}(b_1,b_2,b_3)=\begin{bmatrix}\text{var}(b_1) & \text{cov}(b_1,b_2) & \text{cov}(b_1,b_2)\\ \text{cov}(b_2,b_1) & \text{var}(b_2) & \text{cov}(b_2,b_3)\\ \text{cov}(b_3,b_1) & \text{cov}(b_3,b_2) & \text{var}(b_3)\end{bmatrix}$$

R 中可以用 `vcov(fit_model)` 来获取该矩阵。为了进行区间估计和假设检验，需构造统计量，对于每个系数 b_k，可构造与前面类似的 t 统计量，只不过自由度变成 $N-K$。此时，系数的线性组合也服从 t 分布。

与简单线性回归类似，多元回归可以处理多项式模型（包括幂、交叉项），也可以包含对数。此外，也可以计算其拟合优度。具体例子参见 POE。

6 多元回归的进一步推断

本章主要讨论同时对多个变量进行显著性检验、受限最小二乘法以及多元回归的模型设定问题（重点讨论变量选择）。假设 MR1-MR6 成立。

6.1 联合假设检验

对联合假设的检验采用 F-test。比如：

$$SALES=\beta_1+\beta_2 PRICE+\beta_3 ADVERT+\beta_4 ADVERT^2+e$$

希望检验广告支出是否影响 SALES，需要同时对线性项和平方项进行检验，亦即：

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

$$H_1:\beta_3,\beta_4 0$$

当 H_0 成立时，上面的模型就变成受限模型（restricted model）：

$$SALES = \beta_1 + \beta_2 PRICE + e$$

F 检验基于对受限模型残差平方和（SSE_R）和非受限模型残差平方和（SSE_U）的比较。增加变量个数将减小误差平方和，即 $SSE_R - SSE_U \leq 0$。F 检验就是考察这种差异是不是显著。相应的统计量为：

$$F = \frac{(SSE_R - SSE_U)/J}{SSE_U/(N-K)}$$

其中，J 为限制数，N 为观测数目，K 为非受限模型中的系数个数。若 H_0 成立，该统计量服从 $F(J, N-K)$。当两个平方和之间差异较大时，F 较大，此时拒绝原假设。在上面的例子中，$F = 8.44$，相应的 p 值为 $P(F_{(2,71)} > 8.44) = 0.0005$，对于 0.01 或 0.05 的显著性水平，我们均应拒绝原假设，认为至少有一个系数不等于 0，也就是说，广告支出确实对销售有显著影响。

F 检验的一个重要应用是检验模型的整体显著性。对于模型：

$$y = \beta_1 + x_2\beta_2 + x_3\beta_3 + \cdots + x_K\beta_K + e$$

设定：$H_0 : \beta_2 = \beta_3 = \cdots = \beta_K = 0, H_1 : \beta_K 0$

当原假设成立，受限模型为：

$$y_i = \beta_1 + e_i$$

β_1 的 LS 估计为 $b_1^* = \bar{y}$，受限误差平方和为：

$$SSE_R = \sum_{i=1}^{N}(y_i - b_1^\star)^2 = \sum_{i=1}^{N}(y_i - \bar{y})^2 = SST$$

未受限误差平方和为 $SSE_U = SSE$，限制数为 $J = K - 1$，因此，F 统计量为：

$$F = \frac{(SST - SSE)/(K-1)}{SSE/(N-K)}$$

t 检验和 F 检验的关系：当原假设是一个单等式假设、备择假设为“不等于”时，两者是等价的（这源自 t 分布和 F 分布的关系：服从 df 自由度 t 分布的随机变量的平方，服从自由度为 $(1, df)$ 的 F 分布）。看销售例子中 PRICE 对 SALES 的影响，如果 $H_0 : \beta_2 = 0$，$H_1 : \beta_2 \neq 0$，则相应 t 统计量值为 $7.64/1.046 = 7.3044$，而 $F = 53.355$，正好等于 t^2。但是，这种关

系在单侧检验中并不成立，因为 F 检验并不适用于备择假设为 $>$ 或 $<$ 的情况。而且，这种关系对于多限制原假设也不成立。

F 检验可用于任何不超过 K 个线性等式假设的情况。虽然此时从 H_0 导出限制模型并不容易，但基本原理是不变的。比如，我们做如下检验：

$$H_0 : \beta_3 + 3.8\beta_4 = 1 \qquad H_1 : \beta_3 + 3.8\beta_4 \neq 1$$

当 H_0 成立时，$\beta_3 = 1 - 3.8\beta_4$，代入未受限模型，稍加整理，可得受限模型：

$$(SALES - ADVERT) = \beta_1 + \beta_2 PRICE + \beta_4(ADVERT^2 - 3.8ADVERT) + e$$

令 $y = SALES - ADVERT, x_2 = PRICE, x_3 = ADVERT^2 - 3.8ADVERT$，应用最小二乘法估计，可得受限误差平方和 $SSE_R = 1552.286$，未受限误差平方和与以前一样，$SSE_U = 1532.084$，限制数 $J = 1$，自由度 $N - K = 71$，易得 F 统计量的值为 0.936。对于 $\alpha = 0.05$，$F_c = 3.976$，因此，我们不能拒绝 H_0。换句话说，每月 1900 的广告支出是最优决策。（注意到，这里的原假设只有一个限制，因此也可以用 t 检验，容易计算 $t = 0.9676$，从而 $t^2 = F$。）进一步，如果我们要检验最优支出是否大于 1.9，需要设定如下假设：

$$H_0 : \beta_3 + 3.8\beta_4 \leq 1 \qquad H_1 : \beta_3 + 3.8\beta_4 > 1$$

因为假设中包含了不等号，F 检验不适用，只能用 t 检验。实践中，我们借助软件更方便，上述这些检验都属于一类称为 Wald tests 的检验。

6.2 非样本信息

看一个啤酒需求的例子。设定如下对数模型：

$$\ln(Q) = \beta_1 + \beta_2 \ln(PB) + \beta_3 \ln(PL) + \beta_4 \ln(PR) + \beta_5 \ln(I)$$

其中，Q 为需求量，PB 为 beer 的价格，PL 为 liquor 的价格，PR 为其他商品和服务的价格，I 为收入。采用对数的好处是，排除了负的价格。一个非样本信息是，当所有价格和收入同比例增加时，需求不变（经济主体不受“货币幻觉”影响）。为了描述这个特征，将所有自变量乘上一个常数 λ：

$$\begin{aligned}\ln(Q) &= \beta_1 + \beta_2 \ln(\lambda PB) + \beta_3 \ln(\lambda PL) + \beta_4 \ln(\lambda PR) + \beta_5 \ln(\lambda I) \\ &= \beta_1 + \beta_2 \ln(PB) + \beta_3 \ln(PL) + \beta_4 \ln(PR) + \beta_5 \ln(I) + (\beta_2 + \beta_3 + \beta_4 + \beta_5) \ln \lambda\end{aligned}$$

因此，上面的非样本信息，可以归结为一个特殊限制：$\beta_2+\beta_3+\beta_4+\beta_5=0$。为了得到满足这个限制的参数估计，我们从下面的回归模型开始：

$$\ln(Q)=\beta_1+\beta_2\ln(PB)+\beta_3\ln(PL)+\beta_4\ln(PR)+\beta_5\ln(I)+e$$

样本数据集为 beer.dat。为了引入非样本信息，我们从限制中解出某个参数，比如：

$$\beta_4=-\beta_2-\beta_3-\beta_5$$

代入原来的模型，得到：

$$\begin{aligned}\ln(Q)&=\beta_1+\beta_2\ln(\lambda PB)+\beta_3\ln(\lambda PL)+-\beta_2-\beta_3-\beta_5\ln(\lambda PR)+\beta_5\ln(\lambda I)+e\\&=\beta_1+\beta_2[\ln(PB)-\ln(PR)]+\beta_3[\ln(PL)-\ln(PR)]+\beta_5[\ln(I)-\ln(PR)]+e\\&=\beta_1+\beta_2\ln\left(\frac{PB}{PR}\right)+\beta_3\ln\left(\frac{PL}{PR}\right)+\beta_5\ln\left(\frac{I}{PR}\right)+e\end{aligned}$$

其中，我们消去了 β_4，且构造了几个新变量，得到一个受限模型。对其进行估计，可以得到：

$$b_1^\star=-4.789, b_2^\star=-1.2994, b_3^\star=0.1868, b_5^\star=0.9485$$

再由限制得到，$b_4^\star=0.1668$。

那么，这个受限最小二乘估计过程有什么性质呢？首先，它是有偏的，除非限制是完全正确的。这个结果对于计量经济学意义重大。好的经济学家将找到更可靠的估计，因为他们将引入更好的非样本信息。这点在模型设定方面也是一样。记住：好的经济理论是实证研究中的一个重要组成部分。其次，受限最小二乘估计的方差比通常的最小二乘估计小，无论限制是否成立。也就是说，引入非样本信息，会降低估计过程因随机抽样产生的变异性。注意：这里面有一个 bias-variance 权衡的问题。

6.3 模型设定

遗漏变量。无关变量。模型选择的准则。

6.4 数据不足（Poor Data）、共线性（Collinearity）和不显著（Insignificance）

6.5 预测

7 指示变量（Indicator Variables）

房产估价常用的一种方法是所谓的 hedonic model。