计量经济学

王飞

第1章 绪 论

1.1 什么是计量经济学

1.2 计量经济学的研究方法

1.3 计量模型的应用

1.4 数据类型

1.1 什么是计量经济学

一、计量经济学的定义

二、计量经济学的研究对象

计量经济学起源

经济学的一个分支学科

- 1926年挪威经济学家费瑞希（R.Frish）提出Econometrics
- 1930年成立世界计量经济学会
- 1933年创刊《Econometrica》
- 20世纪40、50年代的大发展和60年代的扩张
- 20世纪70年代以来现代计量经济学的发展

计量经济学与诺贝尔奖

- 1969年的首届诺贝尔经济学奖授予开创现代计量经济学的弗里希和丁伯根。
- 全部诺贝尔奖经济学奖中，共有5届颁给了计量经济学方面的基础性成果。
- 除了首届以外，1980的克莱因，1989的哈威尔莫，2000的赫克曼（Heckman）和麦克法登（McFadden），2003的恩格尔Engel和格兰杰Granger。

什么是计量经济学

“经验表明，统计学、经济理论和数学这三者对于真正了解现代经济生活的数量关系来说，都是必要的，但本身并非是充分条件。三者结合起来，就是力量，这种结合便构成了计量经济学”

“计量经济学可以定义为这样的社会科学：它把经济理论、数学和统计推断作为工具，应用于经济现象的分析”

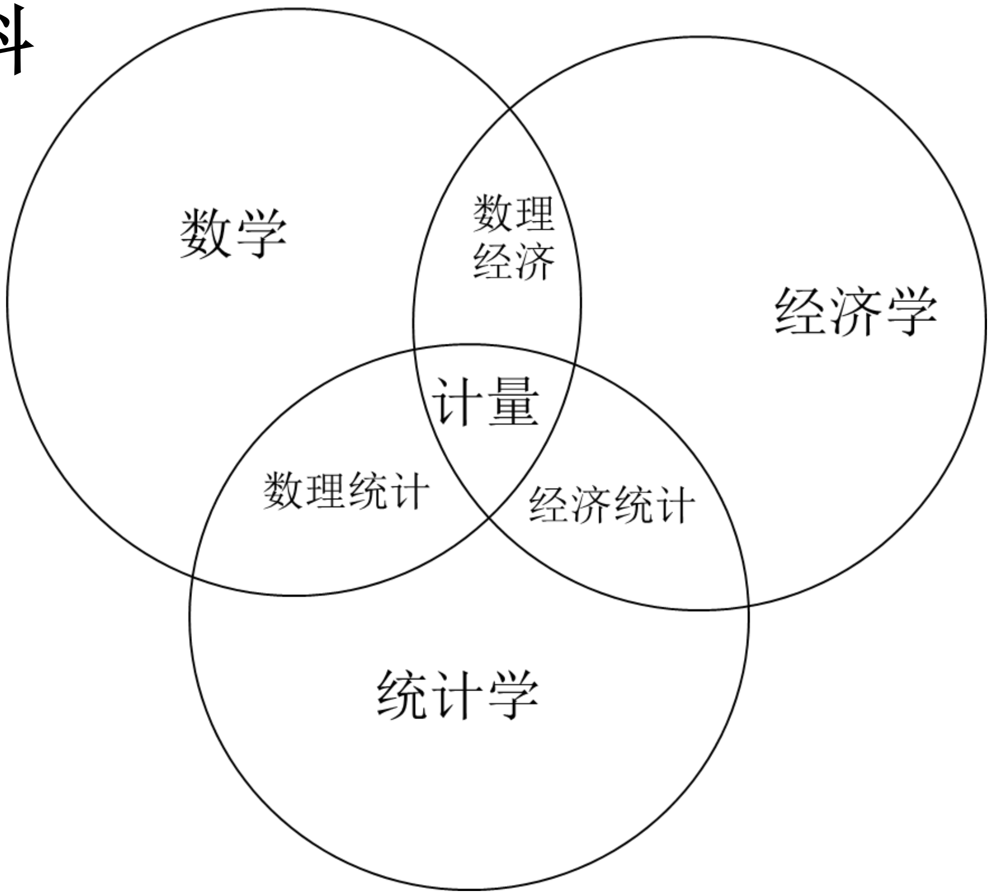
什么是计量经济学

“本质上，计量经济学的研究方法是，利用统计推断的理论和技术作为桥头堡，以达到经济理论和实际测算相衔接的目的。”

经济计量学是以经济理论为基础，以数学方法和统计方法为工具，利用计算机研究经济数量关系的一门学科。

计量经济学的研究对象

计量经济学是研究经济数量关系的一门学科



1. 计量经济学与经济理论

- 经济理论所做的假设或结论大多是定性的。
- 微观经济理论中，在其他条件不变的情况下，一种商品的价格下降会导致对该商品需求量的增加。即经济理论认为商品的价格和商品的需求量之间是一种负向关系。
- 经济理论并没有告诉我们价格和需求量之间任何数量上的联系。
- 研究这种数量上的关系属于计量经济学的任务。换言之，计量经济学对大多数的经济理论赋予经验内容。

2.计量经济学与数理经济学

- 数理经济学是用数学形式（方程式）来表述经济理论。
- 本质上仍属于定性研究，只是用数学符号来代替文字符号，同样没有考虑该理论的可度量性或在经验方面的可验证性。
- 计量经济学常常使用数理经济学家所提出的数学方程式，但是要把这些方程式改造成适合于经验检验的形式。要涉及到一些创造性和实际技巧。这些都属于计量经济学的研究内容。

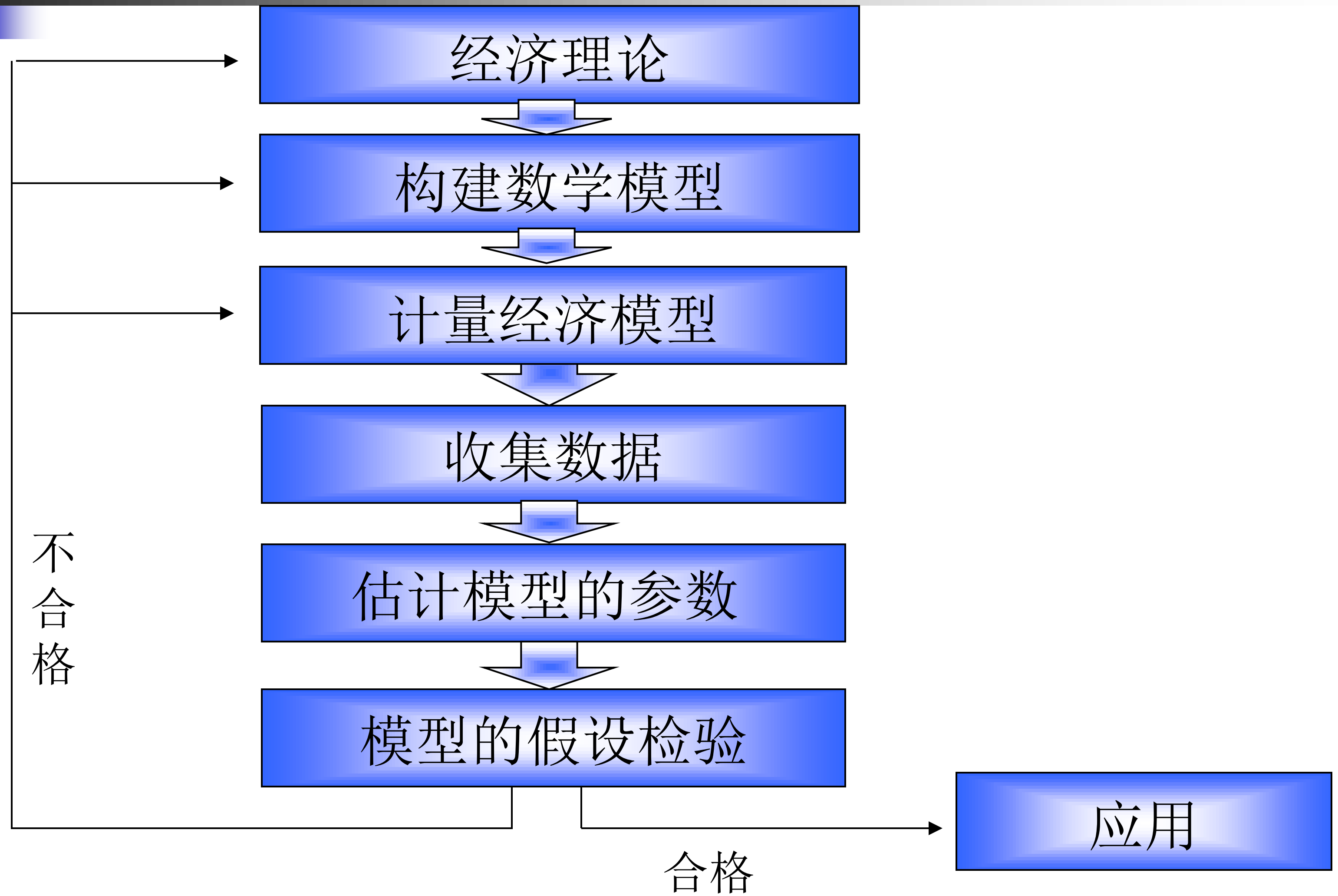
3.计量经济学与经济统计学

- 经济统计学主要是收集、加工并通过图、表等形式以展示、分析经济数据。如收集**GDP**、就业、价格、贸易等数据。
- 经济统计工作较少用这些收集来的数据去检验变量之间的关系，检验经济理论等。
- 这些任务属于计量经济学的工作。

4.计量经济学与数量统计学

- 数理统计学提供了计量分析的许多工具
- 但经济数据大多是非受控条件下的实验结果，计量经济学家常常需要有特殊的方法。
- 在经济社会中，收入、价格、投资、储蓄等数据并不是从实验室的来的，会产生了一些数理统计学家不常遇到的问题。
- 这些问题要求计量经济学家找到某些特殊的方法来处理

1.2 计量经济学的研究方法
计量分析的一般步骤



理论模型的建立

确定模型包含的变量

根据经济学理论和经济行为进行分析。

例如：从供给角度，根据生产要素建立生产函数

例如：根据区域输出基础理论，建立区域经济影响模型

例如：根据享乐价格理论，建立住房价格模型

计量模型的建立

(2) 确定模型的数学形式

利用经济学和数理经济学的成果

根据样本数据作出的变量关系图

选择可能的函数形式

(3) 确定计量模型

例如：

$$Quantity=\beta_0+\beta_1*income+\beta_2*price+\varepsilon$$

其中β_0、β_1和β_2是待估参数

模型参数的估计和检验

(1) 收集数据

(2) 选择模型参数估计方法

课堂教学结合**Eviews**

(3) 对模型和参数进行检验

模型是否合适

参数是否符合理论预期

1.3　计量模型的应用

一、估计变量之间的数量关系

二、经济预测

三、政策评价

四、理论检验与发展

估计变量之间的数量关系

- 解释经济变量之间的数量关系，如弹性、乘数等。
- 对企业而言，最优广告规模是多少
- 对地方政府而言，是否应该投资某项工程

经济预测和控制

- 计量经济学模型作为一类经济数学模型，是从用于经济预测，特别是短期预测而发展起来的
- 企业要预测销售量，合理安排生产
- 城市发展规模的预测
- 住房价格、股票价格的预测
- 为了要实现预定的经济增长目标，政府应当保证多大规模的政府投资

政策评价

- 一项经济政策是否实现了预期效果
- 对企业的**R&D**补贴政策是否促进了企业技术进步，有没有产生挤出效应
- 西部大开发政策是否实现了预期目标

理论检验与发展

- 企业聚集是否能够促进企业效率的提高
  - 企业聚集具有正的外部性
  - 拥挤效应
- 城市规模和环境污染之间的关系怎样的

1.4　数据类型

- 横截面数据
  - 全国30各省市GDP数据
- 时间序列数据
  - 北京市1990-2015年GDP数据
- 合并数据
- 面板数据

其他注意事项

- 参考书
  - 古扎拉蒂.计量经济学精要（第四版），机械工业出版社，**2012**年
  - 高铁梅.计量经济分析方法与建模：**EViews**应用及实例（第**2**版），清华大学出版社，**2009.**
- 软件
  - **Eviews SPSS**
  - **Stata SAS**

第2章 一元线性回归模型

2.1 回归的含义

2.2 回归模型

2.3 估计方法：最小二乘法

2.4 拟合优度

2.5 经典线性回归模型的基本假定

2.6 高斯马尔科夫定理

2.7 回归系数的假设检验

2.8 预测

2.1 回归的含义

回归分析(regression analysis)是研究一个变量（Y）与另一个（一些）变量（X）的依赖关系

比如儿子的身高和父亲的身高，消费和收入，收入和教育水平

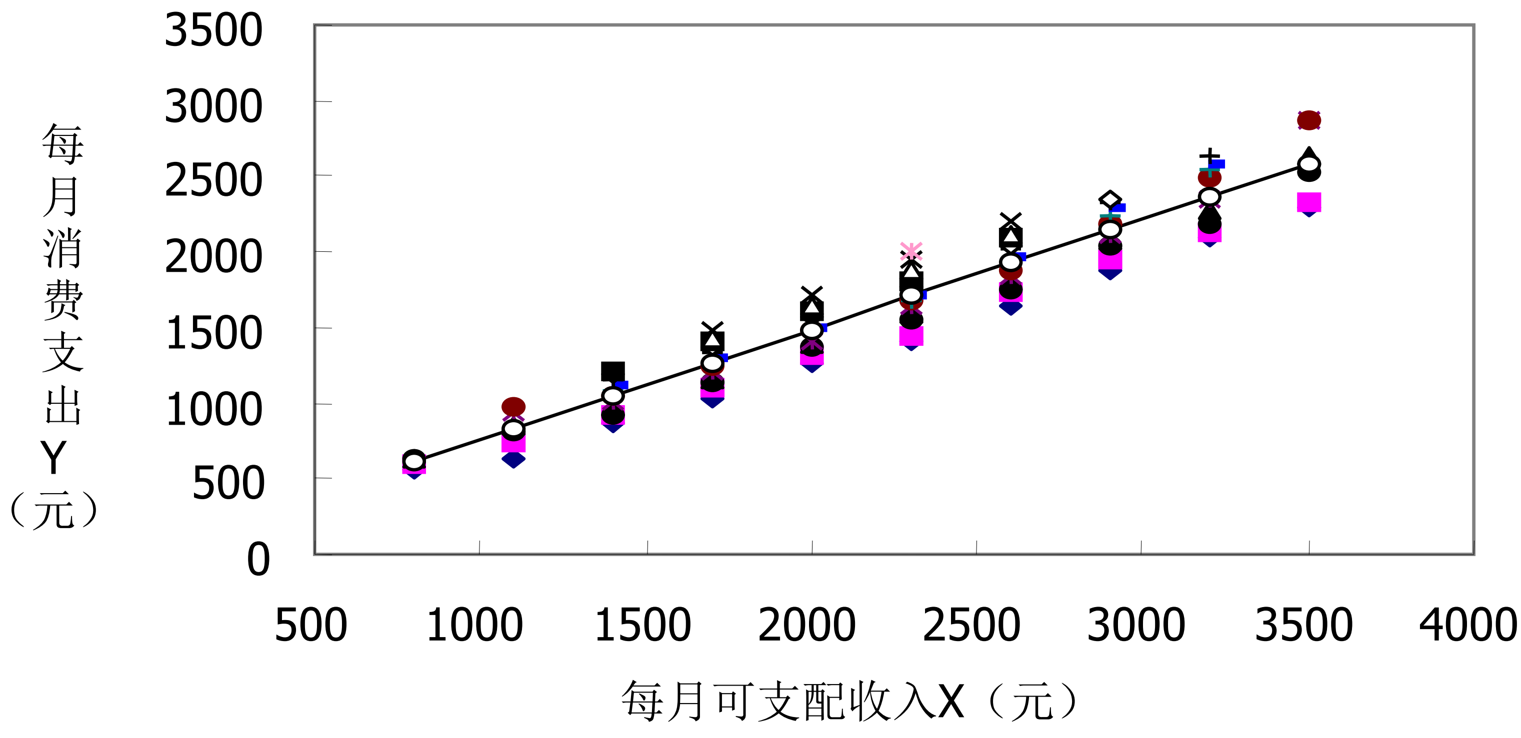
Y被称为**被解释变量（Explained Variable）或因变量（Dependent Variable），X被称为解释变量（Explanatory Variable）或自变量（Independent Variable）**

2.1 回归的含义

- **回归是一种统计关系**
  - **研究的是随机变量，父母所生的孩子的身高不同，同样收入的人消费支出不同**
- **回归与因果关系**
  - **回归分析是研究变量之间的依赖关系，但并非是因果关系。儿子的身高依赖于父母的身高，但其真正的原因并非是父母身高，而是DNA。因此，依赖关系并非真正意义上的因果关系，也可以是间接引起的原因。**

2.2 回归模型

一般而言，随着收入的增加，消费也在增加，



2.2 回归模型

总体回归函数（**PRF**）说明被解释变量**Y**的平均水平随解释变量**X**变化的规律。

例如，可以将居民消费支出看成是其可支配收入的线性函数时：

$$E(Y \mid X_i) = \beta_0 + \beta_1 X_i$$

其中，β_0，β_1是总体未知参数，称为**回归系数。**
其中β_0也称作截距项，β_1也称作斜率系数

2.2 回归模型

对某一个别的家庭，其消费支出可能与该平均水平有偏差。

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

u_i是观察值Y_i围绕它的期望值$E(Y|X_i)$的偏差，是一个不可观测的随机变量，又称为**随机干扰项**或**随机误差项**

2.2 回归模型

随机误差项的意义

1）被忽略的其他影响因素

数据的欠缺

节省的原则

未知的随机因素

2）观测值的观测误差

3）回归模型函数形式设定误差的影响

2.2 回归模型

- 总体回归模型 $Y_i = \beta_0 + \beta_1 X_i + u_i$ 包含两个部分，前一部分是确定性的部分，而后一部分是随机的、不确定的部分。
- 我们感兴趣的是其确定的部分，**Y**的均值的变化。
- 我们可以根据样本观测值估计一条直线

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

- 这就是样本回归方程（函数）

2.2 回归模型

- 样本回归方程 $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
- $\hat{\beta}_0$和$\hat{\beta}_1$分别是$\boldsymbol{\beta}_0$和$\boldsymbol{\beta}_1$的估计值
- $\hat{Y}_i$ 是对 $E(Y \mid X_i)$的估计，也称作$\boldsymbol{Y}$的拟合值或估计值
- 残差$e_i = Y_i - \hat{Y}_i$，是观测值和拟合值之差，也可以看做是对随机误差项$\boldsymbol{u}_{\mathbf{i}}$ 的估计

2.2 回归模型

- 线性的含义
  - 是指参数线性，即β_0和β_1是线性的
- 一元的含义
  - 指一个解释变量，如果有多个解释变量，就称为多元回归模型

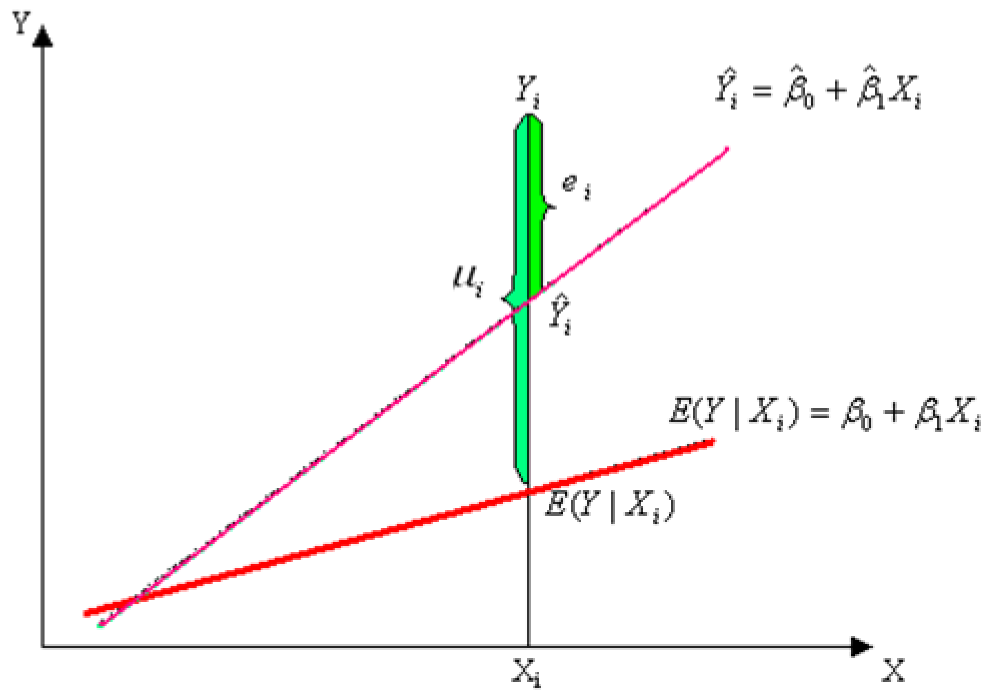
▼回归分析的主要目的：根据数据得到样本回归函数SRF，来估计总体回归函数PRF。

得到 $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

估计 $E(Y \mid X_i) = \beta_0 + \beta_1 X_i$

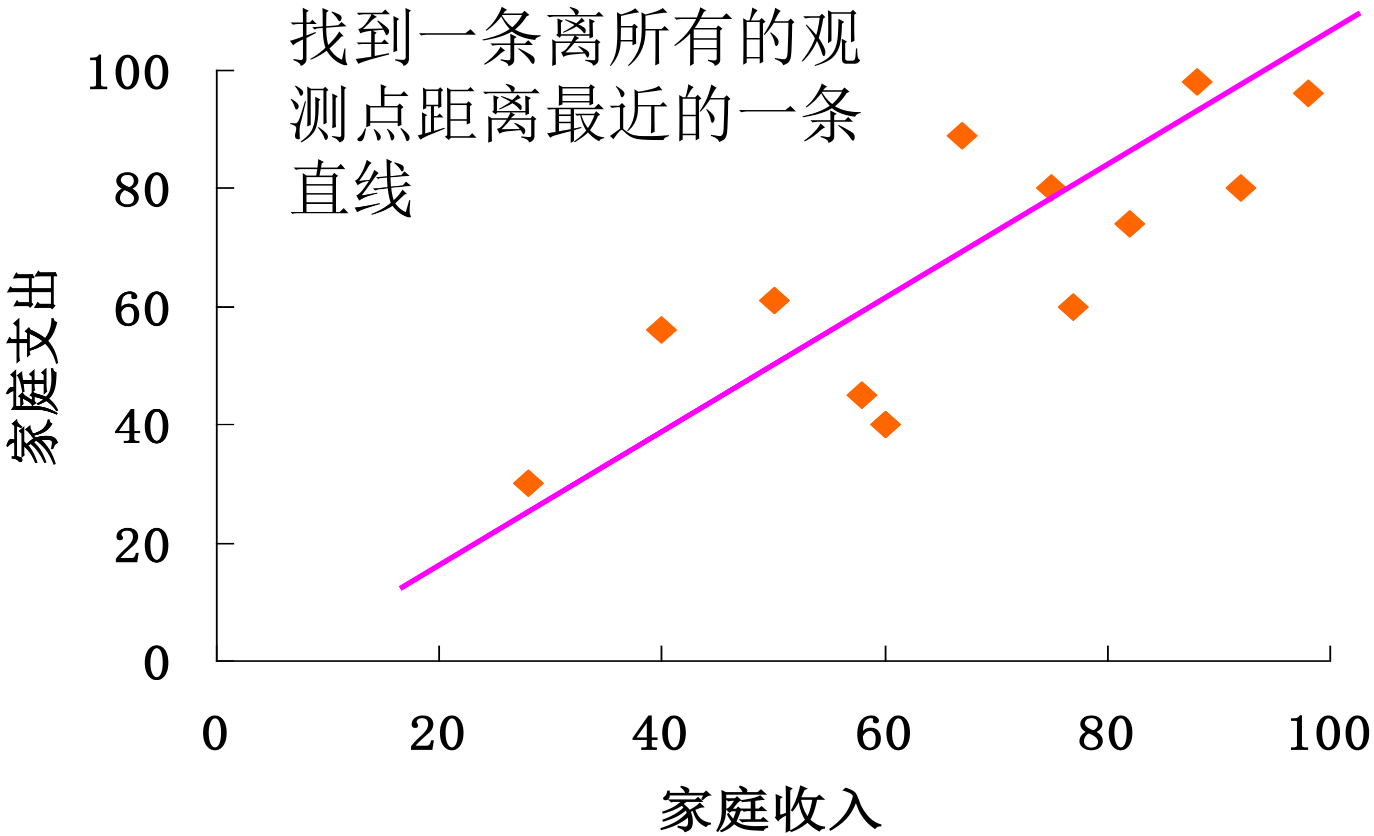
这就要求：

设计一“方法”构造SRF，以使SRF尽可能“接近”PRF，或者说使$\hat{\beta}_i (i = 0,1)$尽可能接近$\beta_i (i = 0,1)$。



注意：这里PRF可能永远无法知道。

2.3 估计方法：最小二乘法 OLS



2.3 估计方法：最小二乘法 OLS

PRF： $Y_i = \beta_0 + \beta_1 X_i + u_i \quad (i = 1,\ 2\cdots, n)$

SRF： $Y_i = \hat{\beta}_0 + \hat{\beta}_0 X_i + e_i = \hat{Y}_i + e_i$

$$e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

OLS:ordinary least square

求解：

$$\min \sum e_i^2 = \min \sum [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)]^2$$

2.3 估计方法：最小二乘法

$$
\text{由}\begin{cases}\dfrac{\partial\sum e_i^2}{\partial\hat{\beta}_0}=2\sum[Y_i-(\hat{\beta}_0+\hat{\beta}_1X_i)]\times(-1)=0\\[2ex]\dfrac{\partial\sum e_i^2}{\partial\hat{\beta}_1}=2\sum[Y_i-(\hat{\beta}_0+\hat{\beta}_1X_i)]\times(-X_i)=0\end{cases}
$$

整理得正规方程组：

$$
\sum Y_i=n\hat{\beta}_0+\hat{\beta}_1\sum X_i \qquad (1)
$$

$$
\sum Y_iX_i=\hat{\beta}_0\sum X_i+\hat{\beta}_1\sum X_i^2 \qquad (2)
$$

2.3 估计方法：最小二乘法

解上述正规方程组得到$\hat{\beta}_0$和$\hat{\beta}_1$的估计值

$$\hat{\beta}_1 = \frac{n\sum Y_i X_i - \sum Y_i \sum X_i}{n\sum X_i^2 - (\sum X_i)^2} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$\hat{\beta}_0 = \frac{1}{n}\sum Y_i - \hat{\beta}_1 \frac{1}{n}\sum X_i = \overline{Y} - \hat{\beta}_1 \overline{X}$$

其中$\overline{X}$和$\overline{Y}$是X和Y的样本均值

定义离差：$x_i = X_i - \overline{X}$；$y_i = Y_i - \overline{Y}$。

小写字母表示对均值的离差

2.3 估计方法：最小二乘法

- 用**OLS**方法估计得到的$\hat{\beta}$又称作**OLS**估计量
- 只要X取值有变异，就可以计算得到**OLS**估计量
- **OLS**估计量是点估计量
- 一旦从样本数据得到**OLS**估计值，就可画出样本回归线。

注意“帽子”的含义

$\hat{\beta}_1$是对β_1的估计；

$\hat{Y}_i$是对$E(Y \mid X_i)$的估计，

而不是对Y_i的估计；

e_i是残差，而u_i是总体随机扰动项。

参数估计举例

- 例2.1 某超市某种苹果的价格(元/千克)与每天销量在过去12天的记录见右表。若需求量仅取决于价格，并且是价格的线性函数。试用OLS估计该需求函数。
- Eviews软件操作

价格X(元/千克)	销售量Y(公斤)
10	55
9	70
8	90
7	100
7	90
7	105
7	80
6.5	110
6	125
6	115
5.5	130
5	130

样本回归方程的性质

性质1：残差e_i的均值等于0

即：$\sum e_i = 0$

证明：由正规方程式(1)

$$2\sum [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)] \times (-1) = 0$$

易知性质成立

样本回归方程的性质

性质2：样本回归方程通过Y和X的样本均值

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$$

证明：由正规方程式(1)可知

$$\sum Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum X_i$$

两边同时除以n得证

样本回归方程的性质

性质3：估计的Y的均值等于观测的Y的均值

即：$\overline{\hat{Y}}=\overline{Y}$

证明：$e_i=Y_i-\hat{Y}_i$

$$\sum e_i=\sum Y_i-\sum \hat{Y}_i$$

等式两边除以n可知$\overline{\hat{Y}}=\overline{Y}$

样本回归方程的性质

性质4： 残差e_i和X_i不相关,即$\sum X_i e_i = 0$

证明：由正规方程式（2）

$$2\sum[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)] \times (-X_i) = 0$$

可知$\sum X_i e_i = 0$

样本回归方程的性质

性质5： 残差e_i和估计的$Y_i(\hat{Y}_i)$值不相关,

即$\sum \hat{Y}_i e_i = 0$

证明：

$$\sum \hat{Y}_i e_i = \sum (\hat{\beta}_0 + \hat{\beta}_1 X_i) e_i = \hat{\beta}_0 \sum e_i + \hat{\beta}_1 \sum X_i e_i = 0$$

2.4 拟合优度：样本决定系数

样本回归线对数据拟合得有多好?

如果全部观测点都落在样本回归线上，则得到的是一个“完美”的拟合。

虽然可以计算得到**OLS**估计量，但并不能保证回归模型具有较高的拟合优度

怎样评价或衡量样本回归方程的拟合优度?

离差平方和的分解

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2$$

- **总平方和（TSS）**

  是观测的Y值围绕其均值的总变异。

$$\sum y_i^2 = \sum (Y_i - \bar{Y})^2$$

- **解释平方和（ESS）**

  是估计的Y值围绕其均值的变异。

$$\sum \hat{y}_i^2 = \sum (\hat{Y}_i - \bar{\hat{Y}})^2$$

- **残差平方和（RSS）**

  是未被解释的Y围绕其均值的变异。

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

2.4 拟合优度：样本决定系数R^2

$$R^2=\frac{ESS}{TSS}=\frac{\sum \hat{y}_i^2}{\sum y_i^2}=\frac{\sum (\hat{Y}_i-\bar{Y})^2}{\sum (Y_i-\bar{Y})^2}$$

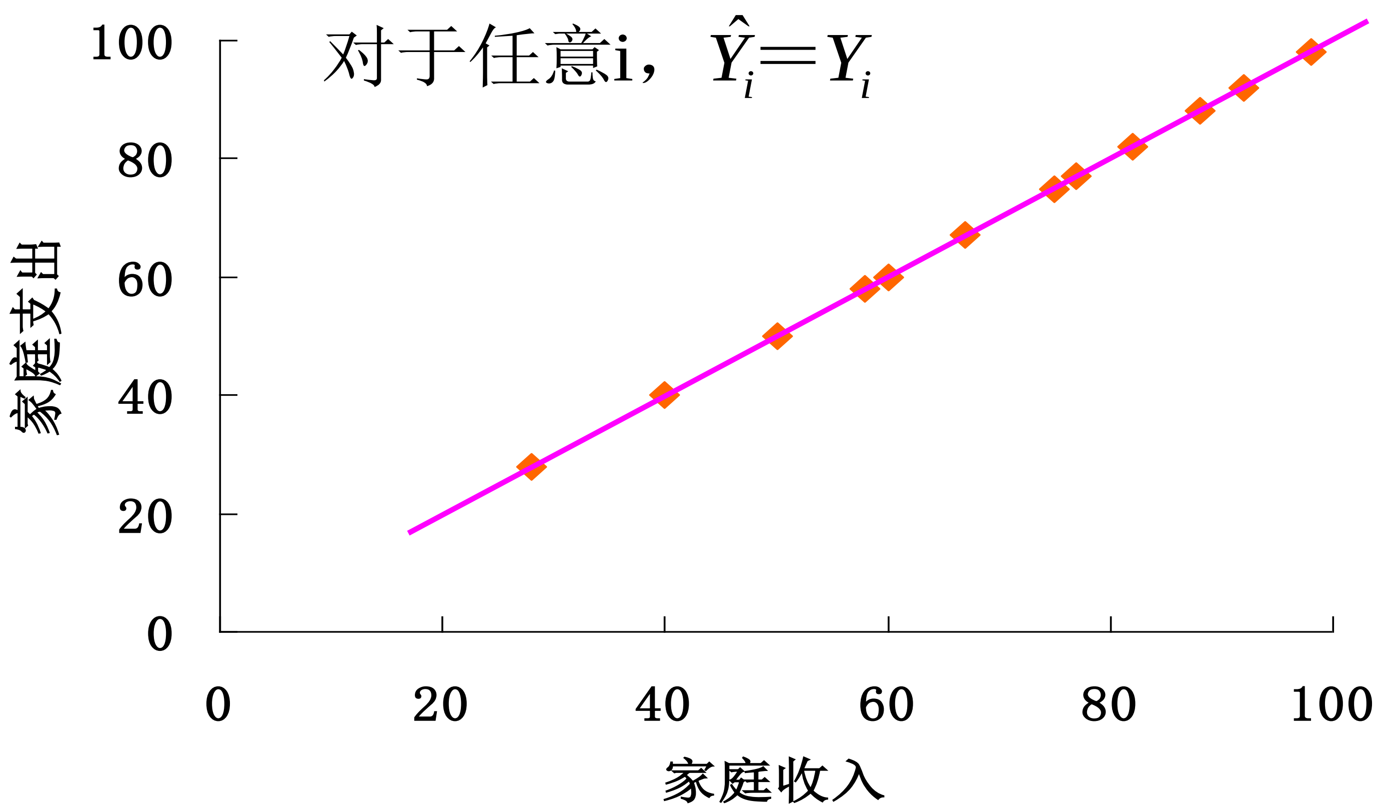
$$\text{或} R^2=1-\frac{RSS}{TSS}=1-\frac{\sum e_i^2}{\sum y_i^2}$$

性质：$0\le R^2\le 1$

问： $R^2=0$ 意味着什么？

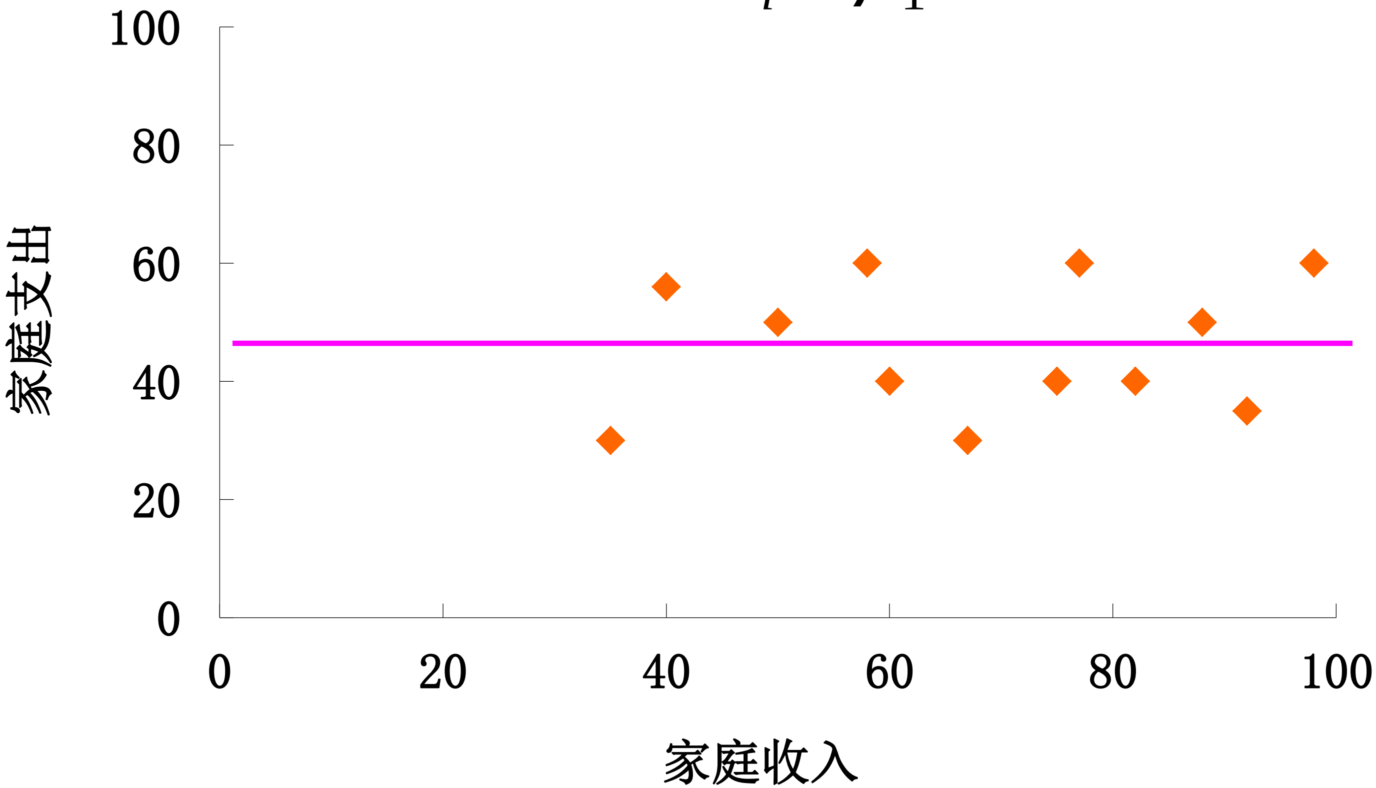
$R^2=1$ 意味着什么？

$R^2=1$



R^2 =0

对于任意i，$\hat{Y}_i = \hat{\beta}_1 = \bar{Y}$



2.4 拟合优度：样本决定系数

- 意义：

  拟合优度越大，自变量对应变量的解释程度越高，自变量引起的变动占总变动的百分比高。观察点在回归直线附近越密集。

  R^2 测度了在Y的总变异中，由回归模型解释的部分所占的比例。 R^2 越高，回归模型拟合的程度就越好。

- R^2又称作样本判定系数、可决系数

$R^2 = 0.86$

- $R^2 = 0.86$
- 表示约有应变量Y的变异的86％能由解释变量X来说明。
- 或者说，解释变量解释了应变量Y变异中的86％。
- 注意：不表示有86％的样本观测点落在了样本回归线上！！

R^2与相关系数r 的区别

样本相关系数r

可以由$r=\pm\sqrt{R^2}$计算

$$r=\frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}=\hat{\beta}_1\frac{\sqrt{\sum x_i^2}}{\sqrt{\sum y_i^2}}。$$

相关系数r 的一些性质

- 取值区间为[-1，+1]
- 用于描述两个变量X与Y之之间线性联系的密切程度
- r等于0，X与Y线性无关
- r大于0， X与Y正相关
- r小于0， X与Y负相关
- r的绝对值越大， X与Y的相关性越强
- 只表示线性关联程度，不表示因果联系

2.5 经典线性回归的基本假定

- 假定1 回归模型参数线性
- 意味着回归模型中自变量的选择是正确的，模型函数形式是正确的

- 假定2 X的取值要有变异
- 否则无法计算**OLS**估计量

2.5 经典线性回归的基本假定

- 假定3 $E(u|X)=0$(零条件均值)
- 意味着1）随机误差项u与X是不相关的，u_i不仅与X_i不相关，与X_j也不相关。
- 对于横截面数据，如果能够做到随机抽样，u_i与X_i不相关就可以保证与X_j也不相关
- 对于时间序列数据，意味着不能存在Y到X的反馈
- 2）$E(u)=0$，随机误差项u对Y的均值没有系统的影响

2.5 经典线性回归的基本假定

- 假定4．$var(u|X)=\sigma^2$ (同方差)
- 对于任意的X取值，u都具有相同的方差，或Y都具有相同的方差

- 假定5．$cov(u_i, u_j|X)=0$（无自相关）
- 随机误差项之间是不相关的

2.6 高斯马尔科夫定理

可以证明在假定1-5下，OLS估计量具有：

（1）线性，即可以表示为Y_i或u_i的线性函数

$$\hat{\beta}_1 = \sum k_i u_i$$

（2）无偏性，即估计量的期望等于总体参数

$$\mathrm{E}(\hat{\beta}_1) = \beta_1$$

（3）有效性，即估计量的方差在所有线性且无偏的估计量中具有最小方差。

$$\mathrm{var}(\hat{\beta}_1) \le \mathrm{var}(\hat{\beta}'_1)$$

因此，OLS估计量可以称作最优线性无偏估计量

BLUE（Best Linear Unbiased Estimator)

2.7 回归系数的假设检验

- 假定6．$u|X \sim N(0,\sigma^2)$ (正态分布假定）
- 对于任意的X取值，u都具有相同的分布，且服从正态分布
- 补充假定6后，我们能够知道回归系数OLS估计值的抽样分布，并进行假设检验
- 正态分布是自然界中常见的一种概率分布
- 随机误差项u包括许多随机变量。可以证明，当省略的随机变量个数较多时，u近似地服从正态分布。

2.7 回归系数的假设检验

- **OLS估计量的抽样分布**

1. $\hat{\beta}_1 \sim N(\beta_1, \dfrac{\sigma^2}{\sum x_i^2})$， $\hat{\beta}_0 \sim N(\beta_0, \dfrac{\sigma^2 \sum X_i^2}{n\sum x_i^2})$

2. $(n-2)\dfrac{\hat{\sigma}^2}{\sigma^2}$服从自由度为 n-2 的卡方分布，其中$\hat{\sigma}^2 = \dfrac{\sum e_i^2}{n-2}$

3. $\hat{\beta}_0$和$\hat{\beta}_1$的分布独立于$(n-2)\dfrac{\hat{\sigma}^2}{\sigma^2}$

4. $\hat{\beta}_0$和$\hat{\beta}_1$在整个无偏估计类中，都具有最小方差（BUE）

2.7 回归系数的假设检验

回归分析是要判断被解释变量Y是否依赖于解释变量X的变化而变化。这就需要对回归系数进行显著性检验。

显著性检验所应用的方法是数理统计学中的假设检验。

假设检验就是对总体参数或总体分布形式作出先验的判断（原假设），然后利用样本信息来判断原假设是否成立。

其原理是小概率不发生的原理。

假设检验中的小概率原理

- **1.** 在一次试验中，一个几乎不可能发生的事件称之为小概率事件。
- **2.** 假定在一次试验中小概率事件不会发生
- **3.**如果H_0成立，则小概率事件**A**不发生

假设检验中的小概率原理

- 如果小概率事件**A**发生了
- H_0不成立
- 如果小概率事件**A**未发生
- H_0成立?
- 小概率由研究者事先确定，通常为**5%**

假设检验的步骤

- 提出原假设和备择假设
- 抽取样本，确定适当的检验统计量
- 计算检验统计量的值
- 确定显著性水平和临界值
- 作出统计决策

假设检验（t检验）

步骤1：设置原假设和备则假设

$$H_0: \beta_1 = \beta_1^*$$

$$H_1: \beta_1 \neq \beta_1^*$$

步骤2：确定检验统计量：$t = \dfrac{\hat{\beta}_1 - \beta_1^*}{se(\hat{\beta}_1)}$

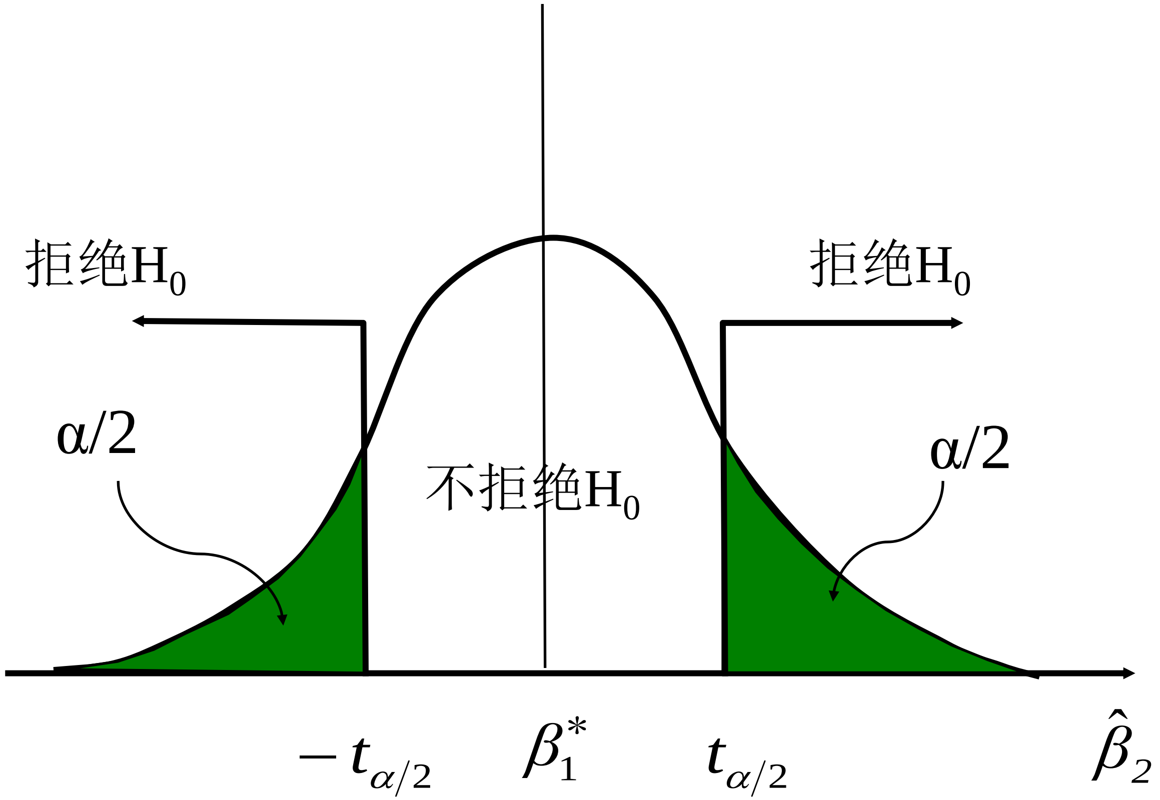
步骤3：计算检验统计量值：$t = \dfrac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\hat{\sigma}^2 / \sum x_i^2}}$

步骤4：确定显著性水平α，得到临界值$t_{\alpha/2}(n-2)$

步骤5：计算的t值是否落在接受域（$-t_{\alpha/2}$，$t_{\alpha/2}$）

如果是，不拒绝H_0；否则拒绝H_0

2.7 回归系数的假设检验



2.7 回归系数的假设检验

假设检验的P值方法

P值是拒绝原假设的最低的显著性水平，可以理解为拒绝原假设犯错误的概率。

当P值小于等于α时，表明在α显著性水平下，可以拒绝H_0

当P值大于α时，表明在α显著性水平下，不能拒绝H_0

Eveiws中的P值法

2.7 回归系数的假设检验

注意事项

上述假设检验过程依赖于一系列的假定。如果假定不成立，则假设检验会失效。

比如有时正态性假定明显不合适

但是**可以证明，大样本下，即使u不服从正态分布，上述t检验方法也同样适用**

2.8 预测

如果得到如下的样本回归方程

$$\hat{Y}_i = 84.4545 + 0.5091X_i$$

其中Y表示广告支出(万元)，X表示利润(万元)

“预测”广告支出100万元的未来利润Y

2.8 预测

1. 点预测

通过样本回归函数 $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1 X$，求得的拟合值为

$$\hat{Y}_0=84.4545+0.5091\times 100=135.3645$$

$\hat{Y}_0$是点预测值，也是条件均值**$E(Y|X=X_0)$**的无偏估计。

2.8 预测

2.区间预测

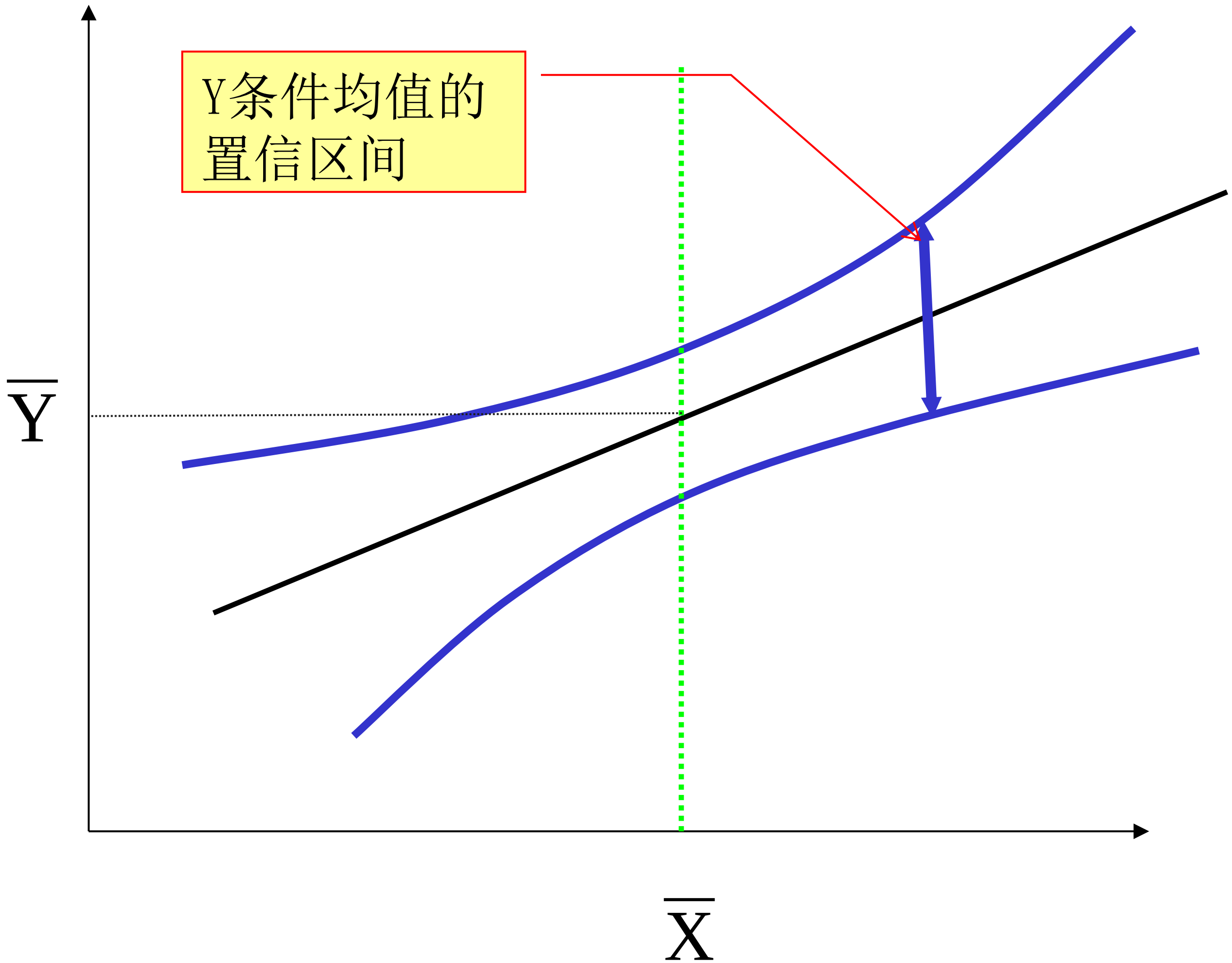
通常的预测区间是

$$(\hat{Y}_0 - 2se(\hat{Y}_0)，\hat{Y}_0 + 2se(\hat{Y}_0))$$

相当于95%的置信区间

用Eviews做预测

2.8 预测



第3章 多元线性回归模型

3.1 为什么用多元回归模型

3.2 多元线性回归模型的基本假定

3.3 估计方法：最小二乘法

3.4 拟合优度

3.5 回归系数的假设检验

3.6 什么时候增加新的解释变量

3.1 为什么用多元线性回归模型

- 现实中对**Y**的影响往往是几个因素共同作用的结果。
  - 产品的需求量不仅取决于产品价格，人们的收入，替代品的价格等等都是会有影响
- 有时我们关注的焦点仅仅是其中某个变量，比如收入的影响，我们可以只构建一元线性回归模型吗？

3.1 为什么用多元线性回归模型

- 偏回归系数

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

- β_i表示的是除去了其他变量$\mathbf{X_j}$的影响后，自变量$\mathbf{X_i}$的变化对因变量**Y**的单独的影响，即自变量$\mathbf{X_i}$对因变量**Y**的直接影响或是净影响

3.2 多元线性回归模型的基本假定

- 假定1. 回归模型参数线性
- 假定2. 诸X之间不存在完全线性相关
- 假定3. $E(u|X)=0$(零条件均值)
- 假定4. $var(u|X)=\sigma^2$ (同方差)
- 假定5. $cov(u_i, u_j|X)=0$（无自相关）
- 假定6. $u|X\sim N(0,\sigma 2)$ (正态分布）

3.2 多元线性回归模型的基本假定

- 假如 $X_{1i}=\lambda X_{2i}$，则变量之间存在完全的线性关系。
- 带入到总体回归模型中，可得

$$Y_i=\beta_0+(\beta_1\lambda+\beta_2)X_{2i}+\cdots+\beta_k X_{ki}+u_i$$

- 可见，我们能够得到的仅是 $(\beta_1\lambda+\beta_2)$ 的估计值，而无法得到$\boldsymbol{\beta_1}$和$\boldsymbol{\beta_2}$ 的估计。
- 现实中完全共线性的情况少见，更多的是两个变量之间不完全（或者说高度、近似）共线性。

3.3 估计方法：最小二乘法

- 多元线性回归模型使用最小二乘法可以得到系数的估计值
- **OLS估计量仍然具有BLUE的性质**

3.4 拟合优度

- Y的离差平方和的分解依然成立
- TSS=ESS+RSS
- $R^2 = \frac{ESS}{TSS}$
- 经验法则，多元时，
  $R^2 \geq 0.7$就比较满意了
- 而一元时，一般要求$R^2 \geq 0.8$

调整后的样本决定系数$\overline{R}^2$

- R^2取值会随着解释变量个数的增加而增加
- $\overline{R}^2 = 1 - \frac{RSS/n-k-1}{TSS/n-1}$
- $\overline{R}^2 < R^2$

3.5 回归系数的假设检验
单个回归系数

步骤1：设置原假设和备则假设

$$H_0:\ \beta_1 = \beta_1^*$$

$$H_1:\ \beta_1 \neq \beta_1^*$$

步骤2：确定检验统计量：$t = \dfrac{\hat{\beta}_1 - \beta_1^*}{se(\hat{\beta}_1)}$

步骤3：计算检验统计量值：$t = \dfrac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\hat{\sigma}^2 \Big/ \sum x_i^2 (1 - R_j^2)}}$

步骤4：确定显著性水平α，得到临界值$t_{\alpha/2}(n-2)$

步骤5：计算的t值是否落在接受域（$-t_{\alpha/2}$，$t_{\alpha/2}$）

如果是，不拒绝H_0；否则拒绝H_0

3.5 回归系数的假设检验
回归系数的线性约束

步骤1：设置原假设和备则假设

$$H_0:\ \beta_1+\beta_2=1$$

$$H_1:\ \beta_1+\beta_2\neq 1$$

步骤2：确定检验统计量：$F=\dfrac{(RSS_{约束}-\mathrm{RSS}_{无约束})/m}{RSS_{无约束}/(n-k-1)}$

步骤3：计算检验统计量值

步骤4：确定显著性水平α，得到临界值$F_\alpha(m,n-k-1)$

步骤5：计算的F值是否落小于F_α

如果是，不拒绝H_0；否则拒绝H_0

3.5 回归系数的假设检验
回归系数的线性约束

假设检验的P值方法

当P值小于等于α时，表明在α显著性水平下，可以拒绝H_0

当P值大于α时，表明在α显著性水平下，不能拒绝H_0

Eveiws中如何检验系数的线性约束

3.6 什么时候增加新的解释变量

- 在大多数的经验研究工作中，研究者对某个模型，是否值得加入某些因变量是没有足够的把握的。
- 根据节简的原则，不应引入一些对因变量缺乏解释力的自变量。
- 但又不应排除一些实质上影响较大的自变量。
- 怎样判断是否值得增加一个或多个解释变量呢？

3.6 什么时候增加新的解释变量

1.做F检验或t检验

只考虑增加一个解释变量时，可以直接使用对单个系数的t检验。$H_0: \beta_j = 0$

考虑增加两个及以上的解释变量，可使用上述的对回归系数约束的F检验

$H_0: \beta_j = \beta_{j+1} = \cdots = 0$

$$F = \frac{(RSS_{约束} - RSS_{无约束})/m}{RSS_{无约束}/(n-k-1)}$$

2.计算调整后的样本决定系数$\bar{R}^2$

当$\bar{R}^2$增加时，可以引入新的解释变量

第4章 回归模型的函数形式

4.1 双对数模型

4.2 半对数模型

4.3 多项式模型

4.4 其他函数形式

4.5 过原点回归模型

4.6 度量单位的影响

4.7 标准化回归系数

4.1 双对数模型

- **CD**生产函数$Y = AK^{\alpha}L^{\beta}$

$$\beta_1 = \frac{d\ln Y}{d\ln K} = \frac{dY/Y}{dK/K}$$

- 可以变换为

$$\ln Y_i = \beta_0 + \beta_1 \ln K_i + \beta_2 \ln L_i + u_i$$

- 变量非线性但是参数线性，可以使用**OLS**方法，**t**检验和**F**检验可以继续使用
- 回归系数**β**表示弹性的含义， **β_1=0.25**表示资本投入**K**增加**1%**，产出**Y**增加**0.25%**。

4.1 双对数模型

- 线性模型VS双对数模型
- 两种模型都可以表示供给（需求）曲线，但实践中应该选择那个模型呢？
- 很可惜经济理论仅仅表明变量之间的变化方向，而没有告诉我们变量之间的函数形式
- 只能从数据本身来决定模型的选择，即模型的选择是一个经验问题。

4.1 双对数模型

- 线性模型VS双对数模型
- 1)直接比较R^2是不可取的，因为两个模型的因变量不同。
- 2)可以将因变量转化成相同的，再比较拟合优度
- 3)要考虑到系数的符号，大小是否符合经济理论或经验
- 4)要考虑系数的统计显著性

4.2 半对数模型

- 因变量是对数形式

$$\ln Y_i = \beta_0 + \beta_1 X_i + u_i$$

- **Y**是收入，**X**是受教育年限
- 系数$\boldsymbol{\beta_1}$表示多增加一年教育，收入相对变动是多少
- $\boldsymbol{\beta_1 = \frac{d\ln Y}{dX} = \frac{dY/Y}{dX}}$

  表示$\boldsymbol{X}$变动一个单位时，$\boldsymbol{Y}$的相对变化

4.2 半对数模型

- 因变量是对数形式

$$\ln Y_i = \beta_0 + \beta_1 X_i + u_i$$

- **Y是FDI，X是GDP的增长率**
- $\boldsymbol{\beta_1 = \frac{dlnY}{dX} = \frac{dY/Y}{dX}}$表示**GDP**增长率每提高一个百分点，**FDI**的相对变动是多少
- 增长**1%**和增长率提高**1**个百分点的区别
- **β_1=0.25**表示**GDP**增长率每提高**1**个百分点，**FDI**增加**0.25%**

4.2 半对数模型

- 指数趋势模型

$$\ln Y_t = \beta_0 + \beta_1 t + u_t$$

- **Y是GDP，t取值是1,2,3……**
- 系数$\boldsymbol{\beta_1}$表示增长速度$\boldsymbol{\beta_1 = \frac{d\ln Y}{dt} = \frac{dY/Y}{dt}}$

4.2 半对数模型

- 自变量是对数形式

$$Y_i = \beta_0 + \beta_1 \ln X_i + u_i$$

- **Y是FDI的增长率，X是GDP**
- 系数**β_1**表示**GDP**每增长一个百分点，**FDI**的增长率变化几个百分点
- $\beta_1 = \frac{dY}{dlnX} = \frac{dY}{dX/X}$
- **β_1=0.25**表示**GDP**每增加**1%**，**FDI**的增长率提高了**0.25**个百分点。

4.3 多项式模型

最常用的多项式模型是二次函数模型

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$$

- **Y**表示某地区企业的平均生产成本，**X**表示聚集在某个地区的企业数量
- 注意$\boldsymbol{\beta_1}$的符号大小不能代表**X**对**Y**的影响
- $\frac{dY}{dX} = \beta_1 + 2\beta_2 X$
- 可以分析聚集经济和不经济的的临界点在哪里

4.4 其他函数形式

- 倒数模型

$$TC = \alpha + \beta Q \qquad AC = \beta + \alpha \frac{1}{Q}$$

- **S**型曲线模型（传染病模型）

$$Y = \frac{1}{\alpha + \beta e^{-t}}$$

4.5 过原点回归

- 通常的线性回归模型都带有截距项
- 如果E(u|X)=a
- $E(Y|X) = \beta_0 + a + \beta_1 X$
- 除非理论上回归模型不带截距项，一般情况下，都带上截距项
- 过原点的回归模型中R^2失效

4.6 度量单位的影响

假设Y表示FDI万美元，X_1表示GDP亿元，X_2表示GDP十亿元($X_1 = 10X_2$)，两个模型有何区别？

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + u_i \qquad Y_i = \beta_0 + \beta_1 X_{2i} + u_i$$

- 自变量单位变化时，截距项不发生变化，斜率系数会变化，如果X度量单位扩大10倍，斜率扩大10倍。
- 但变量之间的真实联系没有因此发生变化

$$Y_i = \alpha_0 + \alpha_1 10 X_{2i} + u_i \qquad 10\alpha_1 = \beta_1, \alpha_0 = \beta_0$$

4.6 度量单位的影响

假设**Y_1表示FDI万美元，Y_2表示FDI百万美元，X表示GDP亿元($Y_1 = 100Y_2$)**，两个模型有何区别？

$$Y_{1i} = \alpha_0 + \alpha_1 X_i + u_i \qquad Y_{2i} = \beta_0 + \beta_1 X_i + u_i$$

- 因变量单位变化时，截距项和斜率系数都会发生变化
- 但变量之间的真实联系没有因此发生变化

$$Y_{2i} = \alpha_0/100 + \alpha_1/100\, X_i + u_i/100$$

$$\alpha_1 = 100\beta_1, \alpha_0 = 100\beta_0$$

4.7 标准化回归系数

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

- 如果Y是房屋价格，X1是PM2.5指数，X2是距市中心距离，哪个因素对Y的影响更大。
- 不能直接比较回归系数，自变量单位不同
- 即使自变量单位相同，也不能直接比较，因为取值范围不同
- 将变量全部标准化后再回归，该回归系数就是标准化的回归系数，可以直接比较。

$$Z_i = \frac{X_i - \overline{X}}{s_X}$$

第5章 虚拟变量回归模型

5.1 什么是虚拟变量

5.2 两分的虚拟变量

5.3 多分的虚拟变量

5.4 多个定性变量

5.5 交互作用

5.6 差别斜率模型

5.7 政策评估

5.1 什么是虚拟变量

- 到目前为止，我们所研究的解释变量和被解释变量都是数值型变量，没有涉及定性的变量
  - 性别　受教育程度
- 能不能在回归方程中引入这样的定性变量呢？
- 在回归分析中，使用定量和定性变量是一样容易的
- 用数值来表示的定性变量在计量经济学中称之为虚拟变量（**dummy**）
- 性别变量D_i（**=1,**男性**;=0,**女性）

5.2 两分的虚拟变量

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + u_i$$

- **Y**表示收入，**X**表示工作年限，**D**表示性别（**=1,**男性**;=0,**女性）
- 回归系数**β**表示什么含义？

$$E(Y_i \mid D=0) = \beta_0 \qquad + \beta_2 X_i \quad \text{女性}$$

$$E(Y_i \mid D=1) = \beta_0 + \beta_1 + \beta_2 X_i \quad \text{男性}$$

- $\boldsymbol{\beta_1}$表示相同年龄下，男性与女性平均收入的差异
- $\boldsymbol{\beta_2}$表示相同性别下，年龄对平均收入的影响
- $\boldsymbol{\beta_1}$和$\boldsymbol{\beta_2}$表示的都是偏影响

5.2 两分的虚拟变量

- 可以继续使用**OLS**法估计参数
- 可以继续使用**t**检验和**F**检验
- 如果$\boldsymbol{\beta_1>0}$**,**且是统计显著的，表明相同年龄下，男性平均收入要高于女性平均收入

虚拟变量陷阱

- 我们用一个虚拟变量**D**表示分类两类的性别变量，能否用两个虚拟变量来表示

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 X_i + u_i$$

- 其中**D1(=1,男;=0,女);D2 (=1,女;=0,男)**

	Intercept	D_1	D_2	X
男性 Y_1	1	1	0	X_1
男性 Y_2	1	1	0	X_2
男性 Y_3	1	1	0	X_3
女性 Y_4	1	0	1	X_4
女性 Y_5	1	0	1	X_5

虚拟变量陷阱

- 我们用两个虚拟变量表示分为两类的性别变量，会导致解释变量之间的完全共线性，无法估计出参数
- 这种情况称作虚拟变量陷阱
- 引入虚拟变量的一般原则是
- 如果一个定性变量分为**m**个类，引入**m-1**个虚拟变量。

5.3 多分的虚拟变量

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 X_i + u_i$$

- 其中Y表示收入，X表示工作年限，D1(=1,本科;=0,其他);D2 (=1,研究生;=0,其他)

$$E(Y_i \mid D_1 = 0, D_2 = 0) = \beta_0 \qquad\qquad + \beta_3 X_i \quad \text{高中} \quad \text{对照组}$$

$$E(Y_i \mid D_1 = 1, D_2 = 0\ ) = \beta_0 + \beta_1 \qquad + \beta_3 X_i \quad \text{本科}$$

$$E(Y_i \mid D_1 = 0, D_2 = 1\ ) = \beta_0 + \quad + \beta_2 + \beta_3 X_i \quad \text{研究生}$$

β_1表示年龄不变情况下，本科毕业与高中毕业在平均收入的差距；**β_2**表示年龄不变情况下，研究生毕业与高中毕业在平均收入上的差距。

5.3 多分的虚拟变量

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 X_i + u_i$$

- **其中Y表示收入，X表示工作年限，D1(=1,高中;=0,其他);D2 (=1,研究生;=0,其他)**

$$E(Y_i \mid D_1 = 0, D_2 = 0) = \beta_0 \qquad + \beta_3 X_i \quad \text{本科} \quad \text{对照组}$$

$$E(Y_i \mid D_1 = 1, D_2 = 0\ ) = \beta_0 + \beta_1 \qquad + \beta_3 X_i \quad \text{高中}$$

$$E(Y_i \mid D_1 = 0, D_2 = 1\ ) = \beta_0 + \quad + \beta_2 + \beta_3 X_i \quad \text{研究生}$$

5.4 多个定性变量

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 X_i + u_i$$

- **其中Y表示收入，X表示工作年限，D1(=1,男性;=0,女性);D2 (=1,高学历;=0,低学历)**

$$E(Y_i \mid D_1 = 0, D_2 = 0) = \beta_0 + \beta_3 X_i \quad \text{女性低学历}$$

$$E(Y_i \mid D_1 = 1, D_2 = 0) = \beta_0 + \beta_1 + \beta_3 X_i \quad \text{男性低学历}$$

$$E(Y_i \mid D_1 = 0, D_2 = 1) = \beta_0 + \beta_2 + \beta_3 X_i \quad \text{女性高学历}$$

$$E(Y_i \mid D_1 = 1, D_2 = 1) = \beta_0 + \beta_1 + \beta_2 + \beta_3 X_i \quad \text{男性高学历}$$

对照组？

5.5 交互作用

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{1i} D_{2i} + \beta_4 X_i + u_i$$

- **其中Y表示收入，X表示工作年限，D1(=1,男性;=0,女性);D2 (=1,高学历;=0,低学历)**

$$E(Y_i \mid D_1 = 0, D_2 = 0) = \beta_0 + \beta_3 X_i \quad \text{女性低学历}$$

$$E(Y_i \mid D_1 = 1, D_2 = 0) = \beta_0 + \beta_1 + \beta_3 X_i \quad \text{男性低学历}$$

$$E(Y_i \mid D_1 = 0, D_2 = 1) = \beta_0 + \quad + \beta_2 + \quad \beta_3 X_i \quad \text{女性高学历}$$

$$E(Y_i \mid D_1 = 1, D_2 = 1) = \beta_0 + \beta_1 + \beta_2 + \beta_4 + \beta_3 X_i \quad \text{男性高学历}$$

β_4表示性别与学历两个变量产生的交互作用

5.6 差别斜率模型

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 X_i + \beta_3 D_i X_i + u_i$$

- 其中**Y**表示收入，**X**表示工作年限，**D(=1,**高学历**;=0,**低学历**)**

$$E(Y_i \mid D=0) = \beta_0 \qquad + \beta_2 X_i \qquad \text{低学历}$$

$$E(Y_i \mid D=1) = \beta_0 + \beta_1 + (\beta_2 + \beta_3) X_i \quad \text{高学历}$$

- $\boldsymbol{\beta_1}$表示差别截距，$\boldsymbol{\beta_3}$表示差别斜率
- 如何检验高学历与低学历收入函数存在差异？
- $\mathbf{H_0:\beta_1=\beta_3=0}$

5.7 政策评估

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

- 其中**Y**表示企业的废品率，**D**表示是否申请了在职培训津贴计划**(=1,**是**)**。**X2**和**x3**分布表示销售额和雇佣人数（都是企业规模的一种度量，企业规模与生产效率存在一定的联系）
- 政府的在职培训津贴计划是否会提高企业员工的生产效率？
- 申请了在职培训津贴的企业成为实验组，而未申请的则构成了对照组，$\boldsymbol{\beta_1}$统计显著就意味着实验组与对照组之间的差异是显著的。

第6章 模型选择

6.1 建模的准则

6.2 遗漏重要变量

6.3 增加无关变量

6.1 建模的准则

- **1．简洁性**
- 建立模型的目的是为了更好地理解这个世界，理解经济变量间复杂的关系，所以并非模型越复杂越好
- 一个好的模型用尽可能少的解释变量来理解因变量。把其他不重要的变量放到随机误差项中。

6.1 建模的准则

- **2. 识别性**
- 一个模型应该是可以识别的
- 所谓识别是指参数具有唯一的估计值
- 当解释变量之间存在完全线性关系时，参数的估计值是不确定的，这就意味着模型是无法识别的。
- 只有价格和销售量两个变量是无法识别需求函数的，因为供给函数和需求函数都涉及这两个变量

6.1 建模的准则

- **3．拟和优度**
- 建立模型的目的用用解释变量来理解被解释量的变化，所以一个模型应该具有较高的解释力，即较高的拟和优度。
- 但不应过分追求高的拟和优度，而忽视了其他方面，比如系数的估计值是否具有预期的符号，是否具有合理的经济意义。
- 其他条件相同时，一个高的拟和优度总是受到欢迎的。

6.1 建模的准则

- **4．理论的一致性**
- 如果一个或多个回归系数的估计值和理论预期的符号相反，即使模型有较高的拟和优度，也不能表明这个模型是好的。
- 比如，价格弹性是正的，价格越高，需求量越大，这种反常的现象需要我们谨慎地对待模型。

6.1 建模的准则

- **5．预测能力**
- 如果一个模型比较准确地反映了经济变量之间的关系的话，就应该能准确预测。
- 较高的拟和优度并不意味着高的预测功效。只表示对现有数据的拟和程度，表示的是对样本数据之内的预测功效。而我们常常需要样本期以外的预测功效。比如未来几年内**GDP**的增长率等等？

6.2 遗漏重要变量

- 如果遗漏变量和模型中的任一变量相关，则一般来说，模型中所有变量前回归系数**OLS**估计量都是有偏的。
- 系数的估计值是不一致的。即随着样本容量的增加，偏差也不会消失。
- 系数估计值的方差是有偏的。即使遗漏变量与模型中变量不相关
- **t**检验和**F**检验是不可靠的。
- 因此，如果经济理论表明一个变量受到其他某些变量的影响，就不应随意地在回归模型中忽略重要的变量。

6.3 增加无关变量

- 系数的**OLS**估计量是无偏的，也是一致的。
- 系数的**OLS**估计量的方差要大于真实的方差。估计值波动大，符号可能不符合理论预期。
- **t**检验中更容易接受原假设
- 因此，也不应随意地在回归模型中增加无关变量

作业1

1. GDP的定义与核算方法
2. 直接消耗系数计算及含义
3. MRIO的基本思路
4. 影响力系数与感应度系数的计算及含义

作业2

1. 构建计量模型有哪些准则？

2. 经典多元线性回归模型的基本假定是什么？

3. 根据数据估计得到某地区**CD**生产函数 $\widehat{lnY}_i = \hat{\beta}_0 + \hat{\beta}_1 K_i + \hat{\beta}_2 L_i$，请叙述如何检验生产函数是否是规模报酬不变的$(\beta_1 + \beta_2 = 1)$。

4. 对于如下回归模型，下标**i**表示城市，**Y**表示劳动生产率，**X**表示城市人口密度**(**衡量聚集程度**)**，**D(=1,**东部地区**;=0,**中西部地区），解释所有斜率系数的经济含义

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 X_i + \beta_3 D_i X_i + u_i$$