第05章 相关和回归分析

# 5.1 变量间关系的度量

## 变量间的关系类型

### 函数关系

* **定义**：两个变量之间存在一一对应的确定关系
* **数学表达**：$Y=f\left(X\right)$，其中$X$为自变量，$Y$为因变量
* **特点**：
	+ 变量间关系是确定的
	+ 从几何学角度看，数据集各观测点会落在一条曲线上
* **示例**：
	+ 商品销售额$Y$与销售量$X$的关系：$Y\_{i}=P\_{i}⋅X\_{i}$
	+ 圆的面积$S$与半径$R$的关系：$S=πR^{2}$

### 相关关系

* **定义**：变量之间存在不确定的依存关系
* **特点**：
	+ 变量间关系是不确定的
	+ 一个变量的取值不能由另一个变量唯一确定
* **示例**：
	+ 父亲身高与子女身高的关系
	+ 收入水平与受教育程度的关系
	+ 粮食产量与施肥量、降雨量、温度的关系

## 相关关系的描述与测度

### 相关分析的基本问题

1. 变量之间是否存在关系？
2. 如果存在关系，它们之间是什么样的关系？
3. 变量之间的关系强度如何？
4. 样本所反映的变量之间的关系能否代表总体变量之间的关系？

### 相关分析的基本假定

* 两个变量之间是线性关系
* 两个变量都是随机变量

### 相关系数

#### 定义与性质

* **定义**：度量变量之间关系强度的统计量
* **记号**：
	+ 总体相关系数：$ρ$
	+ 样本相关系数：$r$
* **性质**：
	1. 取值范围：$r\in \left[−1,1\right]$
	2. 对称性：$r\_{XY}=r\_{YX}$
	3. 与原点及尺度无关
	4. 仅度量线性关系
	5. 不意味着因果关系

#### 计算公式

1. **大FF公式**：

$$\begin{matrix}r=\frac{n∑X\_{i}Y\_{i}−∑X\_{i}∑Y\_{i}}{\sqrt{n∑X\_{i}^{2}−\left(∑X\_{i}\right)^{2}}⋅\sqrt{n∑Y\_{i}^{2}−\left(∑Y\_{i}\right)^{2}}}\end{matrix}$$

1. **小ff公式**：

$$\begin{matrix}r=\frac{∑\left(\left(X\_{i}−\overline{X}\right)\left(Y\_{i}−\overline{Y}\right)\right)}{\sqrt{∑\left(X\_{i}−\overline{X}\right)^{2}∑\left(Y\_{i}−\overline{Y}\right)^{2}}}=\frac{SS\_{XY}}{\sqrt{SS\_{XX}}\sqrt{SS\_{YY}}}=\frac{∑x\_{i}y\_{i}}{\sqrt{∑x\_{i}^{2}∑y\_{i}^{2}}}\end{matrix}$$

其中：

$$\begin{matrix}SS\_{XX}&=\sum\_{i=1}^{n}\left(X\_{i}−\overline{X}\right)^{2}\\SS\_{YY}&=\sum\_{i=1}^{n}\left(Y\_{i}−\overline{Y}\right)^{2}\\SS\_{XY}&=\sum\_{i=1}^{n}\left(X\_{i}−\overline{X}\right)\left(Y\_{i}−\overline{Y}\right)\end{matrix}$$

#### 相关系数的经验解释

* 高度相关：$\left|r\right|\geq 0.8$
* 中度相关：$0.5\leq \left|r\right|<0.8$
* 低度相关：$0.3\leq \left|r\right|<0.5$
* 极弱相关：$\left|r\right|<0.3$

### 偏相关系数

* **定义**：在控制其他变量影响的情况下，两个变量之间的相关系数
* **计算公式**：
	+ 保持$X\_{3i}$不变，$Y\_{i}$和$X\_{2i}$之间的相关系数：

$$\begin{matrix}r\_{12⋅3}=\frac{r\_{12}−r\_{13}r\_{23}}{\sqrt{\left(1−r\_{13}^{2}\right)\left(1−r\_{23}^{2}\right)}}\end{matrix}$$

* 保持$X\_{2i}$不变，$Y\_{i}$和$X\_{3i}$之间的相关系数：

$$\begin{matrix}r\_{13.2}=\frac{r\_{13}−r\_{12}r\_{23}}{\sqrt{\left(1−r\_{12}^{2}\right)\left(1−r\_{23}^{2}\right)}}\end{matrix}$$

## 相关系数的显著性检验

### 检验步骤

1. **提出假设**：
	* $H\_{0}:ρ=0$
	* $H\_{1}:ρ\ne 0$
2. **计算样本统计量**：

$$\begin{matrix}T^{\*}=\left|r\right|\sqrt{\frac{n−2}{1−r^{2}}} ∼t\left(n−2\right)\end{matrix}$$

1. **确定临界值**：
	* 给定显著性水平$α$
	* 查t分布表得$t\_{1−α/2}\left(n−2\right)$
2. **做出决策**：
	* 若$T^{\*}>t\_{1−α/2}\left(n−2\right)$，拒绝$H\_{0}$
	* 若$T^{\*}\leq t\_{1−α/2}\left(n−2\right)$，不拒绝$H\_{0}$

### 注意事项

1. 相关系数的显著性检验必须建立在样本数据的基础上
2. 相关系数的大小并不一定代表关系的实际重要性
3. 相关系数只能反映线性关系，不能反映非线性关系
4. 相关系数不意味着因果关系

# 5.2 回归分析的基本思想

## 相关关系与因果关系

### 相关关系的类型

* **边际相关**：两个变量之间的直接相关关系
* **条件相关**：在控制其他变量影响下的相关关系
* **虚假相关**：看似相关但实际无因果关系的现象

### 相关关系与因果关系的区别

* 相关关系不意味着因果关系
* 因果关系需要满足：
	1. 时间顺序
	2. 理论支持
	3. 排除其他解释

## 回归分析的基本概念

### 无条件概率与无条件期望

* **无条件概率**：
	+ 定义：不受$X\_{i}$变量取值影响下，$Y\_{i}$出现的可能性
	+ 记号：离散变量$P\left(Y\_{i}\right)$；连续变量$g\left(Y\right)$
* **无条件期望**：
	+ 定义：不受$X\_{i}$变量取值影响下，变量$Y\_{i}$的期望值
	+ 记号：$g\left(Y\_{i}\right)$表示连续变量的概率密度函数（cdf）
	+ 计算公式：

$$\begin{matrix}E\left(Y\right)&=\sum\_{1}^{N}Y\_{i}⋅P\left(Y\_{i}\right)&&(离散变量)\\E\left(Y\right)&=∫Y\_{i}⋅g\left(Y\_{i}\right)dY&&(连续变量)\end{matrix}$$

### 条件概率与条件期望

* **条件概率**：
	+ 定义：给定变量$X\_{i}$的取值条件下，$Y\_{i}$出现的可能性
	+ 记号：离散变量$P\left(Y\_{i}|X\_{i}\right)$；连续变量$g\left(Y|X\right)$
* **条件期望**：
	+ 定义：在给定变量$X\_{i}$的取值条件下，$Y\_{i}$的期望值
	+ 记号：$g\left(Y|X\right)$表示连续变量的条件概率密度函数（cdf）
	+ 计算公式：

$$\begin{matrix}E\left(Y|X\_{i}\right)&=\sum\_{1}^{N}\left(Y\_{i}|X\_{i}\right)⋅P\left(Y\_{i}|X\_{i}\right)&&(离散变量)\\E\left(Y|X\_{i}\right)&=∫\left(Y|X\right)⋅g\left(Y|X\right)dY&&(连续变量)\end{matrix}$$

## 总体回归分析

### 总体回归线（PRL）

* **定义**：给定X值时Y的条件期望值的轨迹
* **特点**：
	+ 几何上表现为一条曲线或直线
	+ 统计上就是Y对X的回归
* **类型**：
	+ 总体回归曲线（PRC）：条件期望值的轨迹表现为曲线
	+ 总体回归线（PRL）：条件期望值的轨迹表现为直线

### 总体回归函数（PRF）

* **隐函数形式**：

$$\begin{matrix}E\left(Y|X\_{i}\right)=f\left(X\_{i}\right)&&(PRF)\end{matrix}$$

* **显函数形式**（线性）：

$$\begin{matrix}E\left(Y|X\_{i}\right)=β\_{1}+β\_{2}X\_{i}&&(PRF\\_L)\end{matrix}$$

* $β\_{1},β\_{2}$分别称为截距和斜率系数
* $β\_{1},β\_{2}$为总体参数或回归系数
* $β\_{1},β\_{2}$为未知但固定的参数

### 总体回归模型（PRM）

* **隐函数形式**：

$$\begin{matrix}Y\_{i}&=E\left(Y|X\_{i}\right)+u\_{i}\\&=f\left(X\_{i}\right)+u\_{i}\end{matrix}$$

* **线性形式**：

$$\begin{matrix}Y\_{i}&=E\left(Y|X\_{i}\right)+u\_{i}\\&=β\_{1}+β\_{2}X\_{i}+u\_{i}\end{matrix}$$

* **随机干扰项**$u\_{i}$：
	+ 定义：$u\_{i}=Y\_{i}−E\left(Y|X\_{i}\right)$
	+ 来源：
		1. 理论的含糊性
		2. 数据的不充分
		3. 其他变量的影响
		4. 人类行为的内在随机性
		5. 测量误差
		6. 节省原则
		7. 错误的函数形式

## 样本回归分析

### 样本回归线（SRL）

* **定义**：通过拟合样本数据得到的一条曲线（或直线）
* **特点**：
	+ 由拟合值$\hat{Y}\_{i}$连接而成
	+ $\hat{Y}\_{i}$是对条件期望值$Y|X\_{i}$的拟合
	+ 可以通过不同方法拟合（如OLS）

### 样本回归函数（SRF）

* **线性形式**：

$$\begin{matrix}\hat{Y}\_{i}=\hat{β}\_{1}+\hat{β}\_{2}X\_{i}\end{matrix}$$

* **与PRF的关系**：
	+ $\hat{Y}\_{i}$是对$E\left(Y|X\_{i}\right)$的估计量
	+ $\hat{β}\_{1}$是对$β\_{1}$的估计量
	+ $\hat{β}\_{2}$是对$β\_{2}$的估计量

### 样本回归模型（SRM）

* **隐函数形式**：

$$\begin{matrix}Y\_{i}=g\left(X\_{i}\right)+e\_{i}\end{matrix}$$

* **线性形式**：

$$\begin{matrix}Y\_{i}=\hat{β}\_{1}+\hat{β}\_{2}X\_{i}+e\_{i}&&(SRM\\_L)\end{matrix}$$

* **残差**$e\_{i}$：
	+ 定义：$e\_{i}=Y\_{i}−\hat{Y}\_{i}=Y\_{i}−\left(\hat{β}\_{1}+\hat{β}\_{2}X\_{i}\right)$
	+ 表示样本回归函数与Y的样本观测值之间的离差

## 总体回归与样本回归的比较

### 主要区别

1. **总体回归**：
	* 基于总体数据
	* 参数$β\_{1},β\_{2}$是固定但未知的
	* 随机干扰项$u\_{i}$反映总体中的随机性
2. **样本回归**：
	* 基于样本数据
	* 参数$\hat{β}\_{1},\hat{β}\_{2}$是估计值
	* 残差$e\_{i}$反映样本拟合的不完全性

### 重要结论

1. 随机抽样数据继承了总体的特征
2. 利用随机样本进行数据拟合是对总体规律的”反向追踪”
3. 样本回归模型中的残差是拟合不完全的产物

### 思考问题

1. 怎样判定对随机样本的一次数据拟合是更优的？
2. 是否存在一种”最优”的拟合方法？
3. 如何评估样本回归对总体回归的逼近程度？

# 5.3 OLS方法与参数估计

## 普通最小二乘法（OLS）

### 基本概念

总体回归模型（PRM）：

$$\begin{matrix}Y\_{i}&=β\_{1}+β\_{2}X\_{i}+u\_{i}\end{matrix}$$

样本回归模型（SRM）：

$$\begin{matrix}Y\_{i}&=\hat{β}\_{1}+\hat{β}\_{2}X\_{i}+e\_{i}\end{matrix}$$

### OLS基本原理

OLS的基本原理：残差平方和最小化。

$$\begin{matrix}e\_{i}&=Y\_{i}−\hat{Y}\_{i}\\&=Y\_{i}−\left(\hat{β}\_{1}+\hat{β}\_{2}X\_{i}\right)\end{matrix}$$

$$\begin{matrix}Q&=∑e\_{i}^{2}\\&=∑\left(Y\_{i}−\hat{Y}\_{i}\right)^{2}\\&=∑\left(Y\_{i}−\left(\hat{β}\_{1}+\hat{β}\_{2}X\_{i}\right)\right)^{2}\\&≡f\left(\hat{β}\_{1},\hat{β}\_{2}\right)\end{matrix}$$

$$\begin{matrix}Min\left(Q\right)&=Min\left(f\left(\hat{β}\_{1},\hat{β}\_{2}\right)\right)\end{matrix}$$

## 参数估计

### 回归参数的OLS点估计

正规方程组：

$$\begin{matrix}\left\{\begin{matrix}∑Y\_{i}−n\hat{β}\_{1}−\left(∑X\_{i}\right)\hat{β}\_{2}&=0\\∑X\_{i}Y\_{i}−\left(∑X\_{i}\right)\hat{β}\_{1}−\left(∑X\_{i}^{2}\right)\hat{β}\_{2}&=0\end{matrix}\right.\end{matrix}$$

回归系数的计算公式1（Favorite Five，FF）：

$$\begin{matrix}\left\{\begin{matrix}\hat{β}\_{2}&=\frac{n∑X\_{i}Y\_{i}−∑X\_{i}∑Y\_{i}}{n∑X\_{i}^{2}−\left(∑X\_{i}\right)^{2}}\\\hat{β}\_{1}&=\frac{n∑X\_{i}^{2}Y\_{i}−∑X\_{i}∑X\_{i}Y\_{i}}{n∑X\_{i}^{2}−\left(∑X\_{i}\right)^{2}}\end{matrix}\right.&&(FF solution)\end{matrix}$$

### 回归参数的OLS点估计（离差形式）

离差公式（favorite five，ff）：

$$\begin{matrix}\left\{\begin{matrix}\hat{β}\_{2}&=\frac{∑x\_{i}y\_{i}}{∑x\_{i}^{2}}\\\hat{β}\_{1}&=‾\_{i}−\hat{β}\_{2}‾\_{i}\end{matrix}\right.&&(ff solution)\end{matrix}$$

其中离差计算： $x\_{i}=X\_{i}−‾; y\_{i}=Y\_{i}−‾$。

### 随机干扰项参数的OLS点估计

回归误差方差：

$$\begin{matrix}\hat{σ}^{2}=\frac{∑e\_{i}^{2}}{n−2}\end{matrix}$$

回归误差标准差：

$$\begin{matrix}\hat{σ}=\sqrt{\frac{∑e\_{i}^{2}}{n−2}}\end{matrix}$$

## SRF和SRM的特征

### 基本特征

1. 样本回归线一定会经过样本均值点$\left(‾,‾\right)$：

$$\begin{matrix}‾=\hat{β}\_{1}+\hat{β}\_{2}‾\end{matrix}$$

1. $Y\_{i}$的估计值$\hat{Y}\_{i}$的均值等于Y的样本均值：

$$\begin{matrix}‾=‾\end{matrix}$$

1. 残差的均值$‾$为零：

$$\begin{matrix}∑e\_{i}&=0\\‾&=0\end{matrix}$$

### 离差形式

SRM的离差形式：

$$\begin{matrix}y\_{i}=\hat{β\_{2}}x\_{i}+e\_{i} &&(SRM-dev)\end{matrix}$$

SRF的离差形式：

$$\begin{matrix}\hat{y}\_{i}=\hat{β\_{2}}x\_{i} &&(SRF-dev)\end{matrix}$$

## 估计精度

### 斜率系数的方差

总体方差和标准差：

$$\begin{matrix}Var\left(\hat{β}\_{2}\right)≡σ\_{\hat{β}\_{2}}^{2}&=\frac{σ^{2}}{∑x\_{i}^{2}}\\σ\_{\hat{β}\_{2}}&=\sqrt{\frac{σ^{2}}{∑x\_{i}^{2}}}\end{matrix}$$

样本方差和标准差：

$$\begin{matrix}S\_{\hat{β}\_{2}}^{2}&=\frac{\hat{σ}^{2}}{∑x\_{i}^{2}}\\S\_{\hat{β}\_{2}}&=\sqrt{\frac{\hat{σ}^{2}}{∑x\_{i}^{2}}}\end{matrix}$$

### 截距系数的方差

总体方差和标准差：

$$\begin{matrix}Var\left(\hat{β}\_{1}\right)≡σ\_{\hat{β}\_{1}}^{2}&=\frac{∑X\_{i}^{2}}{n}⋅\frac{σ^{2}}{∑x\_{i}^{2}}\\σ\_{\hat{β}\_{1}}&=\sqrt{\frac{∑X\_{i}^{2}}{n}⋅\frac{σ^{2}}{∑x\_{i}^{2}}}\end{matrix}$$

样本方差和标准差：

$$\begin{matrix}S\_{\hat{β}\_{1}}^{2}&=\frac{∑X\_{i}^{2}}{n}⋅\frac{\hat{σ}^{2}}{∑x\_{i}^{2}}\\S\_{\hat{β}\_{1}}&=\sqrt{\frac{∑X\_{i}^{2}}{n}⋅\frac{\hat{σ}^{2}}{∑x\_{i}^{2}}}\end{matrix}$$

## 区间估计

### 斜率系数的置信区间

$$\begin{matrix}\left[\hat{β}\_{2}−t\_{α/2}⋅S\_{\hat{β}\_{2}}, \hat{β}\_{2}+t\_{α/2}⋅S\_{\hat{β}\_{2}}\right]\end{matrix}$$

### 截距系数的置信区间

$$\begin{matrix}\left[\hat{β}\_{1}−t\_{α/2}⋅S\_{\hat{β}\_{1}}, \hat{β}\_{1}+t\_{α/2}⋅S\_{\hat{β}\_{1}}\right]\end{matrix}$$

### 随机干扰项方差的置信区间

$$\begin{matrix}\left[\left(n−2\right)\frac{\hat{σ}^{2}}{χ\_{1−α/2}^{2}}, \left(n−2\right)\frac{\hat{σ}^{2}}{χ\_{α/2}^{2}}\right]\end{matrix}$$

## 重要注意事项

### 估计量性质

* OLS估计量是纯粹由可观测的样本量(X和Y)表达的，因此容易计算
* 它们是点估计量，即对于给定样本，每个估计量仅提供有关总体参数的一个值
* 一旦从样本数据得到OLS估计值，便容易画出样本回归线

### 估计精度影响因素

* 样本量大小
* 自变量变异程度
* 随机干扰项方差
* 模型设定形式

### 区间估计要点

* 置信区间的宽度反映估计的精确度
* 置信水平的选择影响区间宽度
* 样本量越大，区间越窄
* 随机干扰项方差越小，区间越窄

# 5.4 经典假设与OLS性质

## 经典线性回归模型（CLRM）假设

### 关于模型的假设

**CLRM假设1（模型是正确设置的）**：模型设定正确，这是一切计量分析问题的根本来源。

**CLRM假设2（模型是参数线性的）**：模型应该是参数线性的，具体而言模型中**参数**和**随机干扰项**必须线性，变量可以不是线性。

$$Y\_{i}=β\_{1}+β\_{2}X\_{i}+u\_{i}$$

### 关于自变量X的假设

**CLRM假设3（自变量X是外生的）**：X是固定的（给定的）或**独立于**误差项。也即自变量X**不是**随机变量。

$$\begin{matrix}Cov\left(X\_{i},u\_{i}\right)&=0\\E\left(u\_{i}|X\_{i}\right)&=0\end{matrix}$$

### 关于随机干扰项的假设

**CLRM假设4（随机干扰项条件期望值为零）**：给定$X\_{i}$的情形下，随机干扰项$u\_{i}$的**条件期望**为零。

$$E\left(u|X\_{i}\right)=0$$

**CLRM假设5（随机干扰项的方差为同方差）**：给定$X\_{i}$的情形下，随机干扰项$u\_{i}$的方差处处相等。

$$\begin{matrix}Var\left(u\_{i}|X\_{i}\right)&=E\left[\left(u\_{i}−E\left(u\_{i}\right)\right)^{2}|X\_{i}\right]\\&=E\left(u\_{i}^{2}|X\_{i}\right)\\&=E\left(u\_{i}^{2}\right)\\&≡σ^{2}\end{matrix}$$

**CLRM假设6（随机干扰项之间无自相关）**：给定两个不同的自变量取值情形下，随机干扰项$u\_{i},u\_{j}$的相关系数为0。

$$\begin{matrix}Cov\left(u\_{i},u\_{j}|X\_{i},X\_{j}\right)&=E\left[\left(u\_{i}−E\left(u\_{i}\right)\right)\left(u\_{i}−E\left(u\_{i}\right)\right)\right]\\&=E\left(u\_{i}u\_{j}\right)\\&≡0\end{matrix}$$

### 关于样本数的要求

**CLRM假设7（观测样本数假设）**：观测次数n，要大于待估计参数个数。

## OLS估计量的性质

### 高斯-马尔可夫定理

**高斯-马尔可夫定理**：在给定经典线性回归模型(CLRM)的假定下，最小二乘(OLS)**估计量**是最优线性无偏估计量(BLUE)。

### 线性性

**线性性**（Linearity）：是指$\hat{β}\_{2}$和$\hat{β}\_{1}$对$Y\_{i}$是线性的。

$$\begin{matrix}\hat{β}\_{2}&=∑k\_{i}Y\_{i}&&\leftarrow \left[k\_{i}=\frac{x\_{i}}{∑x\_{i}^{2}}\right]\\\hat{β\_{1}}&=∑w\_{i}Y\_{i}&&\leftarrow \left[w\_{i}=\frac{1}{n}−k\_{i}‾\right]\end{matrix}$$

### 无偏性

**无偏性**(Unbias)：**估计量**期望值等于**参数**的真值。

$$\begin{matrix}E\left(\hat{β}\_{2}\right)&=β\_{2}\\E\left(\hat{β}\_{1}\right)&=β\_{1}\end{matrix}$$

### 方差最小性

**方差最小性**（Best）：在所有线性无偏估计量中，方差为最小。

$$\begin{matrix}Var\left(\hat{β}\_{2}\right)≡σ\_{\hat{β}\_{2}}^{2}&=\frac{σ^{2}}{∑x\_{i}^{2}}\\Var\left(\hat{β}\_{1}\right)≡σ\_{\hat{β}\_{1}}^{2}&=\frac{∑X\_{i}^{2}}{n}⋅\frac{σ^{2}}{∑x\_{i}^{2}}\end{matrix}$$

## 经典正态线性回归模型（N-CLRM）

### N-CLRM假设

在CLRM假设基础上增加干扰项$u\_{i}$服从正态性的假设：

$$u\_{i}∼iid. N\left(0,σ^{2}\right)$$

其中，iid表示独立同分布(Independent Identical Distribution)。

### N-CLRM假设下OLS估计量的性质

1. 无偏性
2. 有效性（方差最小）
3. 一致性（收敛到它们的总体参数上）
4. 估计量$\hat{β}\_{2}$和$\hat{β}\_{1}$服从正态分布
5. 随机变量$Z\_{2}$和$Z\_{1}$服从标准正态分布
6. $X≡\left(n−2\right)\hat{σ^{2}}/σ^{2}$服从自由度为$\left(n−2\right)$的卡方分布
7. 随机变量$\left(\hat{β}\_{2},\hat{β}\_{1}\right)$的分布独立于随机变量$\hat{σ}^{2}$
8. 估计量$\left(\hat{β}\_{2},\hat{β}\_{1}\right)$是最优无偏估计量（BUE）

## 重要注意事项

1. CLRM假设的重要性：
	* 为”从样本推断总体”提供理论基础
	* 确保OLS估计量的BLUE性质
	* 为后续的统计推断提供基础
2. 假设的现实性：
	* 许多假设在现实中可能不完全满足
	* 需要根据实际情况适当放宽或调整假设
	* 违背假设可能影响估计量的性质
3. 正态性假设的作用：
	* 为参数估计量的分布提供理论基础
	* 为构造t统计量、F统计量等提供基础
	* 在实际应用中，中心极限定理和大数定理可以保证估计的有效性

# 5.5 假设检验

## 假设检验的基本原理

### 假设检验的概念

**假设检验**（Hypothesis Testing）：通过制定一套步骤和规则，决定接受或拒绝一个虚拟假设（原假设）。

* **虚拟假设**(null hypothesis) $H\_{0}$：指定或声称的假设，如$H\_{0}:β\_{2}=0$
* **备择假设**(alter hypothesis) $H\_{1}$：
	+ 简单备择假设：$H\_{1}:β\_{2}=1.5$
	+ 复合备择假设：$H\_{1}:β\_{2}\ne 1.5$

### 假设检验的方法

1. **置信区间检验**（confidence interval）
2. **显著性检验**（test of significance）

## 置信区间检验法

### 双侧检验

对于假设：

$$H\_{0}:β\_{2}=0; H\_{1}:β\_{2}\ne 0$$

决策规则： 1. 构造$β\_{2}$的$100\left(1−α\right)\%$置信区间 2. 如果$β\_{2}$在$H\_{0}$假设下落入此区间，就不拒绝$H\_{0}$ 3. 如果它落在此区间之外，就要拒绝$H\_{0}$

## 显著性检验法

### 检验步骤

1. 找到合适的检验统计量（如t统计量、$χ^{2}$统计量、F统计量等）
2. 知道该统计量在$H\_{0}$下的抽样分布
3. 计算样本统计量的值
4. 查表找出给定显著性水平$α$下的临界值
5. 比较样本统计量值和临界值
6. 做出拒绝还是接受$H\_{0}$的判断

### 显著性水平与显著性概率

* **显著性水平**$α$：通常固定在0.01、0.05、0.1水平
* **显著性概率**p值：对给定的样本算出的检验统计量对应的概率

## 回归系数的t检验

### 截距参数的t检验

1. 提出假设：

$$H\_{0}:β\_{1}=0; H\_{1}:β\_{1}\ne 0$$

1. 构造检验统计量：

$$T=\frac{\hat{β}\_{1}−β\_{1}}{S\_{\hat{β}\_{1}}}∼t\left(n−2\right)$$

1. 计算样本统计量：

$$t\_{\hat{β}\_{1}}^{\*}=\frac{\hat{β}\_{1}}{S\_{\hat{β}\_{1}}}$$

### 斜率参数的t检验

1. 提出假设：

$$H\_{0}:β\_{2}=0; H\_{1}:β\_{2}\ne 0$$

1. 构造检验统计量：

$$T=\frac{\hat{β}\_{2}−β\_{2}}{S\_{β\_{2}}}∼t\left(n−2\right)$$

1. 计算样本统计量：

$$t\_{\hat{β}\_{2}}^{\*}=\frac{\hat{β}\_{2}}{S\_{\hat{β}\_{2}}}$$

## 方差分解（ANOVA）

### Y变异的分解

$$\begin{matrix}&&\left(Y\_{i}−‾\right)&&=\left(\hat{Y}\_{i}−‾\right)&&+\left(Y\_{i}−\hat{Y}\_{i}\right)\\&&y\_{i}&&=\hat{y}\_{i}&&+e\_{i}\end{matrix}$$

### 平方和分解

$$\begin{matrix}&&∑y\_{i}^{2}&&=∑\hat{y}\_{i}^{2}&&+∑e\_{i}^{2}\\&&TSS&&=ESS&&+RSS\end{matrix}$$

其中： - $TSS$：总离差平方和 - $ESS$：回归平方和 - $RSS$：残差平方和

## 模型整体显著性F检验

### F检验步骤

1. 提出假设：

$$H\_{0}:β\_{2}=0; H\_{1}:β\_{2}\ne 0$$

1. 构造检验统计量：

$$F=\frac{\left(\hat{β}\_{2}−β\_{2}\right)^{2}∑x\_{i}^{2}}{∑e\_{i}^{2}/\left(n−2\right)}∼F\left(1,n−2\right)$$

1. 计算样本统计量：

$$F^{\*}=\frac{ESS/df\_{ESS}}{RSS/df\_{RSS}}=\frac{MSS\_{ESS}}{MSS\_{RSS}}$$

### F检验与t检验的比较

**联系**： - 在一元回归模型中，t检验与F检验的结论总是一致的 - 对于检验斜率参数$β\_{2}$的显著性，两者可相互替代 - 在一元回归分析中，若假设$H\_{0}:β\_{2}=0$，则$F^{\*}≃\left(t^{\*}\right)^{2}$

**不同**： 1. 检验目的不同： - F检验：检验模型的整体显著性 - t检验：检验各个回归参数的显著性 2. 假设的提出不同 3. 检验原理不同

## 重要注意事项

1. 统计显著性与实际显著性：
	* 不能一味追求统计显著性
	* 需要考虑”实际显著性”的现实意义
2. 置信区间方法与显著性检验方法的选择：
	* 一般来说，置信区间方法优于显著性检验方法
	* 置信区间方法能提供更多信息
3. 假设检验的局限性：
	* 显著性水平的选择具有主观性
	* 样本量会影响检验结果
	* 需要结合实际情况进行判断

# 5.6 拟合优度与残差分析

## 拟合优度

### 基本概念

**拟合优度**（Goodness of fit）：度量样本回归线对一组数据拟合优劣水平。

**判定系数**（coefficient of determination）：一种利用平方和分解，考察样本回归线对数据拟合效果的总度量。

* 一元回归中，一般记为$r^{2}$
* 多元回归中，一般记为$R^{2}$

### 判定系数的计算

判定系数$r^{2}$计算公式1：

$$\begin{matrix}r^{2}&=\frac{ESS}{TSS}\\&=\frac{∑\left(\hat{Y}\_{i}−‾\right)^{2}}{∑\left(Y\_{i}−‾\right)^{2}}\end{matrix}$$

判定系数$r^{2}$计算公式2：

$$\begin{matrix}r^{2}&=1−\frac{RSS}{TSS}\\&=1−\frac{∑e\_{i}^{2}}{∑\left(Y\_{i}−‾\right)^{2}}\end{matrix}$$

判定系数$r^{2}$计算公式3：

$$\begin{matrix}r^{2}&=\frac{ESS}{TSS}\\&=\frac{∑\hat{y}\_{i}^{2}}{∑y\_{i}^{2}}\\&=\frac{∑\left(\hat{β}\_{2}x\_{i}\right)^{2}}{∑y\_{i}^{2}}\\&=\hat{β}\_{2}^{2}\frac{∑x\_{i}^{2}}{∑y\_{i}^{2}}\\&=\hat{β}\_{2}^{2}\frac{S\_{X\_{i}}^{2}}{S\_{Y\_{i}}^{2}}\end{matrix}$$

判定系数$r^{2}$计算公式4：

$$\begin{matrix}r^{2}&=\hat{β}\_{2}^{2}⋅\frac{∑x\_{i}^{2}}{∑y\_{i}^{2}}\\&=\left(\frac{∑x\_{i}y\_{i}}{∑x\_{i}^{2}}\right)^{2}⋅\left(\frac{∑x\_{i}^{2}}{∑y\_{i}^{2}}\right)\\&=\frac{\left(∑x\_{i}y\_{i}\right)^{2}}{∑x\_{i}^{2}∑y\_{i}^{2}}\end{matrix}$$

### 判定系数的性质

1. $r^{2}$是一个非负量
2. $0\leq r^{2}\leq 1$
	* $r^{2}=0$：表示回归线完全不能解释Y的变异
	* $r^{2}=1$：表示回归线完全解释了Y的变异

### 判定系数与相关系数的关系

**总体相关系数**：

$$\begin{matrix}ρ&=\frac{Cov\left(X,Y\right)}{\sqrt{Var\left(X\_{i}\right)Var\left(Y\_{i}\right)}}\\&=\frac{E\left(X\_{i}−EX\right)\left(Y\_{i}−EY\right)}{\sqrt{E\left(X\_{i}−EX\right)^{2}E\left(Y\_{i}−EY\right)^{2}}}\end{matrix}$$

**样本相关系数**：

$$\begin{matrix}r&=\frac{S\_{XY}^{2}}{S\_{X}\*S\_{Y}}\\&=\frac{∑\left(X\_{i}−‾\right)\left(Y\_{i}−‾\right)}{\sqrt{∑(X\_{i}−‾)^{2}∑\left(Y\_{i}−‾\right)^{2}}}\\&=\frac{∑x\_{i}y\_{i}}{\sqrt{∑x\_{i}^{2}∑y\_{i}^{2}}}\end{matrix}$$

**联系与区别**：

* 在一元回归中，判定系数$r^{2}$等于样本相关系数$r$的平方
* 判定系数$r^{2}$表明因变量变异由解释变量所解释的比例
* 相关系数$r$只能表明变量间的线性关联强度
* 在多元回归中，这种区别会更加凸显

## 残差分析

### 残差的定义与作用

**残差**(residual)：是因变量的观测值与根据估计的回归方程求出的估计值之差，用$e\_{i}$表示。

$$\begin{matrix}e\_{i}=Y\_{i}−\hat{Y\_{i}}\end{matrix}$$

残差分析的主要目的： 1. 反映用估计的回归方程去预测而引起的误差 2. 可用于确定有关随机干扰项$μ\_{i}$的假定是否成立 3. 用于检测有影响的观测值

### 标准化残差

**皮尔逊标准化残差**（Pearson residual）：

$$\begin{matrix}e\_{i,sd}^{\*}&=\frac{e\_{i}}{s\_{e\_{i}}}\\&=\frac{\left(Y\_{i}−\hat{Y\_{i}}\right)}{\sqrt{\frac{∑\left(e\_{i}−‾\right)^{2}}{n−1}}}\end{matrix}$$

**学生化标准残差**（Studentized Residuals）：

$$\begin{matrix}e\_{i,st}^{\*}&=\frac{e\_{i}}{\sqrt{MSE\_{\left(i\right)}\left(1−h\_{ii}\right)}}\end{matrix}$$

或

$$\begin{matrix}e\_{i,st}^{\*}&=e\_{i,sd}^{\*}\left(\frac{n−m−2}{n−m−1−e\_{i,sd}^{\*2}}\right)^{2}\end{matrix}$$

其中： - $MSE\_{\left(i\right)}$：删除第$i$个观测值进行建模的均方误差 - $h\_{ii}$：删除第$i$个观测值进行建模的第$i$个影响权重 - $m=k−1$：回归元个数

### 残差图分析

**残差图**(residual plot)：用于呈现残差数据$e\_{i}$的分布情况的统计图图形，主要包括：

1. 关于$X\_{i}$的残差散点图
2. 关于$Y\_{i}$的残差散点图（或关于$\hat{Y\_{i}}$）
3. 关于样本序号的残差散点图或标准化残差散点图

## 重要注意事项

1. 拟合优度的理解：
	* 即使采用OLS方法，对样本数据的拟合也是不完全的
	* 实际数据点在样本回归线附近，而不是在样本回归线上
	* 样本点行为的”变异”可划分为”回归”能解释的部分和”随机”的部分
2. 残差分析的重要性：
	* 残差分析是检验模型假设是否成立的重要手段
	* 标准化残差可以帮助识别异常值和有影响的观测值
	* 残差图可以直观地展示模型的拟合效果和潜在问题
3. 模型诊断的综合性：
	* 不能仅依赖单一指标判断模型的好坏
	* 需要结合拟合优度、残差分析等多种方法
	* 考虑统计显著性的同时，也要关注实际意义

# 5.7 回归预测分析

## 回归预测的基本概念

### 两类预测

一元回归模型下：

$$\begin{matrix}Y\_{i}=β\_{1}+β\_{2}X\_{i}+u\_{i}\end{matrix}$$

**均值预测**(mean prediction)： - 给定$X\_{0}$，预测Y的条件均值$E\left(Y|X=X\_{0}\right)$

**个值预测**(individual prediction)： - 给定$X\_{0}$，预测对应于$X\_{0}$的Y的个别值$\left(Y\_{0}|X\_{0}\right)$

### 预测分析的关键

样本外拟合值$\hat{Y}\_{0}|X=X\_{0}$的性质： - 是均值$E\left(Y|X=X\_{0}\right)$的一个BLUE - 是个值$\left(Y\_{0}|X\_{0}\right)$的一个BLUE

## 均值预测

### 均值预测的分布

在N-CLRM假设和OLS方法下，给定$X\_{0}$下的拟合值$\hat{Y}\_{0}$服从如下正态分布：

$$\begin{matrix}\hat{Y}\_{0}&∼N\left(μ\_{\hat{Y}\_{0}},σ\_{\hat{Y}\_{0}}^{2}\right)\\μ\_{\hat{Y}\_{0}}&=E\left(\hat{Y}\_{0}\right)=E\left(\hat{β}\_{1}+\hat{β}\_{2}X\_{0}\right)=β\_{1}+β\_{2}X\_{0}=E\left(Y|X\_{0}\right)\\var\left(\hat{Y}\_{0}\right)&=σ\_{\hat{Y}\_{0}}^{2}=σ^{2}\left[\frac{1}{n}+\frac{\left(X\_{0}−\overline{X}\right)^{2}}{∑x\_{i}^{2}}\right]\end{matrix}$$

### 均值预测的置信区间

构造t统计量：

$$\begin{matrix}T&=\frac{\hat{Y}\_{0}−E\left(Y|X\_{0}\right)}{S\_{\hat{Y}\_{0}}}∼t\left(n−2\right)\\S\_{\hat{Y}\_{0}}&=\sqrt{\hat{σ}^{2}\left[\frac{1}{n}+\frac{\left(X\_{0}−\overline{X}\right)^{2}}{∑x\_{i}^{2}}\right]}\end{matrix}$$

均值$E\left(Y|X=X\_{0}\right)$的置信区间：

$$\begin{matrix}Pr\left[\hat{Y}\_{0}−t\_{1−α/2}\left(n−2\right)⋅S\_{\hat{Y}\_{0}}\leq E\left(Y|X\_{0}\right)\leq \hat{Y}\_{0}+t\_{1−α/2}\left(n−2\right)⋅S\_{\hat{Y}\_{0}}\right]=1−α\end{matrix}$$

## 个值预测

### 个值预测的分布

在N-CLRM假设和OLS方法下，给定$X\_{0}$下的个别值$Y\_{0}$服从如下正态分布：

$$\begin{matrix}Y\_{0}&∼N\left(μ\_{Y\_{0}},σ\_{Y\_{0}}^{2}\right)\\μ\_{Y\_{0}}&=E\left(Y\_{0}\right)=E\left(β\_{1}+β\_{2}X\_{0}\right)=β\_{1}+β\_{2}X\_{0}\\Var\left(Y\_{0}\right)&=Var\left(u\_{0}\right)=σ^{2}\end{matrix}$$

### 个值预测的置信区间

构造新的随机变量$\left(Y\_{0}−\hat{Y}\_{0}\right)$的分布：

$$\begin{matrix}Y\_{0}−\hat{Y}\_{0}&∼N\left(0,σ^{2}\left[1+\frac{1}{n}+\frac{\left(X\_{0}−\overline{X}\right)^{2}}{∑x\_{i}^{2}}\right]\right)\\Y\_{0}−\hat{Y}\_{0}&∼N\left(0,σ\_{Y\_{0}−\hat{Y}\_{0}}^{2}\right)\end{matrix}$$

构造t统计量：

$$\begin{matrix}T&=\frac{\left(Y\_{0}−\hat{Y}\_{0}\right)}{S\_{\left(Y\_{0}−\hat{Y}\_{0}\right)}}∼t\left(n−2\right)\\S\_{\left(Y\_{0}−\hat{Y}\_{0}\right)}&=\sqrt{\hat{σ}^{2}\left[1+\frac{1}{n}+\frac{\left(X\_{0}−\overline{X}\right)^{2}}{∑x\_{i}^{2}}\right]}\end{matrix}$$

个值$Y\_{0}$的置信区间：

$$\begin{matrix}Pr\left[\hat{Y}\_{0}−t\_{1−α/2}\left(n−2\right)⋅S\_{\left(Y\_{0}−\hat{Y}\_{0}\right)}\leq Y\_{0}\leq \hat{Y}\_{0}+t\_{1−α/2}\left(n−2\right)⋅S\_{\left(Y\_{0}−\hat{Y}\_{0}\right)}\right]=1−α\end{matrix}$$

## 置信带

### 置信带的概念

**置信带**(confidence interval)：对所有的X值，分别进行均值和个值预测，得到：

* 均值预测的置信带——总体回归函数的置信带
* 个值预测的置信带

### 置信带的特点

1. 均值预测比个值预测更准确（置信带更窄）
2. 置信带在中心点$\left(‾,‾\right)$处最窄
3. 样本内置信带用于检验可靠性
4. 样本外置信带用于预测未来值范围

## 重要注意事项

1. 回归预测的基础：
	* 基于OLS估计方法
	* 基于CLRM假设
	* 基于BLUE估计性质
2. 预测的可信度：
	* 均值预测比个值预测更准确
	* 置信带宽度反映预测的不确定性
	* 置信度越高，置信带越宽
3. 预测的局限性：
	* 预测结果依赖于模型假设的合理性
	* 预测精度受样本量和数据质量影响
	* 预测范围受解释变量取值范围的限制

# 5.8 回归报告解读

## 回归报告的基本形式

### 课程要求

* 熟练、正确阅读统计软件给出的各类分析报告
* 理解报告中的关键信息和内涵
* 掌握不同统计软件（如Stata、EViews、R、Excel等）的回归分析报告解读

### 一元回归模型

$$\begin{matrix}Y\_{i}=β\_{1}+β\_{2}X\_{i}+u\_{i}\end{matrix}$$

## 回归报告的呈现形式

### 多行方程表达法

**形式1：多行方程表达法**（精炼报告）：

$$\begin{matrix}\hat{Y}\_{i}&=\hat{β}\_{1}+\hat{β}\_{2}X\_{i}\\\left(t\right)&=\left(t\_{1}\right) \left(t\_{2}\right)\\\left(se\right)&=\left(se\_{1}\right) \left(se\_{2}\right)\\\left(fitness\right)&=R^{2}=…; ‾^{2}=…; F=…; p=…\end{matrix}$$

### 表格列示法

**形式2：表格列示法**（精炼报告）：

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | $\hat{β}\_{1}$ | $se\_{1}$ | $t\_{1}$ | $p\_{1}$ |
| X | $\hat{β}\_{2}$ | $se\_{2}$ | $t\_{2}$ | $p\_{2}$ |

## 统计软件报告解读

### Excel软件报告

**形式3：原始报告**包含以下部分：

1. 参数估计结果
	* 回归系数
	* 标准误差
	* t统计量
	* P值
	* 置信区间
2. 拟合优度信息
	* 判定系数$R^{2}$
	* 调整判定系数$‾^{2}$
	* 标准误差
3. 方差分解（ANOVA表）
	* 回归平方和
	* 残差平方和
	* 总平方和
	* 自由度
	* F统计量
4. 残差分析
	* 残差表
	* 残差图

### EViews软件报告

**形式3：原始报告**包含以下部分：

1. 抬头区域
	* Dependent Variable：因变量
	* Method：分析方法
	* Date/Time：分析时间
	* Sample：样本范围
	* Included observations：样本数
2. 三线表区域
	* Variable：模型变量
	* Coefficient：回归系数
	* Std. Error：标准误差
	* t-Statistic：t统计量
	* Prob.：概率值
3. 指标值区域
	* R-squared：判定系数
	* Adjusted R-squared：调整判定系数
	* S.E. of regression：回归误差标准差
	* Sum squared resid：残差平方和
	* Log likelihood：对数似然值
	* F-statistic：F统计量
	* Prob(F-statistic)：F统计量概率值

### R软件报告

**形式4：原始报告**包含： - 回归系数估计 - 标准误差 - t统计量 - P值 - 拟合优度指标 - 方差分析表

## 重要注意事项

### 报告解读要点

* 关注回归系数的经济含义
* 重视统计显著性检验结果
* 注意模型整体拟合优度
* 检查残差分析结果

### 软件操作要求

* 熟练掌握Excel回归分析操作步骤
* 理解不同软件报告格式的异同
* 能够正确提取和解读关键信息

### 报告应用建议

* 根据研究目的选择合适的报告形式
* 注意报告内容的完整性和准确性
* 结合经济理论和统计检验结果进行综合分析