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**：，其中为自变量，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**：
  + 商品销售额与销售量的关系：
  + 圆的面积与半径的关系：

### 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：
  + 样本相关系数：
* **性质**：
  1. 取值范围：
  2. 对称性：
  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**：
2. **小ff公式**：

其中：

#### 相关系数的经验解释

* 高度相关：
* 中度相关：
* 低度相关：
* 极弱相关：

### 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**：
  + 保持不变，和之间的相关系数：
* 保持不变，和之间的相关系数：

## 相关系数的显著性检验

### 检验步骤

1. **提出假设**：
2. **计算样本统计量**：
3. **确定临界值**：
   * 给定显著性水平
   * 查t分布表得
4. **做出决策**：
   * 若，拒绝
   * 若，不拒绝

### 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：不受变量取值影响下，出现的可能性
  + 记号：离散变量；连续变量
* **无条件期望**：
  + 定义：不受变量取值影响下，变量的期望值
  + 记号：表示连续变量的概率密度函数（cdf）
  + 计算公式：

### 条件概率与条件期望

* **条件概率**：
  + 定义：给定变量的取值条件下，出现的可能性
  + 记号：离散变量；连续变量
* **条件期望**：
  + 定义：在给定变量的取值条件下，的期望值
  + 记号：表示连续变量的条件概率密度函数（cdf）
  + 计算公式：

## 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**：
* **显函数形式**（线性）：
* 分别称为截距和斜率系数
* 为总体参数或回归系数
* 为未知但固定的参数

### 总体回归模型（PRM）

* **隐函数形式**：
* **线性形式**：
* **随机干扰项**：
  + 定义：
  + 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**：
  + 由拟合值连接而成
  + 是对条件期望值的拟合
  + 可以通过不同方法拟合（如OLS）

### 样本回归函数（SRF）

* **线性形式**：
* **与PRF的关系**：
  + 是对的估计量
  + 是对的估计量
  + 是对的估计量

### 样本回归模型（SRM）

* **隐函数形式**：
* **线性形式**：
* **残差**：
  + 定义：
  + 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
   * 参数是固定但未知的
   * 随机干扰项反映总体中的随机性
2. **样本回归**：
   * 基于样本数据
   * 参数是估计值
   * 残差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）：

样本回归模型（SRM）：

### OLS基本原理

OLS的基本原理：残差平方和最小化。

## 参数估计

### 回归参数的OLS点估计

正规方程组：

回归系数的计算公式1（Favorite Five，FF）：

### 回归参数的OLS点估计（离差形式）

离差公式（favorite five，ff）：

其中离差计算： 。

### 随机干扰项参数的OLS点估计

回归误差方差：

回归误差标准差：

## 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：
2. 的估计值的均值等于Y的样本均值：
3. 残差的均值为零：

### 离差形式

SRM的离差形式：

SRF的离差形式：

## 估计精度

### 斜率系数的方差

总体方差和标准差：

样本方差和标准差：

### 截距系数的方差

总体方差和标准差：

样本方差和标准差：

## 区间估计

### 斜率系数的置信区间

### 截距系数的置信区间

### 随机干扰项方差的置信区间

## 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。

### 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。

### 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）**：给定的情形下，随机干扰项的**条件期望**为零。

**CLRM假设5（随机干扰项的方差为同方差）**：给定的情形下，随机干扰项的方差处处相等。

**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，随机干扰项的相关系数为0。

### 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）：是指和对是线性的。

### 无偏性

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

### 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。

## 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服从正态性的假设：

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. 估计量和服从正态分布
5. 随机变量和服从标准正态分布
6. 服从自由度为的卡方分布
7. 随机变量的分布独立于随机变量
8. 估计量是最优无偏估计量（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) ：指定或声称的假设，如
* **备择假设**(alter hypothesis) ：
  + 简单备择假设：
  + 复合备择假设：

### 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：

决策规则： 1. 构造的置信区间 2. 如果在假设下落入此区间，就不拒绝 3. 如果它落在此区间之外，就要拒绝

## 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、统计量、F统计量等）
2. 知道该统计量在下的抽样分布
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. 做出拒绝还是接受的判断

### 显著性水平与显著性概率

* **显著性水平**：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：
2. 构造检验统计量：
3. 计算样本统计量：

### 斜率参数的t检验

1. 提出假设：
2. 构造检验统计量：
3. 计算样本统计量：

## 方差分解（ANOVA）

### Y变异的分解

### 平方和分解

其中： - ：总离差平方和 - ：回归平方和 - ：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：
2. 构造检验统计量：
3. 计算样本统计量：

### 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 - 对于检验斜率参数的显著性，两者可相互替代 - 在一元回归分析中，若假设，则

**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
* 多元回归中，一般记为

### 判定系数的计算

判定系数计算公式1：

判定系数计算公式2：

判定系数计算公式3：

判定系数计算公式4：

### 判定系数的性质

1. 是一个非负量
   * ：表示回归线完全不能解释Y的变异
   * ：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**：

**样本相关系数**：

**联系与区别**：

* 在一元回归中，判定系数等于样本相关系数的平方
* 判定系数表明因变量变异由解释变量所解释的比例
* 相关系数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，用表示。

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）：

**学生化标准残差**（Studentized Residuals）：

或

其中： - ：删除第个观测值进行建模的均方误差 - ：删除第个观测值进行建模的第个影响权重 - ：回归元个数

### 残差图分析

**残差图**(residual plot)：用于呈现残差数据的分布情况的统计图图形，主要包括：

1. 关于的残差散点图
2. 关于的残差散点图（或关于）
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：

**均值预测**(mean prediction)： - 给定，预测Y的条件均值

**个值预测**(individual prediction)： - 给定，预测对应于的Y的个别值

### 预测分析的关键

样本外拟合值的性质： - 是均值的一个BLUE - 是个值的一个BLUE

## 均值预测

### 均值预测的分布

在N-CLRM假设和OLS方法下，给定下的拟合值服从如下正态分布：

### 均值预测的置信区间

构造t统计量：

均值的置信区间：

## 个值预测

### 个值预测的分布

在N-CLRM假设和OLS方法下，给定下的个别值服从如下正态分布：

### 个值预测的置信区间

构造新的随机变量的分布：

构造t统计量：

个值的置信区间：

## 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

## 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）：

### 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) |  |  |  |  |
| X |  |  |  |  |

## 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
   * 调整判定系数
   * 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