第十一章 一元线性回归

11.1 变量间的关系

详见 p234

- 函数关系：一一对应，y 随 x 的变化而变化；
- 相关关系：一个变量不由另一变量确定，不确定性关系。

11.1.2 相关关系的描述与测度

两个假定：①线性关系，②都是随机变量。

1. 散点图：正（负）线性相关，完全正（负）线性相关，非线性相关，不相关。
2. 相关系数：根据样本数据计算的度量两个变量之间线性关系强度。下面的公式表示线性相关系数（Pearson 相关系数）。

$$r=\frac{n\sum xy-\sum x\sum y}{\sqrt{n\sum x^2-(\sum x)^2}\cdot\sqrt{n\sum y^2-(\sum y)^2}}$$

   - $r\in[-1,1]$，越小越负相关，越大越正相关，-1 和 1 表示完全负相关和完全正相关。
   - r 绝对值越小，线性相关性越弱
   - r 具有对称性，$r_{xy}=r_{yx}$
   - r 和 x,y 的尺度无关
   - r 仅用于描述线性关系，不一定代表因果关系
   - $|r|\geq 0.8$ 高度相关，$0.5\leq|r|<0.8$ 中度相关，$|r|<0.5$ 低度相关。

11.3 相关关系的显著性检验

将样本相关系数 r 作为整体相关系数 ρ 的估计值，因为 r 具有随机性需要检验。（使用 t 检验）

1. 假设 $H_0:\rho=0;H_1:\rho\neq 0;$
2. 计算检验的统计量

$$t=|r|\sqrt{\frac{n-2}{1-r^2}}\sim t(n-2)$$

3. 比较 t 与 $t_{\alpha/2}$，双侧检验。

11.2 一元线性回归

描述因变量（y）如何依赖于自变量（x）和误差项 ε 而变化的方程被称为回归模型。

$$y=\beta_0+\beta_1x+\varepsilon$$

几个假定：

1. y 与 x 具有线性关系；
2. 重复抽样中 x 的取值是固定的，与随机误差项线性无关；
3. $\varepsilon\sim\mathcal{N}(0,\sigma^2)\ E(\varepsilon)=0;$
4. 对于所有的 x 值，σ_ε^2 相同；

11.2.1 回归方程

$$\begin{aligned}E(y)&=\beta_0+\beta_1x &&\text{线性回归方程}\\ \hat{y}&=\hat{\beta}_0+\hat{\beta}_1x &&\text{估计的线性回归方程}\end{aligned}$$

11.2.2 参数的最小二乘估计

求解过程看P245，易知回归直线一定经过点 $(\bar{x},\bar{y})$。

$$\begin{cases}\hat{\beta}_1=\frac{n\sum_{i=1}^n x_iy_i-\sum_{i=1}^n x_i\sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i^2-(\sum_{i=1}^n x_i)^2}\\ \hat{\beta}_2=\bar{y}-\hat{\beta}_1\bar{x}\end{cases}$$

11.2.3 回归直线的拟合优度

变差： 因变量取值的波动。

几个残差项

- 总平方和：n 次观察的总变差 $SST=\sum(y_i-\bar{y})^2;$
- 回归平方和：由回归直线来解释的变差部分 $SSR=\sum(\hat{y}_i-\bar{y})^2;$
- 残差平方和/残差平方：除了 x 对 y 的线性影响之外其他因素引起的 y 的变化部分 $SSE=\sum(y_i-\hat{y}_i)^2$

$$SST=SSR+SSE$$

---

1. 判定系数 用于度量估计的回归方程的拟合优度，$R^2 \in [0,1]$，越大表示回归直线与观测点越接近。

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{(y_i - \bar{y})^2} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

相关系数 $r = \sqrt{R} \geq R$（仅在 $|R|$ 取值为 0 或 ± 1 时取等号），慎重考虑使用。

2. 估计标准误差 s_e 度量各实际观测点在直线周围散布情况，时均方残差（MSE）的平方根。

$$s_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SSE}{n-2}} = \sqrt{MSE}$$

11.2.4 显著性检验

1. 线性关系的检验

提出假设：$H_0 : \beta_1 = 0; H_1 : \beta_1 \neq 0$，即两个变量之间的线性关系不显著。计算检验统计量 $F \sim \mathcal{F}(1.n-2)$

$$F \sim \frac{SSR/1}{SSE(n-2)} = \frac{MSR}{MSE} \sim F(1, n-2)$$

进行单侧检验，大于 F_α 就拒绝H_0。

2. 回归系数的检验

提出假设：$H_0 : \beta_1 = 0; H_1 : \beta_1 \neq 0$

用因为 $\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \sigma_{\hat{\beta}_1})$，其中

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}} \approx \frac{s_e}{\sqrt{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}}$$

$$s_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SSE}{n-2}} = \sqrt{MSE}$$

得到方差计算 t 统计检验量

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} \sim t(n-2)$$

做出决策，使用双侧检验。

在一元线性回归中 F 检验和 t 检验是等价的，多元回归中有不同的含义。

如果有用到 Excel 进行回归分析，看书 P254（出这种题的讨论一下有母性吧）

11.2.5 回归分析结果的评价

1. 回归系数 $\hat{\beta}_1$ 的符号（正负性）是否与预期结果一致；
2. y 与 x 的关系（正、负，是否显著）在回归方程和理论上是否一致；
3. 回归模型在多大程度上解释了额因变量取得的差异（R^2），超过 2/3 就算效果还不错；

4. 误差项 ε 的正态性假设是否成立。（残差直方图或正态概率图，但这玩意应该不会手画吧？）

11.3 利用回归方程进行预测

对因变量进行合理的预测。

点估计 包括平均值点估计和个别值点估计，都是给你一个值然后往回归方程里代就好了。

平均值的点估计实际上是对总体参数的估计，个别值的点估计是对因变量的某个具体取值的估计。

11.3.2 区间估计

区间估计包括置信区间估计和预测区间估计，分别是根据一个给定值x_0 计算 y 的平均值的估计区间和根据一个给定值 x_0 求出 y 的个别值的估计区间。

1. 平均值的置信区间估计

先求标准差

$$s_{\hat{y}_0} = s_e\sqrt{\frac{1}{n}+\frac{(x-x_0)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

$$s_e = \sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2}} = \sqrt{\frac{SSE}{n-2}} = \sqrt{MSE}$$

然后用$t_{\alpha/2}$ 得到置信区间上下限为

$$\hat{y}_0 \pm t_{\alpha/2}s_{\hat{y}_0}$$

2. 个别值的预测区间估计

先求标准差，和置信区间的表达式相比，根号下加上一个 1 。

$$s_{ind} = s_e\sqrt{1+\frac{1}{n}+\frac{(x-x_0)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

预测区间比置信区间要宽一点。

11.4 残差分析

11.4.1 残差与残差图

残差是观测值 y_i 与回归方程求出的预测值 $\hat{y}_i$ 之差，即 $e_i = y_i - \hat{y}_i$。

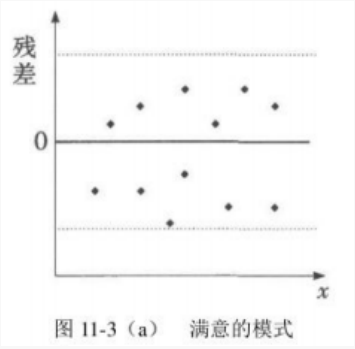

图 11-3（a） 满意的模式

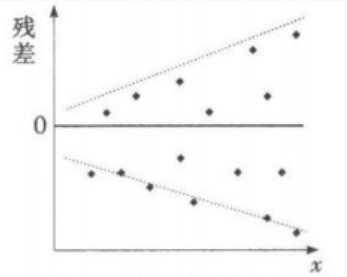

图 11-3（b） 非常数方差

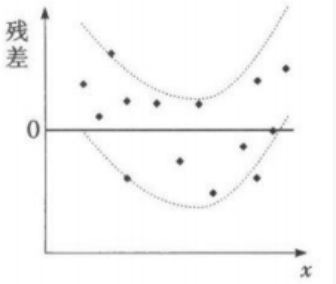

图 11-3（c） 模型形式不合适

11.4.2 标准化残差

$$z_{e_i} = \frac{e_i}{s_e} = \frac{y_i - \hat{y}_i}{s_e}$$

如果误差项 ε 服从正态分布，那么标准化残差的分布也服从正态分布，在标准化残差的图中，大约 95% 的标准化残差在 $[-2, 2]$。