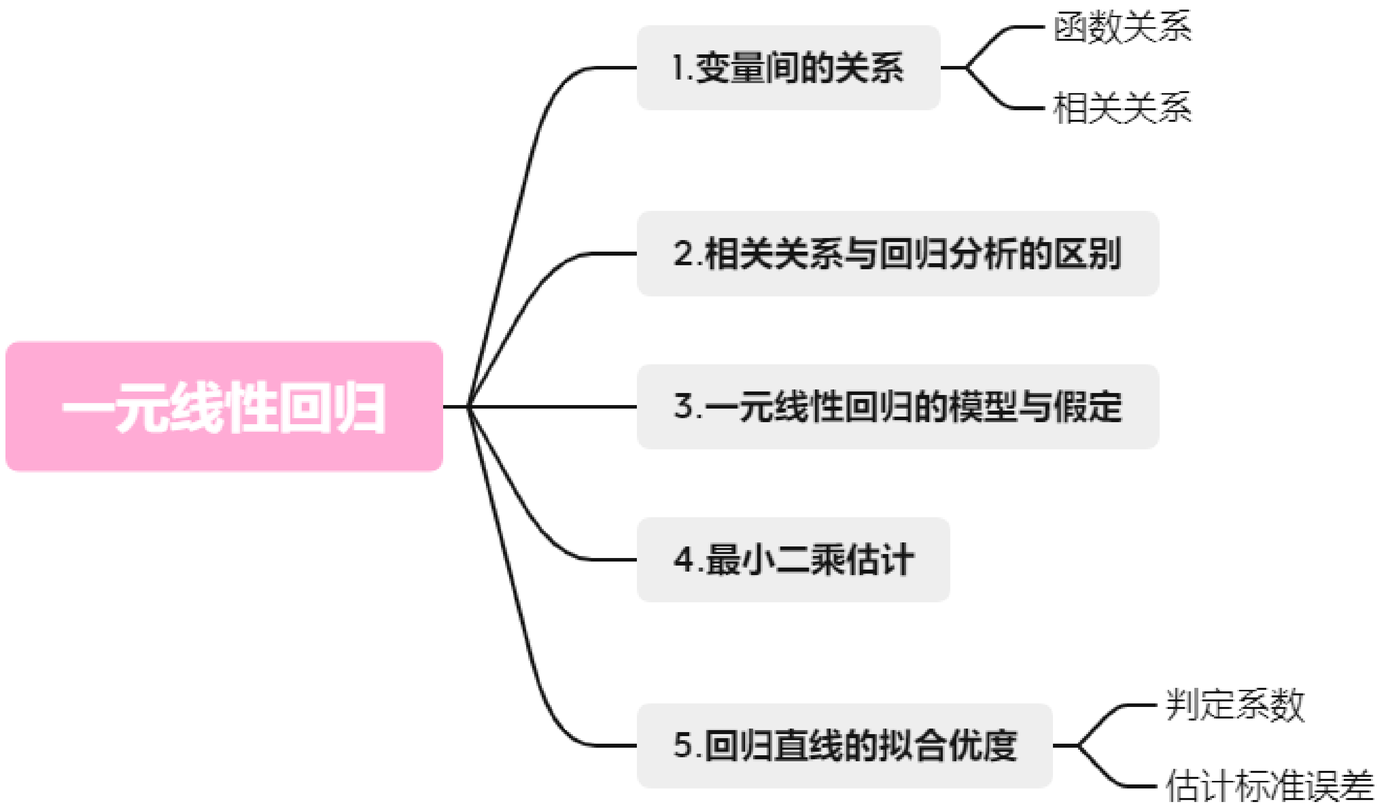
第二十五节：必备模型及其应用（二）

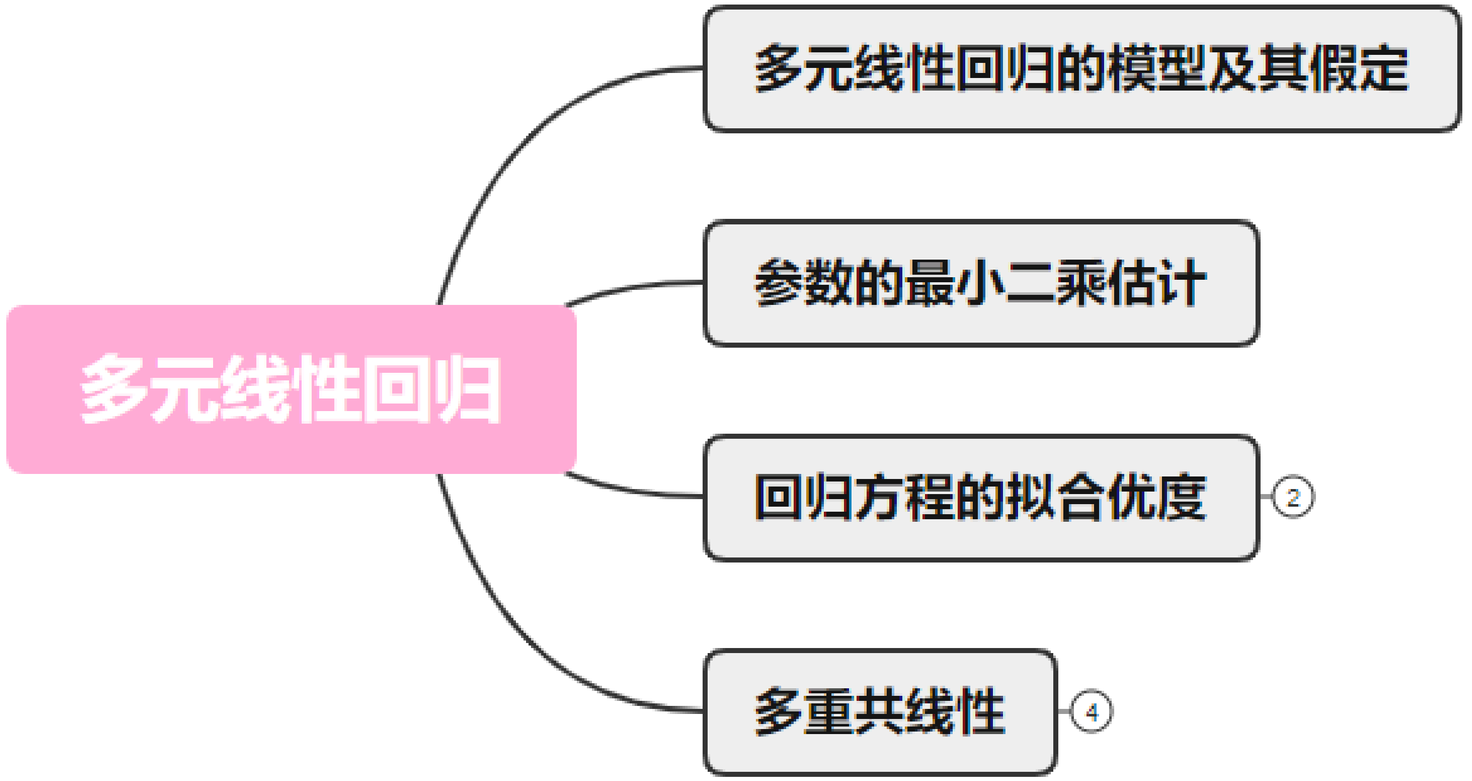
主讲老师 侯梓熙

上节课程回顾


一元线性回归
1.变量间的关系
函数关系
相关关系
2.相关关系与回归分析的区别
3.一元线性回归的模型与假定
4.最小二乘估计
5.回归直线的拟合优度
判定系数
估计标准误差

多元线性回归

本节课程内容



多元线性回归的提出与例子

提出

现实生活中引起被解释变量变化的因素并非仅只一个解释变量,可能有很多个解释变量.

例子

1.销售额与销售价格和广告费;

2.北京市房价与总人口数, 平均工资, 北京地区生产总值;

3.身高与性别,父母身高和营养状况等.

多元线性回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

$\beta_0, \beta_1, \beta_2, \cdots, \beta_k$称为模型的参数, ε称为误差.
其中

1) y是$x_1, x_2 \cdots, x_k$的线性函数($\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$部分)加上误差项.
2) 误差项反映了除$x_1, x_2 \cdots, x_k$对y的线性关系之外的随机因素对y的影响,是不能由$x_1, x_2 \cdots, x_k$与y之间的线性关系所解释的y的变异.

多元线性回归模型的假定

1) 正态性：ε是一个服从正态分布的随机变量，且期望值0，即 $E(\varepsilon) = 0$.

2) 方差齐性：对于自变量$x_1, x_2, \cdots, x_k$的所有值，ε的方差σ^2都相同。

3)独立性：对于自变量$x_1, x_2, \cdots, x_k$一组特定值，它所对应的ε与$x_1, x_2, \cdots, x_k$任意一组其他值所对应的ε相关。

多元线性回归方程

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x_2 + \cdots + \beta_k x_k$$

估计的多元线性回归方程

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

参数的最小二乘估计

根据最小二乘法，使

$$\sum(y_i - \hat{y}_i)^2 = \sum\left(y_i - \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k\right)^2$$

最小。

回归方程的拟合优度

01 多重判定系数

02 估计标准误差

多重判定系数

$$R^2 = \frac{SSR}{SST}$$

与一元回归类似，多元回归中因变量离差平方和的分解也一样：

$$SST = SSR + SSE$$

式中，$SST = \sum(y_i - \bar{y})^2$为总平方和；

$SSR = \sum(\hat{y}_i - \bar{y})^2$为回归平方和；

$SSE = \sum(y_i - \hat{y}_i)^2$为残差平方和。

注：自变量个数的增加将影响到因变量中被估计的回归方程所解释的变差数量，当增加自变量时，会使预测误差变得较小，从而减少残差平方和SSE.

调整的多重判定系数

$$R_a^2 = 1 - (1 - R^2)(\frac{n-1}{n-k-1})$$

称为调整的多重判定系数，R_a^2的解释与R^2类似，不同的是: R_a^2同时考虑了样本量和模型中的自变量个数，这就使得R_a^2的值永远小于R^2，而且R_a^2的值不会由于模型中的自变量个数的增加而越来越接近1.

估计标准误差

$$S_e = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-k-1}} = \sqrt{\frac{SSE}{n-k-1}}$$

式中，k为自变量的个数，其含义是根据自变量$x_1, x_2, \cdots, x_k$来预测因变量y时的平均预测误差。

多重共线性

Why

为什么存在多重共线性?

当回归模型中使用两个或两个以上的自变量时，这些自变量往往会提供多余的信息；也就是说自变量之间彼此相关。

Problem

多重共线性会产生什么问题?

① 变量之间高度相关，可能会使回归的结果混乱，甚至会把分析引入歧途。

② 多重共线性可能对参数的估计值的正负号产生影响。

多重共线性

Distinguish

多重共线性的判别

① 模型中对各自变量之间显著相关。
② 当模型的线性关系检验显著时，几乎所有回归系数β_i的t检验却不显著。
③ 回归系数的正负号与预期相反。

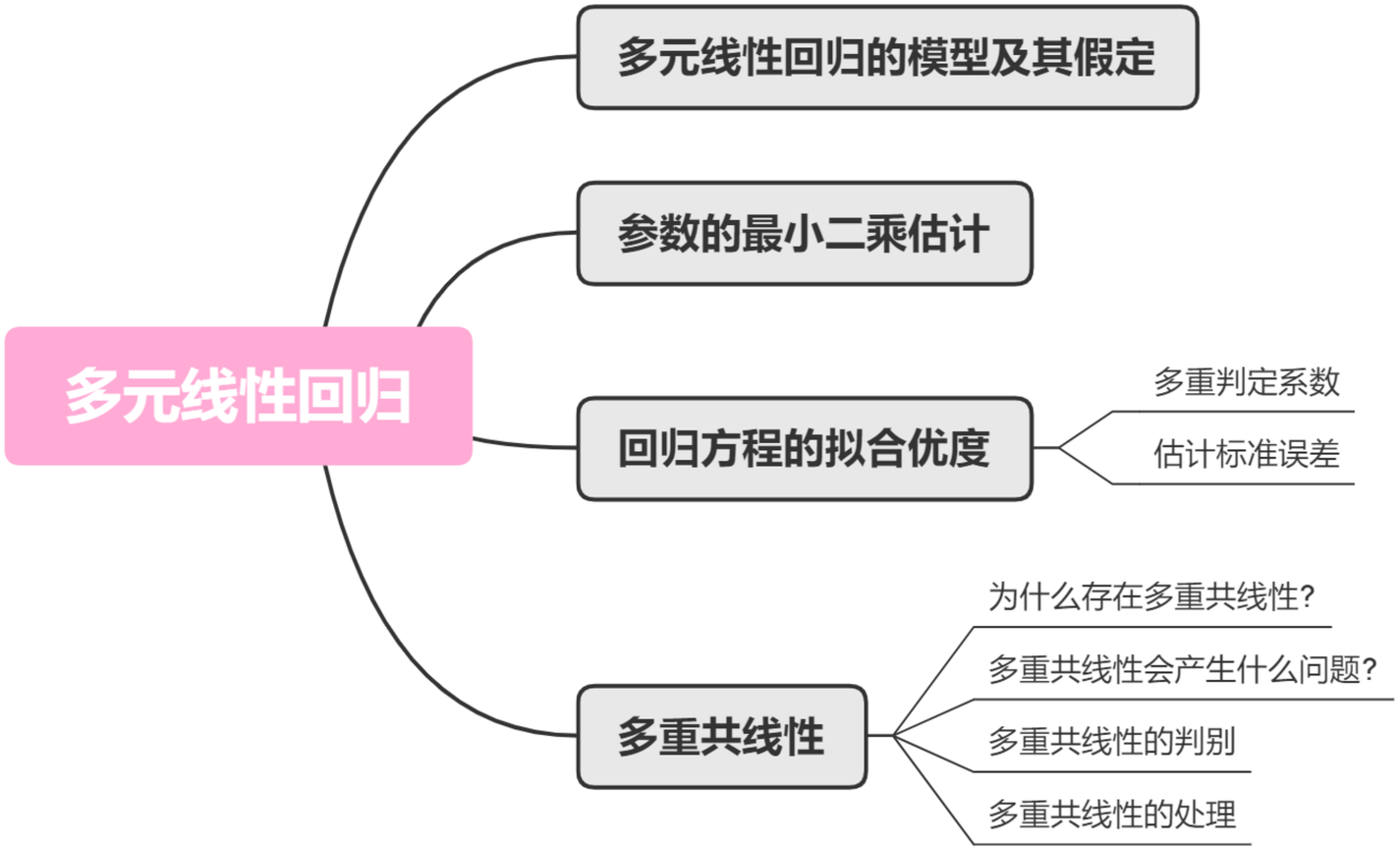
Solve

多重共线性问题的处理

将一个或多个相关的自变量从模型中剔除，使保留的自变量尽可能不相关。

特别提醒：在建立多元线性回归模型时，不要试图引入更多的自变量，除非确实有必要。特别是在社会科学的研究中，由于所使用的大多数数据都是非实验性质的，因此在某些情况下得到的结果往往并不令人满意，但这不一定是因为选择的模型不合适，而是数据的质量不好，或者是引入的自变量不合适。

本节课程回顾


多元线性回归
多元线性回归的模型及其假定
参数的最小二乘估计
回归方程的拟合优度
多重判定系数
估计标准误差
多重共线性
为什么存在多重共线性?
多重共线性会产生什么问题?
多重共线性的判别
多重共线性的处理

THANK YOU

感谢观看

主讲老师 侯梓熙