数模加油站

数学建模模型算法精讲课——

回归分析概述

—— 江北老师

我们各自努力，
高处见

回归分析概述

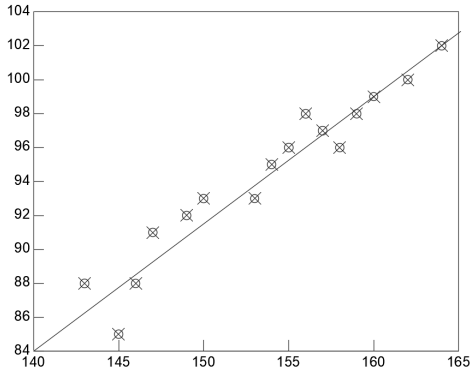
- 模型引出
- 模型原理

回归分析——模型引出

➢ **某团队测了16名成年女子的身高与腿长所得数据如下**

身高	143	145	146	147	149	150	153	154	155	156	157	158	159	160	162	164
腿长	88	85	88	91	92	93	93	95	96	98	97	96	98	99	100	102

- 我们发现，腿长和身高有很强的**正相关性**，身高越高，腿长越长，那么通过这组数据，我们可以预测身高170的成年女性腿长是多长吗？



- 我们还发现，测得的点大致分布在一个直线上，说明身高和腿长为**线性关系**
- 那如果我们找到了这条直线的表达式，是不是就可以预测身高170的腿长了？
- 一元的线性表达式为$y = ax + b$，通过计算我们得到

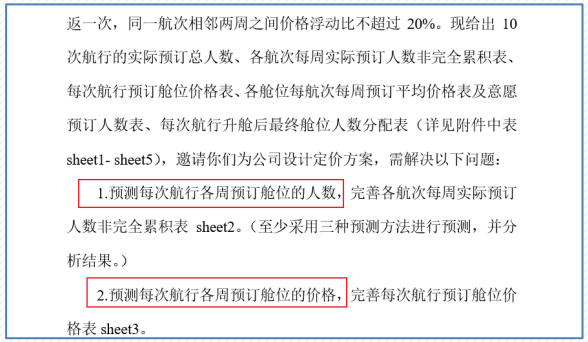
$y = 0.7194x - 16.0730, x = 170$时，$y = 106.225$

- 那么a和b是怎么求得的，这个表达式的准确性该怎么评判？误差到底有多大，就是我们要思考的问题了

回归分析——模型引出

➢ 回归分析

- 在统计学中，回归分析（regression analysis)指的是确定**两种或两种以上变量间相互依赖的定量关系**的一种统计分析方法。回归分析按照涉及的变量的多少，分为一元回归和多元回归分析；按照因变量的多少，可分为简单回归分析和多重回归分析；按照自变量和因变量之间的关系类型，可分为线性回归分析和非线性回归分析。
- 在大数据分析中，回归分析是一种**预测性**的建模技术，它研究的是因变量（目标）和自变量（预测器）之间的关系。这种技术**通常用于预测分析**以及发现变量之间的因果关系。例如，司机的鲁莽驾驶与道路交通事故数量之间的关系，最好的研究方法就是回归。

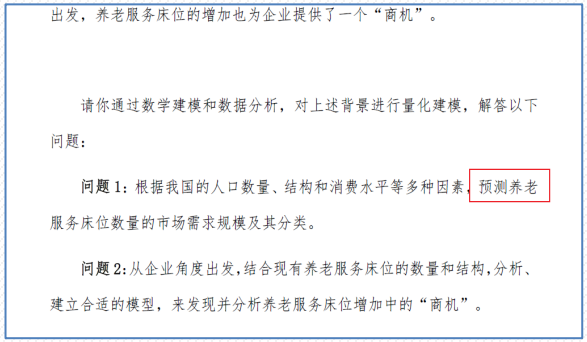


返一次，同一航次相邻两周之间价格浮动比不超过 20%。现给出 10 次航行的实际预订总人数、各航次每周实际预订人数非完全累积表、每次航行预订舱位价格表、各舱位每航次每周预订平均价格表及意愿预订人数表、每次航行升舱后最终舱位人数分配表（详见附件中表 sheet1- sheet5），邀请你们为公司设计定价方案，需解决以下问题：

1.预测每次航行各周预订舱位的人数，完善各航次每周实际预订人数非完全累积表 sheet2。（至少采用三种预测方法进行预测，并分析结果。）

2.预测每次航行各周预订舱位的价格，完善每次航行预订舱位价格表 sheet3。

电工杯赛题



出发，养老服务床位的增加也为企业提供了一个“商机”。

请你通过数学建模和数据分析，对上述背景进行量化建模，解答以下问题：

问题 1：根据我国的人口数量、结构和消费水平等多种因素，预测养老服务床位数量的市场需求规模及其分类。

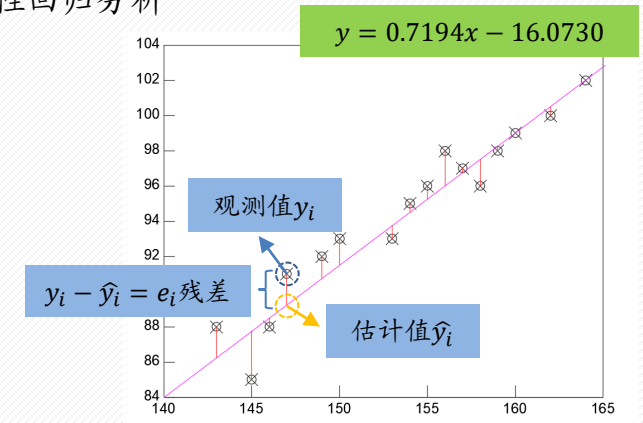
问题 2：从企业角度出发，结合现有养老服务床位的数量和结构，分析、建立合适的模型，来发现并分析养老服务床位增加中的“商机”。

MathorCup赛题

回归分析——模型原理

➢ 回归分析

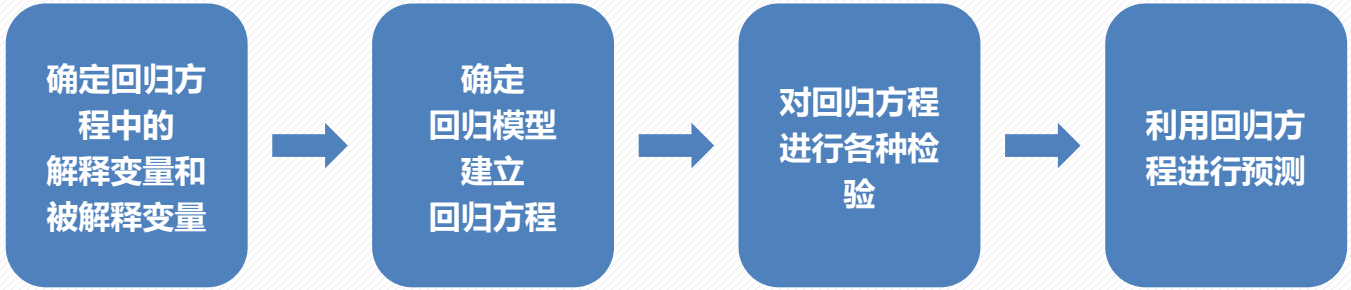
- 相关分析：研究两个或两个以上的变量之间**相关程度及大小**的一种统计方法
- 回归分析：寻找存在**相关关系的变量间**的数学表达式，并进行统计推断的一种统计方法
- 前面也说了，回归分析有两种分类方式
  - ✓ 根据变量的数目可以分为**一元回归、多元回归**
  - ✓ 根据自变量与因变量的表现形式，分为**线性和非线性**
- 所以，回归分析包括四个方向：一元线性回归分析、多元线性回归分析、一元非线性回归分析、多元非线性回归分析



- $y = 0.7194x - 16.0730$为构造的回归方程
- 回归方程计算得到的值为估计值
- 观测值（实际值）与估计值的差为**残差**
- 我们肯定是希望残差**越小越好**

回归分析——模型原理

➢ 回归分析的一般步骤



➢ 回归分析基本概念

- 因变量：被**预测或被解释**的变量，用y表示
- 自变量：**预测或解释**因变量的一个或者多个变量，用x表示
- 对于只有**线性关系**的两个变量，可以用一个方程来表示它们之间的线性关系
- 描述因变量y如何依赖于自变量x和误差项ϵ的方程称为**回归模型**

回归分析——模型原理

➢ 以一元线性回归分析为例

- 对于只涉及一个自变量的**一元线性回归模型**可表示为

$$y = \beta_0 + \beta_1 x + \epsilon$$

- 在这个模型里：
  - ✓ y叫做因变量或被解释变量
  - ✓ x叫做自变量或解释变量
  - ✓ β_0表示**截距**
  - ✓ β_1表示**斜率**
  - ✓ ϵ表示**误差项**，反映除x和y之间的线性关系之外的随机因素对y的影响，是不可避免的
- 一元的例子：
  - ✓ 人均收入是否会显著影响人均食品消费支出
  - ✓ 贷款余额是否影响到不良贷款
  - ✓ 航班正点率是否对顾客投诉次数有显著影响

回归分析——模型原理

➢ 回归方程

- 描述因变量y的期望值如何依赖于自变量x的方程称为回归方程。根据对一元线性回归模型的假设，可以得到它的回归方程为：

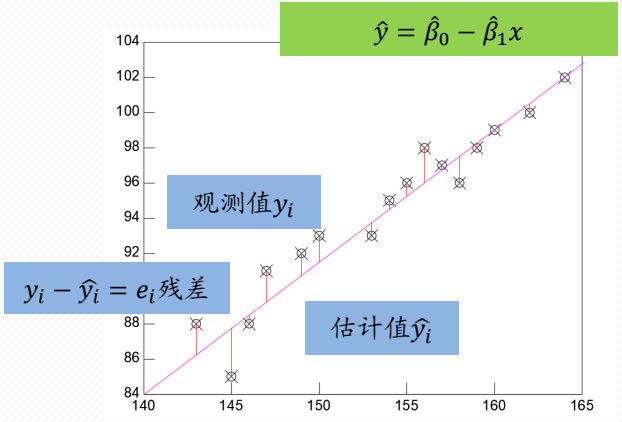
$$E(y) = \beta_0 + \beta_1 x$$

- 如果回归方程中的参数已知，对于一个给定的x值，利用回归方程就能计算出y的期望值
- 用样本统计量代替回归方程中的未知参数，就得到估计的回归方程，简称回归直线

➢ 参数的最小二乘法估计

- 对于回归直线，关键在于求解参数，常用高斯提出的最小二乘法，它是使因变量的观察值y与估计值之间的离差平方和**达到最小**来求解的

$$Q = \sum(y - \hat{y})^2 = \sum\left(y - \hat{\beta}_0 - \hat{\beta}_1 x\right)^2$$



回归分析——模型原理

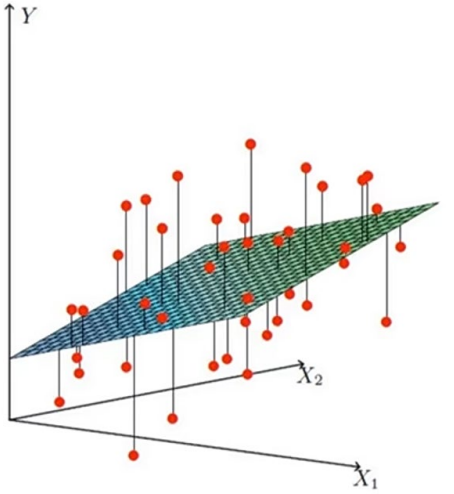
➢ 相关数学知识

- 假设β_1是工资的参数，β_2是存款的参数
- 拟合一个平面：

$$h(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- h可表示银行的信贷评级
- 整合后：

$$h(x) = \sum_{i=0}^{n} \beta_i x_i = \beta^T x$$



➢ 误差

- 真实值和预测值之间肯定是要存在差异的（用ϵ来表示误差）
- 对于每个样本：$y^{(i)} = \beta^T x^{(i)} + \epsilon^{(i)}$

回归分析——模型原理

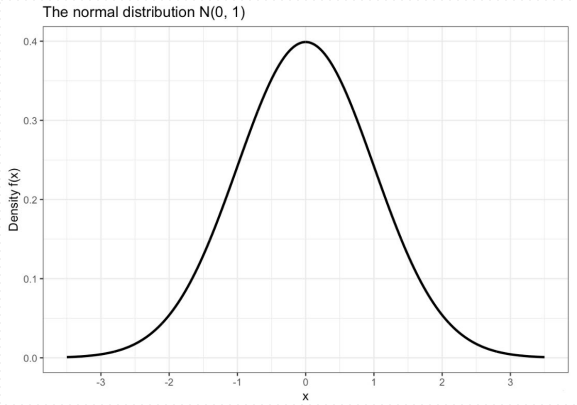
➢ 误差

- 误差$\epsilon^{(i)}$是独立并且具有相同的分布，并且服从均值为0方差为σ^2的**正态分布**

  ✓ 独立：我和好兄弟一起去贷款，无法互相影响

  ✓ 同分布：我们都来得是这家银行

  ✓ 正态分布：银行可能会多贷点，也可能会少贷点，但是绝大多数情况下，这个浮动不会太大，极小情况下浮动会比较大，符合正常的情况



- 预测值与误差：$y^{(i)} = \beta^T x^{(i)} + \epsilon^{(i)}$ (1)
- 正态分布表达式：$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(x-\mu)^2}{2\sigma^2}$
- 因为误差服从正态分布，且期望为0：$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(\epsilon^{(i)})^2}{2\sigma^2}\right)$ (2)
- 将(1)式带入(2)式：$p(y^{(i)} - \beta^T x^{(i)}) = p(y^{(i)}|x^{(i)};\beta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y^{(i)}-\beta^T x^{(i)})^2}{2\sigma^2}\right)$

➢ 引入似然函数的概念

- 统计学中，似然函数（Likelihood function），或简称似然，是一种**关于统计模型参数的函数**。给定输出x时，关于参数θ的似然函数$L(\theta|x)$（在数值上）等于给定参数θ后变量X的概率：$L(\theta|x) = P(X = x|\theta)$. 似然函数在推断统计学（Statistical inference）中扮演重要角色，尤其是在参数估计方法中。

- 设总体X服从分布$P(x;\theta)$（当X是连续型随机变量时为概率密度，当X为离散型随机变量时为概率分布），θ为待估参数，$X_1, X_2, \dots X_n$是来自于总体X的样本，$x_1, x_2 \dots x_n$为样本$X_1, X_2, \dots X_n$的一个观察值，则样本的联合分布

$$L(\theta) = L(x_1, x_2, \dots, x_n; \theta) = \prod P(x_i; \theta)$$称为似然函数。

- 刚刚我们得到了：$p\left(y^{(i)} - \beta^T x^{(i)}\right) = p\left(y^{(i)}|x^{(i)};\beta\right) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-\left(y^{(i)}-\beta^T x^{(i)}\right)^2}{2\sigma^2}\right)$

- 那么可得似然函数：$L(\beta) = \prod_{i=1}^{m} p\left(y^{(i)}|x^{(i)};\beta\right) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-\left(y^{(i)}-\beta^T x^{(i)}\right)^2}{2\sigma^2}\right)$

- 对数似然（乘法转变成加法）：$\log L(\beta) = \log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-\left(y^{(i)}-\beta^T x^{(i)}\right)^2}{2\sigma^2}\right)$

➢ **取对数后得到**

$$\log L(\beta) = \log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-\left(y^{(i)} - \beta^T x^{(i)}\right)^2}{2\sigma^2}\right)$$

- 展开化简：$\log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-\left(y^{(i)} - \beta^T x^{(i)}\right)^2}{2\sigma^2}\right)$

$$= \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-\left(y^{(i)} - \beta^T x^{(i)}\right)^2}{2\sigma^2}\right)$$

$$= m\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} \cdot \frac{1}{2} \sum_{i=1}^{m} \left(y^{(i)} - \beta^T x^{(i)}\right)^2$$

- 我们需要似然函数越大越好，那么$\frac{1}{2\sigma^2} \sum_{i=1}^{m} \left(y^{(i)} - \beta^T x^{(i)}\right)^2$越小越好
- $J(\beta) = \frac{1}{2} \sum_{i=1}^{m} \left(y^{(i)} - \beta^T x^{(i)}\right)^2 = \frac{1}{2} \sum_{i=1}^{m} \left(y^{(i)} - \hat{y}^{(i)}\right)^2$（最小二乘法）

➢ 继续看回归分析

- 前面说到，对于回归直线，关键在于求解参数，常用高斯提出的最小二乘法，它是使因变量的观察值y与估计值之间的离差平方和达到最小来求解的

$$Q = \sum(y - \hat{y})^2 = \sum\left(y - \hat{\beta}_0 - \hat{\beta}_1 x\right)^2$$

- 展开可得:

$$Q = \sum(y - \hat{y})^2 = \sum y^2 + n\hat{\beta}_0^2 + \hat{\beta}_1^2 \sum x^2 + 2\hat{\beta}_0\,\hat{\beta}_1 \sum x - 2\hat{\beta}_0 \sum y - 2\hat{\beta}_1 \sum xy$$

- 求偏导可得:

$$\begin{cases} \sum y = n\hat{\beta}_0 + \hat{\beta}_1 \sum x \\ \sum xy = 2\hat{\beta}_0 \sum x + \hat{\beta}_1 \sum x^2 \end{cases}$$

$$\Rightarrow \begin{cases} \hat{\beta}_1 = \dfrac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

- 即可得到$\hat{\beta}_0$及$\hat{\beta}_1$

➢ 我们来看开头提到的问题

身高	143	145	146	147	149	150	153	154	155	156	157	158	159	160	162	164
腿长	88	85	88	91	92	93	93	95	96	98	97	96	98	99	100	102

- 我们发现，腿长和身高有很强的正相关性，身高越高，腿长越长，那么通过这组数据，我们可以预测身高170的成年女性腿长是多长吗?
- 一元的线性表达式为$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$$\begin{cases} \hat{\beta}_1 = \dfrac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \dfrac{16 \times 232566 - 2458 \times 1511}{16 \times 378220 - 2458^2} = 0.7194 \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 94.4375 - 0.7194 \times 153.625 = -16.08 \end{cases}$$

- $\hat{y} = 0.7194x - 16.08$
- $x = 170$，$\hat{y} = 106.218$
- 预测身高170的成年女性腿长是106.218cm

回归分析——模型原理

➢ 利用回归直线进行估计和预测

- 点估计：利用估计的回归方程，对于x的某一个特定的值，求出y的**一个估计值**就是点估计
- 区间估计：利用估计的回归方程，对于x的一个特定值，求出y的**一个估计值的区间**就是区间估计

➢ 估计标准误差的计算

- 为了度量回归方程的可靠性，通常计算**估计标准误差**，它度量观察值回绕着回归直线的变化程度或分散程度
- 估计平均误差：

$$S_e = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}$$

- 估计标准误差**越大**，则数据点围绕回归直线的**分散程度就越大**，回归方程**代表性就越小**
- 估计标准误差**越小**，则数据点围绕回归直线的**分散程度越小**，回归方程的**代表性越大，可靠性越高**

回归分析——模型原理

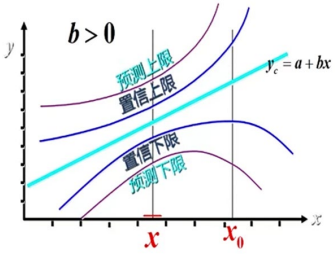
➢ 置信/预测区间估计

- 置信区间：$\hat{y}_0 \pm t_{\frac{\alpha}{2}} s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x - \bar{x})^2}}$
- 预测区间：$\hat{y}_0 \pm t_{\frac{\alpha}{2}} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x - \bar{x})^2}}$

➢ 我们继续来算腿（显著性水平$\alpha = 0.05$，置信水平1-$\alpha = 0.95$）

身高	143	145	146	147	149	150	153	154	155	156	157	158	159	160	162	164
腿长	88	85	88	91	92	93	93	95	96	98	97	96	98	99	100	102

- $\hat{y} = 0.7194x - 16.08$
- $x = 170$，$\hat{y} = 106.218$
- $t_{\frac{\alpha}{2}}(n-k) = t_{0.025}(16-2) = t_{0.025}(14) = 2.145$（查表）
- $s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}} = \sqrt{1.744} = 1.32$



- $\hat{y}_0 \pm t_{\frac{\alpha}{2}} s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x - \bar{x})^2}} = 106.218 \pm 2.007$
- $\hat{y}_0 \pm t_{\frac{\alpha}{2}} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x - \bar{x})^2}} = 106.218 \pm 3.472$

➢ 回归直线的拟合优度

- 回归直线与各观测点的接近程度称为**回归直线对数据的拟合优度**

✓ **总平方和（TSS）**：反映因变量的n个观察值与其均值的总离差

$$TSS = \sum y_i^2 = \sum (y_i - \bar{y}_i)^2$$

✓ **回归平方和（ESS）**：反映了y的总变差中，由于x与y之间的线性关系引起的y的变化部分

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

✓ **残差平方和（RSS）**：反映了除了x对y的线性影响之外的其他因素对y变差的作用，是不能由回归直线来解释的y的变差部分

$$RSS = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

- 总平方和可以分解为**回归平方和、残差平方和**两部分：$TSS = ESS + RSS$

回归分析——模型原理

➢ 判定系数

- **回归平方和占总平方和的比例**，用R^2表示，其值在0到1之间

$$R^2 = \frac{\mathrm{ESS}}{TSS} = 1 - \frac{\mathrm{RSS}}{TSS} = \frac{\sum(\hat{y}_i - \bar{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} = 0.9282$$，是比较接近与1的

- R^2=0：说明y的变化与x无关，x完全无助于解释y的变差
- R^2=1：说明残差平方和为0，拟合是完全的，y的变化只与x有关

➢ 显著性检验

- **显著性检验**的主要目的是根据所建立的估计方程用自变量x来估计或预测因变量y的取值，当建立了估计方程后，还不能马上进行估计或预测，因为该估计方程是根据样本数据得到的，它是否真实的反映了变量x和y之间的关系，**则需要通过检验后才能证实**
- 根据样本数据拟合回归方程时，实际上就已经假定变量x和y之间存在着线性关系，并**假定误差项是一个服从正态分布的随机变量**，且具有相同的方差，但这些假设是否成立需要检验
- 显著性检验包括两方面
  - ✓ 线性关系检验
  - ✓ 回归系数检验

➢ 线性关系检验

- **线性关系检验**是检验自变量x和因变量y之间的线性关系是否显著，或者说，它们之间能否用一个线性模型来表示
- 将**均方回归（MSR）同均方残差（MSE）**加以比较，应用F检验来分析二者之间的差别是否显著
  - ✓ 均方回归（MSR）：回归平方和ESS除以相应的回归自由度（自变量的个数k）
  - ✓ 均方残差（MSE）：残差平方和RSS除以相应的残差自由度（$n-k-1$）
- H_0（原假设）：$\beta_1 = 0$，回归系数与0无显著差异，y与x的线性关系不显著
- H_1：$\beta_1 \neq 0$，回归显著，认为 y 与x 存在 线性关系，所求的线性回归方程有意义
- 计算检验统计量F:

H_0成立时，$F = \dfrac{ESS/1}{RSS/(n-2)} = \dfrac{MSR}{MSE} \sim F(1, n-2)$

若$F > F_{1-\alpha}(1, n-2)$，拒绝H_0，否则接受H_0

$F = 180.9531 > 4.6$，所以线性关系显著

➢ 回归系数的显著性检验

- 回归系数显著性检验的目的是通过检验回归系数β的值与0是否有显著性差异，来判断Y与X之间是否有显著的线性关系，若$\beta = 0$，则总体回归方程中不含X项（即Y不随X变动而变动），因此，变量Y与X之间并不存在线性关系；若$\beta \neq 0$，说明变量Y与X之间存在显著的线性关系
- $\hat{\beta}_1$是根据最小二乘法求出的样本统计量，服从正态分布
- $\hat{\beta}_1$的分布具有如下性质：

  数学期望：$E(\hat{\beta}_1) = \beta_1$

  标准差：$\sigma_{\hat{\beta}_1} = \dfrac{\sigma}{\sqrt{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}}$

- 由于σ未知，需用其估计量S_e来代替得到$\hat{\beta}_1$的估计标准差

$$S_{\hat{\beta}_1} = \frac{S_e}{\sqrt{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}} \qquad S_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-k-1}} = \sqrt{MSE}$$

- t检验的统计量：$t = \dfrac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \sim t(n-2)$

线性关系检验与回归系数检验的区别：

- 线性关系的检验是检验自变量与因变量是否可以用线性来表达，而回归系数的检验是对样本数据计算的回归系数检验是否为0

✓ 在一元线性回归中，自变量只有一个，**线性关系检验与回归系数检验是等价的**

✓ 在多元回归分析中，这两种检验的意义是不同的。线性关系检验只能用来检验**总体回归关系**的显著性，而回归系数检验**可以对各个回归系数**分别进行检验

➢ 多元回归分析

- 经常会遇到某一现象的发展和变化取决于几个影响因素的情况，也就是一个因变量和几个自变量有依存关系的情况，这时需用**多元线性回归分析**
  - ✓ 多元线性回归分析预测法，是指通过对两个或两个以上的自变量与一个因变量的相关分析，建立预测模型进行预测和控制的方法
  - ✓ 多元线性回归预测模型一般式为：

$$y = A + B_1 x_1 + B_2 x_2 + B_3 x_3 + \cdots + B_n x_n + \epsilon$$

- 调整的**多重判定系数**：
  - ✓ 用样本容量n和自变量的个数k去修正R^2得到

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

  - ✓ 避免增加自变量而高估R^2

➢ 曲线回归分析

- 直线关系是两变量间最简单的一种关系，**曲线回归分析**的基本任务是通过两个相关变量x与y的实际观测数据建立曲线回归方程，以揭示x与y间的曲线联系的形式
- 曲线回归分析最困难和首要的工作是确定自变量与因变量间的曲线关系的类型，曲线回归分析的基本过程
  - ✓ 先将x与y进行变量转换
  - ✓ 对新变量**进行直线回归分析**、建立直线回归方程并进行显著性检验和区间估计
  - ✓ 将新变量还原为原变量，由新变量的直线回归方程和置信区间得出原变量的曲线回归方程和置信区间
- 由于曲线回归模型种类繁多，所以没有通用的回归方程可直接使用，但是对**某些特殊的回归模型**，可以通过变量代换、取对数等方法将其线性化，然后使用标准方程求解参数，再将参数带回原方程就是所求
- 如有一组数据x,y数据，画出散点图后显示x与y的变动关系为一条递减的双曲线

  设$\hat{y}=\hat{a}+\hat{b}\frac{1}{x}$

  令$\frac{1}{x}=x'$，原式变为$\hat{y}=\hat{a}+\hat{b}x'$
- 把原始数据转换为$\frac{1}{x},y$，求解出$\hat{a}$、$\hat{b}$，在把$x'=\frac{1}{x}$带入回去得到曲线回归方程

➢ 多重共线性

- 回归模型中**两个或两个以上的自变量彼此相关**的现象
- 多重共线性带来的问题有：
  - ✓ 回归系数**估计值的不稳定**增强
  - ✓ 回归系数**假设检验的结果不显著**等

➢ 多重共线性检验的主要方法

- 容忍度

$$Tol_i = 1 - R^2$$

  - ✓ R_i是解释变量x_i与方程中其他解释变量间的负相关系数
  - ✓ 容忍度在0~1之间，越接近与0，表示多重共线性越强，越接近于1，表示多重共线性越弱
- 方差膨胀因子（VIF）

$$VIF_i = \frac{1}{1 - R_i^2}$$

  - ✓ VIF_i越大，特别是大于等于10，说明解释变量x_i与方程中其他解释变量之间有严重的多重共线性
  - ✓ VIF_i越接近1，表明解释变量x_i和其他解释变量之间的多重共线性越弱

欢迎关注数模加油站

THANKS