应用统计学II第3讲 多元线性回归模型

Instructor: 郝壮

haozhuang@buaa.edu.cn
School of Economics and Management
Beihang University

May 18, 2022

多元线性回归模型 Multiple Regression Model

本节对应教材第十章.
本章推荐参考教材: 伍德里奇. 计量经济学导论: 现代观点 (第六版). 人民大学出版社.

或

英文原版 Wooldridge, J. M. (2016). Introductory econometrics: A modern approach. Nelson Education.

多元线性回归模型

- 因变量只受单一自变量影响的情况非常少见
- 通常影响一个变量的变量有多个
- 一元线性回归 $\Rightarrow$ 多元线性回归

两个自变量

我们先假设有两个自变量:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, i = 1, \cdots, n$$

其中 $\varepsilon_i \sim N\left(0, \sigma^2\right)$.

同时要求随机误差项ε_i无序列相关 $Cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j$, 且自变量间无严重多重相关(multicolinearity)

同样可以用OLS估计位置参数.

最小化残差平方和:

$$f\left(\beta_0,\beta_1,\beta_2\right)=\sum_{i=1}^{n}\left(y_i-\left(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}\right)\right)^2$$

然后根据优化理论:

$$\frac{\partial f}{\partial \beta_0}=0$$
$$\frac{\partial f}{\partial \beta_1}=0$$
$$\frac{\partial f}{\partial \beta_2}=0$$

多元线性回归模型

一般化地, 多元线性回归模型总体模型可写为：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

样本模型：

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, ..., n$$

其中随机误差$\varepsilon_i \sim N\left(0, \sigma^2\right)$

- p 个自变量
- $p + 2$ 个待估参数

多元线性回归模型-矩阵形式

采用矩阵的写法, 可以把上述线性回归模型写成矩阵形式, 记为

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

或

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

其中, $\mathbf{Y}_{n\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$, $\mathbf{X}_{n\times(p+1)} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$,

$\boldsymbol{\beta}_{(p+1)\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$, $\boldsymbol{\varepsilon}_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$

式中, Y 为因变量观测值向量; β 为总体参数向量; X 为自变量观测值矩阵, 它是一个 $n \times (p+1)$ 维的数据表; ε 为随机误差向量.

高斯—马尔科夫假定: 对于随机误差项 ε, 假设它是一个多元正态且独立的随机向量, 其期望值 $E(\varepsilon) = 0$, 方差-协方差矩阵

$$\operatorname{cov}(\varepsilon) = E\left(\varepsilon\varepsilon^{\mathrm{T}}\right) = \sigma^2 \boldsymbol{I}_{n\times n}$$

简写为

$$\varepsilon \sim N\left(0, \sigma^2 \boldsymbol{I}_{\boldsymbol{n}\times\boldsymbol{n}}\right)$$

上式假设随机误差项之间是互不相关的. 此外, 还假设随机误差项与自变量之间也是互不相关的(在固定设计假设下显然).

根据高斯—马尔科夫假定, 随机向量 $\boldsymbol{Y}$ 的条件期望值为

$$E(\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$$

一般最小二乘法(OLS)

OLS估计是求使得残差平方和最小化的β的估计量$\hat{\beta}$, 即

$$f\left(\beta_0, \beta_1, \beta_2, \cdots, \beta_p\right) = \sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\right)\right)^2$$

一阶条件(FOC):b

$$\begin{cases} -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}\right) = 0 \\ -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}\right) x_{i1} = 0 \\ \vdots \\ -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}\right) x_{ip} = 0 \end{cases}$$

以上一阶条件化简, 可写成矩阵形式:

$$X^\top X\hat{\beta} = X^\top Y$$

然后我们可以得到OLS参数估计:

$$\hat{\beta}_{OLS} = \left(X^\top X\right)^{-1} X^\top Y$$

拓展知识： 计量经济学中学到多元回归**矩估计**时, 用$E(X'\varepsilon) = 0$的矩条件也可推出上述估计量表达形式 (注意上页表达式中括号中的$y_i - x_i\beta$即为ε_i). 所以, 矩估计和一般最小二乘估计在线性回归模型中是等价的.

Gauss-Markov 定理

Gauss-Markov 定理: 最小二乘估计量$\hat{\beta}$是总体参数 β的线性最小方差无偏估计量.

sketch of proof:

1. 用矩阵语言可轻易给出无偏性证明:

$$E(\hat{\beta}_{OLS}) = E[\left(X^\top X\right)^{-1} X^\top Y] = \beta + E[\left(X^\top X\right)^{-1} X^\top \varepsilon] = \beta$$

2. 最小方差的证明由独立同正态分布假设可以推出OLS方差估计和MLE方差估计相同, 达到C-R下界.

OLS残差$e_i = y_i - \hat{y}_i$的性质

记残差

$$e_i = y_i - \hat{y}_i$$

(1) 残差和为零

残差 $e_i = y_i - \hat{y}_i, \quad i = 1, 2, \cdots, n$, 由FOC的第 1 个等式, 得

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip} \right) = 0$$

(2) 残差的样本均值为0

$$\bar{e} = \frac{1}{n} \sum_{i=1}^{n} e_i = 0$$

估计误差项的方差

误差项方差 σ^2 可以通过残差的样本方差(均方误差, MSE) s_e^2 来估计.

$$s_e^2 \equiv \text{MSE} = \frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-p-1} = \frac{\sum_{i=1}^{n} e_i^2}{n-p-1} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p-1}$$

$(n-p-1)$ 是自由度, 表示数据系列 $e_1, e_2, \cdots, e_n$ 中共有 n 个元素, 但是由于存在约束 $\sum_{i=1}^{n} e_i = 0$ 和 $\sum_{i=1}^{n} e_i x_{ij} = 0$, $j = 1, 2, \cdots, p$ (p个OLS的FOC条件), 自由取值的元素个数只有 $(n-p-1)$ 个.

残差越小, 拟合值与观测值越接近, 各观测点在拟合直线周围聚集程度越高, 也就是说, 拟合方程解释 y 的能力就越强.

记残差的样本标准差为

$$s_e = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n} e_i^2}$$

s_e 又被称为估计标准误差(standard error of estimate), 当 s_e 越小时, 说明残差值 e_i 的变异程度越小.
由于残差的样本均值为零, 所以, 其离散范围越小, 拟合的模型就越为精确.

标准误差的性质

可以证明, 如果基本假设 $\varepsilon_i \sim N\left(0, \sigma^2\right)$成立, 则

- s_e^2 是总体随机误差方差 σ^2 的无偏估计量:$E\left(s_e^2\right) = \sigma^2$
- s_e 是总体标准差 σ 的一致估计量 (但一般是有偏估计量):$plim(s_e) = \sigma$

证明过程见伍德里奇教材(Appendix E).

拓展知识：证明$E\left(s_e^2\right)=\sigma^2$

PROOF: Write
$\hat{\mathbf{u}}=\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}=\mathbf{y}-\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}=\mathbf{M}\mathbf{y}=\mathbf{M}\mathbf{u}$, where $\mathbf{M}=\mathbf{I}_n-\mathbf{X}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$, and the last equality follows because $\mathbf{M}\mathbf{X}=\mathbf{0}$. Because $\mathbf{M}$ is symmetric and idempotent,

$$\hat{\mathbf{u}}'\hat{\mathbf{u}}=\mathbf{u}'\mathbf{M}'\mathbf{M}\mathbf{u}=\mathbf{u}'\mathbf{M}\mathbf{u}$$

Because $\mathbf{u}'\mathbf{M}\mathbf{u}$ is a scalar, it equals its trace.

拓展知识：证明$E\left(s_e^2\right)=\sigma^2$

Therefore,

$$\begin{aligned}\mathrm{E}\left(\mathbf{u}'\mathbf{M}\mathbf{u}\mid\mathbf{X}\right)&=\mathrm{E}\left[\mathrm{tr}\left(\mathbf{u}'\mathbf{M}\mathbf{u}\right)\mid\mathbf{X}\right]=\mathrm{E}\left[\mathrm{tr}\left(\mathbf{M}\mathbf{u}\mathbf{u}'\right)\mid\mathbf{X}\right]\\&=\mathrm{tr}\left[\mathrm{E}\left(\mathbf{M}\mathbf{u}\mathbf{u}'\mid\mathbf{X}\right)\right]=\mathrm{tr}\left[\mathbf{M}\mathbf{E}\left(\mathbf{u}\mathbf{u}'\mid\mathbf{X}\right)\right]\\&=\mathrm{tr}\left(\mathbf{M}\boldsymbol{\sigma}^2\mathbf{I}_n\right)=\sigma^2\,\mathrm{tr}(\mathbf{M})=\sigma^2(n-k-1)\end{aligned}$$

The last equality follows from $\mathrm{tr}(\mathbf{M})=\mathrm{tr}\left(\mathbf{I}_n\right)-\mathrm{tr}\left[\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]=n-\mathrm{tr}\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\right]=n-\mathrm{tr}\left(\mathbf{I}_{k+1}\right)=n-(k+1)=n-k-1$. Therefore,

$$\mathrm{E}\left(\hat{\boldsymbol{\sigma}}^2\mid\mathbf{X}\right)=\mathrm{E}\left(\mathbf{u}'\mathbf{M}\mathbf{u}\mid\mathbf{X}\right)/(n-k-1)=\sigma^2$$

拟合优度

通过方差分解我们有:

$$\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR}$$

拟合优度 (Goodness of Fit), 又叫测定系数 (Coefficient of Determination) :

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

思考： 1. 拟合优度越大模型就越好吗? 2. 拟合优度小模型就一定差吗?

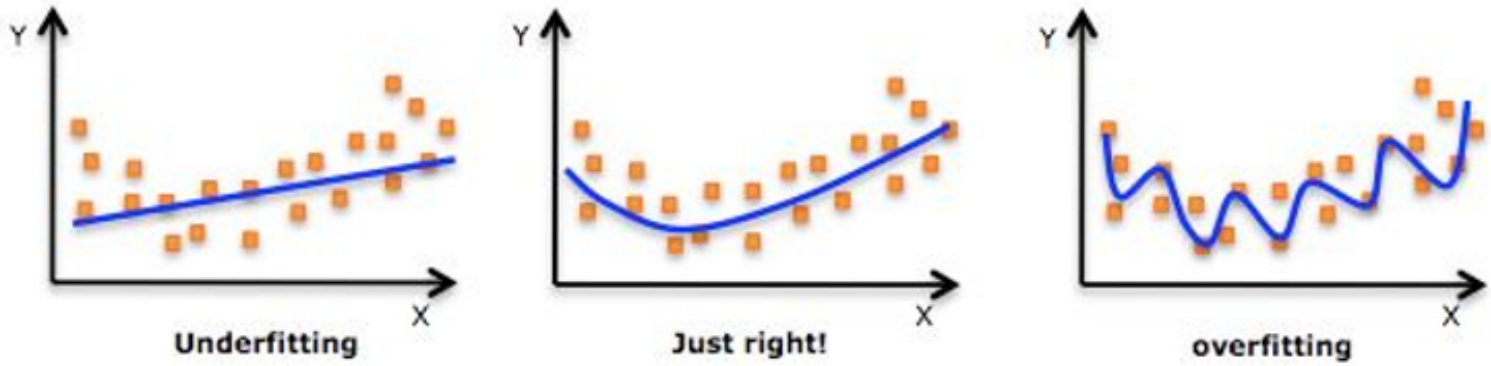
思考： 1. 拟合优度越大模型就越好吗? 2. 拟合优度小模型就一定差吗?

1. over-fitting, 自变量越多, R^2越大
2. fixed effect model vs model with dummy variables, different R^2 but same estimate

过度拟合(overfitting)

当增加变量个数, 而样本容量过小时, 会出现过度拟和现象.

- 样本噪音干扰过大, 将部分噪音认为是特征, 从而扰乱了建模规则
- 参数太多, 模型复杂度过高
- 线性回归中内插的拟合效果好, 但外推预测效果不好



调整的拟合优度

$$R^2_{\text{adj}} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

$\hat{\beta}$ 的分布

我们可以得到:

$$\hat{\beta} \sim N\left(\beta, \sigma^2 \left(X^\top X\right)^{-1}\right)$$

所以:

$$E\left[\hat{\beta}_j\right] = \beta_j, \mathrm{Var}\left[\hat{\beta}_j\right] = \sigma^2 C_{jj}$$

其中 $C = \left(X^\top X\right)^{-1}$

拓展知识：$\hat{\beta}$ 的方差推导

$$
\begin{aligned}
Var(\hat{\beta}) &= E\left[(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}}))(\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}}))'\right] \\
&= E\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mid \boldsymbol{X}\right] \\
&= E\left[(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}'\varepsilon\varepsilon'\boldsymbol{X}(\boldsymbol{X}\boldsymbol{X})^{-1}\right] \\
&= (\boldsymbol{X}\boldsymbol{X})^{-1}\boldsymbol{X}'E\left(\varepsilon\varepsilon'\right)\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2\boldsymbol{I}_n\boldsymbol{X}(\boldsymbol{X}\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}\boldsymbol{X})^{-1}
\end{aligned}
$$

$\hat{\beta}$ 的标准误 (Standard Error)

$\hat{\beta}$ 的标准误为:

$$\mathrm{SE}[\hat{\beta}] = s_e \sqrt{(X^\top X)^{-1}}$$

对每一个 $\hat{\beta}_j$,

$$\mathrm{SE}\left[\hat{\beta}_j\right] = s_e \sqrt{C_{jj}}$$

对 β 的统计推断

因为:

$$\hat{\beta}_j \sim N\left(\beta_j, \sigma^2 C_{jj}\right)$$

我们得到:

$$\frac{\hat{\beta}_j - \beta_j}{s_e\sqrt{C_{jj}}} \sim t(n-p-1)$$

β 的区间估计

对每一个 β_j, 其区间估计为:

$$\hat{\beta}_j \pm t_{1-\alpha/2}(n-p-1) \cdot s_e \sqrt{C_{jj}}$$

拓展内容: 预测-置信区间

给定未知的 x_0, 点预测为:

$$\hat{y}\left(x_0\right)=x_0^{\top}\hat{\beta}=\hat{\beta}_0+\hat{\beta}_1x_{01}+\hat{\beta}_2x_{02}+\cdots+\hat{\beta}_px_{0p}$$

标准误为: $\mathrm{SE}\left[\hat{y}\left(x_0\right)\right]=s_e\sqrt{x_0^{\top}\left(X^{\top}X\right)^{-1}x_0}$

$\hat{y}$的置信区间:

$$\hat{y}\left(x_0\right)\pm t_{1-\alpha/2}(n-p-1)\cdot s_e\sqrt{x_0^{\top}\left(X^{\top}X\right)^{-1}x_0}$$

拓展内容: y的预测区间

计算y的预测区间, 由于$y = \hat{y} + \varepsilon$ 我们需要加一个随机误差项, 因此标准误为

$$\mathrm{SE}\left[\hat{y}\left(x_0\right) + \varepsilon\right] = s_e\sqrt{1 + x_0^{\top}\left(X^{\top}X\right)^{-1}x_0}$$

预测区间为:

$$\hat{y}\left(x_0\right) \pm t_{1-\alpha/2}(n-p-1) \cdot s_e\sqrt{1 + x_0^{\top}\left(X^{\top}X\right)^{-1}x_0}$$

F-test 回归模型的显著性检验

通过方差分解我们有: $\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR}$. 多元回归中, 回归模型的显著性检验的零假设为:

$$\mathrm{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

备择假设为: H_1 : 至少存在一个 $\beta_j \neq 0, j = 1, 2, \cdots, (p-1)$

F-test 回归模型的显著性检验

F检验统计量为

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

其中, $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$, $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

可选$\alpha = 0.05$
如果 $\mathrm{F} > \mathrm{F}_{1-\alpha}(p, n-p-1)$, 拒绝 H_0 (F检验通过)
如果 $\mathrm{F} \leq \mathrm{F}_{1-\alpha}(p, n-p-1)$, 不拒绝 H_0

t-test 单参数检验

我们要检验:

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0$$

统计量

$$t = \frac{\hat{\beta}_j - \beta_j}{SE\left[\hat{\beta}_j\right]} = \frac{\hat{\beta}_j - 0}{s_e\sqrt{C_{jj}}}$$

在零假设成立的情况下, 服从 $t(n-p-1)$ 分布.

可选$\alpha = 0.05$

如果 $|t| > t_{1-\alpha/2}(n-p-1)$, 拒绝 H_0 (t-检验通过)

如果 $|t| \leq t_{1-\alpha/2}(n-p-1)$, 不拒绝 H_0

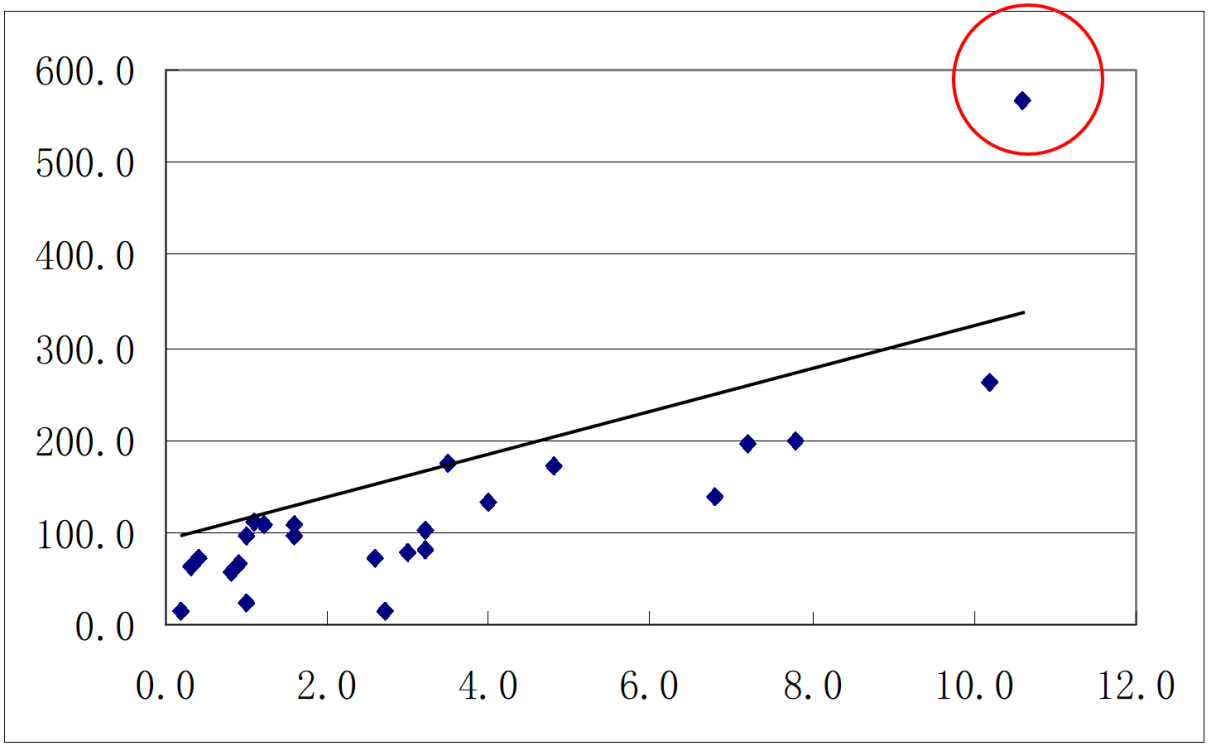
不正常点

除了检查模型假设之外, 还应该注意“不正常点" (unusual observations).
因为有时少量的不正常点对回归的影响是非常大的.
常见的不正常点:

- 异常点 (Outliers)
- 高杜杆点 (Points with high leverage)

异常点

异常点是没有被模型很好的拟合的点, 通常是有很大的标准化残差(standardized residual) 的观测值.
可能原因: 发生突发事件; 统计口径变化; 数据错误.

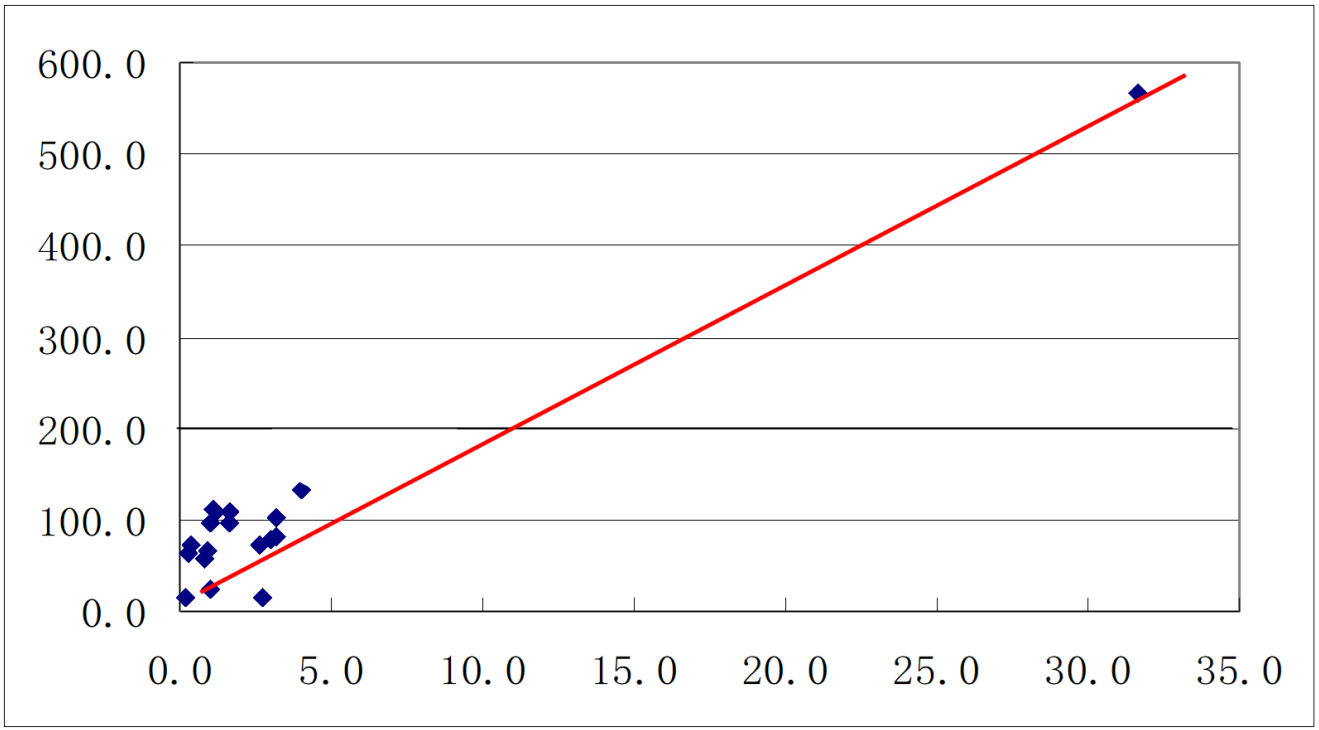


判断标准如: 标准化残差的绝对值大于3.

高杠杆点

高杜杆点, 即杠杆值很大点.
杠杆值 (一元线性回归):

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2}$$



高杠杆值通常认为是大于二倍的平均杠杆值.

变量选择

变量选择通常是一个迭代过程, 在每一步中, 以某种预先定好的标准来决定是否加入或提出某个自变量. 这个标准可以是:

- 假设检验, 如 F 检验或 t 检验
- R^2_{adj}
- Akaike information criterion (AIC) 或 Bayesian information criterion (BIC)
- Mallows's C_p
- 其他标准

变量筛选方法

- 1.向后筛选法 (Backward Elimination)
- 2.向前选择法 (Foreword Selection)
- 3.逐步回归法 (Stepwise Regression)

向后筛选法 (Backward Elimination)

在一开始假设模型中包含所有自变量, 然后依据某种标准逐渐易除不显著的变量, 重复直到现存变量均不符合剔除条件. 以 t 检验为例, 向后筛选法为:

1. 所有自变量 $X_1, X_2, \cdots, X_p$ 均包含在模型中
   - 如果 t 检验都显著, 则 $X_1, X_2, \cdots, X_p$ 均包含在模型中
   - 若存在若干t检验不通过的参数, 则先把p值最大($|t|$值最小的)的变量删除
2. 对剩余的 $p-1$ 个变量做回归方程, 删除 t 检验不通过中 p -值最大的变量
3. 重复以上步骤, 直到模型中所以变量均通过 t 检验

向前选择法 (Forward Selection)

1. 起始. 模型中没有任何变量. 分别计算 Y 与每一个 X 的一元线性回归模型.
   - 选择 t -test值最大的变量首先进入模型
2. 对剩余的$p-1$个变量分别做二元回归方程
   - 在所有通过 t -test 的变量中, 选择 $|t_j|$ 值最大的进入方程
3. 重复以上步骤, 直到模型外所有变量均不能通过t-test

逐步回归法 (Stepwise Regression)

- **前进法的问题:**一旦某自变量进入模型后, 它就永远留在模型中. 然而, 随着其他自变量的引入, 一些先进入模型的变量的作用会变得不再显著.
- **向后法的问题:**一旦某自变量被删除后, 就永远不再进入模型. 然而, 随着其他自变量被删除, 它的作用有可能会显著起来.

逐步回归法 (Stepwise Regression)

- 对于模型外部的变量, 只要还能提供显著的解释作用, 则可以再次进入模型. 而在模型内部的变量, 只要它的t检验不再显著, 则可以从模型中删除.
- 方法: 边进边退
- 起始: 同前进法
- 结束: 模型外所有变量均不能通过t检验.

Stata实现-变量筛选

help stepwise

自变量多重相关性问题(Multicollinearity)

1. **现象:**自变量之间存在相关

2. **危害:** $\hat{\beta} = (X'X)^{-1} X'Y$
这时, 相关系数矩阵不满秩, 行列式 $\to 0$. 回归系数的方差 $\to \infty$

3. **常见的表象:**
   - 结果不稳健: 增加一个变量后, 回归系数变化非常大; 回归系数的符号无法解释.
   - R^2很大, F检验通过, 但t检验却均不通过;

很多情况下自变量都会相关, 只有相关特别强的时候多重共线性才需要担心. 多重共线性危害较大, 不过解决方法相对简单. 最直接的方法可以丢掉多重共线的变量. 在实际工作中, Stata 会自动丢掉变量.

如果y和x的总体关系是非线性

思考多元线性回归模型

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

实际上, **线性**是对参数$\beta_0, \beta_1, \beta_2, \cdots, \beta_p$而言, 而并非对自变量而言. 只要参数是线性的, 所有OLS回归的良好性质均成立.

我们可以利用自变量高次幂, 交叉项, 指数等变换作为新的自变量去拟合y和x的非线性关系

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i2} + \beta_4 x_{i1} x_{i2} + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

对于结构性已知的总体关系, 也可以用其他变换形式将非线性关系转换为线性关系. (下面举例)

举例

1. **指数模型** Exponential model

$$Y = a \cdot e^{bX+\varepsilon}$$

变换方法:取对数

$$\ln Y = \ln a + bX + \varepsilon$$

2. **幂指数模型** Multiplicative model (如Cobb-Douglas production function)

$$Y = a \cdot X_1^{b_1} \cdot X_2^{b_2} \cdot \varepsilon$$

变换方法:取对数

$$\ln Y = \ln a + b_1 \ln X_1 + b_2 \ln X_2 + \ln \varepsilon$$

虚拟变量/哑变量(dummy variable), 类别变量(categorical variable)

1. **虚拟自变量 (dummy variable):**只有两个可能结果的定性自变量 (如:性别)

可转换为取值为0,1的数值变量进入回归.

2. **类别变量(categorical variable):**有多个可能结果的定性自变量 (如:职业)

可转换为多个取值为0,1的数值变量进入回归. (注意: 要丢掉一个变量作为参照组, 否则完全共线)

(stata: tab x, gen())

拓展知识: 交叉验证(Cross Validation)

泛化能力 (generalizability)/外部效度(external validity): 是模型对新样本的预测能力.

- 交叉有效性 (cross validation)(Stone, 1974): 将数据集分成两部分, 一部分作为训练集; 另一部分作为测试集. 用训练集建立模型, 然后将测试集带入模型, 验证模型的表现.

模型的预测精度:可以用验证集的均方误差来测量

- PRESS (predicted residual error sum of squares):$\mathrm{PRESS} = \sum_{i=1}^{n_2} (y_i - \hat{y}_i)^2$

拓展知识: 留一交叉验证法 (Leave-one-out CV, LOOCV)

由于诸多实际情况下并不存在足够的数据留作验证数据(比如小样本, 10个观察), 这个时候通常用留一交叉验证法

- 假设有 n 个原始观察 $(y_1, x_1), \cdots, (y_n, x_n)$, 将第$k$个观察作为验证样本, 对剩下的 $n-1$ 个观察估计模型, 然后计算第k个观察上的预测误差: $e_i^* = y_i - \hat{y}_i$
- 对于每个观察 $k = 1, \cdots, n$, 重复第1步;
- 基于 $e_1^*, \cdots, e_n^*$ 计算均方误差, 做为该模型的预测均方误差的估计值. $PRESS = \sum_{i=1}^{n} (e_i^*)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

拓展知识: 其他交叉验证法

由于LOOCV的计算量较大, 它可以被拓展成K交叉验证.

- **K折交叉验证(K-fold cross-validation)**: 将初始样本随机分割成 K 个子样本集合, 每次将第K个样本集合作为验证样本, 其他 K - 1个样本集合做为训练样本. 交叉验证模型重复 K次, 将 K 次的结果计算平均数, 做为最终模型选择的依据.
- **十折交叉验证(10-fold cross-validation)**: 将数据集分成10份. 轮流将其中9份作为训练数据, 1份作为测试数据, 进行试验. 10次结果的模型精度的平均值, 作为对模型精度的估计.
- 有时还可以进行多次10折交叉验证(例如10次10折交叉验证), 再求其均值, 作为对算法准确性的估计.

本章小结:回归建模的意义

在确定变量之间存在相关关系和因果关系的前提下, 使用回归模型研究变量之间的因果关系.

1、分析哪些因素可能对因变量有解释作用

2、分析每一个自变量在解释因变量时的作用程度

3、用于对因变量的预测分析

注意: 预策分析和因果分析都可以用回归分析的技术, 然而两类分析的目的是截然不同的. 工作中要做好区分, 哪些回归是有预测目的的, 那些是有因果推断目的的, 切忌将预策解释成因果. 如何做好两类分析在未来还有大量更高级的课程供同学们学习.

应用统计学II 作业3暨上机实验9

运用“污染数据”(2014)开展分析

- 计算和分析变量之间的相关关系
- 建立回归模型,完成多元线性回归的建模、推断、进行模型诊断(主要针对回归假设)及预测过程. 并对模型含义进行解释.