第 3 章　多元线性回归

李　杰

数据科学学院，浙江财经大学

2019 年 12 月 4 日

内容提要

3.1 多元线性回归模型

☞ **多元线性回归模型 (multiple linear regression model) 的一般形式**

$$y = \underline{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p} + \varepsilon \tag{1}$$

- 被解释变量: y
- 解释变量: $x_1, x_2, \cdots, x_p$
- 回归截距: β_0
- 回归斜率: $\beta_1, \cdots, \beta_p$
- 误差项: ε: (error term) 或扰动项 (disturbance).
- 重要假设 (多元线性回归模型的 G-M 条件):

$$\mathrm{E}(\varepsilon|x_1, \cdots, x_p) = 0, \quad \mathrm{Var}(\varepsilon|x_1, \cdots, x_p) = \sigma^2 \tag{2}$$

- 总体回归方程

$$\mathrm{E}(y|x_1, \cdots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{3}$$

- "线性"的涵义: 式 (1) 对所有参数 β_j 是线性函数.

- 多元线性回归模型的矩阵表达式 对于 n 个观测 $(x_{i1}, x_{i2}, \cdots, x_{ip}; y_i)\ (i = 1, 2, \cdots, n)$, 有样本回归模型

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

记

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \ \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

则多元线性回归模型的矩阵表达式为

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4}$$

其中 $\mathbf{X}$ 称为**设计矩阵**或**资料矩阵**

☞ 多元线性回归模型的基本假设

① 解释变量 $x_1, x_2, \cdots, x_p$ 是确定性的, 非随机变量, 且 $\mathrm{rank}(\mathbf{X}) = p + 1 << n$

② G-M 条件: 零均值、同方差、不相关

$$\begin{cases} \mathrm{E}(\varepsilon_i) = 0, & i = 1, 2, \cdots, n \\ \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases} & i, j = 1, 2, \cdots, n \end{cases}$$

③ 正态分布假设

$$\begin{cases} \varepsilon_i \sim N(0, \sigma^2), \ i = 1, 2, \cdots, n \\ \varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n \text{ 相互独立} \end{cases}$$

④ 正态分布假设的矩阵表示

$$\begin{aligned} &\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n) \\ &\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} \\ &\mathrm{Var}(\mathbf{y}) = \sigma^2 \mathbf{I}_n \\ &\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n) \end{aligned}$$

☞ **多元线性回归方程的解释**;

- 对于多元线性回归模型

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

可知

$$\mathrm{E}(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

- 偏回归系数

$$\frac{\partial \mathrm{E}(y)}{\partial x_i} = \beta_i$$

- β_i 的涵义: 在影响因变量 y 的其他所有因素 (包括误差项) 不变的情况下, x_i 增加 (或减少) 一单位, y 增加 (或减少) β_i 个单位.

3.2 回归参数的估计

☞ 回归参数的普通最小二乘估计

- **基本思想**
  - 选择估计量 $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$ 使得残差平方和 $\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2$ 最小，即

$$\min_{\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \min_{\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip})^2$$

- 求**一阶条件** (first order conditions, FOCs), 得到**正规方程组** (regular equations)

$$\begin{cases} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}) = 0 \\ \sum_{i=1}^{n} x_{i1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}) = 0 \\ \sum_{i=1}^{n} x_{i2} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}) = 0 \\ \quad \vdots \\ \sum_{i=1}^{n} x_{ik} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}) = 0 \end{cases} \tag{5}$$

- **最小二乘估计**
  - ✧ 若方程组 (5) **存在唯一解**，则称解 $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$ 为模型 (1) 的最小二乘估计 (OLS)。其中 $\hat{\beta}_0$ 为 **OLS 截距估计**，$\hat{\beta}_1, \cdots, \hat{\beta}_p$ 为 **OLS 斜率估计**。称

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p \tag{6}$$

    为**样本回归函数** (sample regression function, SRF)(也称经验回归方程).

- **正规方程组和 OLSE 的矩阵表示形式**

$$\mathbf{X}'\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) = \mathbf{0} \Rightarrow \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

若 $\mathbf{X}'\mathbf{X}$ 可逆, 则最小二乘估计为

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

☞ 回归值与残差

- **回归值**
  - ✧ 由式 (6) 可知, y_i 在 $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ 时的**回归值** (拟合值, fitted value) 为

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$$

称

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

为因变量向量 $\mathbf{y} = (y_1, y_2, \cdots, y_n)'$ 的回归值.

  - ✧ **帽子矩阵**

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \underline{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}$$

称 $H = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ 为帽子矩阵, 即 $\hat{\mathbf{y}} = H\mathbf{y}$.

① 帽子矩阵又称为投影矩阵;

② 帽子矩阵是对称矩阵: $H' = H$;

③ 帽子矩阵是幂等矩阵: $H^2 = H$.

④ 令帽子矩阵主对角线元素为 h_{ii}, 则有 $\sum_{i=1}^{n} h_{ii} = p + 1$.

【证】 $\sum_{i=1}^{n} h_{ii} = tr(H) = tr((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}) = tr(I_{p+1}) = p + 1.$

- 残差
  - ✧ 残差: i 次观测的残差 (residual) 是实际观测值 y_i 与其拟合值之差

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}$$

$e_i > 0$ 表明预测值偏低，$e_i < 0$ 表明预测值偏高。

  - ✧ 残差矩阵

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \mathbf{y} - \hat{\mathbf{y}} = (I - H)\mathbf{y}$$

① 称 $I - H$ 为残差矩阵;
② 残差矩阵是对称矩阵: $(I - H)' = (I - H)$
③ 残差矩阵为幂等矩阵: $(I - H)^2 = (I - H)$
④ 残差方差为

$$\begin{aligned} \mathrm{Cov}(\mathbf{e}, \mathbf{e}) &= \mathrm{Cov}[(I - H)\mathbf{y}, (I - H)\mathbf{y}] \\ &= (I - H)\mathrm{Cov}(\mathbf{y}, \mathbf{y})(I - H)' \\ &= \sigma^2(I - H) \end{aligned}$$

⑤ $\mathrm{Var}(e_i) = (1 - h_{ii})\sigma^2 \quad i = 1, 2, \cdots, n.$

⑥ 残差满足

$$\begin{cases} \sum_{i=1}^{n} e_i = 0 \\ \sum_{i=1}^{n} e_i x_{i1} = 0 \\ \cdots\cdots \\ \sum_{i=1}^{n} e_i x_{ip} = 0 \end{cases}$$

即 $\mathbf{X}'\mathbf{e} = \mathbf{0}$.

⑦ 误差项方差估计量

$$\hat{\sigma}^2 = \frac{1}{n-(p+1)} SSE = \frac{1}{n-(p+1)} (\mathbf{e}'\mathbf{e})$$

且 $\mathrm{E}(\hat{\sigma}^2) = \sigma^2$.

⑧ $\sum_{i=1}^{n} e_i \hat{y}_i = \mathbf{e}'\hat{\mathbf{y}} = \mathbf{e}'H\mathbf{y} = \mathbf{y}'(I-H)H\mathbf{y} = 0$

☞ 回归参数的最大似然估计

- 模型: $\mathbf{y} = \mathbf{X}\beta + \varepsilon$
- 假设: $\varepsilon \sim N(\mathbf{0}, \sigma^2 I_n)$
- 因变量 $\mathbf{y}$ 的分布: $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2 I_n)$
- 似然函数

$$L(\beta, \sigma^2) = (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)\right]$$

- 对数似然函数

$$\ln(L) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)\right]$$

- 最大似然估计量

$$\hat{\beta}_{MLE} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$
$$\hat{\sigma}^2_{MLE} = \frac{1}{n}SSE = \frac{1}{n}\mathbf{e}'\mathbf{e}.$$

☞ 对 OLS 回归方程的解释

- OLS 回归方程

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

- $\hat{\beta}_0$: 因变量 y 在 $x_1 = 0, x_2 = 0, \cdots, x_k = 0$ 时的预测值;
- $\hat{\beta}_j$ $(j = 1, 2, \cdots, k)$: **偏效应** (partial effect) 或**其他条件不变** (ceteris paribus). 譬如，在 $x_1 = (x_{i1}, x_{i2}, \cdots, x_{ik})$ 处的拟合值为

$$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}$$

在 $x_2 = (x_{i1} + \Delta x_1, x_{i2} + \Delta x_2, \cdots, x_{ik} + \Delta x_k)$ 处的拟合值为

$$\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 (x_{i1} + \Delta x_1) + \cdots + \hat{\beta}_k (x_{ik} + \Delta x_k)$$

拟合值的改变量为

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \cdots + \hat{\beta}_k \Delta x_k$$

斜率系数，譬如 $\hat{\beta}_1$，度量的是，在其他所有条件不变的情况下 $(\Delta x_2 = \Delta x_3 = \cdots = \Delta x_k = 0)$，因 $\Delta x_1 = 1$ 而导致的 y 的变化:

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 = \hat{\beta}_1$$

☞ 案例：工资方程

```
install.packages(foreign)
library(foreign)

mydata <- read.dta("E:/statafiles/WAGE1.DTA", convert.factors = FALSE)

my_lm <- lm(log(wage)~educ+exper+tenure,data=mydata)

my_slm <- summary(my_lm)

r = cor(log(mydata$wage),fitted(my_lm)) # 求拟合值与因变量的相关系数 cor()
r^2 == my_slm$r.squared          # 比较相关系数的平方与 R^2
```

3.3 参数估计量的性质

☞ **性质 1:** $\hat{\boldsymbol{\beta}}$ 是随机向量 $\mathbf{y}$ 的线性变换.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

☞ **性质 2:** $\hat{\boldsymbol{\beta}}$ 是 $\boldsymbol{\beta}$ 的无偏估计.

$$\begin{aligned}
\mathrm{E}(\hat{\boldsymbol{\beta}}) &= \mathrm{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}(\mathbf{y}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\
&= \boldsymbol{\beta}
\end{aligned}$$

☞ **性质 3:** $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \mathrm{Cov}\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\beta}}\right) = \mathrm{Cov}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\right] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{Cov}(\mathbf{y},\mathbf{y})\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]' \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

☞ **性质 4 (G-M 定理): 在 G-M 条件下, 最小二乘估计 $\hat{\beta}$ 是 β 的 BLUE.**

- ✧ 在 G-M 条件下, 每个分量 $\hat{\beta}_j$ 是 β_j $(j = 0, 1, 2, \cdots, p)$ 的 BLUE.
- ✧ 可能存在 β 的非线性估计量, 其方差小于最小二乘估计 $\hat{\beta}$.
- ✧ 可能存在 β 的有偏估计量, 在某种意义上比最小二乘估计 $\hat{\beta}$ 更优.
- ✧ 在正态分布假设下, $\hat{\beta}$ 是 β 的最小方差无偏估计.

☞ **性质 5:** $\mathrm{Cov}\left(\hat{\beta}, \mathbf{e}\right) = \mathbf{0}$

【证】 $\mathrm{Cov}\left(\hat{\beta}, \mathbf{e}\right) = \mathrm{Cov}\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}, (I - H)\mathbf{y}\right]$

$$= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(I - H) = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left[(I - H)\mathbf{X}\right]' = \mathbf{0}$$

☞ **性质 6:** 在正态分布假设 $(\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I}_n^2))$ 下, 有

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

② $\dfrac{SSE}{\sigma^2} \sim \chi^2(n - p - 1)$.

3.4 回归方程的显著性检验

☞ F-检验

- **检验的问题:** 多元线性回归模型整体上是显著的吗?
- **原假设与备择假设**

$$H_0: \ \beta_1 = 0, \beta_2 = 0, \cdots, \beta_p = 0$$

- **检验统计量**

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

- **方差分析表**

方差来源	自由度	平方和	均方	F-值	p-值
回归	p	SSR	SSR/p	$\frac{SSR/p}{SSE/(n-p-1)}$	$p = P(F > F\_value)$
残差	$n-p-1$	SSE	$SSE/(n-p-1)$		
总和	$n-1$	SST			

- **检验:**对给定的显著性水平 α
  - ✧ 若 $F > F_\alpha(p, n-p-1)$ 时, 拒绝原假设;
  - ✧ 若 $F \leqslant F_\alpha(p, n-p-1)$ 时, 接受原假设;
  - ✧ 当 $\alpha > p$ 时, 拒绝原假设.

☞ t-检验

- **为什么要进行 t 检验?**
  - ✧ 在多元线性回归模型中, 回归方程的整体显著性不代表单个自变量对 y 的影响也是显著的, 从建模的 KISS (keep it sophisticatedly simply, KISS) 准则出发, 需要删除.
  - ✧ t-检验是对每个自变量的显著性进行检验, 如果自变量 x_j 对 y 的影响不显著, 则体现在其回归系数绝对值 $|\beta_j|$ 接近 0.
- **原假设:** 对于解释变量 x_j, 原假设是 x_j 对 y 的影响不显著, 即

$$H_{0j}: \beta_j = 0, \quad H_{1j}: \beta_j \neq 0$$

- **检验统计量**

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

$$(\mathbf{X}'\mathbf{X})^{-1} = (c_{ij}), \quad i, j = 0, 1, 2, \cdots, p$$

$$\hat{\beta}_j \sim N(\beta_j, c_{jj}\sigma^2), \quad j = 0, 1, 2, \cdots, p$$

$$t_j = \frac{\hat{\beta}_j}{\sqrt{c_{jj}}\hat{\sigma}} \sim t(n-p-1)$$

$$\hat{\sigma} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n} e_i^2} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- **检验:** 对给定的显著性水平 α
  - ✧ 若 $|t| > t_{\frac{\alpha}{2}}(p, n-p-1)$ 时, 拒绝原假设;
  - ✧ 若 $|t| \leqslant t_{\frac{\alpha}{2}}(p, n-p-1)$ 时, 接受原假设;
  - ✧ 当 $\alpha > p$ 时, 拒绝原假设.
- **后退法:** 在多元线性回归模型中, 如果某些自变量不显著, 则用"**后退法**"剔除不显著的变量, **不能一次性剔除所有不显著的变量, 原则上只能剔除一个自变量**, 先剔除其中 $|t|$ 值最小 (或 p 值最大的) 自变量, 然后再估计新的回归模型, 有不显著的自变量再剔除, 直至所有保留的自变量对 y 都有显著影响为止.

☞ **回归系数的置信区间:** 因为

$$t_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\hat{\sigma}} \sim t(n-p-1)$$

故 β_j 的置信度为 $1-\alpha$ 的置信区间为

$$\left(\hat{\beta}_j - t_{\frac{\alpha}{2}}\sqrt{c_{jj}}\hat{\sigma}, \hat{\beta}_j + t_{\frac{\alpha}{2}}\sqrt{c_{jj}}\hat{\sigma}\right)$$

☞ 拟合优度

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$R^2 \in [0, 1]$$

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

$$= \text{Corr}(\mathbf{y}, \hat{\mathbf{y}})$$

称 R 为 y 关于 $x_1, x_2, \cdots, x_p$ 的**样本复相关系数**.

- 案例: 国际旅游收入

```
instal.packages(foreign)
library(foreign)

travel <- read.spss("D:\\documents\\MyDoc\\JobInZufe\\Courseware
          \\应用回归分析\\数据\\例3.1 国际旅游收入.sav",to.data.frame =T)
travel_lm <- lm(Y~.,data=travel)
summary(travel_lm)
```

- 回归方程的报告

$$\hat{Y} = \underset{}{-205.552} \underset{(116.965)}{} - \underset{(22.887)}{1.495}\, X_1 + \underset{(18.592)}{2.6488} X_2 + \underset{(2.467)}{3.291} X_3 - \underset{(1.298)}{0.944} X_4$$

$$- \underset{(4.512)}{5.512} X_5 + \underset{(3.957)}{4.060} X_6 + \underset{(5.077)}{4.162} X_7 - \underset{(10.829)}{15.436} X_8$$

$$+ \underset{(8.368)}{17.373} X_9 + \underset{(10.160)}{9.127}\, X_{10} - \underset{(5.619)}{10.537} X_{11} + \underset{(5.005)}{1.360} X_{12}$$

$$n = 31,\ R^2 = 0.875,\ \bar{R}^2 = 0.792$$

3.5 中心化与标准化

☞ 数据标准化的理由

① 在多元线性回归分析中，涉及的变量都带有量纲，且变量的单位往往不同，单位的变化给回归分析带来一定的困难.

② 在同一数据集中，不同变量的数据，小数点位差别可能很大，可能导致很大的舍入误差.

☞ 数据中心化及中心化回归

- **中心化经验回归方程**

  多元回归模型: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$

  经验回归方程: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$

  在样本中心点: $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 + \cdots + \hat{\beta}_p \bar{x}_p$

  中心化经验回归方程: $\hat{y} - \bar{y} = \hat{\beta}_1 (x_1 - \bar{x}_1) + \hat{\beta}_2 (x_2 - \bar{x}_2) + \cdots + \hat{\beta}_p (x_p - \bar{x}_p)$

- **数据中心化:**

$$x'_{ij} = x_{ij} - \bar{x}_j, \quad i = 1, 2, \cdots, n; \ j = 1, 2, \cdots, p$$

$$y'_i = y_i - \bar{y}, \quad i = 1, 2, \cdots, n;$$

- **无截距回归**

  ✧ 基于中心化后的数据做无截距回归.

☞ 中心化回归

```
cent_travel <- scale(travel, center=T, scale=F)
            # scale 函数返回的结果包含一个矩阵，以及各变量的样本均值
cent_travel <- as.data.frame(cent_travel)
            # 矩阵不能用来做回归，故将矩阵强制转换成数据框
cent_travel_lm <- lm(Y~.-1,data=cent_travel)
            # 利用中心化后的数据做做回归
summary(cent_travel_lm)
```

☞ 标准化回归

- **数据标准化**

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{\sqrt{L_{jj}}}, \quad i = 1, 2, \cdots, n;\ j = 1, 2, \cdots, p$$

$$y_i^* = \frac{y_i - \bar{y}}{\sqrt{L_{yy}}}, \quad i = 1, 2, \cdots, n;$$

其中 $L_{jj} = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 = SST_j$.

- **标准化回归方程**
  - ✧ 基于标准化数据, 用最小二乘法估计无截距回归模型, 得到**标准化回归方程**

$$\hat{y}^* = \hat{\beta}_1^* x_1^* + \hat{\beta}_2^* x_2^* + \cdots + \hat{\beta}_p^* x_p^*$$

- **标准化回顾系数与最小二乘估计的关系**

$$\hat{\beta}_j^* = \frac{\sqrt{L_{jj}}}{\sqrt{L_{yy}}}\hat{\beta}_j, \quad j = 1, 2, \cdots, p$$

- **标准化回归系数的解释:**
  - ① **在其他因素不变的情况下**, x_j 增加 1%, y 增加 $\hat{\beta}_j^*$%.
  - ② **在其他因素不变的情况下**, x_j 增加 1 个标准差, y 增加 $\hat{\beta}_j^*$ 个标准差.

☞ 标准化回归

```
scale_travel <- scale(travel, center=T, scale=T)

scale_travel <- as.data.frame(scale_travel)

stand_travel_lm <- lm(Y~.-1,data=scale_travel)

summary(scale_travel_lm)
```

☞ 样本相关系数

- **复相关系数:** 复相关系数 R 体现的是因变量 y 与所有自变量 $x_1, x_2, \cdots, x_p$ 之间的相关系, 是**整体和共性指标**.
- **样本相关矩阵:** 简单相关系数体现是的两个变量之间的相关系, 是**局部和个性指标**. 对于变量 $x_1, x_2, \cdots, x_p$, 样本相关矩阵为

$$\mathbf{r} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

记中心化后的资料矩阵为 $\mathbf{X}^* = (x_{ij})^*_{n\times p}$, 则样本相关矩阵可表示为

$$\mathbf{r} = (\mathbf{X}^*)' \mathbf{X}^*$$

- **增广样本相关矩阵:** 对于观测变量 $y, x_1, x_2, \cdots, x_p$, 其**增广样本相关矩阵**为

$$\tilde{\mathbf{r}} = \begin{bmatrix} 1 & r_{y1} & r_{y2} & \cdots & r_{yp} \\ r_{1y} & 1 & r_{12} & \cdots & r_{1p} \\ r_{2y} & r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{py} & r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}$$

- **代码**

```
cor(travel) # 求相关矩阵

install.packages(psych)
library(psych)
corr.test(travel)          # 相关性检验
```

☞ 偏决定系数

① **偏决定系数:** 对于变量 $y, x_1, x_2, \cdots, x_p$, 任意固定其中 $p-1$ 个变量后, 另外两个变量之间的相关系数.

② 两个自变量时的偏决定系数:

✧ 二元线性回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

✧ $SSE(x_2)$ 表示只有自变量 x_2 时的残差平方和; $SSE(x_1, x_2)$ 表示自变量为 x_1, x_2 时的残差平方和. 此时的偏决定系数为

$$R^2_{y1;2} = \frac{SSE(x_2) - SSE(x_1, x_2)}{SSE(x_2)}$$

同理, 在已含有 x_1 的条件下, y 与 x_2 的偏决定系数为

$$R^2_{y2;1} = \frac{SSE(x_1) - SSE(x_1, x_2)}{SSE(x_1)}$$

③ **一般情形的偏相关系数**

当模型含有 $x_2, \cdots, x_p$ 时, y 与 x_1 的偏相关系数为

$$R^2_{y1;2,3,\cdots,p} = \frac{SSE(x_2, \cdots, x_p) - SSE(x_1, x_2, \cdots, x_p)}{SSE(x_2, \cdots, x_p)}$$

☞ 偏相关系数

- **偏相关系数:** 偏决定系数的平方根为**偏相关系数**, 其符号与相应回归系数的符号相同.

作业

- **P88**
  - 3.7
  - 3.8
  - 3.9
  - 3.10
  - 3.11
  - 3.12