第三章：多元线性回归

主讲人：黄磊

数学学院

西南交通大学

September 26, 2019

Outline

1. 多元线性回归模型
   - 随机向量补充知识
   - 多元线性回归模型
   - 回归参数的估计
   - 参数估计量的性质
   - 回归方程的显著性检验
   - 中心化和标准化
   - 相关阵和偏相关系数
   - 总结与评注

随机向量

在第二章已经看到，当我们把响应变量和解释变量向量化或者矩阵化以后，一元线性回归模型的形式变得非常简单。这一章将进入多元线性回归模型的研究，有必要对随机向量的一些基本性质，作一些介绍和总结。
设$\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ 为$n \times 1$的随机向量，称

$$E(\boldsymbol{Y}) = (E(Y_1), \ldots, E(Y_n))^\top$$

为$\boldsymbol{Y}$的均值向量。

Theorem 1

设$\boldsymbol{A}$ 为$m \times n$的非随机矩阵，$\boldsymbol{Y}$和$\boldsymbol{Z}$分别为$n \times 1$ 和$m \times 1$的随机向量，
记$\eta = \boldsymbol{A}\boldsymbol{Y} + \boldsymbol{Z}$, 则

$$E(\eta) = \boldsymbol{A}E(\boldsymbol{Y}) + E(\boldsymbol{Z}).$$

随机向量

n维随机向量$\boldsymbol{Y}$的协方差阵定义为

$$Cov(\boldsymbol{Y}) \overset{def}{=} D(\boldsymbol{Y}) = E\left[(\boldsymbol{Y} - E\boldsymbol{Y})(\boldsymbol{Y} - E\boldsymbol{Y})'\right],$$

其中$E\boldsymbol{Y} = E(\boldsymbol{Y})$，为了方便起见。

Proposition 1

$trD(\boldsymbol{Y}) = \sum_{i=1}^{n} Var(Y_i)$, 这里$tr$ 表示方阵的迹，即对角线元素之和。

Theorem 2

设$\boldsymbol{Y}$是任意任意$n \times 1$随机向量，则它的协方差阵$D(\boldsymbol{Y})$是半正定的对称阵。

Proof 考察任意非随机$n \times 1$向量$\boldsymbol{c}$, 令$Z = \boldsymbol{c}'\boldsymbol{Y}$,

$$\begin{aligned} 0 \le Var(Z) = Var(\boldsymbol{c}'\boldsymbol{Y}) &= E(\boldsymbol{c}'\boldsymbol{Y} - E\boldsymbol{c}'\boldsymbol{Y})^2 \\ &= E(\boldsymbol{c}'\boldsymbol{Y} - E\boldsymbol{c}'\boldsymbol{Y})(\boldsymbol{c}'\boldsymbol{Y} - E\boldsymbol{c}'\boldsymbol{Y})' \\ &= \boldsymbol{c}'E\left[(\boldsymbol{Y} - E\boldsymbol{Y})(\boldsymbol{Y} - E\boldsymbol{Y})'\right]\boldsymbol{c} = \boldsymbol{c}'D(\boldsymbol{Y})\boldsymbol{c}. \end{aligned} \tag{1.1}$$

随机向量

Theorem 3

设$\boldsymbol{A}$为$m\times n$非随机阵，$\boldsymbol{Y}$为$n\times 1$随机向量，$\boldsymbol{Z}=\boldsymbol{A}\boldsymbol{Y}$, 则

$$D(\boldsymbol{Z})=\boldsymbol{A}D(\boldsymbol{Y})\boldsymbol{A}'$$

Proof 根据协方差阵的定义，

$$\begin{aligned} D(\boldsymbol{Z}) &= E\left[(\boldsymbol{Z}-E\boldsymbol{Z})(\boldsymbol{Z}-E\boldsymbol{Z})'\right] \\ &= E\left[(\boldsymbol{AY}-E\boldsymbol{AY})(\boldsymbol{AY}-E\boldsymbol{AY})'\right] \\ &= \boldsymbol{A}E\left[(\boldsymbol{Y}-E\boldsymbol{Y})(\boldsymbol{Y}-E\boldsymbol{Y})'\right]\boldsymbol{A}' \\ &= \boldsymbol{A}D(\boldsymbol{Y})\boldsymbol{A}'. \end{aligned} \tag{1.2}$$

随机向量

设$\boldsymbol{Y}_1$和$\boldsymbol{Y}_2$分别为$n\times 1$, $m\times 1$的随机向量，定义

$$Cov(\boldsymbol{Y}_1,\boldsymbol{Y}_2)=E\left[(\boldsymbol{Y}_1-E\boldsymbol{Y}_1)(\boldsymbol{Y}_2-E\boldsymbol{Y}_2)'\right]$$

Theorem 4

设$\boldsymbol{Y}_1$和$\boldsymbol{Y}_2$分别为$n\times 1$, $m\times 1$的随机向量, $\boldsymbol{A}$和$\boldsymbol{B}$分别为$p\times n$和$q\times m$的非随机矩阵，则

$$Cov(\boldsymbol{A}\boldsymbol{Y}_1,\boldsymbol{B}\boldsymbol{Y}_2)=\boldsymbol{A}Cov(\boldsymbol{Y}_1,\boldsymbol{Y}_2)\boldsymbol{B}'$$

Proof

$$\begin{aligned}Cov(\boldsymbol{A}\boldsymbol{Y}_1,\boldsymbol{B}\boldsymbol{Y}_2)&=E\left[(\boldsymbol{A}\boldsymbol{Y}_1-E\boldsymbol{A}\boldsymbol{Y}_1)(\boldsymbol{B}\boldsymbol{Y}_2-E\boldsymbol{B}\boldsymbol{Y}_2)'\right]\\&=\boldsymbol{A}E\left[(\boldsymbol{Y}_1-E\boldsymbol{Y}_1)(\boldsymbol{Y}_2-E\boldsymbol{Y}_2)'\right]\boldsymbol{B}'\\&=\boldsymbol{A}Cov(\boldsymbol{Y}_1,\boldsymbol{Y}_2)\boldsymbol{B}'.\end{aligned}\tag{1.3}$$

矩阵的偏导

Theorem 5

对矩阵$\boldsymbol{A} \in R^{n\times p}$, 以及$\boldsymbol{X} \in R^p$, 有以下两个结论

$$\frac{\partial \boldsymbol{AX}}{\partial \boldsymbol{X}} = \boldsymbol{A}', \quad \frac{\partial \boldsymbol{X}'\boldsymbol{X}}{\partial \boldsymbol{X}} = 2\boldsymbol{X}, \tag{1.4}$$

Theorem 6

复合函数求导— $\boldsymbol{Z}(\boldsymbol{Y})$是$\boldsymbol{Y}$的向量函数，$\boldsymbol{Y}(\boldsymbol{X})$是$\boldsymbol{X}$的向量函数，则

$$\frac{\partial \boldsymbol{Z}}{\partial \boldsymbol{X}} = \frac{\partial \boldsymbol{Y}}{\partial \boldsymbol{X}}\frac{\partial \boldsymbol{Z}}{\partial \boldsymbol{Y}} \tag{1.5}$$

于是，若$\boldsymbol{A}$是$n \times p$的非随机矩阵，$\boldsymbol{X} \in R^p$, 则

$$\frac{\partial\{(\boldsymbol{AX})'\boldsymbol{AX}\}}{\partial \boldsymbol{X}} = \boldsymbol{A}'2\boldsymbol{AX} = 2\boldsymbol{A}'\boldsymbol{AX}.$$

多元线性回归模型的一般形式

响应变量y与解释变量$x_1, x_2, \ldots, x_p$的线性回归模型为：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon \tag{1.6}$$

$\beta_j, j = 0, \ldots, p$是$p+1$个未知参数，$\beta_0$是回归常数，$\beta_j, j = 1, \ldots, p$是回归系数。$p = 1$时就是一元线性回归模型；$p \geq 2$时，就称为多元线性回归模型。与一元线性回归模型一样，对误差项ε，通常假定

$$\begin{cases} E(\varepsilon) = 0 \\ Var(\varepsilon) = \sigma^2 \end{cases} \tag{1.7}$$

称

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \tag{1.8}$$

为理论回归方程。

多元线性回归模型的一般形式

对一个实际问题，响应变量和解释变量往往可以获得n组观测数据$\{y_i; x_{1i}, \ldots, x_{pi}\}(i = 1, \ldots, n)$, 则通过向量形式，线性回归模型(1.6)可以表示为

$$\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon \tag{1.9}$$

其中$\boldsymbol{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{X} = (\boldsymbol{1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$ 称之为$n \times (p + 1)$维的设计矩阵，$\boldsymbol{1} = (1, \ldots, 1)'$是一个$n \times 1$的列向量，$\boldsymbol{x}_j = (x_{j1}, \ldots, x_{jn})'$是对应第$j$个解释变量的观测向量,
$j = 1, \ldots, p$。$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$是$(p + 1) \times 1$维的未知参数向量，$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$是$n \times 1$维的误差项向量。

基本假定

为了更方便地去估计模型的参数以及检验回归方程，需要如下基本假定

1. 解释变量$x_j, j = 1, \ldots, p$是确定性变量（非随机），且$rank(\boldsymbol{X}) = p + 1 < n$, 即要求$\boldsymbol{x}$是一个满秩矩阵，其列向量线性无关，样本量个数大于解释变量个数。
2. (Gauss-Markov)随机误差项零均值、等方差，即

$$\begin{cases} E(\varepsilon_i) = 0, i = 1, \ldots, n \\ Cov(\varepsilon_i, \varepsilon_j) = 0, i \neq j; \quad Var(\varepsilon_i) = \sigma^2, \end{cases} \tag{1.10}$$

3. 正态分布假设

$$\begin{cases} \varepsilon_i \sim N(0, \sigma^2), i = 1, \ldots, n \\ \varepsilon_1, \ldots, \varepsilon_n \quad 相互独立 \end{cases} \tag{1.11}$$

对于多元线性回归的矩阵形式(1.9), 这个条件也可以表示为$\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$. 由此假设，$\boldsymbol{y}$服从$n$维正态分布，$E(\boldsymbol{y}) = \boldsymbol{X}\beta$, $Var(\boldsymbol{y}) = D(\boldsymbol{y}) = \sigma^2 \boldsymbol{I}_n$, $\boldsymbol{y} \sim N(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}_n)$

回归方程的解释

以$p=2$的一个微观经济问题为例子，给出回归方程的几何解释和经济意义。y表示空调机销售量，x_1表示空调机的价格，x_2表示消费者可支配收入，建立的二元线性回归模型如下

$$\begin{cases} y=\beta_0+\beta_1x_1+\beta_2x_2+\varepsilon \\ E(y)=\beta_0+\beta_1x_1+\beta_2x_2 \end{cases} \tag{1.12}$$

在式(1.12)里面，假如保持x_2不变，则有

$$\frac{\partial E(y)}{\partial x_1}=\beta_1 \tag{1.13}$$

因此，β_1可解释为在消费者收入x_2保持不变时，空调机价格x_1每增加一个单位，空调机销售量y的平均增加幅度。类似的，可以解释β_2为在空调机价格x_1保持不变时，消费者收入x_2每增加一个单位，空调机销售量y平均增加的幅度。

当$p>2$时，各个回归系数（也称为偏回归系数）也可以类似地解释。

LSE

根据回归方程矩阵的形式(1.9), 最小二乘估计就是为了寻找使得如下损失函数$Q(\beta)$最小的$\hat{\beta}$,

$$Q(\beta) = (\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta) \tag{1.14}$$

根据Theorem 5、6的结论，为了找到$\hat{\beta}$，我们可对$Q(\beta)$求关于β的一阶偏导，解方程

$$\frac{\partial Q(\beta)}{\partial \beta} = -\boldsymbol{X}'2(\boldsymbol{y} - \boldsymbol{X}\beta) = -2\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\beta) = 0 \tag{1.15}$$

整理得$\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{X}\beta$, 当$\boldsymbol{X}'\boldsymbol{X}$的逆存在时，就可以得到最小二乘估计$\hat{\beta}$为

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, \tag{1.16}$$

另外可求Hessian 矩阵$\partial^2 Q(\beta)/\partial\beta\partial\beta' = 2\boldsymbol{X}'\boldsymbol{X}$, 说明$\hat{\beta}$是$Q(\beta)$的最小值点。

回归值

由以上结论，得经验回归方程

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \overset{def}{=} \boldsymbol{H}\boldsymbol{y} \tag{1.17}$$

称$\hat{\boldsymbol{y}}$为因变量观测值$\boldsymbol{y}$的回归拟合值，简称回归值或者拟合值。由式子(1.17)看到，$\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ 的作用就是把$\boldsymbol{y}$变为$\hat{\boldsymbol{y}}$, 形式上就是戴了一顶帽子。帽子矩阵$\boldsymbol{H}$有以下几个特征

1. $\boldsymbol{H}' = \boldsymbol{H}$，对称阵
2. $\boldsymbol{H}^2 = \boldsymbol{H}$, 幂等阵
3. v, $\boldsymbol{H}$ 投影阵
4. $diag(\boldsymbol{H}) = (h_{11}, \ldots, h_{nn})$,
   $tr(H) = \sum_{i=1}^{n} h_{ii} = tr(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = tr((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}) = p + 1$

残差

称$\boldsymbol{e}=\boldsymbol{y}-\hat{\boldsymbol{y}}$为$\boldsymbol{y}$的残差向量，将$\hat{\boldsymbol{y}}=\boldsymbol{H}\boldsymbol{y}$ 代入，得

$$\boldsymbol{e}=\boldsymbol{y}-\boldsymbol{H}\boldsymbol{y}=(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y},$$

同理，同学们可以验证，$(\boldsymbol{I}-\boldsymbol{H})$ 也是对称阵，幂等阵。因此，

$$\begin{aligned} D(\boldsymbol{e}) &= Cov(\boldsymbol{e},\boldsymbol{e}) \\ &= Cov((\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y},(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{y}) \\ &= (\boldsymbol{I}-\boldsymbol{H})Cov(\boldsymbol{y},\boldsymbol{y})(\boldsymbol{I}-\boldsymbol{H})' \\ &= (\boldsymbol{I}-\boldsymbol{H})\sigma^2\boldsymbol{I}(\boldsymbol{I}-\boldsymbol{H})' \\ &= \sigma^2(\boldsymbol{I}-\boldsymbol{H}), \end{aligned} \tag{1.18}$$

于是，$Var(e_i)=(1-h_{ii})\sigma^2, i=1,\dots,n$。

残差

根据式子(1.15), 残差向量还满足以下关系式，

$$\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{0},$$

根据矩阵的乘法法则，就得到$\sum_{i=1}^{n} e_i = 0$ 以及$\sum_{i=1}^{n} e_i x_{ji} = 0, j = 1, \ldots, p$的关系式。解释为，残差的平均值为零，残差关于每个自变量的加权平均也为零。

误差项方差σ^2的无偏估计为$\hat{\sigma}^2$

$$\hat{\sigma}^2 = \frac{1}{n-p-1} SSE = \frac{1}{n-p-1} \boldsymbol{e}'\boldsymbol{e} = \frac{1}{n-p-1} \sum_{i=1}^{n} e_i^2,$$

课堂练习1：请同学们推导其无偏性。

影响分析(补充知识)

类似于一元回归分析，回归诊断首先要考残差是否基本满足Gauss-Markov假设，然后还需要做影响分析，即研究那些样本点$(y_i, \boldsymbol{x}_i^\top)$对估计或者预测有异常大的影响。由此，我们考虑用以下记号表示剔除第i个样本点的线性回归模型

$$\boldsymbol{y}_{(i)} = \boldsymbol{X}_{(i)}\boldsymbol{\beta} + \boldsymbol{e}_{(i)},$$

将从这个模型求到的$\boldsymbol{\beta}$的最小二乘估计记为$\hat{\boldsymbol{\beta}}_{(i)}$, 则

$$\hat{\boldsymbol{\beta}}_{(i)} = (\boldsymbol{X}_{(i)}^\top \boldsymbol{X}_{(i)})^{-1} \boldsymbol{X}_{(i)}^\top \boldsymbol{y}_{(i)},$$

显然向量$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}$的模反映了第$i$个样本点对回归系数估计的影响大小，但是简单的模会受到$\hat{\boldsymbol{\beta}}$的协方差矩阵影响，因此需要考虑它的一种标准的数量化函数，Cook 统计量就是其中应用最广泛的一种, 其定义如下，

$$D_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^\top \boldsymbol{X}^\top \boldsymbol{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{p\hat{\sigma}^2}, i = 1, \ldots, n, \tag{1.19}$$

影响分析(补充知识)

但是Cook统计量的定义式(1.19)十分不方便计算，要计算n个Cook统计量，需要做n+1次回归，计算量大。下面的定理不仅提供了计算D_i的简便公式，也从理论上推导了D_i与学生化残差r_i还有杠杆值h_{ii}之间的关系。根据(1.18), 学生化残差定义为

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}, i = 1, \ldots, n.$$

Theorem 7

式(1.19)定义的统计量D_i与学生化残差r_i和杠杆值h_{ii}恰有如下关系

$$D_i = \frac{1}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right) r_i^2, i = 1, \ldots, n, \tag{1.20}$$

这里h_{ii}就是帽子矩阵$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$的对角元。

影响分析(补充知识)

定理(7)的证明需要用到下面的引理

Lemma 1

设$\boldsymbol{A}$为$n \times n$可逆矩阵，$\boldsymbol{u}, \boldsymbol{v}$都是$n \times 1$列向量，若$(\boldsymbol{A} - \boldsymbol{u}\boldsymbol{v}')$可逆，则有下面恒等式

$$(\boldsymbol{A} - \boldsymbol{u}\boldsymbol{v}')^{-1} = \boldsymbol{A}^{-1} + \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}'\boldsymbol{A}^{-1}}{1 - \boldsymbol{u}'\boldsymbol{A}^{-1}\boldsymbol{v}}.$$

请同学们验证该引理成立。

定理(7)的证明：

$X'X$的讨论

求解$\hat{\beta}$的时候，要求$(X'X)^{-1}$必须存在，这就是要求以下充分必要条件存在

1. $|X'X| \neq 0$, 非奇异矩阵
2. $rank(X'X) = p + 1$, 满秩矩阵
3. $rank X = p + 1$

以及必要条件$n \geq p + 1$。

在本课程里，其实还要假设p为常值，而n可以趋于无穷大。

MLE

如果正态分布假设满足，

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(\mathbf{0}, \sigma^2 I_n)$$

则y的概率分布为：$y \sim N(X\beta, \sigma^2 I_n)$, 这是似然函数为

$$L(\beta) = (2\pi\sigma^2)^{-n/2} exp\{-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\}$$

对其做“$-\ln$”变换，记为$\ell(\beta)$，可以证明β的极大似然估计$\hat{\beta}_{MLE}$与最小二乘估计等价。而误差项的方差σ^2的极大似然估计

$$\hat{\sigma}^2_{MLE} = \frac{1}{n} SSE = \frac{1}{n} e'e,$$

是有偏估计，在$n \to \infty$的情况下，它是渐近无偏估计。

性质

Property 1

$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$是一个线性估计，即它是$\boldsymbol{y}$的线性变换。

Property 2

$\hat{\beta}$是β的无偏估计。

证明：

$$\begin{aligned} E(\hat{\beta}) &= E\{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}\} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E\boldsymbol{y} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\beta = \beta, \end{aligned} \tag{1.21}$$

性质

Property 3

$D(\hat{\beta}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$

证明：

$$
\begin{aligned}
D(\hat{\beta}) &= Cov(\hat{\beta}, \hat{\beta}) \\
&= Cov((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}) \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2\boldsymbol{I}_n\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}, \qquad (1.22)
\end{aligned}
$$

当$p = 1$的时候，就是一元线性回归模型的情况，此时

$$
\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}
$$

性质

同上，$p = 1$时，对$\boldsymbol{X}'\boldsymbol{X}$求逆，就可以得到

$$D(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{bmatrix} Var(\hat{\beta}_0) & Cov(\hat{\beta}_0, \hat{\beta}_1) \\ Cov(\hat{\beta}_0, \hat{\beta}_1) & Var(\hat{\beta}_1) \end{bmatrix}$$

随堂测验一

题目1：对

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-p-1}\boldsymbol{e}'\boldsymbol{e},$$

(1). 证明其无偏性。
(2). 在正态分布假定下，计算$Cov(\boldsymbol{e}, \hat{\boldsymbol{\beta}})$，从而证明$\hat{\sigma}^2$ 与$\hat{\boldsymbol{\beta}}$的独立性。

题目2: 在正态分布假定下，对一元线性回归模型，构造假设检验 $H_0: 2\beta_0 = \beta_1$ v.s. $H_1: 2\beta_0 \neq \beta_1$ 的检验统计量。(提示：$H_0: 2\beta_0 - \beta_1 = 0$)

讲解一下答案

性质

Property 4

高斯-马尔科夫(G-M)定理: 假设$E(\boldsymbol{y}) = \boldsymbol{X}\beta$, $D(\boldsymbol{y}) = \sigma^2 \boldsymbol{I}_n$成立，则$\beta$的任一线性函数$\boldsymbol{c}'\beta$的最小方差线性无偏估计(BLUE)为$\boldsymbol{c}'\hat{\beta}$, 其中，$\boldsymbol{c}$是任一$p+1$维常值列向量，$\hat{\beta}$是$\beta$的最小二乘估计。

证明*: 假设$\boldsymbol{a} = (a_1, \ldots, a_n)^\top \in R^n$, $\boldsymbol{a}'\boldsymbol{y}$为$\boldsymbol{c}'\beta$的一个线性无偏估计，

性质

此性质说明了最小二乘估计$\hat{\beta}$是理想的估计量，注意以下四点

1. 取常值向量$\boldsymbol{c}$的第j个分量为1，其余全是零，这时G-M 定理表明$\hat{\beta}_j$是β_j的最小方差线性无偏估计，以后简记为*BLUE*。

$$\boldsymbol{c} = (0, \ldots, 1, \ldots, 0)^\top$$

2. 当然可能存在$\boldsymbol{y}$的非线性函数，其作为$\boldsymbol{c}'\beta$的无偏估计，比最小二乘估计$\boldsymbol{c}'\hat{\beta}$有更小的方差。
3. 也可能存在$\boldsymbol{c}'\beta$的有偏估计，在某种意义上（比如后面要学的均方误差MSE）比最小二乘估计$\boldsymbol{c}'\hat{\beta}$更好。
4. 在正态假设下，$\boldsymbol{c}'\hat{\beta}$是$\boldsymbol{c}'\beta$的最小方差无偏估计。这时，既不存在$\boldsymbol{y}$的非线性函数，也不存在$\boldsymbol{y}$的其他线性函数，作为$\boldsymbol{c}'\beta$的无偏估计，比最小二乘估计$\boldsymbol{c}'\hat{\beta}$ 的方差更小。

性质

Property 5

在G-M假设下，最小二乘估计$\hat{\beta}$与残差向量$\boldsymbol{e}$不相关，即$Cov(\hat{\beta}, \boldsymbol{e}) = \mathbf{0}$。进一步，在正态假设下，$\hat{\beta}$与$\boldsymbol{e}$独立，从而$\hat{\beta}$与$SSE = \boldsymbol{e}'\boldsymbol{e} = ||\boldsymbol{e}||^2$独立。

证明：随堂测验

Property 6

当$\boldsymbol{y} \sim N(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I}_n)$时，则

$$(1)\hat{\beta} \sim N(\beta, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}); \qquad (2) SSE/\sigma^2 \sim \chi^2(n-p-1).$$

证明：

F检验

在得到多元线性回归方程后，还需要对回归方程进行显著性检验，首先来看回归方程显著性F检验，也就是考察$x_1, \ldots, x_p$从整体上对随机变量y是否有显著的影响，提出原假设

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

H_0被接受和拒绝分别代表什么含义。为了建立关于H_0检验的F统计量，利用总离差平方和分解式，即

$$||\boldsymbol{y} - \bar{y}\mathbf{1}_n||^2 = ||\boldsymbol{y} - \hat{\boldsymbol{y}} + \hat{\boldsymbol{y}} - \bar{y}\mathbf{1}_n||^2 = ||\boldsymbol{y} - \hat{\boldsymbol{y}}||^2 + ||\hat{\boldsymbol{y}} - \bar{y}\mathbf{1}_n||^2, \quad (1.23)$$

简写为

$$SST = SSE + SSR$$

F检验

构造统计量如下

$$F = \frac{SSR/p}{SSE/(n-p-1)} \tag{1.24}$$

正态假设下，当H_0成立时，$F \sim F(p, n-p-1)$分布。F检验还可以推广为对任意的如下H_0做检验

$$H_0 : \beta_{j1} = \beta_{j2} = \ldots = \beta_{jr} = 0,$$

其中$r < p$, $\{j_1, j_2, \ldots, j_r\}$是$\{1, 2, \ldots, p\}$的任意真子集(或看成任何一个排列)。那么其实我们是在比较如下两个模型

$$Model1 : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon \tag{1.25}$$

$$Model0 : y = \beta_0 + \beta_{j_{r+1}} x_{j_{r+1}} + \ldots + \beta_{j_p} x_{j_p} + \varepsilon \tag{1.26}$$

思考这时H_0 被接受或者拒绝代表什么含义。

F检验

这时只需要分别计算Model1的残差平方和SSE_1以及Model0的残差平方和SSE_0, 构造统计量

$$F=\frac{(SSE_0-SSE_1)/r}{SSE_1/(n-p-1)}\sim F(r,n-p-1) \tag{1.27}$$

前面提到的回归方程的显著性检验$H_0:\beta_1=\ldots=\beta_p=0$就相当于在比较如下两个模型

$$Model1: y=\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_px_p+\varepsilon \tag{1.28}$$

$$Model0: y=\beta_0+\varepsilon \tag{1.29}$$

此时Model0的SSE_0就是式(1.23)中Model1的SST(同学们证明一下), 于是在(1.27)中，$F=\frac{SSR/p}{SSE/(n-p-1)}$，与(1.24)等价。因此(1.24)其实是(1.27)的特例。且这里r就是从Model1到Model0减少的变量个数。

F检验

检验$H_0: \beta_{j_1} = \beta_{j_2} = \ldots = \beta_{j_r} = 0$,其实可以用矩阵形式表达$H_0: A\beta = \mathbf{0}$,其中$A$是一个$r \times p$维的矩阵，其行向量的第$j_k$个位置为1其余全是零$\{(0, \ldots, 1, \ldots, 0)\}$，$k = 1, \ldots, r$.

Example 1

对如下两个模型

$$Model1: y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$
$$Model0: y = \beta_0 + \beta_2 x_2 + \beta_4 x_4 + \varepsilon$$

做F检验，就等于对$H_0: \beta_1 = \beta_3 = 0$做检验，也等价于对$H_0: A\beta = \mathbf{0}$做检验，其中$\beta = (\beta_1, \beta_2, \beta_3, \beta_4)'$,

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

注意，$rank(A) = 2$。

推广的F检验

对任意$r \times p$维矩阵A，且$r \leq p$，如果$rank(A) = r$, 则对$H_0 : A\beta = \mathbf{0}$做检验，其统计量也为

$$F = \frac{(SSE_0 - SSE_1)/r}{SSE_1/(n-p-1)} \sim F(r, n-p-1) \tag{1.30}$$

其中SSE_0就是在零假设H_0成立下的简化模型的残差平方和，SSE_1就是全模型的残差平方和。r就是矩阵A的秩, 也就是从Model1到Model0减少的参数个数。当$r = p$, 即A是可逆矩阵时候，$H_0 : A\beta = \mathbf{0}$即等价于$H_0 : \beta = \mathbf{0}$, 回到了回归方程显著性检验的问题。

推广的F检验

以一个例子来说明如何处理一般的$H_0 : A\beta = \mathbf{0}$检验问题。

Example 2

对犯罪率y, 老师收集到了x_1, x_2, x_3三个解释变量，样本量为n。现在对如下全模型做了回归

$$Model1 : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

请问如何检验$H_0 : \beta_1 = -\beta_2 = -\beta_3$？

推广的F检验

1 第一步，将$H_0: \beta_1 = -\beta_2 = -\beta_3$写成$H_0: A\boldsymbol{\beta} = \mathbf{0}$的形式，其中

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}, \quad \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)', \quad \mathbf{0} = (0, 0)'$$

2 第二步，计算$rank(A) = 2$, 计算SSE_1。在假设H_0成立的条件下，即

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 - \beta_1 x_2 - \beta_1 x_3 + \varepsilon \\ &= \beta_0 + \beta_1 (x_1 - x_2 - x_3) + \varepsilon \\ &= \beta_0 + \beta_1 z_1 + \varepsilon \end{aligned}$$

对y关于z_1做回归，并计算SSE_0.

3 第三步，构造统计量

$$F = \frac{(SSE_0 - SSE_1)/2}{SSE_1/(n-3-1)},$$

最后做检验并分析结果。

t检验

多元线性回归中，回归方程显著不代表每个解释变量对y显著，通过对每个解释变量进行显著性检验，可以剔除一些次要的、可有可无的变量。$H_0: \beta_j = 0$ 其实也是上面讲的$H_0: A\boldsymbol{\beta} = \mathbf{0}$的特例，因此t检验本来就和广义F检验是等价的。由性质6，

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

记$(\boldsymbol{X}'\boldsymbol{X})^{-1} = (c_{ij}), i, j = 0, 1, 2, \ldots, p$, 就有

$$E(\hat{\beta}_j) = \beta_j, \quad Var(\hat{\beta}_j) = c_{jj}\sigma^2, \quad \hat{\beta}_j \sim N(\beta_j, c_{jj}\sigma^2)$$

于是可以构造t统计量

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

其中$\hat{\sigma} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^n e_i^2} = \sqrt{\frac{1}{n-p-1}||\boldsymbol{e}||^2} = \sqrt{\frac{1}{n-p-1}SSE}$. 并且可以构造$\beta_j$的置信水平为$1-\alpha$的置信区间

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

t检验

由推广的F检验，$H_0 : \beta_j = 0$可以通过以下三个步骤

1. 第一步，$H_0 : \beta_j = 0$等价于$H_0 : A\beta = 0$， A这里只是一个行向量。
2. 第二步，$rank(A) = 1$, 计算SSE_1。然后在H_0成立的情况下，相当于在Model1中剔除x_j后，做回归，并计算SSE_0。
3. 第三步，构造统计量$F_j = \frac{(SSE_0 - SSE_1)/1}{SSE_1/(n-p-1)}$, 最后做检验并分析结果。$F_j$也称为偏F统计量。

拟合优度

多元线性回归中，也可以定义样本决定系数为

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

其取值在[0,1]之间，越接近1，拟合效果越好；越接近0，拟合效果越差。与F检验比，R^2更清楚直观地反应了回归拟合的效果，但并不严谨。另外，称

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

为y关于$x_1,\ldots,x_p$的样本复相关系数。也是一种拟合程度的度量，整体地衡量$x_1,\ldots,x_p$与y的线性关系。

中心化和标准化

为什么要中心化和标准化？

1. 多元线性回归中，自变量之间单位不同
2. 数据量大，可能因舍入误差导致计算不准确。舍入误差的产生有两个原因：1.数量级的差别；2.病态矩阵的求逆。看如下R程序结果：

```
> b<-matrix(c(1,0.995,0.995,1),nrow=2)
> a<-matrix(c(1,0.99,0.99,1),nrow=2)
> a
     [,1] [,2]
[1,] 1.00 0.99
[2,] 0.99 1.00
> b
      [,1]  [,2]
[1,] 1.000 0.995
[2,] 0.995 1.000
> solve(a)
          [,1]      [,2]
[1,]  50.25126 -49.74874
[2,] -49.74874  50.25126
> solve(b)
          [,1]      [,2]
[1,] 100.25063 -99.74937
[2,] -99.74937 100.25063
```

中心化

多元线性回归模型的经验回归方程为

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p, \tag{1.31}$$

此方程经过样本中心$(\bar{x}_1, \ldots, \bar{x}_p; \bar{y})$, 做坐标变换如下,

$$\begin{aligned} x_{ji}^c &= x_{ji} - \bar{x}_j, i = 1, \ldots, n; j = 1, \ldots, p \\ y_i^c &= y_i - \bar{y}, i = 1, \ldots, n, \end{aligned} \tag{1.32}$$

其中$\bar{y} = 1/n \sum_{i=1}^n y_i, \bar{x}_j = 1/n \sum_{i=1}^n x_{ji}, j = 1, \ldots, p$。然后，经验方程变为

$$\hat{y}^c = \hat{\beta}_1 x_1^c + \ldots + \hat{\beta}_p x_p^c, \tag{1.33}$$

中心化

备注：

1. 中心化经验回归方程的常数项为0，而回归系数的最小二乘估计不变，即$\hat{\beta}_1, \ldots, \hat{\beta}_p$不变。
2. 中心化经验回归方程只包含p个参数，比一般经验回归方程少了一个，在变量较多时，可减少计算工作量（尤其是手工计算）。因此，往往手工计算时，先中心化再估计中心化经验回归方程，再根据
$$\hat{\beta}_0 = \hat{y} - (\hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p)$$
求出常数项估计值β_0。
3. 思考题：(1.31)与中心化后的(1.33) 的回归系数最小二乘估计不变，但他们的t检验统计量也保持不变吗？试用R程序模拟验证或实证。

标准化

为了消除量纲的不同和数量级的差异所带来的影响，还需要对样本数据作标准化处理，然后用最小二乘估计未知参数。标准化的公式为：

$$x_{ji}^* = \frac{x_{ji} - \bar{x}_j}{\hat{\sigma}_{xj}}; \quad y_i^* = \frac{y_i - \bar{y}}{\hat{\sigma}_y}, \quad i = 1, \ldots, n; \quad j = 1, \ldots, p \tag{1.34}$$

其中$\hat{\sigma}_{xj}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ 是自变量x_j的样本方差。标准化后的经验回归方程为，

$$\hat{y}^* = \hat{\beta}_1^* x_1^* + \ldots + \hat{\beta}_p^* x_p^*, \tag{1.35}$$

标准化包括中心化，因此标准化的回归常数项为零。另外，请同学们验证，

$$\hat{\beta}_j^* = \frac{\hat{\sigma}_{xj}}{\hat{\sigma}_y}\hat{\beta}_j, j = 1, \ldots, p$$

中心化

备注：

1. 普通最小二乘估计$\hat{\beta}_j$的解释
2. 标准化回归系数$\hat{\beta}_j^*$的解释
3. 例子，

$$\hat{y} = \pi + 300x_1 + 3x_2$$

其中x_1是以米为单位；x_2是以厘米为单位。

标准化回归系数，增加了比较变量间相对重要性的可能，但仍需谨慎使用。

样本相关阵

复相关系数R反应的是y与一组解释变量整体的相关性，是整体和共性指标；简单相关系数反应的是两个变量间的相关性，是局部和个性指标。实践中，应该整体与局部结合、共性与个性结合。

关于复相关系数R，其实就是$(y_i, \hat{y}_i), i = 1, \ldots, n$的样本相关系数，例子

```
> mod1<-lm(y~M+Ed+U1, data=UScrime)
> y.hat<-fitted(mod1)
> results<-summary(mod1)
> results$r.squared^0.5
[1] 0.3385233
> cor(UScrime$y,y.hat)
[1] 0.3385233
```

Figure 2: 复相关系数R的意义（同学们证明）

复相关系数R的意义

课堂练习一

证明：　若$\{\hat{y}_i, i = 1, \ldots, n\}$是$\{y_i, i = 1, \ldots, n\}$做线性回归后的拟合值，则$y$与$\hat{y}$的样本相关系数$r_{y\hat{y}}$和回归模型的决定系数$R^2$有如下关系

$$r_{y\hat{y}}^2 = \frac{SSR}{SST} = R^2$$

样本相关阵

由样本观测值$\{x_{1i}, x_{2i}, \ldots, x_{pi}, i = 1, \ldots, n\}$，分别计算$x_j$与$x_k$之间的简单相关系数$r_{jk}$, 得自变量样本相关阵

$$\boldsymbol{r} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix} \tag{1.36}$$

相关阵是对称阵。若记$\boldsymbol{X}^* = (\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_p^*)_{n\times p}$为标准化后的设计矩阵，则

$$\boldsymbol{r} = (\boldsymbol{X}^*)'\boldsymbol{X}^*/(n-1)$$

样本相关阵

进一步求出y和每个解释变量x_j的相关系数r_{yj}, 得增广样本相关阵

$$\boldsymbol{r}=\begin{bmatrix} 1 & r_{y1} & r_{y2} & \cdots & r_{yp} \\ r_{1y} & 1 & r_{12} & \cdots & r_{1p} \\ r_{2y} & r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \vdots & \vdots & \\ r_{py} & r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix} \tag{1.37}$$

偏决定系数

若干概念:

1. 复相关系数R是y和$\hat{y}$之间的简单相关系数。
2. 偏相关系数，是在多元线性回归中，固定其他变量后，给定两个变量之间的相关系数。
3. 研究偏相关系数，首先研究偏决定系数
4. 复决定系数$R^2 = \frac{SST - SSE}{SST}$是回归中一组解释变量$x_1, \ldots, x_p$使得响应变量$y$的变差被解释掉的相对量。
5. 偏决定系数则是在回归方程已包含若干解释变量时，再引入某一个新的变量后，y的剩余变差被解释掉的相对量，衡量的是解释变量对y变差解释的边际贡献。

以两个解释变量的偏决定系数为例

二元线性回归模型

$$Model1: y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, i = 1, \ldots, n$$
$$Model0: y_i = \beta_0 + \beta_2 x_{2i} + \varepsilon_i, i = 1, \ldots, n$$

以$SSE(x_2)$表示$Model0$的残差平方和，也就是未被x_2解释的y的变差。$SSE(x_1, x_2)$则是$Model1$的残差平方和，也是未被x_1, x_2解释的y的变差。因此，当模型$Model0$ 已经有x_2时，再加入x_1使得y的剩余变差被解释掉的相对量为：

$$r^2_{y1;2} = \frac{SSE(x_2) - SSE(x_1, x_2)}{SSE(x_2)} \stackrel{?}{=} \frac{SSR(x_1, x_2) - SSR(x_2)}{SSE(x_2)}$$

偏决定系数一般情况

由此，当模型中已有$x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p$时，$y$与$x_j$的偏决定系数为：

$$
\begin{aligned}
r^2_{yj;-j} &= \frac{SSE(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p) - SSE(x_1, \ldots, x_p)}{SSE(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p)} \\
&= \frac{SSR - SSR(-j)}{SSE(-j)}
\end{aligned}
$$

偏相关系数

Definition 1

偏决定系数的平方根称为偏相关系数，其符号与相应的回归系数的符号相同，偏相关系数与回归系数的显著性检验的t值是等价的。

对任意p个变量$x_1, \ldots, x_p$可定义他们之间的偏相关系数。记

$$r_{jk} = \frac{S_{jk}}{\sqrt{S_{jj} S_{kk}}}$$

为x_j, x_j之间的简单相关系数，$\boldsymbol{r} = (r_{jk})_{p \times p}$为相关阵，则在固定$x_3, \ldots, x_p$不变时，$x_1$与$x_2$之间的偏相关系数为：

$$r_{12;3,\ldots,p} = \frac{-\Delta_{12}}{\sqrt{\Delta_{11} \Delta_{22}}}$$

其中Δ_{jk}表示$\boldsymbol{r}$第j行第k列元素的代数余子式，

总结与评注

1. 满足多元线性模型的基本假定时，普通最小二乘估计才可靠
2. R^2的大小与n和p有关，n和p接近时，R^2容易接近1，此时有虚假成分
3. 回归方程的显著性检验与推广的F检验的区别和联系。回归方程的显著性检验，当拒绝H_0时仍需谨慎, 此时可能漏掉了一些重要的自变量；另外，有的回归系数可能并不通过$H_0 : \beta_j = 0$的检验（回归系数的t检验）
4. 回归系数的解释问题，理想和现实两种情况
5. 实际中，难以避免的问题有，多重共线性，非线性，交互作用。 后面的章节专门解决这些问题。