《线性回归》
—线性回归(3)

杨　　瑛

清华大学　数学科学系
Email: yangying@mail.tsinghua.edu.cn
Tel: 62796887

2019.03.14

主要内容：线性模型(3)

- （LSE）残差的性质
- R中的回归: lm
- 如何衡量拟合度？
- 方差分析
- r：样本相关系数

1 多重线性回归
- 统计误差
- 估计与残差
- 计算估计
- 残差的性质
- R^2和调整的$\tilde{R}^2$

(LSE)残差的性质

- $\sum_{i=1}^{n} \widehat{\epsilon_i} = 0$, 因为回归直线经过$(\overline{X}, \overline{Y})$.
- $\sum_{i=1}^{n} \mathbf{X}_i \widehat{\epsilon_i} = 0$, $\widehat{Y}_i \widehat{\epsilon_i} = 0. \Longrightarrow$ 残差与自变量$\mathbf{X}_i$不相关，同时与拟合值$\widehat{Y}_i$不相关。
- 只要协变量的取值不全相等，则最小二乘估计是唯一定义的。如果$\sum_{i=1}^{n} (\mathbf{X}_i - \overline{X})^2 = 0$, 则最小二乘估计不唯一。
  为什么？试解释原因。
  上面的求和等于零意味着什么？

R中的回归: lm

- model $<-$ lm$(y \sim x)$
- summary(model)
- 回归系数：model$coef 或者coef(model)
- 拟合值: model$fitted 或者fitted(model),或者fitted.values(model)
- 残差(residuals): model$resid 或者resid(model) 或者residuals(model)

R中的回归: 模拟数据例子

- R中lm的例子（体重和身高）[以下是模拟的数据]
- 模型数据：$h = runif(30, 165, 185)$,
  $w = 1.1(h - 100) + rnorm(30, 0, 2)$.
- 对（模拟的）身高和体重数据$(h_i, w_i), i = 1, \cdots, 30$, 建立线性模型

$$w_i = \alpha + \beta h_i + \epsilon_i, \epsilon_i \sim N(0, \sigma^2), i = 1, \cdots, 30.$$

- 利用R函数lm可以实现w对h的回归（普通最小二乘估计）

$$L = lm(w \sim h).$$

R中的回归: 模拟数据例子(续)

- summary(L)可以得到如下结果：【电脑演示】
- R的结果对应的模型是【黑板】
- 如何解释回归系数？$\widehat{\beta}$是什么意思？$\hat{\alpha}$是什么意思？解释合理吗？
- $\widehat{\beta}$的解释是合理的。但是$\hat{\alpha}$的解释有点问题？问题处在哪里？如何解决？
- 我们在观测范围之外的变量关系做解释时要特别的小心。
- 解决方案之一：将$h_i(i = 1, \cdots, n)$**进行中心化**
- 注意原始数据和中心化数据之后，β的系数及其检验的值没有发生变化，而α的估计及其检验值发生了变化。试解释其中的原因。

残差标准差

- 残差标准差: $\hat{\sigma} = \sqrt{\text{SSE}/(n-2)} = \sqrt{\frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{n-2}}$.
- $n-2$个自由度(因为我们估计两个参数α和β, 故我们失去了两个自由度)。
- 对于模拟数据,$\hat{\sigma} \approx 2$。解释如下:
  - ✓ 平均而言，使用最小二乘回归线从报告（模拟的）数据预测体重，导致大约2公斤的误差。

R^2

- 对于数据$(\mathbf{X}_i, \mathbf{Y}_i), i = 1, \cdots, n$, 可以建立两个模型：

$$\begin{aligned} M_0 : \mathbf{Y}_i &= \alpha' + \epsilon_i', i = 1, \cdots, n \\ M_1 : \mathbf{Y}_i &= \alpha + \mathbf{X}_i\beta + \epsilon_i, i = 1, \cdots, n \end{aligned}$$

  模型M_0是不包含协变量$\mathbf{X}_i$的模型，称为**空模型**，模型M_1中包含了协变量$\mathbf{X}_i$，称为**简单线性模型**.

- 利用普通的最小二乘方法可以得到模型M_0和M_1中参数的估计：$\widehat{\alpha}' = \overline{Y}$,
  $\widehat{\beta} = \sum_{i=1}^n (\mathbf{X}_i - \overline{X})(\mathbf{Y}_i - \overline{Y}) / \sum_{i=1}^n (\mathbf{X}_i - \overline{X})^2$和$\widehat{\alpha} = \overline{Y} - \overline{X}\widehat{\beta}$.
- M_0的拟合值为：$Y_i' = \overline{Y}, i = 1, \cdots, n$,
  M_1的拟合值为：$Y_i = \widehat{\alpha} + \mathbf{X}_i\widehat{\beta}, i = 1, \cdots, n$.
- M_0的残差为：$\widehat{\epsilon}_i' = \mathbf{Y}_i - \overline{Y}, i = 1, \cdots, n$,
  M_1的残差为：$\widehat{\epsilon}_i = \mathbf{Y}_i - \widehat{\alpha} - \mathbf{X}_i\widehat{\beta} = \mathbf{Y}_i - \widehat{\mathbf{Y}}_i, i = 1, \cdots, n$.

R^2(续)

- M_0的残差平方和：$\mathsf{TSS} = \sum_{i=1}^{n}(\mathbf{Y}_i - \overline{Y})^2$,[又称为**总平方和**]
  M_1的残差平方和：$\mathsf{SSE} = \sum_{i=1}^{n}(\mathbf{Y}_i - \widehat{\mathbf{Y}}_i)^2$.
  注意到

$$\begin{aligned}\mathsf{TSS} &= \sum_{i=1}^{n}([\mathbf{Y}_i - \widehat{\mathbf{Y}}_i] - [\widehat{\mathbf{Y}}_i - \overline{Y}])^2 \\ &= \sum_{i=1}^{n}[\mathbf{Y}_i - \widehat{\mathbf{Y}}_i]^2 + \sum_{i=1}^{n}[\widehat{\mathbf{Y}}_i - \overline{Y}]^2 \\ &= \mathsf{SSE} + \mathsf{RegSS},\end{aligned}$$

  其中

$$\mathsf{RegSS} = \sum_{i=1}^{n}[\widehat{\mathbf{Y}}_i - \overline{Y}]^2 = \sum_{i=1}^{n}(\mathbf{X}_i - \overline{X})^2\widehat{\beta}^2,$$

  称之为**回归平方和**。

R^2(续)

- 从上面的式子，很容易看出：

$$\text{SSE} \leq \text{TSS} \text{ (为什么？)}$$

- $\text{TSS} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$为总平方和, 即空模型$M_0$中的残差平方和。
- SSE为残差平方和，即模型M_1中的残差平方和。
- 回归平方和: $\text{RegSS} = \text{TSS} - \text{SSE} = \sum_{i=1}^{n}(\mathbf{X}_i - \overline{X})\widehat{\beta}^2$。
- 定义确定系数：$R^2 = \text{RegSS}/\text{TSS} = 1 - \text{SSE}/\text{TSS}$
  总平方和中回归平方和所占比例。
- 也可以解释为R^2是由线性回归中能够解释Y的变化比例。

R^2(续)

- 从TSS $=$ SSE $+$ RegSS, 或者R^2的表达式可以看出来，如果SSE/TSS的值越靠近1, 则$\mathbf{X}$对$\mathbf{Y}$的（线性）的影响越弱，反之则越强。即，R^2越大，$\mathbf{X}$对$\mathbf{Y}$的影响可能就越大。
- R^2是无量纲的$\Longrightarrow$ 不随着尺度或者量纲的变化而改变。
- R^2的“好”值在不同的应用领域有很大差异。

方差分析

- $\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)(\widehat{Y}_i - \overline{Y}) = 0$ (几何解释！)
- RegSS $= \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$ (证明！)
- 因此，

TSS	=	SSE	RegSS
$\sum_{i=1}^{n}(Y_i - \overline{Y})^2$	=	$\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$	$\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$
总离差平方和		残差平方和	回归平方和
$n-1$		$n-1$	1

这个分解称为方差分析（**analysis of variance**）.

说明：

不同的教材中这些平方和的表示方法有可能不一样。更多方差分析的内容请阅读Draper和Smith 的《Applied Regression Analysis》中的1.3节中内容. 后面还要学习ANOVA的内容.

r

- 相关系数$r = \pm R^2$ (如果$\widehat{\beta} > 0$, 则取正号，如果$\widehat{\beta} < 0$, 则取负号).
- r表示$\mathbf{X}$ 和$\mathbf{Y}$ 的关系的强度和方向。
- 公式: $r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$.
- 利用这个公式，可以得到$\widehat{\beta} = r\frac{\mathrm{SD}_Y}{\mathrm{SD}_X}$ (推导!)。
- 在目测回归中，陡峭的直线有斜率$\frac{\mathrm{SD}_Y}{\mathrm{SD}_X}$,而另一条直线的斜率$\widehat{\beta} = r\frac{\mathrm{SD}_Y}{\mathrm{SD}_X}$是正确的。
- r 在X和Y中是对称的。
- r 没有单位$\Longrightarrow$不随单位的改变而改变。

多个独立协变量

- $Y = \alpha + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \epsilon$.
- 这个模型描述了三维空间$\{\mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}\}$中的平面.
  - ✓ α是截距
  - ✓ 当X_2保持恒定时，β_1是X_1增加一个单位时$\mathbf{Y}$相应增加的量。
  - ✓ 当X_1保持恒定时，β_2是X_2增加一个单位时$\mathbf{Y}$相应增加的量。

统计误差

- 数据: $(\mathbf{X}_{11}, \mathbf{X}_{12}, \mathbf{Y}_1), \cdots, (\mathbf{X}_{n1}, \mathbf{X}_{n2}, \mathbf{Y}_n)$.
- 建立模型： $\mathbf{Y}_i = \alpha + \beta_1 \mathbf{X}_{i1} + \beta_2 \mathbf{X}_{i2} + \epsilon_i$, 其中$\epsilon_i$是第$i$个观测的统计误差。
- 因此观测值$\mathbf{Y}_i$等于$\alpha + \beta_1 \mathbf{X}_{i1} + \beta_2 \mathbf{X}_{i2}$,但是要相差$\epsilon_i$是未知的随机量。
- 我们对ϵ_i做出与以前相同的假设(预先假设了ϵ_i与$\mathbf{X}_i$独立)：
  - ✓ $\mathbf{E}[\epsilon_i] = 0, i = 1, \cdots, n$
  - ✓ $\mathsf{Var}(\epsilon_i) = \sigma^2, i = 1, \cdots, n$
  - ✓ $\mathsf{Cov}(\epsilon_i, \epsilon_j) = 0, i \neq j$

估计与残差

- 总体参数$\alpha, \beta_1, \beta_2, \sigma$是未知的
- 我们可以计算总体参数的估计：$\widehat{\alpha}, \widehat{\beta}_1, \widehat{\beta}_2, \widehat{\sigma}$
- $\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}_1 \mathbf{X}_{i1} + \widehat{\beta}_2 \mathbf{X}_{i2}$称为拟合值。
- $\widehat{\epsilon}_i = \mathbf{Y}_i - (\widehat{\alpha} + \widehat{\beta}_1 \mathbf{X}_{i1} + \widehat{\beta}_2 \mathbf{X}_{i2})$称为残差
- 残差是可观察的，可用于检验关于统计误差ϵ_i的假设。
- 平面上方的点具有正残差，平面下方的点具有负残差。
- 适合数据的平面具有较小的残差。

计算估计

- 最小化SSE$(\alpha, \beta_1, \beta_2) = \sum_{i=1}^{n} \widehat{\epsilon}_i^2 = \sum_{i=1}^{n} (\mathbf{Y}_i - \alpha - \beta_1 \mathbf{X}_{i1} - \beta_2 \mathbf{X}_{i2})^2$得到估计量$\widehat{\alpha}, \widehat{\beta}_1, \widehat{\beta}_2$.
- 我们可以对SSE$(\alpha, \beta_1, \beta_2)$求偏导并让它们等于0。
- 这样就给出了三个方程关于未知数α, β_1, β_2的三个方程。解这些**正规方程** (normal equations)得到回归系数的估计：$\widehat{\alpha}, \widehat{\beta}_1, \widehat{\beta}_2$.
- 最小二乘估计是唯一的，除非其中一个自变量是不变的，或者自变量是完全共线的。
- 对p个协变量$\mathbf{X_1}, \cdots, \mathbf{X_p}$的多重线性回归模型也是一样的。但是，使用矩阵表示法更容易(我们可以在这里具体推导！)。
- 在R中: model= lm(y∼ x1+x2); 如果拟合没有截距项的回归模型，则model= lm(y∼ 0+x1+x2);

残差的性质

- $\sum_{i=1}^{n} \widehat{\epsilon_i} = 0.$
- 残差$\widehat{\epsilon_i}$与拟合值$\widehat{Y_i}$不相关，与协变量$\mathbf{X_1}, \cdots, \mathbf{X_p}$中的任何一个不相关。
- 残差的标准误差$\widehat{\sigma} = \sqrt{\sum_{i=1}^{n} \widehat{\epsilon_i^2}/(n-p-1)}$给出了残差的“平均”大小。
- $n-p-1$ 为自由度(因为我们估计$p+1$个参数$\alpha, \beta_1, \cdots, \beta_p$，故我们失去$p+1$自由度.

R^2和调整的$\tilde{R}^2$

- $\mathsf{TSS} = \sum_{i=1}^{n}(\mathbf{Y}_i - \overline{Y})^2$.
- $\mathsf{SSE} = \sum_{i=1}^{n}(\mathbf{Y}_i - \widehat{\mathbf{Y}}_i)^2 = \sum_{i=1}^{n}\widehat{\epsilon}_i^2$ (残差平方和)
- $\mathsf{RegSS} = \mathsf{TSS} - \mathsf{SSE} = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$.
- $R^2 = \mathsf{RegSS}/\mathsf{TSS} = 1 - \mathsf{SSE}/\mathsf{TSS}$
  表示变异中由$\mathbf{Y}$对$\mathbf{X_1}, \cdots, \mathbf{X_p}$的线性回归可以解释的比例。
- 当我们向模型中添加额外的变量时，R^2永远不会减少。为什么？
- 调整R^2:
$$\tilde{R}^2 = 1 - \frac{\mathsf{SSE}/(n-p-1)}{\mathsf{TSS}/(n-1)},$$
  当模型中有额外变量时，惩罚原来的R^2.
- 如果样本量很大的话，R^2和$\tilde{R}^2$相差很小。

【第七讲结束】