《线性回归》
—线性回归(4)

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(4)

- 臭氧例子
- 臭氧数据
- R的计算结果
- 标准化系数
- 使用hinge spread
- (Hinge Speaad标准化)系数的解释
- 使用标准差(st. dev)
- (st. dev. 标准化)系数的解释
- 附加变量图(added variable plot)
- 总结

臭氧例子

- 数据来自Sandberg, Basso, Okin (1978)：
  - ✓ SF = San Francisco 夏季小时平均臭氧读数的最大值，单位为百万分之一
  - ✓ SJ = 同上，但是在San Jose
  - ✓ YEAR = 臭氧测量年
  - ✓ RAIN = 旧金山湾区前两个冬季平均冬季降水量，以厘米为单位
- 研究问题: SF如何依赖于年份YEAR和降雨量RAIN?
- 关于假设:哪个假设可能被违反?

J. S. SANDBERG, M. J. BASSO, B. A. OKIN (1978), Winter Rain and Summer Ozone: A Predictive Relationship, *Science*, 200,1051-1054

臭氧数据

YEAR	RAIN	SF	SJ
1965	18.9	4.3	4.2
1966	23.7	4.2	4.8
1967	26.2	4.6	5.3
1968	26.6	4.7	4.8
1969	39.6	4.1	5.5
1970	45.5	4.6	5.6
1971	26.7	3.7	5.4
1972	19.0	3.1	4.6
1973	30.6	3.4	5.1
1974	34.1	3.4	3.7
1975	23.7	2.1	2.7
1976	14.6	2.2	2.1
1977	7.6	2.0	2.5

R的计算结果

model=lm(SF∼YEAR+RAIN,data=dat)
summary(model)
Call:
lm(formula=SF ∼ YEAR + RAIN, data = dat)
Residuals:

Min	1Q	Median	3Q	Max
-0.61072	-0.20317	0.06129	0.16329	0.51992

Coefficients:

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	388.412083	49.573690	7.835	1.41e-05 $\star\star\star$
YEAR	-0.195703	0.025112	-7.793	1.48e-05$\star\star\star$
RAIN	0.034288	0.009655	3.551	0.00526 $\star\star$

R的计算结果(续)

Signif. codes: 0 ‘⋆⋆⋆’ 0.001 ‘⋆⋆’ 0.01 ‘⋆’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3224 on 10 degrees of freedom

Multiple R-squared: 0.9089, Adjusted R-squared: 0.8906

F-statistic: 49.87 on 2 and 10 DF, p-value: 6.286e-06

说明：

1. 要正确理解输出结果的每一个部分
2. 进阶：

✓ 利用ls(model)查看model中所包含的部分。

ls(model)

”assign” ”call” ”coefficients” ”df.residual” ”effects” ”fitted.values” ”model” ”qr” ”rank” ”residuals” ”terms” ”xlevels”

✓ 用model$residuals 进一步了解每一个项的含义。

✓ 也可以赋值：res=model$residuals, 做进一步的分析和计算。验证：$\sum_{i=1}^{n} \widehat{\epsilon_i} = 0$: sum(res). 也可以验证：‘Residual standard error: 0.3224’

标准化系数

- 我们经常要比较不同自变量的系数。
- 当自变量用相同的单位测量时，系数的比较直接的。
- 如果自变量是用不同的单位测量的（例如，有的是重量单位，有的是长度单位），我们可以通过使用变异度
  - ✓ hinge spread（分散度）
    [数据的75%分位数减去25%的分位数]
  - ✓ 标准偏差

  来重新调整回归系数来进行有限的比较。

说明：hinge 的定义和来源

John W. Tukey, (1977). **Exploratory Data Analysis.** Page 32: ” ... of 13 values appears as follows: -3.2,-1.7, -0.4, 0.1, 0.3,1.2,1.5,1.8, 2.4, 3.0, 4.3, 6.4, 9.8. The five summary numbers are, in order, -3.2, 0.1,1.5, 3.0 and 9.8, one ar each folding point. ... the 5 numbers (extremes, hinges, median) that make up a **5-number summary**”

使用hinge spread

- Hinge spread = interquartille range (IQR)[四分位数间距]
- 设$\mathrm{IQR}_1, \cdots, \mathrm{IQR}_p$是$\mathbf{X_1}, \cdots, \mathbf{X_p}$的IQR。
- 注意到：$\mathbf{Y}_i = \widehat{Y}_i + (\mathbf{Y}_i - \widehat{Y}_i) = \widehat{Y}_i + \widehat{\epsilon}_i$，我们从等式$Y_i = \widehat{\alpha} + \widehat{\beta}_1 \mathbf{X}_{i1} + \cdots + \widehat{\beta}_p \mathbf{X}_{ip} + \widehat{\epsilon}_i$出发。
- 这个等式可以改写为：$Y_i = \widehat{\alpha} + \left(\widehat{\beta}_1 \mathrm{IQR}_1\right) \frac{\mathbf{X}_{i1}}{\mathrm{IQR}_1} + \cdots + \left(\widehat{\beta}_p \mathrm{IQR}_p\right) \frac{\mathbf{X}_{ip}}{\mathrm{IQR}_p} + \widehat{\epsilon}_i$
- 令$Z_{ij} = \frac{\mathbf{X}_{ij}}{\mathrm{IQR}_j}, j = 1, \cdots, p, \ i = 1, \cdots, n.$
- 令$\widehat{\beta}_j^{\star} = \widehat{\beta}_j \mathrm{IQR}_j, j = 1, \cdots, p.$
- 则我们有$\mathbf{Y}_i = \widehat{\alpha} + \widehat{\beta}_1^{\star} Z_{i1} + \cdots + \widehat{\beta}_p^{\star} Z_{ip} + \widehat{\epsilon}_i.$
- $\widehat{\beta}_j^{\star} = \widehat{\beta}_j \mathrm{IQR}_j$ 称为**标准化的回归系数**。它们可以进行比较.

说明：也可以先对每个协变量用中位数中心化，然后再标准化。

(Hinge Speaad标准化)系数的解释

- 解释: 保持$Z_\ell(\ell \neq j)$不变，将Z_j增加1个单位，$\widehat{\beta}_j^\star$是响应变量$\mathbf{Y}$平均增加的量。
- Z_j增加1，意味着X_j增加X_j的一个IQR。
- 所以在保持$\mathbf{X}_\ell(\ell \neq j)$不变的条件之下，$X_j$增加$\mathbf{X}_j$的一个IQR之后，响应变量$\mathbf{Y}$将平均增加$\beta_j^\star$.
- 对于臭氧的例子，有

变量	系数	Hinge Spread	标准化的系数
YEAR	-0.196	6	-1.176
Rain	0.034	11.6	0.394

使用标准差(st. dev.)

- 令S_Y表示$\mathbf{Y}$的标准差，$S_1, \cdots, S_p$分别表示$\mathbf{X_1}, \cdots, \mathbf{X_p}$的标准差。
- 注意到：$\mathbf{Y}_i = \widehat{\mathbf{Y}}_i + (\mathbf{Y}_i - \widehat{\mathbf{Y}}_i) = \widehat{\mathbf{Y}}_i + \widehat{\epsilon}_i$，我们从等式$\mathbf{Y}_i = \widehat{\alpha} + \widehat{\beta}_1 \mathbf{X}_{i1} + \cdots + \widehat{\beta}_p \mathbf{X}_{ip} + \widehat{\epsilon}_i$出发。
- 这个等式可以改写为：$\frac{\mathbf{Y}_i - \overline{\mathbf{Y}}}{S_\mathbf{Y}} = \left(\widehat{\beta}_1 \frac{S_1}{S_\mathbf{Y}}\right) \frac{\mathbf{X}_{i1} - \overline{X}_1}{S_1} + \cdots + \left(\widehat{\beta}_p \frac{S_p}{S_Y}\right) \frac{\mathbf{X}_{ip} - \overline{X}_p}{S_p} + \frac{\widehat{\epsilon}_i}{S_Y}$.
- 令$Z_{i\mathbf{Y}} = \frac{Y_i - \overline{Y}}{S_Y}$, $Z_{ij} = \frac{\mathbf{X}_{ij} - \overline{X}_j}{S_j}$, $j = 1, \cdots, p$, $i = 1, \cdots, n$.
- 令$\widehat{\beta}_j^\star = \widehat{\beta}_j \frac{S_j}{S_\mathbf{Y}}$, $\widehat{\epsilon}_i^\star = \frac{\widehat{\epsilon}_i}{S_\mathbf{Y}}$, $j = 1, \cdots, p$, $i = 1, \cdots, n$.
- 则我们有$Z_{i\mathbf{Y}} = \widehat{\beta}_1^\star Z_{i1} + \cdots + \widehat{\beta}_p^\star Z_{ip} + \widehat{\epsilon}_i^\star$.
- $\widehat{\beta}_j^\star = \widehat{\beta}_j \frac{S_j}{S_\mathbf{Y}}$ 称为**标准化的回归系数**。它们可以进行比较.

注意：请注意两种标准化过程的异同！

(st. dev. 标准化)系数的解释

- 解释: 保持$Z_\ell (\ell \neq j)$不变，将Z_j增加1个单位，$\widehat{\beta}_j^\star$是响应变量$Z_{\mathbf{Y}}$平均增加的量。
- Z_j增加1，意味着X_j增加S_j (X_j的一个SD)。
- $Z_{\mathbf{Y}}$增加1，意味着$\mathbf{Y}$增加$S_{\mathbf{Y}}$ ($\mathbf{Y}$的一个SD)。
- 所以在保持$\mathbf{X}_\ell (\ell \neq j)$不变的条件之下，$X_j$增加$S_j$ ($\mathbf{X}_j$的一个SD之后)，响应变量$\mathbf{Y}$将平均增加$\beta_j^\star \times S_{\mathbf{Y}}$.
- 对于臭氧的例子，结果是:

变量	系数	$\frac{\text{St.dev(variable)}}{\text{St.dev}(\mathbf{Y})}$	标准化的系数
YEAR	-0.196	3.99	-0.783
Rain	0.034	10.39	0.353

- 两种方法（使用hinge spread或标准偏差）仅允许非常有限的比较，都假定具有较大差异的预测因子更为重要，但事实并非总是如此。

附件变量图
【阅读这一页的内容时，请逐行运行R去理解结果】

- 假设我们从SF ~ YEAR开始。
- 我们想知道添加变量RAIN是否有助于SF.
- 我们想对那些未被变量YEAR解释的SF部分建立模型（查看lm（SF~YEAR）的残差），主要是利用RAIN中不能被YEAR解释的部分（lm(RAIN ~YEAR)的残差）
- 将这些残差相互绘制图形，称为RAIN对SF影响的附加变量图（added variable plot），控制YEAR。
- lm(SF~YEAR)的残差对lm(RAIN~YEAR)的残差进行回归，给出RAIN的系数。

上述过程在R中的实现

- M1=lm(SF∼YEAR); summary(M1) #查看结果
- M2= lm(RAIN∼YEAR); summary(M2) #查看结果
- 提取M1和M2的残差：
  ResM1=M1$resid;
  ResM2=M2$resid;
  然后ResM1对ResM2进行回归，
  Mres12=lm(ResM1∼ResM2); summary(Mres12) #查看结果
- M0=lm(FS∼YEAR+RAIN);
  试比较M0$coef中RAIN的系数与Mres12$coef中RAIN 的系数。
  事实上，这两个对应于RAIN的系数是相同的，试说明这样做的理由。【作业！】

总结：

- 线性统计模型：$\mathbf{Y} = \alpha + \beta_1 \mathbf{X}_1 + \cdots + \beta_p \mathbf{X}_p + \epsilon$.
- 我们假设统计误差ϵ的均值为0，恒定的标准偏差为σ，并且是不相关的。
- 总体的参数$\alpha, \beta_1, \cdots, \beta_p$和$\sigma$是不可观测的。此外，统计误差$\epsilon$也是无法观察的。
- 我们定义**拟合值** $\widehat{\mathbf{Y}}_i = \widehat{\alpha} + \widehat{\beta}_1 \mathbf{X}_{i1} + \cdots + \widehat{\beta}_p \mathbf{X}_{ip}$
  和**残差**$\widehat{\epsilon}_i = \mathbf{Y}_i - \widehat{\mathbf{Y}}_i$.
  我们可以使用残差来检查有关统计误差的各种假设。
- 我们通过最小化残差平方和
  $\text{SSE} = \sum_{i=1}^n \widehat{\epsilon}_i^2 = \sum_{i=1}^n (\mathbf{Y}_i - (\widehat{\alpha} + \widehat{\beta}_1 \mathbf{X}_{i1} + \cdots + \widehat{\beta}_p \mathbf{X}_{ip}))^2$
  来求出$\alpha, \beta_1, \cdots, \beta_p$的估计$\widehat{\alpha}, \widehat{\beta}_1, \cdots, \widehat{\beta}_p$.
  但是, σ^2的估计要通过残差来估计。

总结

- 系数的解释？【要得到系数合理的解释，对协变量要中心化或者标准化。注意，如果协变量时分类变量，例如，性别，则不需要中心化和标准化。】
- 为了衡量模型拟合的好坏程度，我们可以使用:
  - ✓ 残差标准误：$\hat{\sigma} = \sqrt{\text{SSE}/(n-p-1)}$【有截距项的模型】
  - ✓ 多重相关系数R^2
  - ✓ 调整后的多重相关系数$\tilde{R}^2$
  - ✓ 相关系数r
- 方差分析(ANOVA): TSS = SSE + RegSS
- 标准化回归系数
- 附加变量图(偏回归图)

【第八章结束】