

《线性回归》 —线性回归

杨 瑛

清华大学 数学科学系

Email: yangying@mail.tsinghua.edu.cn

Tel: 62796887

2019.02.24

主要内容：线性模型

1 线性回归

- 概要

2 简单线性回归

- 线性回归
- 线性模型
- 小的残差
- 极小化 $\sum_{i=1}^n \hat{\epsilon}_i^2$
- 估计参数 α 和 β 的其它方法
- 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

3 多重线性回归

- 统计误差
- 估计与残差

概要

- ♠ 我们已经看到线性回归有其局限性。然而，线性回归还是值得研究的，因为：
 - ✓ 有时数据(几乎)满足假设。
 - ✓ 有时，通过转换数据【后面的一个专题】可以(几乎)满足假设。
 - ✓ 线性回归有许多有用的扩展：
加权回归、稳健回归、非参数回归和广义线性模型。
- ♠ 线性回归是如何进行的？我们从一个自变量的回归模型开始。

线性模型

- ♠ 线性统计模型: $\mathbf{Y} = \alpha + \beta\mathbf{X} + \epsilon$.
- ♠ α 是直线的截距, β 是直线的斜率。 \mathbf{X} 增加一个单位, 则 \mathbf{Y} (平均) 增加 β 个单位。
- ♠ ϵ 称为统计误差, 很多时候称为随机误差 (random error)。它解释了统计模型没有给出数据精确拟合这一事实。
- ♠ 统计误差有随机的部分, 同时可以包括固定的部分。
 - ✓ 固定部分: 当真实关系不是线性的时候出现 (也称为缺乏拟合误差, 偏差)
 - ✓ 随机部分: 由于 \mathbf{Y} 中有测量误差, 由模型中未包含的变量和随机变化组成

线性模型

- ♠ 数据 $(\mathbf{X}_1, \mathbf{Y}_1), \dots, (\mathbf{X}_n, \mathbf{Y}_n)$.
- ♠ 然后模型假定为:
$$\mathbf{Y}_i = \alpha + \beta \mathbf{X}_i + \epsilon_i, i = 1, \dots, n$$

其中 ϵ_i 是第 i 个个体的统计误差。
- ♠ 因此, 观测值 \mathbf{Y}_i 几乎等于 $\alpha + \beta \mathbf{X}_i$, 除了添加一个未知的随机量 ϵ_i 。
- ♠ 统计误差 ϵ_i 是无法观测到的。为什么?
- ♠ 我们假定 ϵ_i 与 \mathbf{X}_i 独立, 且
 - ✓ $\mathbf{E}[\epsilon_i] = 0, i = 1, \dots, n;$
 - ✓ $\text{Var}[\epsilon_i] = \sigma^2, i = 1, \dots, n;$
 - ✓ $\text{Cov}(\epsilon_i, \epsilon_j) = 0$, 对于任意的 $i \neq j$ 。
- ♠ 请仔细体会这里面蕴含着的各种假设。

线性模型

- ♠ 总体参数 α, β 和 σ 是未知的, 是不可以观测的。我们用小写的希腊字母表示总体参数。
- ♠ 我们的目标是估计总体参数: $\hat{\alpha}, \hat{\beta}, \hat{\sigma}$ 。
- ♠ $\hat{\mathbf{Y}}_i = \hat{\alpha} + \hat{\beta}\mathbf{X}_i$ 称为拟合值(fitted value)。
- ♠ $\hat{\epsilon}_i = \mathbf{Y}_i - \hat{\mathbf{Y}}_i = \mathbf{Y}_i - (\hat{\alpha} + \hat{\beta}\mathbf{X}_i)$ 称为残差(residual)。
- ♠ 残差是可观测的, 可以用来检验关于统计误差 ϵ_i 的假设。
- ♠ 直线上的点的残差为正, 直线下的点的残差为负。
- ♠ 符合数据的直线有较小的残差。

小的残差

- ♠ 我们希望残差的幅度很小，因为大的负残差和大的正残差一样糟糕。
- ♠ 所以我们不能简单地要求： $\sum_{i=1}^n \hat{\epsilon}_i = 0$.
- ♠ 事实上，任何通过点 (\bar{X}, \bar{Y}) 的直线都满足 $\sum_{i=1}^n \hat{\epsilon}_i = 0$ (需要仔细的推导！)
- ♠ 两个直接的解决方案：
 - ✓ 要求 $\sum_{i=1}^n |\hat{\epsilon}_i|$ 小；
 - ✓ 要求 $\sum_{i=1}^n \hat{\epsilon}_i^2$ 小；
- ♠ 我们考虑第二个选择，因为从数学上讲，使用平方比使用绝对值更容易处理(例如，求导更容易)。然而，第一个选项对异常值更有抵抗力。
- ♠ 目测回归(从图形上直接看出和猜出回归结果)。

第四讲结束。

极小化 $\sum_{i=1}^n \hat{\epsilon}_i^2$ 极小化 $\sum_{i=1}^n \hat{\epsilon}_i^2$

- ♠ SSE代表平方误差和(Sum of Squared Error)
- ♠ 我们欲求出 $(\hat{\alpha}, \hat{\beta})$ 极小化 $\text{SSE}(\alpha, \beta) = \sum_{i=1}^n (\mathbf{Y}_i - \alpha - \beta \mathbf{X}_i)^2$.
- ♠ 因此，我们对 $\text{SSE}(\alpha, \beta)$ 关于 α 和 β 求偏导数并令它们等于零：

$$\frac{\partial \text{SSE}(\alpha, \beta)}{\partial \alpha} = \sum_{i=1}^n -2(\mathbf{Y}_i - \alpha - \beta \mathbf{X}_i) = 0$$

$$\Rightarrow \sum_{i=1}^n (\mathbf{Y}_i - \alpha - \beta \mathbf{X}_i) = 0,$$

$$\frac{\partial \text{SSE}(\alpha, \beta)}{\partial \beta} = \sum_{i=1}^n -\mathbf{X}_i(\mathbf{Y}_i - \alpha - \beta \mathbf{X}_i) = 0$$

$$\Rightarrow \sum_{i=1}^n \mathbf{X}_i(\mathbf{Y}_i - \alpha - \beta \mathbf{X}_i) = 0,$$

- ♠ 我们现在有两个关于 α 和 β 的正规方程。其解是

$$✓ \quad \hat{\beta} = \frac{\sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{Y}_i - \bar{\mathbf{Y}})}{\sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})^2},$$

$$✓ \quad \hat{\alpha} = \bar{\mathbf{Y}} - \hat{\beta} \bar{\mathbf{X}}.$$

估计参数 α 和 β 的其它方法

- ♠ 对于简单线性回归模型： $\mathbf{Y}_i = \alpha + \beta \mathbf{X}_i + \epsilon_i, i = 1, \dots, n$, 寻求参数 α 和 β 的估计就是找一条直线可以拟合观测数据 $(\mathbf{Y}_i, \mathbf{X}_i), i = 1, \dots, n$, 使得这 n 个点尽可能的与所找的直线尽可能的“近”。
- ♠ 从几何直观上如何刻画点与直线的‘近’?
- ♠ 不同的‘近’, 会导出不同的方法!
- ♠ MLE: 如果对随机误差的分布有更多的认识的话, 可以采用!

残差的性质

- ♠ $\sum_{i=1}^n \hat{\epsilon}_i = 0$, 因为回归直线经过 (\bar{X}, \bar{Y}) .
- ♠ $\sum_{i=1}^n \mathbf{X}_i \hat{\epsilon}_i = 0, \hat{Y}_i \hat{\epsilon}_i = 0 \implies$ 残差与自变量 \mathbf{X}_i 不相关, 同时与拟合值 \hat{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 $\sum_{i=1}^n (\mathbf{X}_i - \bar{X})^2 = 0$, 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 ~ 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lm的例子 (体重和身高)
- ♠ 模型数据: $h = \text{runif}(20, 165, 185)$,
 $w = 1.1(h - 100) + \text{rnorm}(20, 0, 2)$.

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

- ♠ 比较拟合模型和空的模型 $\mathbf{Y} = \alpha' + \epsilon'$, 空模型中不包含任何的协变量 \mathbf{X} .
- ♠ 对于空模型, 我们定义拟合值 $\hat{\mathbf{Y}}'_i = \hat{\alpha}'$ 和残差 $\hat{\epsilon}'_i = \mathbf{Y}_i - \hat{\mathbf{Y}}_i$.
- ♠ 其中 $\hat{\alpha}'$ 通过极小化 $\sum_{i=1}^n (\hat{\epsilon}'_i)^2 = \sum_{i=1}^n (\mathbf{Y}_i - \alpha)^2$ 得到, 这样 $\hat{\alpha}' = \bar{Y}$.
- ♠ 注意
$$\sum_{i=1}^n (\mathbf{Y}_i - \hat{\mathbf{Y}}_i)^2 = \sum_{i=1}^n \hat{\epsilon}'_i \leq \sum_{i=1}^n (\hat{\epsilon}'_i)^2 = \sum_{i=1}^n (\mathbf{Y}_i - \bar{Y})^2$$
(为什么?)

R^2

- ♠ $TSS = \sum_{i=1}^n (\hat{\epsilon}_i')^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2$ 为总平方和: 模型中不使用协变量的误差平方和。
- ♠ SSE是线性模型中误差的平方和(或者为残差平方和)。
- ♠ 回归平方和: $RegSS = TSS - SSE$ 回归平方和。
- ♠ $R^2 = RegSS/TSS = 1 - SSE/TSS$ 平方误差中由线性回归减少比例。
- ♠ 因此, R^2 是由线性回归中能够解释的 Y 变化的比例。
- ♠ R^2 是无量纲的 \implies 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

♠ $\sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 0$ (几何解释!)

♠ $\text{RegSS} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ (证明!)

♠ 因此,

TSS	=	SSE	RegSS
$\sum_{i=1}^n (Y_i - \bar{Y})^2$	=	$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$
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 $\hat{\beta} > 0$, 则取正号, 如果 $\hat{\beta} < 0$, 则取负号).
- ♠ r 表示 **X** 和 **Y** 的关系的强度和方向。
- ♠ 公式:
$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}.$$
- ♠ 利用这个公式, 可以得到 $\hat{\beta} = r \frac{SD_Y}{SD_X}$ (推导!).
- ♠ 在目测回归中, 陡峭的直线有斜率 $\frac{SD_Y}{SD_X}$, 而另一条直线的斜率 $\hat{\beta} = r \frac{SD_Y}{SD_X}$ 是正确的。
- ♠ r 在 X 和 Y 中是对称的。
- ♠ r 没有单位 \Rightarrow 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: $(\mathbf{X}_{11}, \mathbf{X}_{12}, \mathbf{Y}_1), \dots, (\mathbf{X}_{n1}, \mathbf{X}_{n2}, \mathbf{Y}_n)$.
- ♠ 建立模型: $\mathbf{Y}_i = \alpha + \beta_1 \mathbf{X}_{i1} + \beta_2 \mathbf{X}_{i2} + \epsilon_i$, 其中 ϵ_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 ϵ_i 是未知的随机量。
- ♠ 我们对 ϵ_i 做出与以前相同的假设:
 - ✓ $\mathbf{E}[\epsilon_i] = 0, i = 1, \dots, n$
 - ✓ $\text{Var}(\epsilon_i) = \sigma^2, i = 1, \dots, n$
 - ✓ $\text{Cov}(\epsilon_i, \epsilon_j) = 0, i \neq j$

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： $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}$
- ♠ $\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 \mathbf{X}_{i1} + \hat{\beta}_2 \mathbf{X}_{i2}$ 称为拟合值。
- ♠ $\hat{\epsilon}_i = \mathbf{Y}_i - (\hat{\alpha} + \hat{\beta}_1 \mathbf{X}_{i1} + \hat{\beta}_2 \mathbf{X}_{i2})$ 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 $SSE(\alpha, \beta_1, \beta_2) = \sum_{i=1}^n \hat{\epsilon}_i^2 = \sum_{i=1}^n (\mathbf{Y}_i - \alpha - \beta_1 \mathbf{X}_{i1} - \beta_2 \mathbf{X}_{i2})^2$ 得到估计量 $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2$.
- ♠ 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: $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2$.
- ♠ 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 $\mathbf{X}_1, \dots, \mathbf{X}_p$ 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

- ♠ $\sum_{i=1}^n \hat{\epsilon}_i = 0$.
- ♠ 残差 $\hat{\epsilon}_i$ 与拟合值 \hat{Y}_i 不相关，与协变量 $\mathbf{X}_1, \dots, \mathbf{X}_p$ 中的任何一个不相关。
- ♠ 残差的标准误差 $\hat{\sigma} = \sqrt{\sum_{i=1}^n \hat{\epsilon}_i^2 / (n - p - 1)}$ 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 $\alpha, \beta_1, \dots, \beta_p$, 故我们失去 $p + 1$ 自由度).

R^2 和调整的 \tilde{R}^2

♠ $TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2.$

♠ $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n \hat{\epsilon}_i^2$ (残差平方和)

♠ $RegSS = TSS - SSE = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2.$

♠ $R^2 = RegSS/TSS = 1 - SSE/TSS$

表示变异中由 \mathbf{Y} 对 $\mathbf{X}_1, \dots, \mathbf{X}_p$ 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 :

$$\tilde{R}^2 = 1 - \frac{SSE/(n-p-1)}{TSS/(n-1)},$$

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。

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 ***
YEAR	-0.195703	0.025112	-7.793	1.48e-05 ***
RAIN	0.034288	0.009655	3.551	0.00526 **

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： $\sum_{i=1}^n \hat{\epsilon}_i = 0$: `sum(re)`. 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 at each folding point. ... the 5 numbers (extremes, **hinges**, median) that make up a **5-number summary**"

使用hinge spread

♠ Hinge spread = interquartile range (IQR)[四分位数间距]

♠ 设 IQR_1, \dots, IQR_p 是 $\mathbf{X}_1, \dots, \mathbf{X}_p$ 的IQR。

♠ 注意到: $\mathbf{Y}_i = \hat{Y}_i + (\mathbf{Y}_i - \hat{Y}_i) = \hat{Y}_i + \hat{\epsilon}_i$, 我们从等式 $Y_i = \hat{\alpha} + \hat{\beta}_1 \mathbf{X}_{i1} + \dots + \hat{\beta}_p \mathbf{X}_{ip} + \hat{\epsilon}_i$ 出发。

♠ 这个等式可以改写

$$\text{为: } Y_i = \hat{\alpha} + \left(\hat{\beta}_1 IQR_1 \right) \frac{\mathbf{X}_{i1}}{IQR_1} + \dots + \left(\hat{\beta}_p IQR_p \right) \frac{\mathbf{X}_{ip}}{IQR_p} + \hat{\epsilon}_i$$

♠ 令 $Z_{ij} = \frac{\mathbf{X}_{ij}}{IQR_j}, j = 1, \dots, p, i = 1, \dots, n$.

♠ 令 $\hat{\beta}_j^* = \hat{\beta}_j IQR_j, j = 1, \dots, p$.

♠ 则我们有 $\mathbf{Y}_i = \hat{\alpha} + \hat{\beta}_1^* Z_{i1} + \dots + \hat{\beta}_p^* Z_{ip} + \hat{\epsilon}_i$.

♠ $\hat{\beta}_j^* = \hat{\beta}_j IQR_j$ 称为标准化的回归系数。它们可以进行比较。

(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, $\hat{\beta}_j^*$ 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 β_j^* 。
- ♠ 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, \dots, S_p 分别表示 $\mathbf{X}_1, \dots, \mathbf{X}_p$ 的标准差。
- ♠ 注意到: $\mathbf{Y}_i = \hat{\mathbf{Y}}_i + (\mathbf{Y}_i - \hat{\mathbf{Y}}_i) = \hat{\mathbf{Y}}_i + \hat{\epsilon}_i$, 我们从等式 $\mathbf{Y}_i = \hat{\alpha} + \hat{\beta}_1 \mathbf{X}_{i1} + \dots + \hat{\beta}_p \mathbf{X}_{ip} + \hat{\epsilon}_i$ 出发。
- ♠ 这个等式可以改写为: $\frac{\mathbf{Y}_i - \bar{\mathbf{Y}}}{S_Y} = \left(\hat{\beta}_1 \frac{S_1}{S_Y} \right) \frac{\mathbf{X}_{i1} - \bar{\mathbf{X}}_1}{S_1} + \dots + \left(\hat{\beta}_p \frac{S_p}{S_Y} \right) \frac{\mathbf{X}_{ip} - \bar{\mathbf{X}}_p}{S_p} + \frac{\hat{\epsilon}_i}{S_Y}$.
- ♠ 令 $Z_{i\mathbf{Y}} = \frac{\mathbf{Y}_i - \bar{\mathbf{Y}}}{S_Y}$, $Z_{ij} = \frac{\mathbf{X}_{ij} - \bar{\mathbf{X}}_j}{S_j}$, $j = 1, \dots, p$, $i = 1, \dots, n$.
- ♠ 令 $\hat{\beta}_j^* = \hat{\beta}_j \frac{S_j}{S_Y}$, $\hat{\epsilon}_i^* = \frac{\hat{\epsilon}_i}{S_Y}$ $j = 1, \dots, p$, $i = 1, \dots, n$.
- ♠ 则我们有 $Z_{i\mathbf{Y}} = \hat{\beta}_1^* Z_{i1} + \dots + \hat{\beta}_p^* Z_{ip} + \hat{\epsilon}_i^*$.
- ♠ $\hat{\beta}_j^* = \hat{\beta}_j \frac{S_j}{S_Y}$ 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, $\hat{\beta}_j^*$ 是响应变量 Z_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_Y 增加1, 意味着 Y 增加 S_Y (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 Y 将平均增加 $\beta_j^* \times S_Y$ 。
- ♠ 对于臭氧的例子, 结果是:

变量	系数	$\frac{\text{St.dev(variable)}}{\text{St.dev}(Y)}$	标准化的系数
YEAR	-0.196	6	-0.783
Rain	0.034	11.6	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, 但是事实并非总是如此。

添加变量图

【阅读这一页的内容时，要逐行结合逐行运行R去理解结果】

- ♠ 假设我们从 $SF \sim YEAR$ 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 $\text{lm}(SF \sim YEAR)$ 的残差），主要是利用 RAIN 中不能被 YEAR 解释的部分（ $\text{lm}(RAIN \sim YEAR)$ 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，控制 YEAR。
- ♠ $\text{lm}(SF \sim YEAR)$ 的残差对 $\text{lm}(RAIN \sim YEAR)$ 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 σ 是不可观测的。此外，统计误差 ϵ 也是无法观察的。
- ♠ 我们定义拟合值 $\hat{\mathbf{Y}}_i = \hat{\alpha} + \hat{\beta}_1 \mathbf{X}_{i1} + \cdots + \hat{\beta}_p \mathbf{X}_{ip}$ 和残差 $\hat{\epsilon}_i = \mathbf{Y}_i - \hat{\mathbf{Y}}_i$. 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
$$\text{SSE} = \sum_{i=1}^n \hat{\epsilon}_i^2 = \sum_{i=1}^n (\mathbf{Y}_i - (\hat{\alpha} + \hat{\beta}_1 \mathbf{X}_{i1} + \cdots + \hat{\beta}_p \mathbf{X}_{ip}))^2$$
来求出 $\alpha, \beta_1, \cdots, \beta_p$ 的估计 $\hat{\alpha}, \hat{\beta}_1, \cdots, \hat{\beta}_p$. 但是， σ^2 的估计要通过残差来估计。
- ♠ 系数的解释？

总结

♠ 为了衡量模型拟合的好坏程度，我们可以使用：

- ✓ 残差标准误： $\hat{\sigma} = \sqrt{\text{SSE}/(n - p - 1)}$
- ✓ 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): $\text{TSS} = \text{SSE} + \text{RegSS}$

♠ 标准化回归系数

♠ 添加变量图(偏回归图)