线性回归分析及 Bootstrap 应用

授课教师：吴翔

邮箱：wuhsiang@hust.edu.cn

March 16, 2019

Section 1

线性回归分析概述

简单回归模型

考虑由数据生成过程（**data generating process**, DGP）$y = -5 + 2 \cdot x$ 得到的样本。

```
# generate dataset
x <- rnorm(n = 200, mean = 10, sd = 8)
beta <- c(-5, 2)
y <- beta[1] + beta[2] * x + rnorm(n = 200, mean = 0, sd = 2)
dat <- data.frame(x = x, y = y)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -4.8      0.234     -21  1.7e-51
## x                2.0      0.019     106 1.4e-176
```

线性模型 $R^2 = 0.98$，预测值 $\hat{\beta} = (-4.84, 1.99)$ 接近实际值 $\beta = (-5, 2)$。

虚假效应

考虑变量 z，它受 x 影响，但不受 y 影响。在模型设定错误下，

```
# another variable
z <- 6 - 5 * x + rnorm(n = 200, mean = 0, sd = 4)
dat2 <- cbind(dat, z)
# linear regression
fit2 <- lm(y ~ z, data = dat2)
summary(fit2)$coef
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2.18     0.2637    -8.3  1.9e-14
## z              -0.39     0.0046   -86.2 6.1e-159
```

回归模型显示，$y = -2.18 + -0.39z$，且 $R^2 = 0.97$。

真实效应

我们考虑真实模型 $y = \beta_0 + \beta_1 x + \beta_2 z$。

```
# linear regression
fit3 <- lm(y ~ x + z, data = dat2)
summary(fit3)$coef
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -4.674      0.329  -14.23  4.7e-32
## x              1.857      0.185   10.03  2.2e-19
## z             -0.027      0.037   -0.74  4.6e-01
```

回归模型显示，$y = -4.67 + 1.86x$，且 $R^2 = 0.98$。

正效应 vs 负效应?

考虑增加一个样本 $c(164, -500)$，重新运行模型。

```
# add a sample
dat1 <- rbind(dat, c(164, -500))
# linear regression
fit1 <- lm(y ~ x, data = dat1)
summary(fit1)$coef
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     29.6       3.01     9.8  7.1e-19
## x               -1.6       0.18    -9.1  9.8e-17
```

线性模型 $R^2 = 0.29$，预测值 $\hat{\beta} = (29.64, -1.61)$ 大幅偏离实际值 $\beta = (-5, 2)$。

如何学习线性回归?

图 1: Master & PhD students who are learning regression models

课程存储地址

- 课程存储地址：https://github.com/wuhsiang/Courses
- 资源：课件、案例数据及代码

图 2: 课程存储地址

参考教材

- 谢宇．回归分析．北京：社会科学文献出版社．2010.
- 威廉·贝里．理解回归假设．上海: 格致出版社．2012.

Section 2

线性回归分析原理

遗传与变异

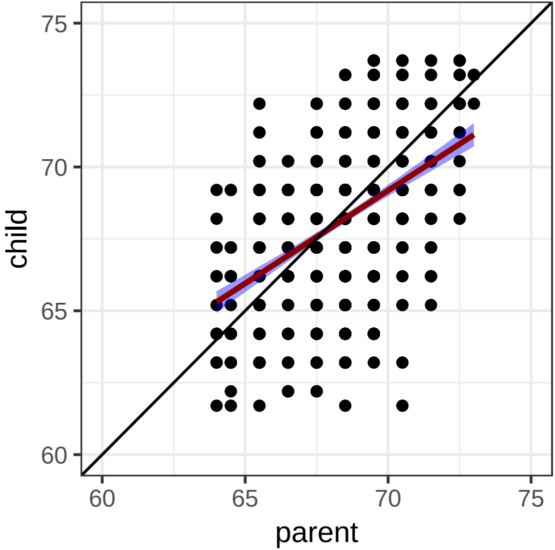
Francis Galton（以及 Karl Pearson）研究

- 个体差异：确立了社会科学研究与自然科学研究的根本区别
- 遗传与个体差异的关系：倡导“优生学”
- 双生儿法（twin method）：匹配方法（matching）之先河

变异与个体差异

- 随着物种的变异，其个体差异是否会一直增大?
- 个体差异上的两极分化是否是一般规律?

Galton 的身高研究



什么是“回归”?

Galton 的身高研究发现:

- 父亲的身高增加时，儿子的身高也倾向于增加
- 当父亲高于平均身高时，儿子身高比他更高的概率要小于比他更矮的概率；父亲矮于平均身高时，儿子身高比他更矮的概率要小于比他更高的概率。

回归效应:

- 向平均数方向的回归 (regression toward mediocrity)
- 天之道，损有余而补不足

回归分析原理：模型设定

考虑教育程度 x 与收入 y 的关系，回归模型为：

$$y_i = \alpha + \beta x_i + \epsilon_i.$$

暗含的假设：

- A1. 线性假设（$E(y|x) = \beta x$）：非线性模型、结构模型
- A2. 同质性假设：随机参数/效应模型、分层线性模型

总体回归方程

给定 x_i，在的 ϵ_i i.i.d $\sim N(0, \sigma^2)$ 假定下，对回归模型求条件期望得到如下**总体回归方程**，

$$E(y|x = x_i) = \mu_{y|x_i} = \alpha + \beta x_i.$$

含义:

- 给定任意 x_i，对应的 $y_i \sim N(\mu_{y|x_i}, \sigma^2)$。
- 回归线穿过 $(x_i, \mu_{y|x_i})$。
- 参数 β 刻画了 x 的变化对 y 的期望的影响。

总体回归线

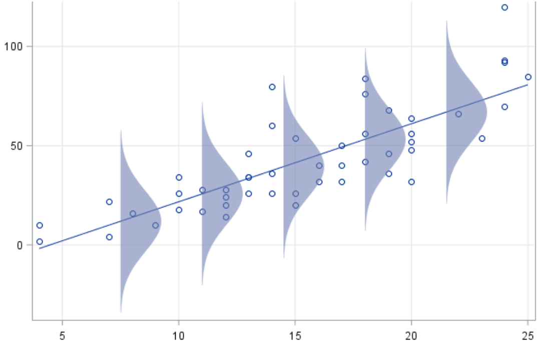


图 3: 总体回归线

暗含的假设

- A3. 独立同分布假设：
  - $E(\epsilon_i) = 0$：随机效应模型中的随机截距参数
  - $Cov(\epsilon_i, \epsilon_j) = 0$：时间序列模型、空间计量模型、嵌套模型
  - $\sigma_i = \sigma$：异方差问题
- A4. 关于 y 的假设：
  - y 应是连续变量：广义线性模型
  - y 的条件期望 $\mu_{y|x_i} = E(y|x = x_i)$ 符合正态分布：分位数回归
- A5. 正交（严格外生）假设
  - 误差项 ϵ 和 x 不相关，即 $Cov(x, \epsilon) = 0$
  - 内生性问题

Gauss–Markov 定理

Section 3

线性回归诊断