第三章线性回归分析及 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 $\beta = (-5, 2)$。

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 $\beta = (-5, 2)$。

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 $R^2 = 0.98$。

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.

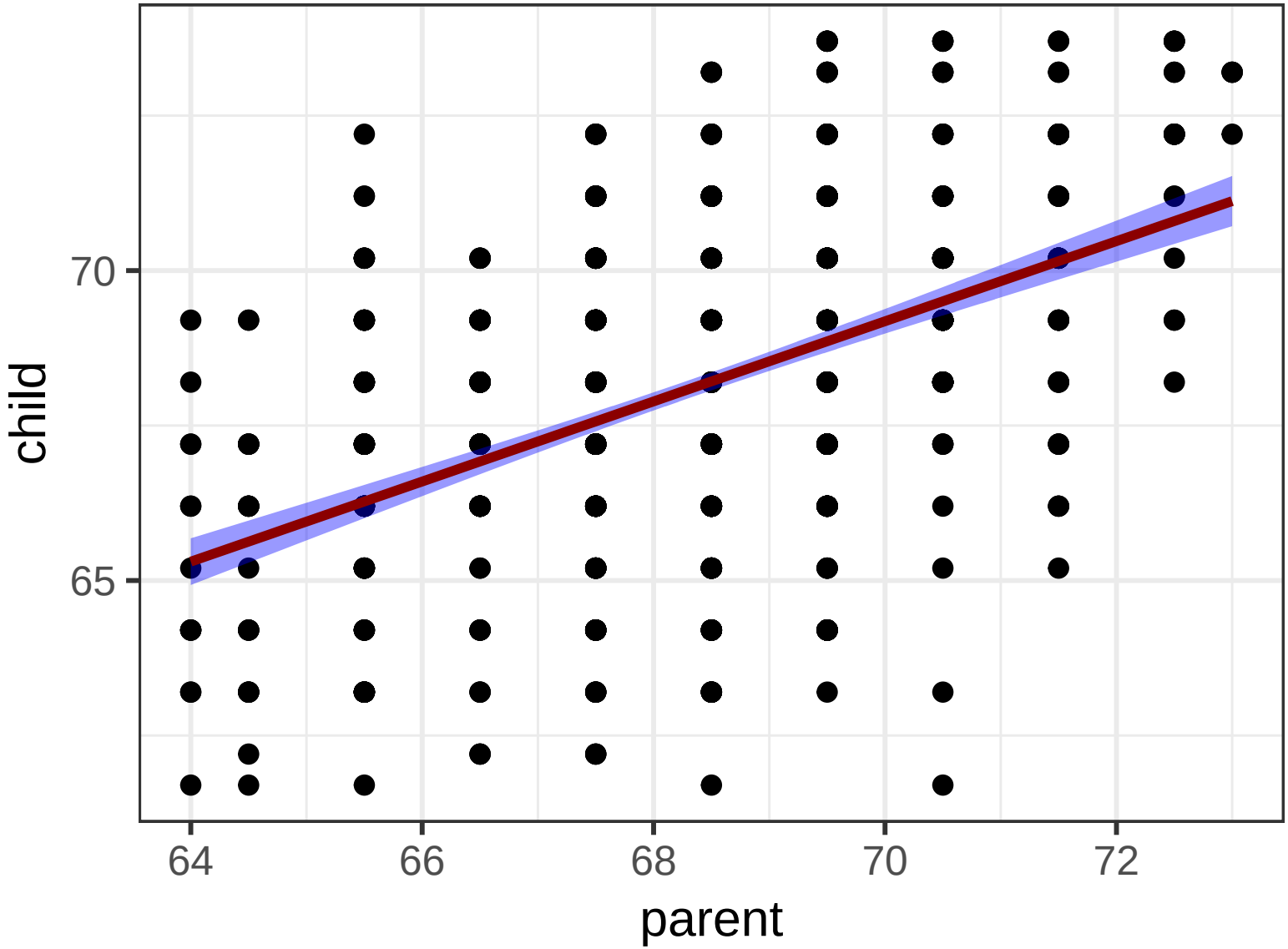
Section 2

线性回归分析原理

遗传与变异

什么是“回归”?

高尔顿的身高研究



身高数据及回归结果

```
# linear regression
fit <- lm(child ~ parent, data = galton)
summary(fit)$coef
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     23.94      2.811     8.5  6.5e-17
## parent           0.65      0.041    15.7  1.7e-49
```

回归分析原理：简单案例

考虑教育程度 x 与收入 y 的关系，回归模型为：

$$y_i = \alpha + \beta x_i + \epsilon_i, \epsilon_i \sim N(0, \sigma^2).$$

暗含的假设：

- A.0.1. 线性假设：
- A.0.2. 同质性假设：
- A.0.3. 同方差假设：

Section 3

线性回归假设与诊断

Section 4

线性回归的贝叶斯估计