线性回归分析

授课教师：吴翔
邮箱：wuhsiang@hust.edu.cn

March 16, 2019

线性回归概述

简单案例

考虑由数据生成过程（**data generating process**, DGP）$y = -5 + 2 \cdot x$ 得到的样本。另外有变量 z，它受 x 影响，但不受 y 影响。

```
# 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)
z <- 6 - 5 * x + rnorm(n = 200, mean = 0, sd = 4)
dat <- data.frame(x = x, y = y, z = z)
```

回归分析

```
##               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
```

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

```
##               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
```

考虑 z 对 y 的效应，$y = -2.18 + -0.39z$，且 $R^2 = 0.97$。

虚假 vs 真实效应

```
# linear regression
fit3 <- lm(y ~ x + z, data = dat)
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
```

考虑模型 $y = \beta_0 + \beta_1 x + \beta_2 z$。结果显示，$y = -4.67 + 1.86x$，且 $R^2 = 0.98$。

Q1: z 对 y 的效应，是否显著?

正效应 vs 负效应?

```
# add a sample
dat1 <- rbind(dat, c(164, -500, 200))
fit4 <- lm(y ~ x, data = dat1)
summary(fit4)$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
```

增加一个样本 $c(164, -500, 200)$，重新考虑 x 对 y 的效应，$R^2 = 0.29$，预测值 $\hat{\beta} = (29.64, -1.61)$ 大幅偏离实际值 $\beta = (-5, 2)$。

Q2: x 对 y 的效应，到底是正还是负?

如何学习线性回归?

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

理念

- 方便有多门，归元无二路
- 挽弓当挽强，用箭当用长

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.
- 欧文·琼斯. R 语言的科学编程与仿真. 西安：西安交通大学出版社. 2014.

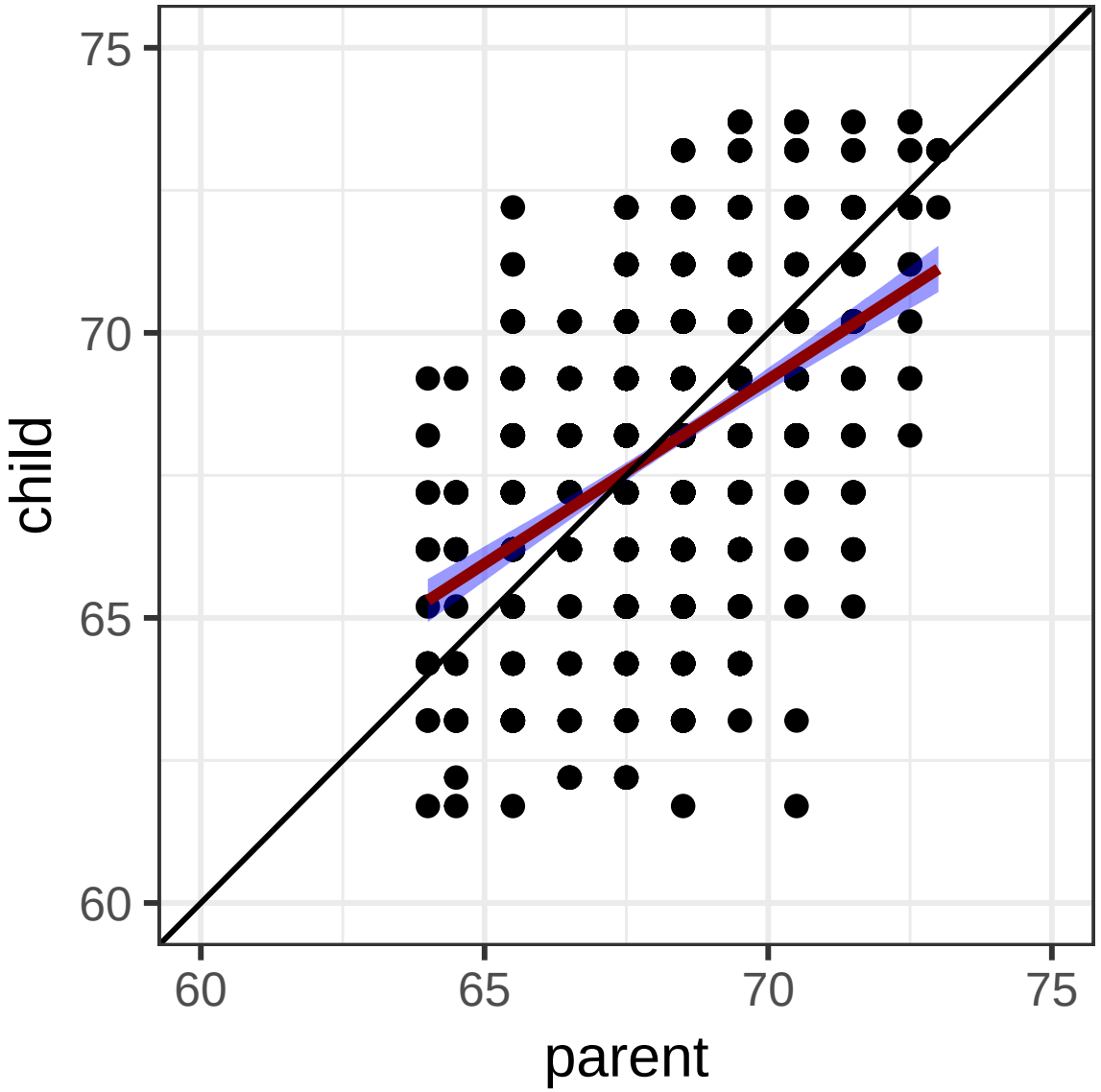
线性回归原理

缘起

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:

- 父代的身高增加时，子代的身高也倾向于增加
- 当父代高于平均身高时，子代身高比他更高的概率要小于比他更矮的概率；父代矮于平均身高时，子代身高比他更矮的概率要小于比他更高的概率。
- 同一族群中，子代的身高通常介于其父代的身高和族群的平均身高之间。

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

Galton 的开创性研究

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

理解回归的三种视角

考虑回归模型

$$y_i = f(X_i) + \epsilon_i = \beta X_i + \epsilon_i$$

- **因果性**（计量经济领域）：观测项 = 机制项 + 干扰项
- **预测性**（机器学习领域）：观测项 = 预测项 + 误差项
- **描述性**（统计领域）：观测项 = 概括项 + 残差项

回归模型设定

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 $x = x^k$，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**，

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

含义：

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

总体回归线

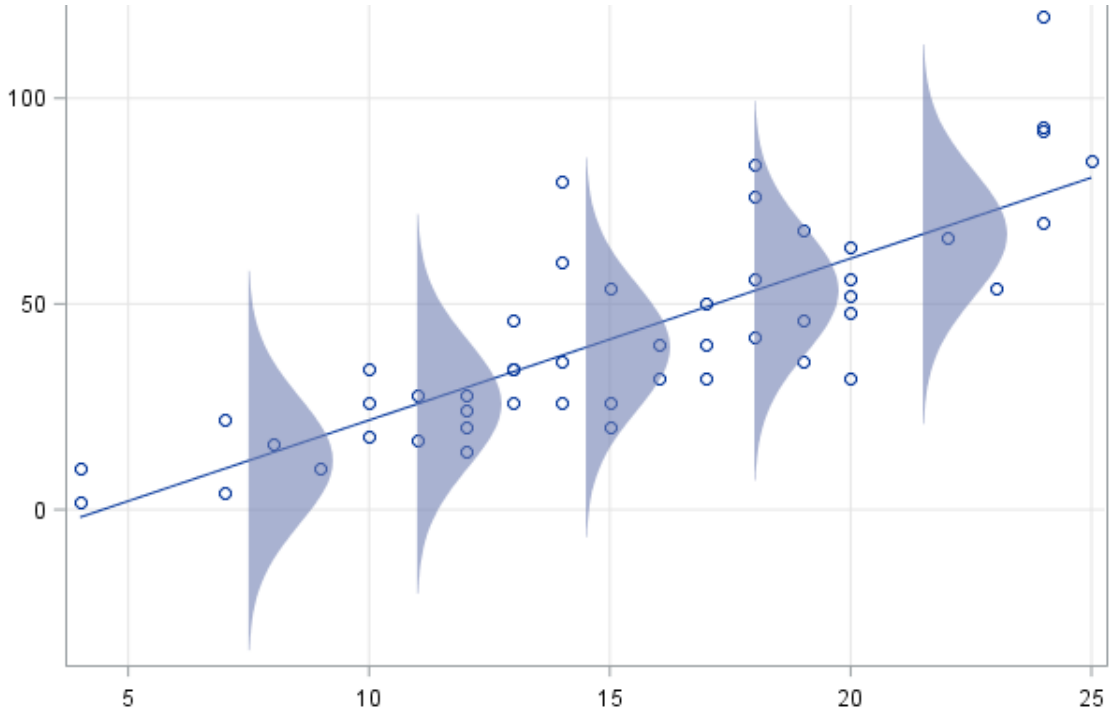


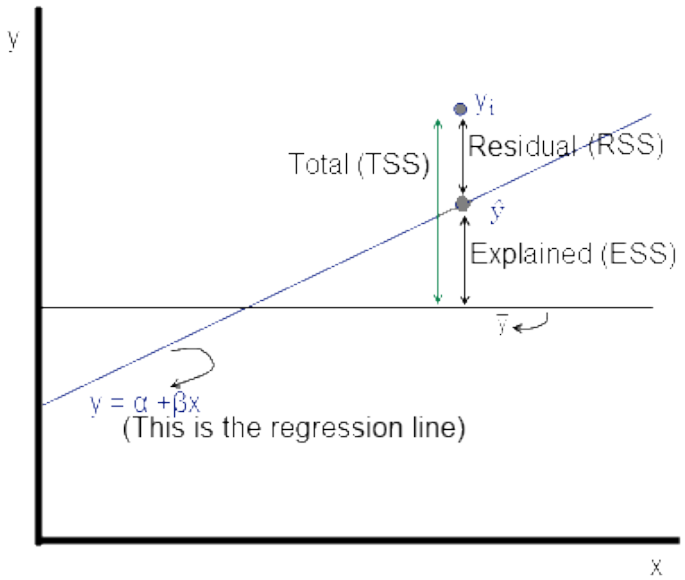
图 3: 总体回归线

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 $\mu_{y|x^k} = E(y|x = x^k)$ 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

变异的分解

样本观测值 y_i、均值 $\bar{y}$、预测值 $\hat{y}$ 之间的关系



$\hat{y}$ is the predicted value of y given x, using the equation $y = \alpha + \beta x$.

y_i is the actual observed value of y

$\bar{y}$ is the mean of y.

The distances that RSS, ESS and TSS represent are shown in the diagram to the left - but remember that the actual calculations are squares of these distances.

$$TSS = \Sigma (y_i - \bar{y})^2$$

$$RSS = \Sigma (y_i - \hat{y})^2$$

$$ESS = \Sigma (\hat{y} - \bar{y})^2$$

图 4：变异的分解

变异分解公式

总平方和（sum of squares total, SST）可以分解为回归平方和（sum of squares regression, SSR）和残差平方和（sum of squares error, SSE）之和，

具体而言：

$$\begin{aligned} SST &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \\ &= \sum_{i=1}^{n}[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \\ &= SSE + SSR \end{aligned}$$

$R^2 = SSE/SST.$

参数估计

最小化残差平方和（扩展到多元回归的情境 $y = \beta X + \epsilon$）：

$$\min\ SSE = \min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta X_i)^2$$

由偏导公式

$$\frac{\partial SSE}{\partial \beta} = 0$$

得到参数估计值

$$\hat{\beta} = (X'X)^{-1}X'y.$$

Q3：如何在熟悉的编程语言中，撰写函数估计多元线性模型?

Gauss–Markov 定理

线性回归诊断