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

考虑智力测验成绩 x、教育年限 z 和年收入 y（万元）之间的关系。数据生成过程（**data generating process**, DGP）$y = -0.5 + 0.2 \cdot x$ 得到的样本。

```
# generate dataset
x <- rnorm(n = 200, mean = 110, sd = 10)
beta <- c(-0.5, 0.2)
y <- beta[1] + beta[2] * x + rnorm(n = 200, mean = 0, sd = 0.5
z <- round(-2 + 0.1 * x + rnorm(n = 200, mean = 0, sd = 0.4))
dat <- data.frame(x = x, y = y, z = z)
```

回归分析

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.32     0.4138   -0.77  4.4e-01
## x                 0.20     0.0038   52.93 8.0e-119
```

考虑 x 对 y 的效应，线性模型 $R^2 = 0.93$，预测值 $\hat{\beta} = (-0.32, 0.2)$ 接近实际值 $\beta = (-0.5, 0.2)$。

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.6       0.63      12  3.1e-25
## z                  1.5       0.07      22  1.6e-55
```

考虑 z 对 y 的效应，$y = 7.58 + 1.55z$，且 $R^2 = 0.71$。

虚假 vs 真实效应

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

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.356     0.4332   -0.82  4.1e-01
## x              0.201     0.0078   25.71  7.3e-65
## z             -0.021     0.0697   -0.30  7.6e-01
```

考虑模型 $y = \beta_0 + \beta_1 x + \beta_2 z$。结果显示，$y = -0.36 + 0.2x$，且 $R^2 = 0.93$。

课堂思考：z 对 y 的效应，是否显著?

正效应 vs 负效应?

```r
# add a sample
dat1 <- rbind(dat, c(160, -100, 10))
fit4 <- lm(y ~ x, data = dat1)
summary(fit4)$coef
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    34.76      6.784     5.1  7.1e-07
## x              -0.13      0.061    -2.1  4.2e-02
```

增加一个样本 $c(160, -100, 10)$，重新考虑 x 对 y 的效应，$R^2 = 0.02$，预测值 $\hat{\beta} = (34.76, -0.13)$ 大幅偏离实际值 $\beta = (-0.5, 0.2)$。

课堂思考：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.

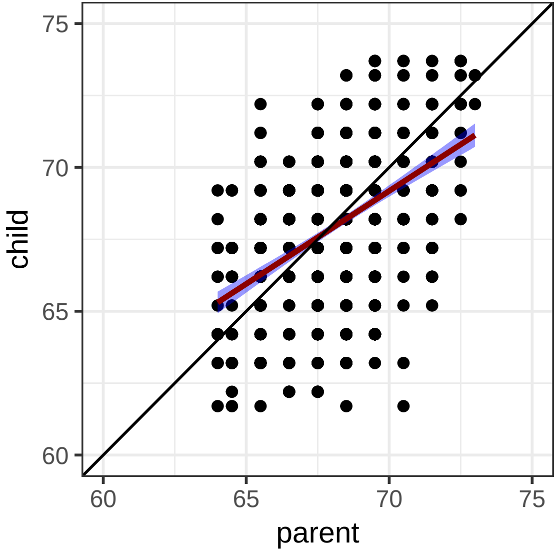
线性回归原理

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

社会科学定量研究逻辑

社会科学定量研究与自然科学定量研究的区别：

- 核心区别：变异（variation）vs 共相（universal，相对应的是殊相 particular）
- 结论：或然性 vs 必然性
- 方法：归纳法 vs 演绎法
- 特征：普适规律 vs 特定**情境**下的规律

因而，社会科学定量研究即是，在特定的**社会（或管理）情境**，选取合宜的解释变量，以尽可能能理解总体中结果变量的变异的来源。

理解回归的三种视角

回归模型考虑解释变量 x 与结果变量 y 的关系，

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

将观测值 y_i 分为结构部分 $f(X_i)$ 和随机部分 ϵ_i，并可以从**三个视角**来理解：

- **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

考虑收入 x 与中老年人抑郁水平 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

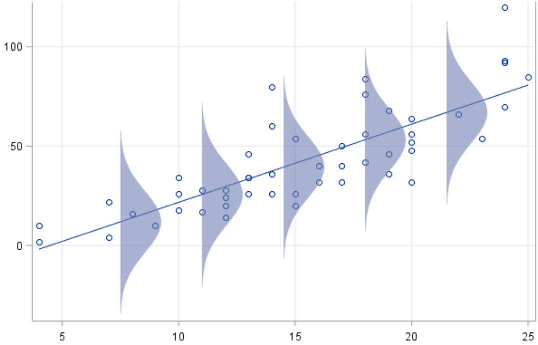


图 3: 总体回归线

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

参数估计

普通最小二乘法（ordinary least squares, OLS）通过最小化残差平方和（扩展到多元回归的情境 $y = \beta X + \epsilon$）估计参数：

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

由偏导公式

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

得到参数估计值

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

课堂思考：(1) 如何在熟悉的编程语言中，撰写函数估计多元线性模型？(2) 在实践中，OLS 会造成什么缺陷？

衡量估计方法

评判估计的黄金准则（Fisher）:

- **无偏性**：在总体中进行 M 次抽样，$E[\hat{b}_m] = \beta$。
- **有效性**：在众多估计量中，b 的抽样分布的方差最小。
- **一致性**：样本量增大时，b 趋近于 β。

课堂思考：统计显著性与样本量有无关系?

变异分解逻辑

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

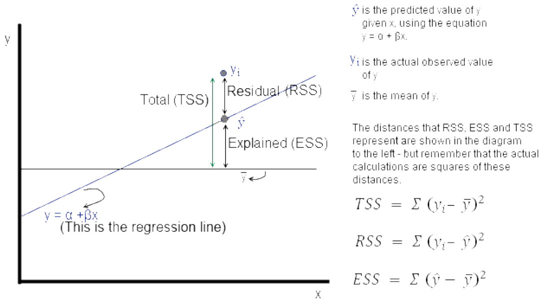


图 4: 变异的分解

板书演示：变异分解逻辑

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：

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

判定系数（coefficient of determination）$R^2 = SSE/SST$.

板书演示：变异分解推导

多元线性回归与方差分析

假定多元线性模型中，待估计的参数个数为 p，那么方差和自由度的分解如下：

- SST: 自由度为 $n-1$
- SSE: 自由度为 $n-p$
- SSR: 自由度为 $p-1$

因而，自由度的分解为：

$$n-1=(n-p)+(p-1)$$

课堂思考：假设模型有两个解释变量，其中 x_1 是连续变量，x_2 是包含 5 个分类的分类变量，SSR 的自由度为多少？

方差分析表

表 1: 多元线性回归的方差分析表

变异来源	平方和	自由度	均方
回归模型	SSR	$p-1$	$\mathrm{MSR}=\mathrm{SSR}/(p-1)$
误差	SSE	$n-p$	$\mathrm{MSE}=\mathrm{SSE}/(n-p)$
总变异	SST	$n-1$	$\mathrm{MST}=\mathrm{SST}/(n-1)$

相应地，可以构造 F 检验：

$$F(\mathrm{df}_{\mathrm{SSR}},\mathrm{df}_{\mathrm{SSE}})=\frac{\mathrm{MSR}}{\mathrm{MSE}}?>F_\alpha$$

延伸内容：聚类分析

模型选择

- 模型选择：**精确性原则** vs **简约性原则**
- 情境：假定在线性回归模型 A 的基础上，加了几个变量得到模型 B，应当如何在模型 A 和 B 之间选择?

构造 F 检验：

$$F(\Delta \mathrm{df}, \mathrm{df}_{\mathrm{SSE}}) = \frac{\Delta \mathrm{SSR}/\Delta \mathrm{df}}{\mathrm{MSE}_{\mathrm{B}}} ? > F_{\alpha}$$

线性回归案例

中老年精神健康案例

从 CHARLS 数据中随机抽取样本 $n = 488$，考虑中老年抑郁水平。`income` 为个人收入，以万元计；`educ` 表示教育水平是否在初中及以上，`hukou` 表示是否是城市户口。

表 2: 描述性统计量

变量	均值	标准差	最小值	最大值
cesd10	6.62	5.95	0	30
income	2.12	2.12	0.01	20
educ	0.76	0.43	0	1
hukou	0.28	0.45	0	1

收入与精神健康

表 3: 不同线性回归模型比较

变量	模型 1	模型 2	模型 3
常数项	2.18 (0.05)	2.47 (0.11)	2.49 (0.11)
log(income)	-0.18 (0.04)	-0.14 (0.04)	-0.12 (0.05)
educ	-	-0.39 (0.13)	-0.34 (0.13)
hukou	-	-	-0.23^{ns} (0.12)
R^2	0.04	0.06	0.06

课堂讨论：应选择哪个模型?

变异分解与模型选择

```
##                    Df Sum Sq Mean Sq F value  Pr(>F)    
## I(log(income))      1     26   25.79   19.60 1.2e-05 ***
## educ                1     12   11.61    8.82  0.0031 ** 
## hukou               1      4    4.48    3.41  0.0656 .  
## Residuals         484    637    1.32                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

最终模型

权衡**精确性原则**与**简约性原则**，选择模型 2。

```
> summary(fit2)

Call:
lm(formula = y ~ I(log(income)) + educ, data = charlswh)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9758 -0.7034  0.0318  0.7732  3.0278

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.4686     0.1098   22.47   <2e-16 ***
I(log(income))  -0.1375     0.0446   -3.08   0.0022 **
educ            -0.3875     0.1308   -2.96   0.0032 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.1 on 485 degrees of freedom
Multiple R-squared:  0.0551,	Adjusted R-squared:  0.0512
F-statistic: 14.1 on 2 and 485 DF,  p-value: 1.07e-06
```

图 5: 模型估计结果

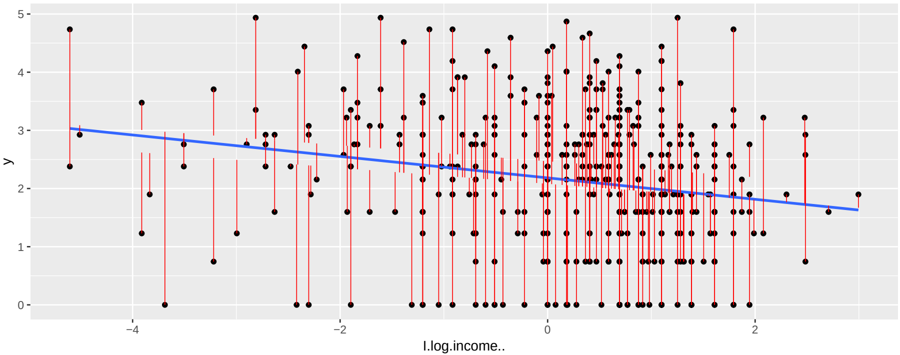
显著性与效应大小

- 统计显著性（statistical significance）
- 效应大小（effect size）

请参阅 Github 上的完整案例

课堂讨论：二者有何区别?

回归结果图示



变异分解：编程计算

```r
# calculate predicted values
yhat <- predict.lm(fit2)
# calculate and print SST, SSR, and SSE
ybar <- mean(charlswh$y)
sst <- sum((charlswh$y - ybar) ^ 2)
ssr <- sum((yhat - ybar) ^ 2)
sse <- sum((charlswh$y - yhat) ^ 2)
c(sst, ssr, sse)
```

```
## [1] 679  37 641
```

变异分解：系统输出

```r
# variation decomposition
summary.aov(fit2)
```

```
##                  Df Sum Sq Mean Sq F value  Pr(>F)
## I(log(income))    1     26   25.79   19.50 1.2e-05 ***
## educ              1     12   11.61    8.78  0.0032 **
## Residuals       485    641    1.32
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

线性回归诊断

因变量分布与 Box-Cox 变换

当因变量不服从正态分布时，Box & Cox (1964) 建议采用如下 Box-Cox 变换

$$y_i = \begin{cases} [(y_i+\lambda_2)^{\lambda_1}-1]/\lambda_1 & \text{if } \lambda_1 \neq 0, \\ ln(y_i+\lambda_2) & \text{if } \lambda_1 = 0. \end{cases}$$

将非正态的分布转换为正态分布。

课堂思考：(1) 对数变换或 Box-Cox 变换是否合适？(2) 如何推导出“变化比例”这一含义？

多重共线性

参数估计值

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

要求 $X'X$ 是**可逆（非奇异）**的。

- 完全多重共线性：模型无法识别
- 严重多重共线性：不影响估计的无偏性和一致性，损害参数估计的**有效性**，及标准误会增大
- 判断标准：**方差膨胀因子**（variance inflation factor，VIF）最大值超过 10，平均值明显大于 1

消除共线性

- k 水平分类变量：虚拟变量（dummy variable）化后，只能有 $k-1$ 个虚拟变量
- 减少解释变量个数
- 维度规约：因子分析
- 变量选择：如 lasso 等统计机器学习方法，尤其是 $n < p$ 时模型无法识别的情形

```r
suppressMessages(library(car))
# calculate VIF
vif(fit2)
```

```
## I(log(income))           educ
##            1.1            1.1
```

异方差

通常将违背残差分布假定的

- 自相关：$\mathrm{Cov}(\epsilon_i, \epsilon_j) \neq 0$
- 异方差：$\mathrm{Var}(\epsilon_i) \neq \mathrm{Var}(\epsilon_j)$

统称为**异方差。**异方差不影响估计的无偏性和一致性，但会损害估计的**有效性。**

处理异方差的方法包括：

- 调整标准误的计算，采用稳健标准误
- 采用广义最小二乘法（generalized least squares, GLS）估计模型

处理非线性

- 纳入二次项：处理 U 型关系
- 采用对数项：处理比例关系
- 纳入交互项：处理调节作用

高影响点及异常值处理

OLS 采用最小化误差**平方和**的方式，使估计值对异常值非常敏感

- **高影响点/高杠杆点** (influential/leverage points)：观测案例 i 对**回归系数**影响较大的点，通常可由 Cook 距离等统计量衡量
- **异常值**：模型拟合失败的观测点，它们大幅**偏离回归线**，通常由标准化残差来衡量 (其绝对值不宜大于 5)

因而需要识别高影响点和异常值，并**谨慎判断**是否要排除这些观测样本。

实践中的回归假设

1. 模型设定假设
   - 线性模型假设：$E(y|X) = \beta X + \epsilon$ （可检验）
   - 同质效应假设：$\beta_i = \beta$ （可检验，高阶议题 *）
2. 正交假设（OLS 自动保证，不必检验）
   - 误差项均值为 0：$E(\epsilon) = 0$
   - 误差项与解释变量不相关：$\text{Cov}(X, \epsilon) = 0$
3. 独立同分布假设
   - 误差项相互独立：$\text{Cov}(\epsilon_i, \epsilon_j) = 0$
   - 误差项方差相同：$\text{Var}(\epsilon_i) = \sigma^2$ （可检验）
4. 正态分布假设（大样本时，不必要）
   - 误差项服从正态分布：$\epsilon_i \sim N(0, \sigma^2)$

实践中的回归假设（续）

由回归模型设定、OLS 估计衍生出来的问题：

1. 结果变量 y 的分布（可检验，Box-Cox 变换）
2. 多重共线性（可检验）
3. 异常值（可检验）

线性回归高阶议题（*）

内生性与异质性

考虑是否上大学（$D_i = 0, 1$）和收入的关系

$$y_i = \alpha_i + \beta_i D_i.$$

- **内生性**：匹配法（matching）vs 随机控制试验法（RCT）
- **异质性**：分层线性模型

贝叶斯视角

背景:

- Efron 提出的 bootstrapping 方法
- 大数据时代的统计推断
- 频率学派 vs 贝叶斯学派

案例思考:

- 射击选手 B，999/1000
- 射击选手 A，100/100

案例思考

图 6: 灵犀一指 (999/1000) vs 小李飞刀 (100/100)

先验的作用

情境一（先验 8/10）：

- A：先验 (8/10) + 数据（999/1000）-> 后验（1007/1010 = 0.9997）**[获胜]**
- B：先验 (8/10) + 数据（100/100）-> 后验（108/110 = 0.9818）

情境二（先验 9999/10000）：

- A：先验 (9999/10000) + 数据（999/1000）-> 后验（10998/11000 = 0.9998）
- B：先验 (9999/10000) + 数据（100/100）-> 后验（10099/10100 = 0.9999）**[获胜]**

回归分析总结

1. 回归假设与诊断：如何得到可靠的结论？
2. 变异及其分解：社会科学定量研究的核心
3. 高阶议题：计量经济应用的前沿