Chapter 9: 回归分析方法

授课教师：吴翔

Section 1

线性回归设定

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 1：课程存储地址

经典的现实问题

请思考面临的一些现实问题：

- Galton 的身高研究：遗传因素如何影响人类的身高？换言之，父代的身高 如何影响子代的身高？
- 兰德医疗保险实验（RAND Health Insurance Experiment）：更慷慨而全面的医疗保险 是否会影响患者的医疗使用行为 及健康水平？
- 法定饮酒年龄 如何影响地区的死亡率？
- 教育年限 如何影响工资收入？
- 过去三十年里，民众的经济收入 如何（随着时间）变化？
- ...

我们希望理解这些议题，其共通之处是什么？

解释变量（explanatory variable, X）–> 结果变量（outcome variable, Y）

回归模型设定

我们先考虑最简单的情形：一元线性回归模型。我们希望探讨解释变量 X 如何影响结果变量 Y，且在总体中随机抽取了 n 个观测样本，由此得到 (X_i, Y_i)。那么，一元线性回归模型设定为：

$$Y_i = \alpha + \beta X_i + \epsilon_i, i \in [1, n].$$

其中 ϵ_i 为误差项，它代表了未被纳入模型的因素，亦即除 X 以外的因素，对结果变量的影响。因而，给定 X_i 时，我们估计或预测结果变量为：

$$\hat{Y}_i = Y_i - \epsilon_i.$$

提示：这一步骤，我们应当仔细思考两个问题：(一) 从常识、逻辑和理论上分析，解释变量 X 到底如何影响结果变量 Y？亦即，二者是否真的存在作用关系？(二) 模型的误差项到底包含了哪些变量？

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，

$$Y_i = f(X_i) + \epsilon_i = \alpha + \beta X_i + \epsilon_i$$

将观测值 Y_i 分为结构部分$\hat{Y}_i = f(X_i)$ 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

本课程结合因果性和描述性的视角：在给定情境下采取因果性视角，在一般情形下采取描述性视角。

Galton 的身高研究

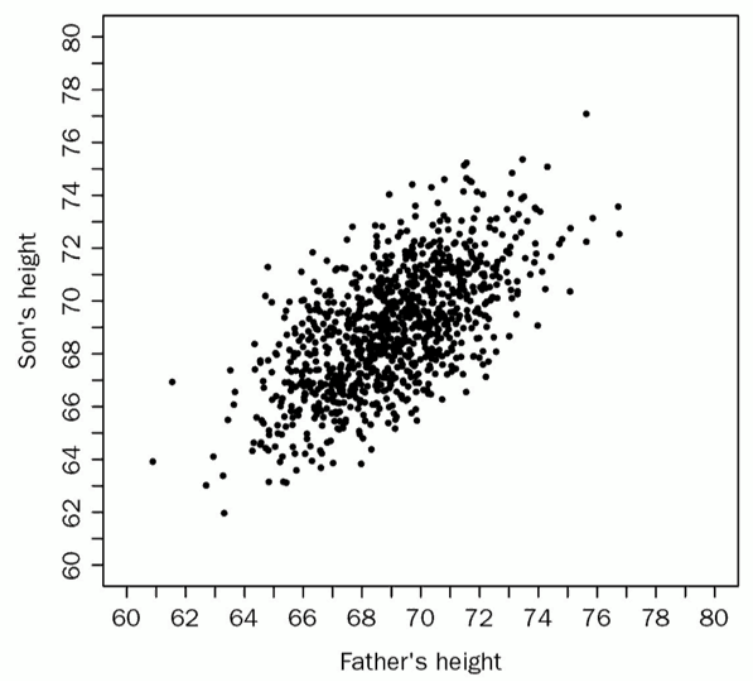


图 2: Galton 的身高研究

- 探讨”父代的身高如何影响子代的身高?“，到底意味着什么?
- 给定父代的身高，我们如何估计或预测子代的身高? 换言之，如何定义预测值 $\hat{Y}$? 如何建立预测值 $\hat{Y}$ 与解释变量 X 之间的关系?

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：特定情境下的规律 vs 普适规律

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，以尽可能理解总体中结果变量的变异的来源。变异性是社会科学研究的真正本质。

条件分布

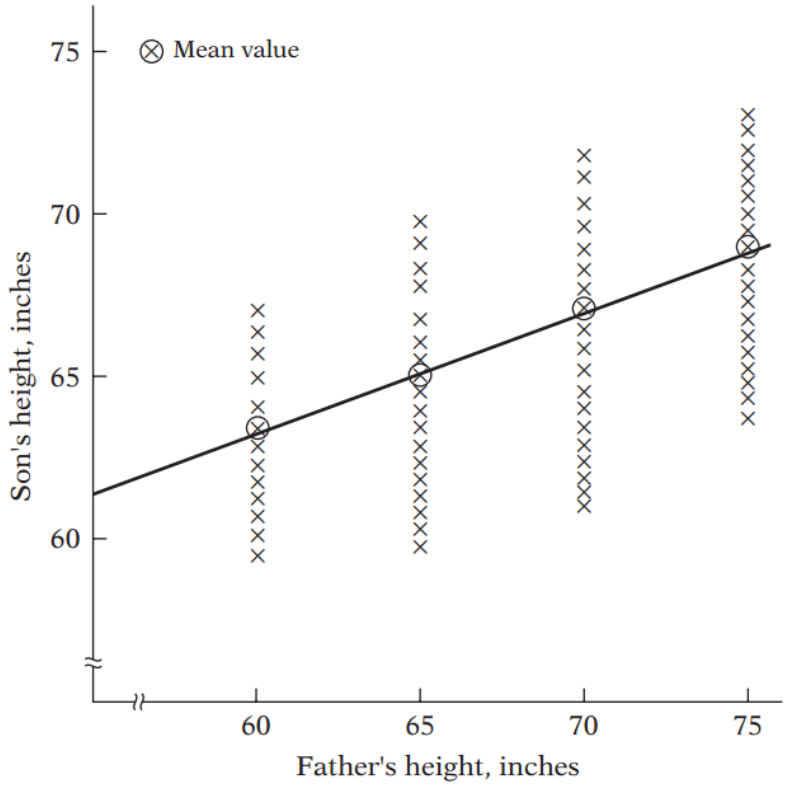


图 3: 条件分布

给定 $X = X_k$，对应的**多个**Y 值，应当如何给出其估计或预测值 $\hat{Y}\big|X = X_k$?

线性回归核心假设-1：（误差）独立同分布假设

给定 $X = X_k$ 时，结果变量存在多个值，进一步假定其分布为

$$Y|X = X_k \sim N(\mu_k, \sigma^2).$$

从而给定 $X = X_k$ 时，结果变量的合理预测值 $\hat{Y}$ 是其条件均值或条件期望，

$$\hat{Y} = E(Y|X = X_k) = \mu_k.$$

这一假定也意味着，误差项服从

$$\epsilon_i \text{ i.i.d. } \sim N(0, \sigma^2).$$

讨论：严重偏态分布下，应如何合理选择预测值 $\hat{Y}$？例如，北美社会的收入变化。

线性回归核心假设-2：线性模型假设

尽管模型设定上，我们容易直观理解线性假设的含义。但更本质地，线性模型假设可以从两个方面理解：

- 关于变量的线性 (linearity in the variables)：结果变量的条件期望 $E(Y|X)$ 是解释变量 X 的线性函数，即 $E(Y|X) = \alpha + \beta X$。这一视角下，$E(Y|X) = \alpha + \beta X^2$ 不是线性模型。
- 关于参数的线性 (linearity in the parameters)：结果变量的条件期望 $E(Y|X)$ 是参数 β 的线性函数，因而 $E(Y|X) = \alpha + \beta X^2$ 是线性模型。

两个方面的理解都是合理的。从建模的角度来看，第二个解释更合适。因为使用新变量 $X' = X^2$ 替换原始变量 X 之后，即可变换为通常形式的线性模型。

线性模型假设：图示

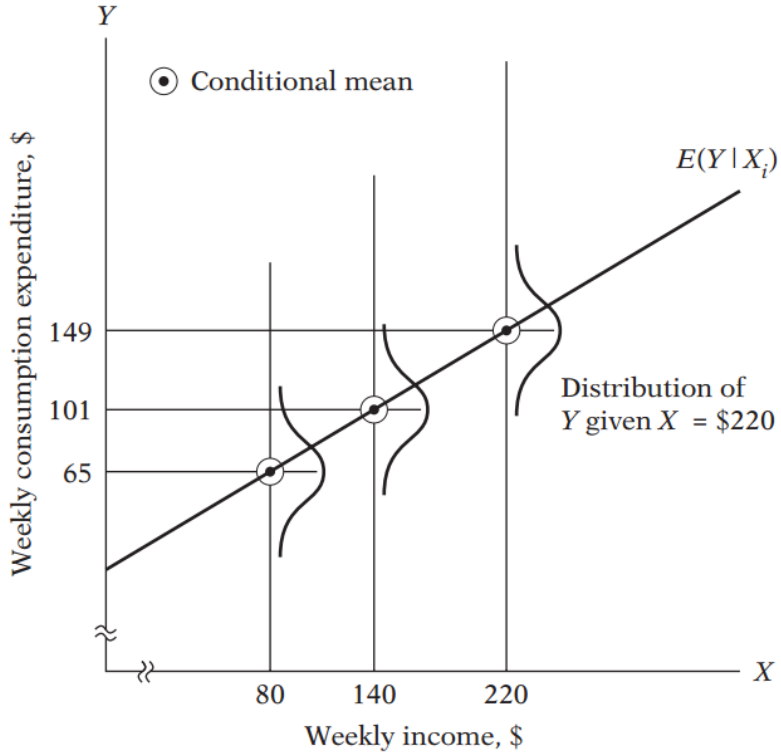


图 4: 总体回归线

总体回归线穿过 $(X_k, \mu_{Y|X_k})$ 这一系列点。参数 β 刻画了 X 的变化对 Y 的条件期望的影响。

$$E(Y|X = X_k) = \mu_k = \alpha + \beta X_k.$$

线性模型假设：建模层面

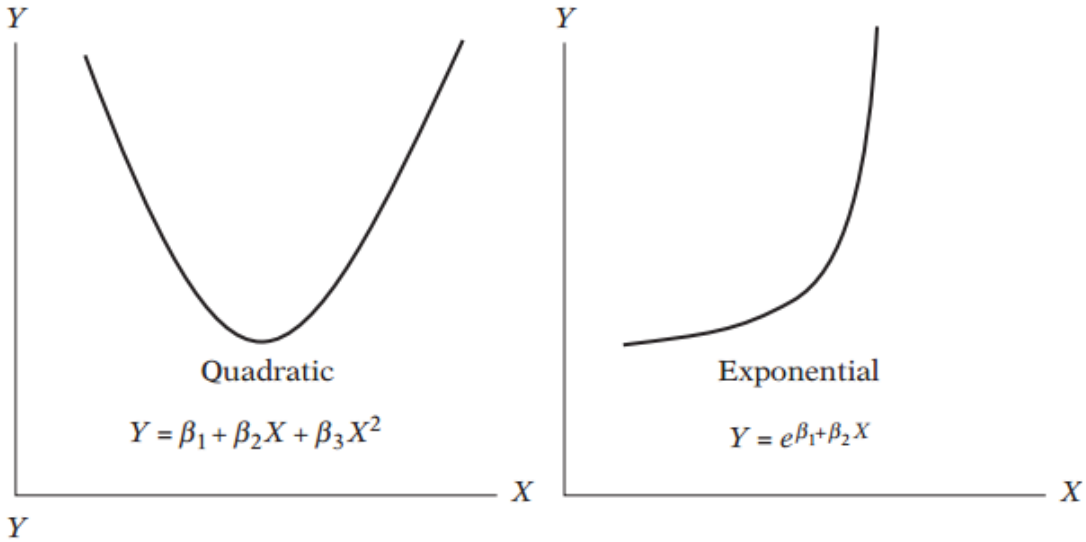


图 5：线性建模示例

以上两个模型，是否属于线性模型的范畴？

线性回归核心假设-3：正交或外生假设

解释变量 X 是外生的（exogenous），或者说是确定性的（deterministic）。例如在随机对照实验中，解释变量 X 是事先确定的实验处理方案。这意味着，解释变量 X 不受误差项 ϵ（亦即没有纳入模型的、遗漏的变量）的影响。因此，外生假设也称为正交假设：

$$X \perp \epsilon.$$

或者，

$$\text{Cov}(X, \epsilon) = 0.$$

注：随机对照试验（RCT）严格符合外生假设，从而能够提供准确的因果效应估计。但 RCT 只是满足外生假设的一种形式，观察研究只要满足外生假设，也能够提供准确的因果效应估计。

线性回归核心假设

核心假设包括三条：

1. (误差) 独立同分布假设：ϵ_i i.i.d. $\sim N(0, \sigma^2)$
2. 线性模型假设：$E(Y|X) = \alpha + \beta X$
3. 正交或外生假设：$X \perp \epsilon$

核心假设如何作用于线性回归模型?

在回归模型的经典设定下，

$$Y_i = \alpha + \beta X_i + \epsilon_i,$$

在误差独立同分布假设下，对等式两边同时取条件期望，

$$E[Y|X] = \alpha + \beta X + E[\epsilon|X],$$

由正交假设$X \perp \epsilon$ 以及误差独立同分布假设ϵ_i i.i.d. $\sim N(0, \sigma^2)$，可知$E[\epsilon|X] = 0$，从而:

$$E[Y|X] = \hat{Y} = \alpha + \beta X.$$

这意味着，线性模型假设与我们最初的经典设定是一致的。换言之，只有符合线性模型假设（及其它两个核心假设），最初的经典设定才是正确的，才不存在模型错误设定(model misspecification) 问题。

回归的含义

Regression analysis is concerned with the study of the dependence of one variable, the dependent variable, on one or more other variables, the explanatory variables, with a view to estimating and/or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) values of the latter. – By Gujarati & Porter (2008), "Basic Econometrics".

回归的现代解释，包含了如下核心概念：

- 条件期望 (conditional expectation)：给定解释变量 $X = X_k$ 时，如何估计或者预测结果变量的总体均值$E(Y)$?

Section 2

线性回归估计与解释

抽样分布

先考虑从总体中抽取 n 个样本，样本均值 $\bar{x}$ 的分布：

Sampling distribution of the sample mean

We take many random samples of a given size *n* from a population with mean μ and standard deviation σ.

Some sample means will be above the population mean μ and some will be below, making up the sampling distribution.

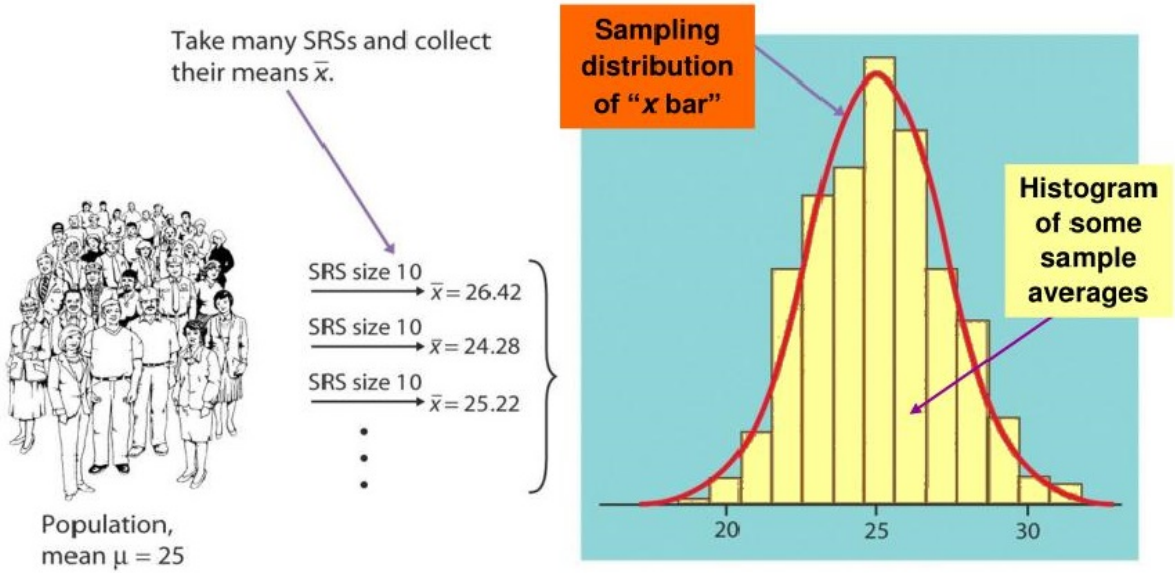


图 6：样本均值的抽样分布

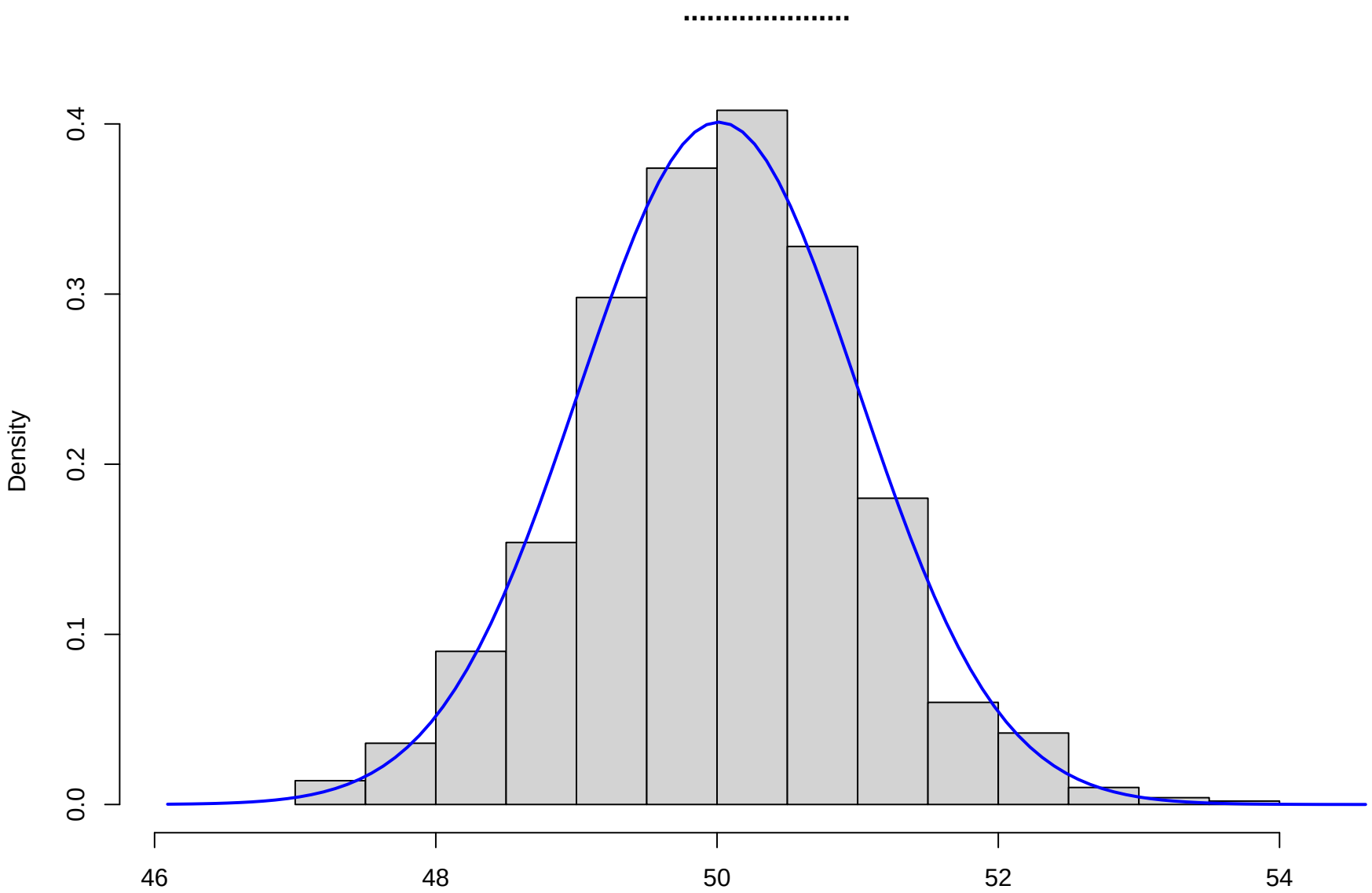
抽样分布（续）

```r
# 1. 创建总体数据
population <- rnorm(1e5, mean = 50, sd = 10)

# 2. 创建模拟抽样的函数
simulate_sampling <- function(population, n, k) {
  sample_means <- numeric(k)
  for(i in 1:k) {
    sample <- sample(population, n)
    sample_means[i] <- mean(sample)
  }
  return(sample_means)
}

# 3. 模拟抽样过程
sample_means <- simulate_sampling(population, n = 100, k = 100
```

抽样分布（续）



展示（或解读）回归结果

估计线性回归模型之后，需要以表格形式展示回归结果。阅读文献时，也需要理解它们展示的回归结果。

Table 7. Stratification Analysis (Dependent Variable Is *Quality*)

	Hospital size			Hospital location	
	Small	Medium	Large	Rural	Urban
Adoption	0.467	0.045	0.093	0.280	0.100
	(0.460)	(0.156)	(0.126)	(0.341)	(0.134)
MU1	0.751***	0.371***	0.223***	0.716***	0.327***
	(0.250)	(0.108)	(0.086)	(0.207)	(0.085)
MU2	0.465	0.135	0.089	0.547**	0.080
	(0.312)	(0.106)	(0.084)	(0.276)	(0.079)
Hospital fixed effects	Yes	Yes	Yes	Yes	Yes
Year fixed effects	No	No	No	No	No
Adjusted R^2	0.570	0.539	0.605	0.586	0.551
Number of hospitals	662	1,237	608	693	1,814

coefficients

standard errors

Notes. Robust standard errors are shown in parentheses (clustered on hospital). Control variables are included in the estimations, but the results are not shown here because of space limitations. The results show significant quality benefits from MU1 achievement across all strata, but small and rural hospitals exceed their counterparts in the degree of effect realized. On the other hand, only rural hospitals see a significant effect of MU2 achievement. This suggests substantial heterogeneity in the MU1/MU2 effects.
***$p < 0.01$; **$p < 0.05$.

图 7：如何展示或解读回归结果？

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

评判估计量（estimator）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（i.e.，标准误最小）。
- **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

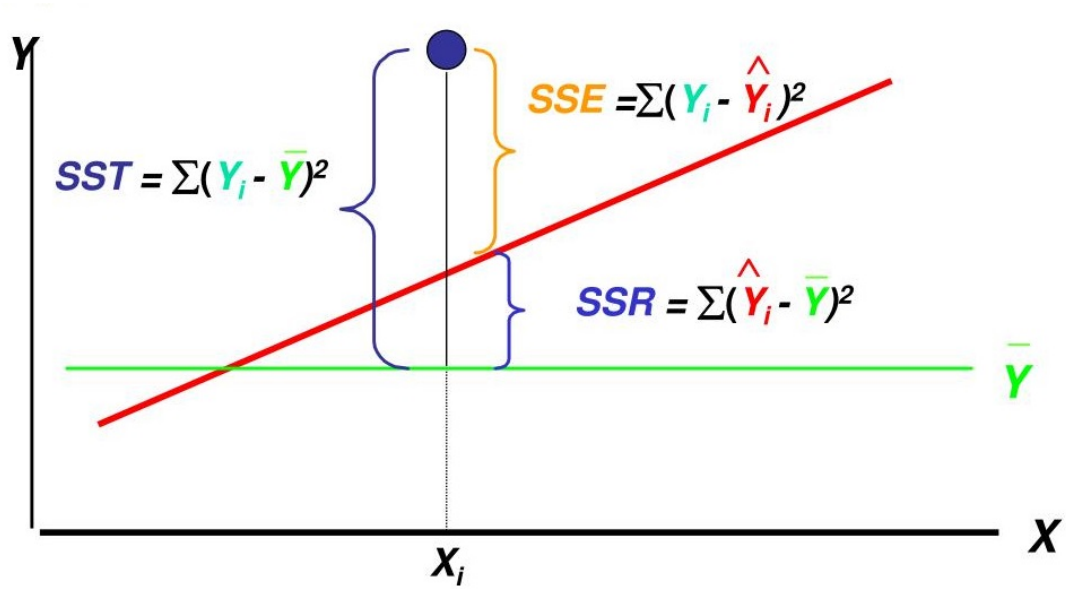


图 8：变异的分解

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）$R^2 = SSR/SST$.

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，样本量为 n，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	$p-1$	$\mathsf{MSR} = \mathsf{SSR}/(p-1)$
误差	SSE	$n-p$	$\mathsf{MSE} = \mathsf{SSE}/(n-p)$
总变异	SST	$n-1$	$\mathsf{MST} = \mathsf{SST}/(n-1)$

相应地，可以构造 F 检验：

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

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：

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

遗漏变量问题

遗漏变量（omitted variables）问题是指，真实模型为：

$$y \sim \beta_1 x_1 + \beta_2 x_2 . (1)$$

但遗漏了变量 x_2，且 $Cov(x_1, x_x) \neq 0$，从而模型被错误设定为：

$$y \sim \beta_1 x_1 . (2)$$

记模型 (1) 的系数估计分别为 $\hat{\beta}_1$ 和 $\hat{\beta}_2$，模型 (2) 的系数估计为 $\tilde{\beta}_1$，模型 $x_2 \sim \gamma x_1$ 的系数估计为 $\hat{\gamma}$，那么，可以证明：

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \cdot \hat{\gamma}.$$

遗漏变量偏误

遗漏变量偏误（omitted variable bias, OVB）即为：

$$\text{OVB} = \hat{\beta}_2 \cdot \hat{\gamma}.$$

案例：辛普森悖论、大学招生、教育的经济回报、药物治疗效果

课堂讨论：（1）哪些变量一定要纳入模型，作为控制变量？（2）哪些变量不必纳入模型作为控制变量？（3）出现了遗漏变量，如何判断回归系数偏误的方向？

辛普森悖论

患者	处理组（服药）		对照组（未服药）	
	痊愈人数	痊愈率	痊愈人数	痊愈率
女性	81/87	93%	234/270	87%
男性	192/263	73%	55/80	69%
总数	273/350	78%	289/350	83%

图 9: Simpson's Paradox

- 不论男性患者还是女性患者，处理组（服药）的痊愈率更高
- 就整体人群而言，处理组（服药）的痊愈率更低
- 该药物到底是否能够提高痊愈率?

性别歧视?

学生	历史系		机械系		总数	
	录取人数	录取率	录取人数	录取率	录取人数	录取率
女性	12/40	30%	6/10	60%	18/50	36%
男性	1/10	10%	20/40	50%	21/50	42%

图 10: Sex discrimination

- 就整个大学来看，女性的录取率低于男性
- 就各专业来看，女性的录取率均高于男性
- 该大学是否在招生时存在针对女性的性别歧视?

Section 3

线性回归建模

线性回归建模

本部分请参阅 slides!

Section 4

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

```
suppressMessages(library(car))
# calculate VIF
vif(fit)
```

异方差

通常将违背残差分布假定的

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

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。

回归分析总结

1. 回归建模与诊断：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