思睿教育
UTSG校区

WINTER 20/21

# STA 303

Review

面条



思 睿 教 育 集 团 荣 誉 出 品

# Disclaimers

# Winter 2021 STA 303

Mixed assessment

February 19, 2021

## Contents

# 1 Review of Linear Regression

- Model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i$$

  - $y$: response variable
  - $x_1, \ldots, x_p$: Explanatory variable/predictor
  - $\varepsilon$: Error term

- Key Assumptions:

  - **Linear** relationship: All the $\beta$'s enter the model in a linear way. The predictor can be in any form.
    * $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}^2 + \varepsilon_i \quad \checkmark$
    * $y_i = \beta_0 + \beta_1 \beta_2 x_{1i} + \varepsilon_i \quad \times$

    总而言之，不要让不同的 $\beta$ 出现在同一项里。
  - Errors are **independent** (Satisfied if observations are independent)
    比如说对于这一个病人，连续测量了 100 天的体温，就不能算作是一组 independent observations。
  - Errors are **normally** distributed with zero expected value, i.e., $E(\varepsilon_i) = 0$
    如果 response variable 是一个 categorical variabel，那就不能用 linear regression 来做。
  - **Equal/constant** variance (**homoscedasticity**) , i.e., $\text{var}(\varepsilon_i) = \sigma^2$.

# 2 Common Statistical Tests

## 2.1 Introduction

| | Parametric | Non-Parametric |
|---|---|---|
| 主要的区别 | Require distributional assumption | Purely based on data(Distribution Free) |
| 与某组数字对比差异 | One-sample t-test | Wilcoxon signed rank |
| 两组数据之间的差异 | Two-sample t-test | Mann-Whitney-U |
| 多组数据之间的差异 | One-way ANOVA | Kruskal-Wallace |
| 配对数据之间的差异 | Paired t-test | Paired Wilcoxon signed rank |

这门课我们主要 focus 在 Parametric Tests 上，因为大多数时候我们对于样本都假设服从正态分布 – Normal distribution.

## 2.2 One-Sample t-test

- **Assumptions:**
  1. The data are continuous
  2. The data are normally distributed
  3. The sample is a simple random sample from its population. （意味着 population 里每一个个体都有相同的概率被包含在样本里，同时 random sampling 保证了 observation 之间的独立性）
- **Hypotheses:**

$$H_0 : \mu = \text{hypothesized value}$$
$$H_1 : \mu \neq \text{hypothesized value}$$

- **Test Statistic:**

$$t = \frac{\bar{x} - \text{hypothesized value}}{s/\sqrt{n}}$$

  - $\bar{x}$: sample mean
  - $s$: sample standard deviation

  Under $H_0$, the test statistic follows the $t$-distribution with degree of freedom $\text{df} = n-1$.

- **p-value:**

$$\text{p-value} = \Pr(|t_{n-1}| > |t|) = 2\Pr(t_{n-1} > |t|)$$

- **Relationship with Linear Regression:**

$$y = \beta_0 + \varepsilon \qquad \text{(Intercept-only model)}$$

  Then,

$$H_0 : \mu = \text{hypothesized value} \Leftrightarrow H_0 : \beta_0 = \text{hypothesized value}$$
$$H_1 : \mu \neq \text{hypothesized value} \Leftrightarrow H_1 : \beta_0 \neq \text{hypothesized value}$$

## 2.3 Dummy Variables

这一部分内容是为了帮助大家理解后面介绍的多样本检验。

- **Model matrix/Design matrix:** Consider a multiple linear regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

and

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}.$$

Here, $\mathbf{X}$ is the so-called model matrix.

  - Usually, if we have $p$ predictors, we will have a $n \times (p+1)$ matrix.
  - However, if one of the predictor is a categorical variable (also called factor) with $k$ different levels. Then we may have another story.

- **How R deals with factor?**

  - **Case 1:** When the predictor is numerical data, one column for each predicator.
  - **Case 2:** When the predictor is categorical data with $k$ levels, $k-1$ columns will be allocated to this predictor with meaningful names. In this case, if $k > 2$, the model matrix is no longer tidy.

```
> y = rnorm(6)
> #Case 1: X is numerical data
> x = rnorm(6)
> mod1 = lm(y~x)
> model.matrix(mod1)
  (Intercept)           x
1           1 -0.36779526
2           1  0.80742972
3           1  0.31876245
4           1  1.14942357
5           1 -0.08720758
6           1 -0.25540711
attr(,"assign")
[1] 0 1
>
> #Case 2: X is categorical data, suppose X have three levels ("Low","Medium","High")
> x = rep(c("Low","Medium","High"),2)
> mod2 = lm(y~x)
> model.matrix(mod2)
  (Intercept) xLow xMedium
1           1    1       0
2           1    0       1
3           1    0       0
4           1    1       0
5           1    0       1
6           1    0       0
attr(,"assign")
[1] 0 1 1
attr(,"contrasts")
attr(,"contrasts")$x
[1] "contr.treatment"
```

- **Dummy variable**
  - What R is doing for a categorical variable(factor)?
    * Drop the first level (Alphabetically), the dropped level becomes the reference level.
    * Create dummy variables for the other levels.
  - How to interpret dummy variables?
    * 0: The observation not belongs to that level
    * 1: The observation belongs to that level
  - Why we have to drop one level?
    * Mathematically, we have to make the model matrix $\mathbf{X}$ an invertible matrix to get estimates of $\boldsymbol{\beta}$.
    * Intuitively, we should have linearly independent predictors.

## 2.4 Two-sample t-test

- **Assumptions:**

  1. The data are continuous
  2. The data are normally distributed in each group
  3. Each group is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample
  4. ★ The variances for the groups are equal.

- **Hypotheses:**

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

- **Test Statistic:**

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

  - $\bar{x}_1, \bar{x}_2$: sample mean for each group
  - $s$: pooled sample standard deviation

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

  Under $H_0$, the test statistic follows the $t$-distribution with degree of freedom

$$\text{df} = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2.$$

- **p-value:**

$$\text{p-value} = \Pr(|t_{n_1+n_2-2}| > |t|) = 2\Pr(t_{n_1+n_2-2} > |t|)$$

- **Relationship with Linear Regression:**

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (x \text{ is a factor with two levels})$$

  Then,

$$H_0 : \mu_1 = \mu_2 \Leftrightarrow H_0 : \beta_1 = 0$$
$$H_1 : \mu_1 \neq \mu_2 \Leftrightarrow H_1 : \beta_1 \neq 0$$

## 2.5 One-way ANOVA (F-test)

可以看作是 independent two-sample t-test 的延伸，用于检验多组样本（>2）之间的差异。

- **Assumptions:**
  1. The data are continuous
  2. The data are normally distributed in each group
  3. Each group is a simple random sample from its population. Each individual in the population has an equal probability of being selected in the sample
  4. ★ The variances for the groups are equal.

- **Hypotheses:**

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_n$$

$$H_1 : \text{at least one } \mu \text{ differ from the others.}$$

- **Relationship with Linear Regression:**

$$y = \beta_0 + \beta_1 D_1 + \ldots + \beta_{n-1} D_{n-1} + \varepsilon$$

Here, we assume predictor $x$ is a factor with $n$ levels (corresponding to $n$ different groups), therefore R will allocate $n-1$ dummy variables in the model.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_n \Leftrightarrow H_0 : \beta_1 = \beta_2 = \cdots = \beta_{n-1} = 0$$

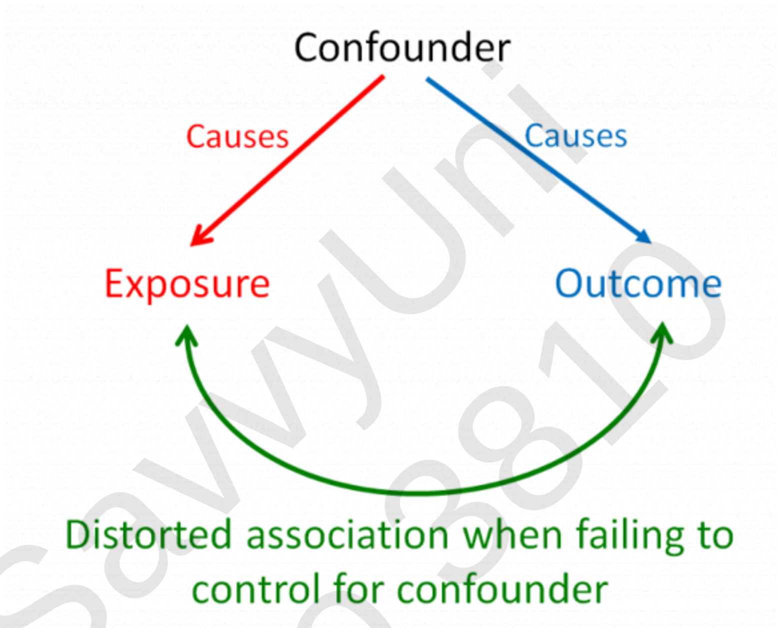
$$H_1 : \text{at least one } \mu \text{ differ from the others.} \Leftrightarrow H_1 : \text{at least one } \beta \text{ differ from 0.}$$

所以 linear regression 里的 F-test 其实在比较的是一个 full model 和 intercept-only model 之间的 fit performance。

# 3 Confounding and Study Design

## 3.1 Confounding

- Confounder: 干扰因素，会同时影响 dependent variable ($Y$) 和 independent variable ($X$)，忽略它的存在会使得我们得出错误的结论 (spurious association)



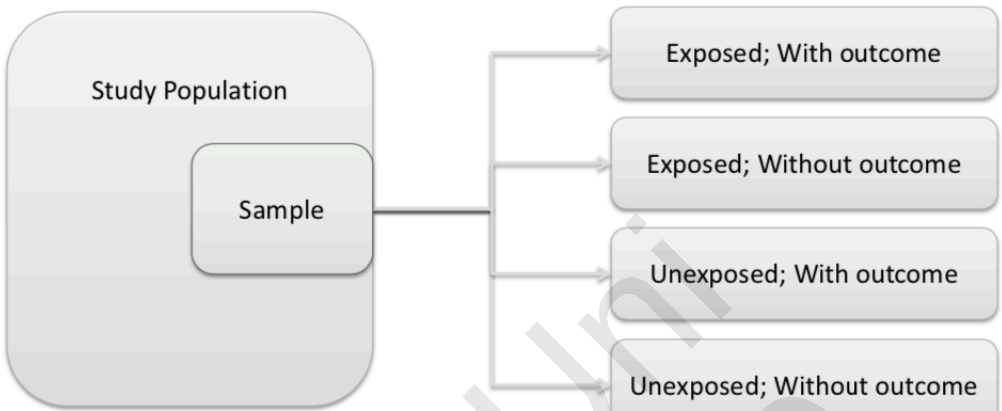

## 3.2 Association and causation

- **Association:** 相关性, 来自于调查的数据或者观察到的现象 (Observational Study)。
  - Confounder 的存在可能会使得我们得出错误的 association
- **Causation:** 因果性, 需要严格的科学实验来得出结论 (Experimental Study)
  - 会对 confounder 进行控制，消除 confounder 的影响

## 3.3 Observational vs Experimental Study

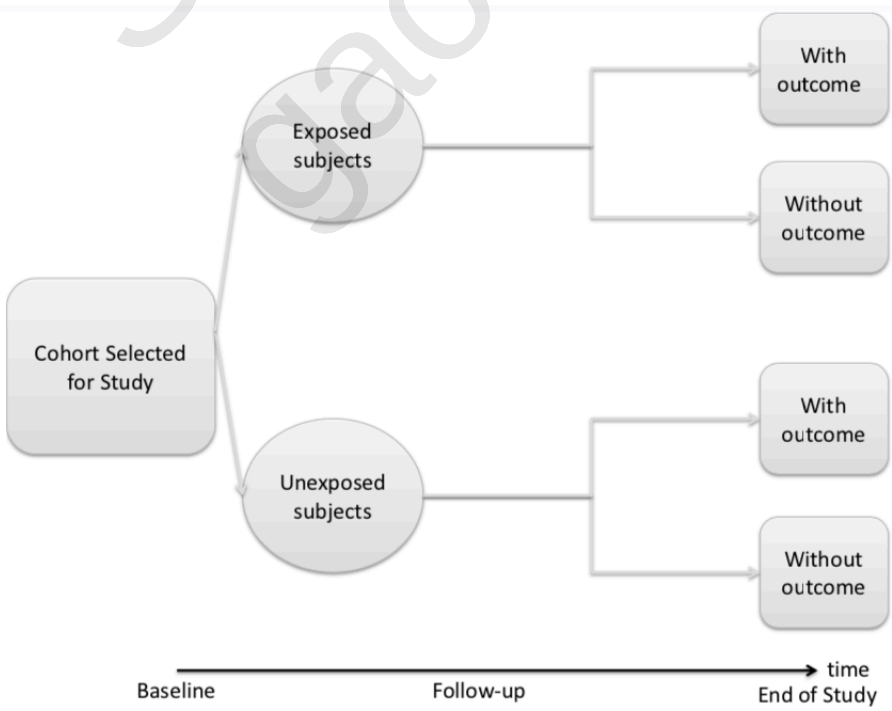
| | **Observational Study** | **Experimental Study** |
|---|---|---|
| **Methods** | Survey/Cross Sectional Study<br>Cohort Study, Case-Control Study | Randomized Control Trial |
| **Difference** | Without control over the exposure | With control over the exposure |
| **Conclusion** | Association | Causation |
| **Limitation** | Susceptible to confounding | Not always feasible (impratical/**unethical**/costly) |

## Survey/Cross-Sectional Study (Observational)




- 根据调查同时得到每一个 subject 的 exposure 和 outcome
- 样本需要有代表性，否则会面临 selection bias

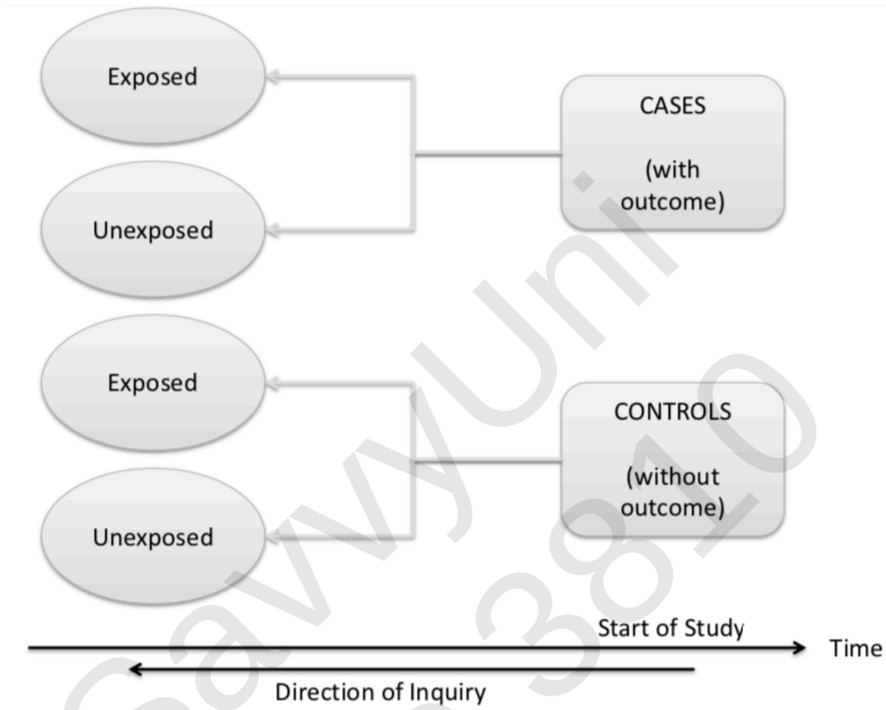
## Cohort Study (Observational)




- 先按照是否有 exposure 把研究对象分成两组，跟踪调查一段时间看看是否有 outcome 出现

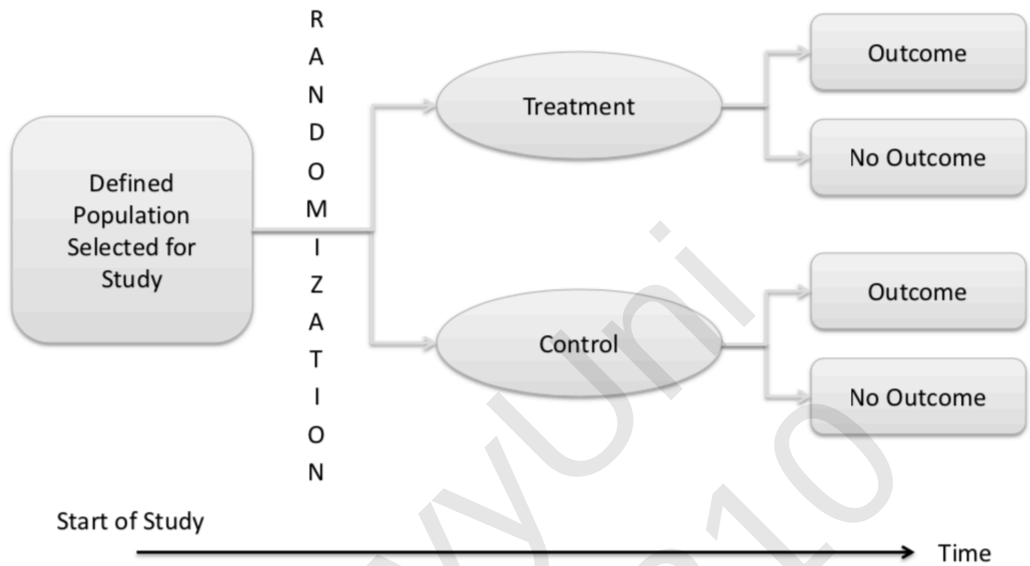
- 比较费时

## Case-Control Study (Observational)




- 先按照是否有 outcome 把研究对象分成两组，回溯过去一个阶段是否存在 exposure
- 比较适合潜伏期长或者罕见的疾病

### Randomized Controlled Trial (RCT) (Experimental)




- 先挑选没有 exposure 也没有 outcome 的群体，然后随机分配是否接受 exposure。最后跟踪调查看看哪些人会出现 outcome
- 因为实验是随机分配的，所以两个 group 之前唯一的差异就是是否存在 exposure/intervention，其他差异都是因为偶然，这样子就消除掉了 confounder 的影响。

# 4 Random Effect vs. Fixed Effect

- **Fixed effect:** 一般针对我们实验中直接想要去研究的未知参数(variable of interest)。通常来说我们的数据里包含了这个 factor 所有可能的类别。

$$\boxed{y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, i = 1, \ldots, a, j = 1, \ldots, n}$$

  - $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$
  - $\mu + \alpha_i$ denotes group mean for level $i$.

- **Random effect:** 一般存在于 correlated data 里，由于 dependence 的影响导致我们的分析存在了偏差，我们需要引入 random effect 来刻画 group 的差异。
  - 当一个 factor 有很多类别，但是数据里只包含了随机选择的一部分类别，为了避免以偏概全，我们会把它当作一个 random effect/variable
  - 通常 group 的差异对于结论的影响并不是我们直接关心的事。让我们把它放在 regression 里考虑的唯一原因是不然就违背了 independence assumption。

$$\boxed{y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, i = 1, \ldots, a, j = 1, \ldots, n}$$

  - $\alpha_i \sim N(0, \sigma_\alpha^2), \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$
  - Observations within same group share common term $\alpha_i$, therefore they are correlated.

- **Mixed effect:** 如果一个模型里既考虑了 fixed effect 又考虑了 random effect，那么就叫做 mixed effect model。
  - 有的时候我们甚至会考虑 interaction between fixed and random effect，注意 interaction term 也仍然要当作一个random variable。

$$\boxed{y_{ijk} = \mu + \alpha_i + b_j + (\alpha b)_{ij} + \varepsilon_{ijk}, i = 1, \ldots, a, j = 1, \ldots, n}$$

  - $b_j \sim N(0, \sigma_b^2), (\alpha b)_{ij} \sim N(0, \sigma_{\alpha b}^2), \varepsilon_{ijk} \overset{iid}{\sim} N(0, \sigma^2)$

按照老师的意思，只要涉及到 random effect model，我们列的 model equation 不会出现 $\beta_i$ 的结构。

**如何区分 fixed/random effect？**

- 如果是 grouping unit，都当作 random effect，比如说 group id/group name/group number
- 除此之外，都当作 fixed。尤其注意那些 group 的属性，都是 fixed。

**如何估计 $\sigma_b^2, \sigma_{\alpha b}^2, \sigma^2$?**

假设我们有 $I$ 个 treament level (fixed effect), $J$ 个 groups (random effect), 每一个 group 每一种 treatment repeat $K$ 次。

- Aggregate (Main) model:

$$\texttt{Response} \sim \texttt{Treatment+Group}$$

- Interaction model:

$$\texttt{Response} \sim \texttt{Treatment*Group}$$

- Group model:

$$\texttt{Average Response over each group} \ \sim 1$$

- $\hat{\sigma}^2 \Leftarrow$ `summary(Interaction model)$sigma2`
- $\hat{\sigma}_b^2 \Leftarrow$ `summary(group model)$sigma2` $-$ `summary(Aggregate model)$sigma2`$/I$
- $\hat{\sigma}_{\alpha b}^2 \Leftarrow$ `summary(Aggregate model)$sigma2`$-$`summary(Interaction model)$sigma2`$/K$

## 5 Nested design vs. Crossed effect design

- Nested design: 个体嵌套于群体/组别。通常每一个组别只会接受一种 treamtnet
- Crossed effect design: 每一个个体会接受所有的 treament，通常是个体的重复性实验。

对于 nested design 来说，考虑 group 和 treatment 之间的 interaction 没有任何意义。因为每一个 group 里只接受一种 treatment，不存在重叠。

## 6 Likelihoods

- 相比于 regression，有以下几点好处:
  - 不需要假设样本是正态分布
  - 不需要假设样本之间是独立的
- Likelihood: 表示的是在给定参数下能够获取到当前观测数据的数据可能性
- Likelihood Ratio Test: 适用于 nested model 之间的比较。通常来说会有一个 reduced model，一个 full model。full model 里包含了所有 reduced model 里面的 parameters。

$$\text{LRT} = 2\log\left(\frac{\max(\text{Lik(Full model)})}{\max(\text{Lik(Reduced model)})}\right)$$

- $H_0$ : Simpler model explains the data just as well as the more complicated
- 比较方法：利用 likelihood ratio test
- 结论：
  - 如果 p value >0.05，no evidence against, should use the simpler model
  - 如果 p value <0.05，strong evidence against, should use the complicated model