复制论文作业说明

南开大学金融学院　赵博

2021 秋季

1　为什么？

“可复制” (reproducible) 是科学研究的最基本、也是最关键的要求。尤其在自然科学领域，“可复制性” 是检验研究结果是否可靠的最重要的标准之一[1]。研究结果不可复制有可能产生于：

1. 学术不端
2. 研究方法有错误
3. 偶然性的因素使得某次实验结果成功，对于认识世界并没有普遍意义

无论是哪种情况，检验研究成果是否可以复制，都可以帮助人们分辨谬误，厘清事实，接近真理。复制别人的工作同时也可以带来很多好处：

1. 可以熟悉编程和数据，并且通过对比检验自己（或原作者）的结果是否正确
2. 可以迅速熟悉一个领域的研究方法

在社会科学领域，由于可控实验比较困难，对于 “可复制” 的要求没有自然科学领域中那么严苛。这也使得社会科学的研究中，谬误成为一种常

[1]近期的由于 “不可复制” 而引发的学术争议作者包括韩春雨，小保方晴子，以及经济学领域的 C. Reinhart, K. Rogoff (2010, AER), A. Rampini, S. Viswanathan G. Vuillemey(2020, JF，已撤稿)。

见的情况[2]。有鉴于此，the *American Economic Review* 自 1986 年起，要求作者提供数据和程序代码给有需求的读者，并且逐渐要求所有发表文章的作者上传程序代码和数据至指定的服务器。类似的政策也可见于其他一些经济学和金融学期刊上。

社会科学领域的复制工作通常采取两种形式：1. 直接检验原实证研究的结果；2. 检验一个实证研究的结论是否在其他地区、其他时间、子样本中成立。

这个作业要求的是第 1 种复制，也就是用原论文采用的数据，复制原论文的实证结果。

2 怎么做？

2.1 作业分组

两人一组。如果总人数为单数，则落单的 1 人可选择自己完成（相对而言打分会更加宽松），或加入到其他组。特殊情况下也可选择 3 人一组，但必须经过我的同意。在最终的报告中，请仔细说明每一个小章节由哪位同学完成。例如，**1. 研究概述**（李雷），**2.2 Table 1** (韩梅梅)，**2.3 Table 2**（共同完成），等等。

2.2 挑选论文建议

1. 挑选一篇你感兴趣的领域的论文，方向尽量和自己毕业论文的研究方向接近

2. 从好的期刊挑选（“好的”期刊列表见附录），时间在 1990 年以后

3. 谷歌学术引用量在 200 次以上

[2] …inadvertent errors in published empirical articles are a commonplace rather than a rare occurrence. (Dewald, W., Thursby, J., & Anderson, R. (1986). Replication in Empirical Economics: The Journal of Money, Credit and Banking Project. *The American Economic Review*, 76(4), 587-603.)

4. 数据必须可公开获得。也就是说，不要挑选数据集是作者自行收集的论文，因为这样的数据已被作者预处理过。公开的数据库比如 CRSP, Compustat, Federal Reserve Economic Data，万得，国泰安，等。(或者作者已提供数据，但数据同时是可公开获得的)

5. 必须是实证论文，也即有数据、计量方法检验的论文。也可以是理论 + 实证的论文。不能选择纯理论论文

6. 请尽量选择英文论文进行复制

7. 也可以选择利用交叉学科方法的论文进行复制

8. 多读，多看，多想，多问

当你有了候选论文之后，下载论文并做简要概述并发给我。概述可用一两句话讲清楚：

1. 论文研究的问题

2. 为什么需要研究这个问题

3. 文章用到的计量方法

4. 文章的基本结论

请将论文的引用信息（例如：Dewald, W., Thursby, J., & Anderson, R. (1986). Replication in Empirical Economics: The Journal of Money, Credit and Banking Project. *The American Economic Review*, 76(4), 587-603.），**以及概述**，发至我的邮箱：zhaobo@nankai.edu.cn。得到我的批准以后，才可进行接下来的工作。

2.3 复制的内容

不必完整复制论文的所有表格和图表。挑选论文最核心的部分进行复制。请根据自己的时间合理安排。

2.4 提交的格式

文字和图表**请以 PDF 文档**的方式提交。图表的格式请**以标准格式**调整，而不要直接粘贴程序回归结果。以下是正确和不正确的格式[3]：

Table 1: Regression table

	(1) Price	(2) Price
Weight (lbs.)	1.747**	3.465***
	(2.72)	(5.49)
Mileage (mpg)	−49.51	21.85
	(−0.57)	(0.29)
Car type		3673.1***
		(5.37)
Constant	1946.1	−5853.7
	(0.54)	(−1.73)
Observations	74	74

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

图 1: A good example

```
Call:
lm(formula = Sepal.Length ~ Sepal.Width + Petal.Length + Petal.Width,
    data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-0.82816 -0.21989  0.01875  0.19709  0.84570

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.85600    0.25078   7.401 9.85e-12 ***
Sepal.Width   0.65084    0.06665   9.765  < 2e-16 ***
Petal.Length  0.70913    0.05672  12.502  < 2e-16 ***
Petal.Width  -0.55648    0.12755  -4.363 2.41e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3145 on 146 degrees of freedom
Multiple R-squared:  0.8586, Adjusted R-squared:  0.8557
F-statistic: 295.5 on 3 and 146 DF,  p-value: < 2.2e-16
```

(a) Copy from R output directly

```
. sysuse auto
(1978 Automobile Data)

. reg price weight mpg foreign

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(3, 70)        =     23.29
       Model |   317252881         3  105750960   Prob > F        =    0.0000
    Residual |   317812515        70  4540178.78   R-squared       =    0.4996
-------------+----------------------------------   Adj R-squared   =    0.4781
       Total |   635065396        73  8699525.97   Root MSE        =    2130.8

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   3.464706    .630749     5.49   0.000     2.206717    4.722695
         mpg |    21.8536   74.22114     0.29   0.769    -126.1758    169.883
     foreign |    3673.06   683.9783     5.37   0.000     2308.909    5037.212
       _cons |  -5853.696   3376.987    -1.73   0.087    -12588.88    881.4934
------------------------------------------------------------------------------
```

(b) Copy from Stata output directly

图 2: Bad examples

3 时间表

以下是各个阶段作业进度的时间表：

[3]请自行查找如何制作表格。例如，Stata 可以参考 estout, outreg2，R 可以参考 xtable 等等

- 10 月 30 号：确定论文题目（请尽量提早确定，以免选题不妥，导致后续时间不够）
- **11 月 15 号**：请于此日期前给我发送邮件确认题目。同时请收集好数据，计算描述性统计量。（不要晚于此日期确定题目，此日期后更改题目将扣除总得分的 30%）
- **12 月 30 号**（终稿提交日期）：最终作业提交。

做好时间规划！论文题目确定和数据收集会占据很多时间。**最晚请在 11 月 15 号之前将开题确定。换句话说，11 月 15 日以前，必须弄清楚文章的基本内容，以及数据是否可得。否则将扣除最终得分的 30%。12 月 30 号为最终期限，过期作业将不予收取。**

开题和终稿请提交至我的邮箱：zhaobo@nankai.edu.cn。命名格式："1-姓名-姓名-选题.pdf"，"2-姓名-姓名-终稿.pdf"，"2-姓名-姓名-代码.txt"。

最终作业提交包括两份文件：

1. PDF 格式的文字、图表
2. 源代码文件

请将两个文件分开 (无需打包压缩) 并按照以上类似格式进行命名。终稿中无须包括开题时已提交的文件，只需提交一份完整的报告和代码。

4 常见问题

- 实在复制不出原文的结果

  有几种可能：1. 你的代码出错了。2. 原文作者做错了。3. 原文作者藏着掖着一些步骤没有说清楚，让你无法知道自己哪里做的不对。

  解决办法：1. 仔细检查自己的代码。这个过程中你能进一步熟悉模型，熟悉软件，熟悉代码。2. 给原文作者发邮件询问。

  如果成功复制出来了，恭喜。如果还是不能成功，把你们尝试过的、做过的、想过的，尽量写清楚，打分不会苛求完整复制。也许未来多读多做了一些以后，能回头搞清楚是怎么回事。

- 论文我很喜欢，但找不到国外的数据、能找到中国的数据，怎么办?

  如果你真的非它不选，那么用中国的数据复制。这样的缺点是：你无法确切知道自己的代码是否正确。

5 附录: 部分可选期刊建议（只包括经济学和金融学专业）

保险和精算专业的同学也请选择和计量经济学方法相关的文章。可以选择在此列表之外的期刊或者研究报告，但必须和我讨论确定。

American Economic Review

Econometrica

Journal of Political Economy

Quarterly Journal of Economics

Review of Economic Studies

Journal of Finance

Journal of Financial Economics

Review of Financial Studies

Journal of Financial Quantitative Analysis

Management Science

Economic Journal

International Economic Review

Journal of Econometrics

Journal of Economic Theory

Journal of Labor Economics

Journal of Monetary Economics

Review of Economics and Statistics

Review of Finance

Journal of Banking and Finance

Journal of Financial Markets

...

经济研究

金融研究
管理世界
世界经济