计量考卷：去年（2024年）和这套卷子（2022年）的区别是就删了一道题。

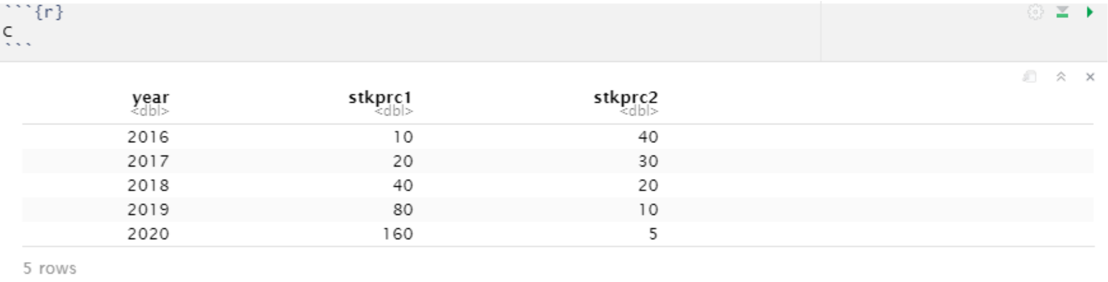
里面的答案不一定对，因为是当时学长和学姐在做的过程中打印出来写的，具体答案还需要让ds和豆包辩论得出。

今年是不是还一样的试卷只能祈祷

Final Exam of Financial Econometrics, Fall 2022

December 26, 2022
8:30-10:30 am
Total Points: 120

1. (8 points, 2 points each)
For each of the following R codes, describe what the code is trying to do, and write down the output:



~~~
```{r}
C
```
~~~

year <dbl>	stkprc1 <dbl>	stkprc2 <dbl>
2016	10	40
2017	20	30
2018	40	20
2019	80	10
2020	160	5

5 rows

a) C[1,]

输出第一行 标题!
year stkprc1 stkprc2
2016 10 40

b) C[c(1,2),c(1,3)]

输出第一二行、第一三列
year stkprc2
2016 40
2017 30

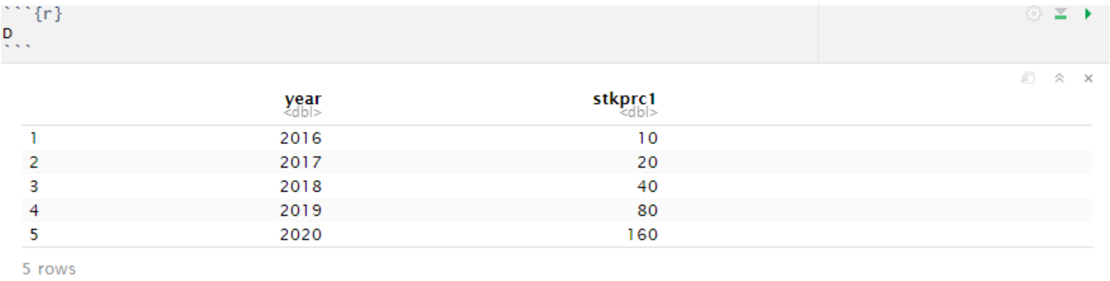
c) C[year>=2019]

输出年份大于等于2019的数据
year stkprc1 stkprc2
2019 80 10
2020 160 5

d) library(dplyr)
(C %>% select(year, stkprc2) %>% filter(year<2018))

输出年份小于2018、year stkprc2两列
year stkprc2
2016 40
2017 30

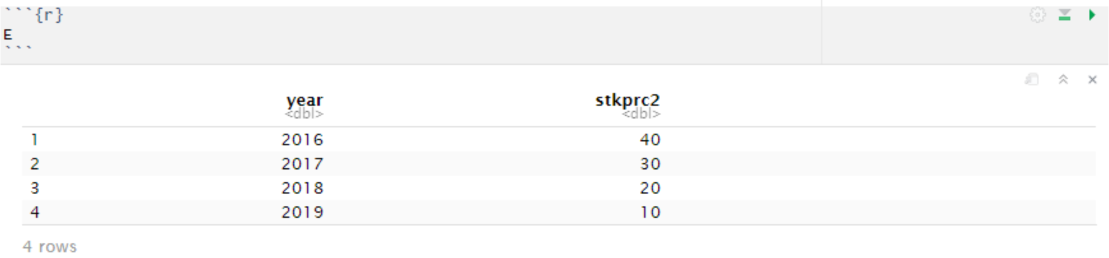
2. (9 points, 3 points each) Write R codes to perform the tasks required.



~~~~
```{r}
D
```
~~~~

	year <dbl>	stkprc1 <dbl>
1	2016	10
2	2017	20
3	2018	40
4	2019	80
5	2020	160

5 rows



~~~~
```{r}
E
```
~~~~

	year <dbl>	stkprc2 <dbl>
1	2016	40
2	2017	30
3	2018	20
4	2019	10

4 rows

a) Using D and E, assemble a new data frame F.
F should have three variables: year, stkprc1 and stkprc2.
F should have 5 rows, corresponding to year 2016, 2017, 2018, 2019, 2020.
Your code should take only one line.
Just write down the code. No need to write down the output.

```
F <- merge(D, E, by = "year", all.x = TRUE)
```

b) Using D and the ggplot2 package, draw a scatter plot of stkprc1 (y) on year (x) with a linear regression line and the associated confidence intervals of the linear regression.
Just write down the code. No need to draw the figure yourself.

```
ggplot(D, aes(x = year, y = stkprc1)) + geom_point() + geom_smooth(method = "lm", se = TRUE)
```

c) Using C and the ggplot2 package, draw a line plot of stkprc1 (y1) and stkprc2 (y2) on year (x).
Just write down the code. No need to draw the figure yourself.

```
ggplot(C, aes(x = year)) +
  geom_line(aes(y = stkprc1), color = "blue") +
  geom_line(aes(y = stkprc2), color = "red")
```

3. (6 points) You are studying the relationship between a firm’s return on equity (ROE) and CEO salary. In the dataset ceosal, you find that the average salary of a CEO (salary) is 1,281 thousand USD, and the average ROE of a firm (roe) is 17.18%. The covariance between salary and roe is 1342.54. The standard deviation of salary is 1,372 thousand USD, and the standard deviation of roe is 8.52%. The regression equation is:

salary_i = b0 + b1 * roe_i + u_i

What is the OLS estimate of b1?

b1 = cov(salary_i , roe_i)/var(roe_i) = 1342.54 / 0.0852 ^ 2 = 184947.32

[分母是8.52还是8.52%?]

4. (5 points) In the linear regression model, we have defined the total sum of squares (TSS), the explained sum of squares (SSE), and the residual sum of squares (SSR). They are calculated as follows:

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 \,, SSE = \sum_{i=1}^{n} (\,\hat{y}_i - \bar{y})^2 \,, SSR = \sum_{i=1}^{n} \hat{u}_i^2$$

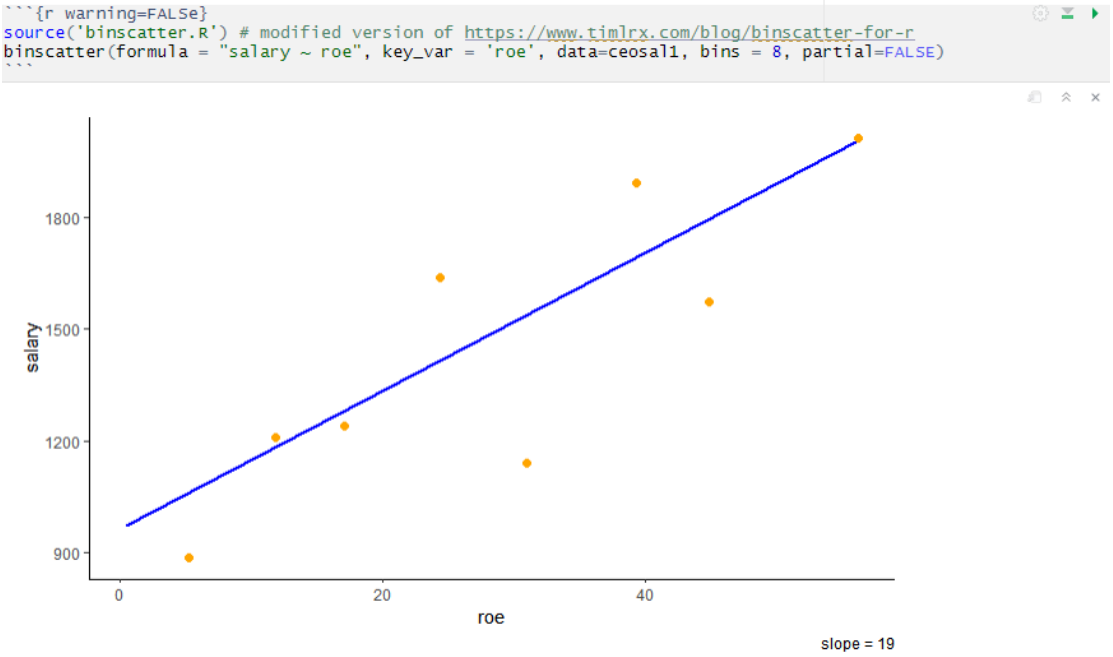
where $\hat{y}_i = b_0 + \hat{b}_1 x_{1i} + \cdots + + \hat{b}_k x_{ki}$ and $\hat{u}_i = y_i - \hat{y}_i$.

Question: Which of the following is the R-squared?

(A) $\left(\frac{\mathrm{SSE}}{\mathrm{SSR}}\right)$ (B) $\left(\frac{\mathrm{SSE}}{\mathrm{SSR}}\right)^2$ [✓ (C) $\left(\frac{\mathrm{SSE}}{\mathrm{TSS}}\right)$] (D) $1 - \left(\frac{\mathrm{SSE}}{\mathrm{TSS}}\right)$

解释平方和/总平方和
R方是模型解释的因变量方差占总方差的比例
也可以表述为 1-SSR/TSS

5. (5 points, 1 point each) True or False (binscatter)



~~~~
```{r warning=FALSE}
source('binscatter.R') # modified version of https://www.timlrx.com/blog/binscatter-for-r
binscatter(formula = "salary ~ roe", key_var = 'roe', data=ceosal1, bins = 8, partial=FALSE)
```
~~~~

A) The code generates a scatter plot of salary (y) on roe (x) using the ceosal1 dataset.

B) The code generates a scatter plot of grouped averages of salary (y) on roe (x) using the ceosal1 dataset.

C) The figure indicates that the homotheticity assumption likely does not hold.

D) The figure indicates that the relationship between salary and roe may not be as simple as a linear relationship.

E) The binscatter graph is a helpful tool, but it is not useful for exploring the relationship between y and x when y contains measurement errors.

a) False

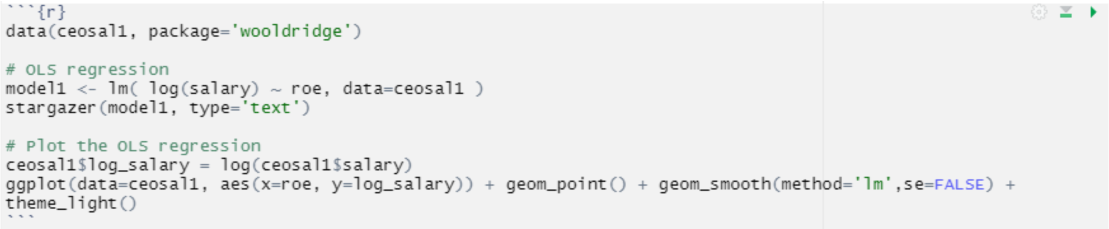
b) True
bin=8 不是绘制原始的每个数据点的散点图，而是对每个分箱内的数据进行汇总，绘制的是分箱后的组平均数据的散点图

c) True

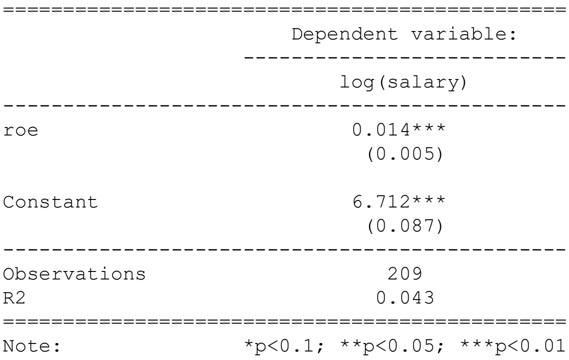
d) False

e) False

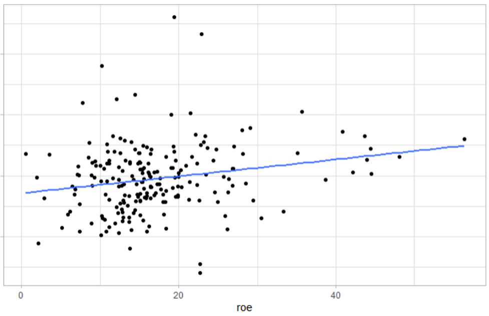
6. (9 points, 3 point each) Interpret the regression results in the following table. The dependent variable is the logarithm of the salary of a firm’s CEO, and the independent variable is the firm’s return on equity (ROE).



~~~
```{r}
data(ceosal1, package='wooldridge')

# OLS regression
model1 <- lm( log(salary) ~ roe, data=ceosal1 )
stargazer(model1, type='text')

# Plot the OLS regression
ceosal1$log_salary = log(ceosal1$salary)
ggplot(data=ceosal1, aes(x=roe, y=log_salary)) + geom_point() + geom_smooth(method='lm',se=FALSE) +
theme_light()
```
~~~



```
===============================================
                        Dependent variable:
                    ---------------------------
                            log(salary)
-----------------------------------------------
roe                          0.014***
                              (0.005)

Constant                     6.712***
                              (0.087)

-----------------------------------------------
Observations                    209
R2                             0.043
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```



a) First, look at the regression results on the left. Interpret the economic size and the statistical significance of the regression coefficient of log(salary) on roe.

经济意义：ROE增加一个单位，工资的百分比增加一个单位
统计意义：ROE对log（salary）的回归高度显著，因为t统计量=0.014/0.005=2.8 比较大，p值也小于0.01***

b) Now, look at the figure on the right. Is the standard error in part a) correct? Why?

同方差的假设并不满足。看图残差的方差在 x 的范围的两端更小，在 x 的中间更大，判断可能不满足同方差假设
如果满足同方差假设，散点围绕回归线的分布在整个自变量取值范围内的离散程度应该大致相同，不会出现某些区域的点离回归线特别远或特别近，也不会出现随着自变量值的变化，点的离散程度明显增大或减小的情况。
1128计量 01：17：18

c) A friend recommended you to run the following code in addition to the code above. What does the code do and what is the meaning of “HC” in “HC0”?

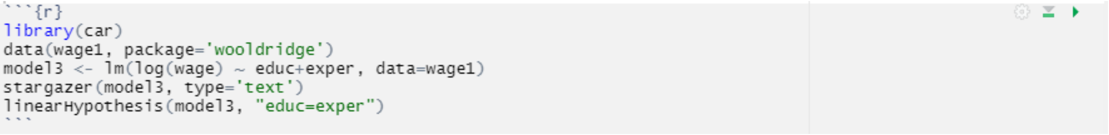
coeftest(model1, vcov=vcovHC, type=HC0)

coeftest 函数是用于检验回归系数的统计显著性，它会根据指定的协方差矩阵估计方法

这里通过 vcov = vcovHC 指定使用 vcovHC 函数来估计协方差矩阵来重新计算回归系数的标准误差和相应的统计量，进而得到更稳健的统计检验结果

“HC” 代表 “Heteroskedasticity-Consistent”，即异方差一致性。
“HC0” 是异方差一致性估计的一种具体类型。它是对协方差矩阵进行估计的一种方法，用于修正可能存在的异方差问题，使得回归系数的标准误差估计更加稳健，即使在误差项存在异方差的情况下，也能提供更可靠的统计推断。

7. (6 points, 3 point each) Interpreting the following code and result. The data set contains the wage level (wage), the number of years of education (educ), and the number of years of work experience (exper) for 526 workers.



~~~
```{r}
library(car)
data(wage1, package='wooldridge')
model3 <- lm(log(wage) ~ educ+exper, data=wage1)
stargazer(model3, type='text')
linearHypothesis(model3, "educ=exper")
```
~~~

```
===============================================
                        Dependent variable:    
                    ---------------------------
                             log(wage)         
-----------------------------------------------
educ                         0.098***          
                              (0.008)          
                                               
exper                        0.010***          
                              (0.002)          
                                               
Constant                      0.217**          
                              (0.109)          
                                               
-----------------------------------------------
Observations                    526            
R2                             0.249           
Adjusted R2                    0.246           
Residual Std. Error      0.461 (df = 523)      
F Statistic          86.862*** (df = 2; 523)   
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
Linear hypothesis test

Hypothesis:
educ - exper = 0

Model 1: restricted model
Model 2: log(wage) ~ educ + exper

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1    524 141.92                                  
2    523 111.34  1    30.576 143.62 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

a) Look at the last line of code that starts with “linearHypothesis(…” What does this line of code do?

linearHypothesis()线性假设检验函数
这一行代码是表示要进行的线性假设为教育程度（educ）和工作经验（exper）的系数相等

b) What can you say about the effects of one more year of education versus one more year of work experience, respectively, on the worker’s wage?

education：在其他条件不变的情况下,工人多接受一年教育，会导致他的工资水平平均增加0.098个百分点

exper： 在其他条件不变的情况下,工人多一年工作经验，会导致他的工资水平平均增加0.010个百分点

8. (6 points, 3 points each)

Interpret the effect of x variables on wage:

```
===============================================
                        Dependent variable:
                    ---------------------------
                             log(wage)
-----------------------------------------------
married                      0.213***
                              (0.055)

female                       -0.110**
                              (0.056)

educ                         0.079***
                              (0.007)

exper                        0.027***
                              (0.005)

I(exper2)                    -0.001***
                             (0.0001)

married:female               -0.301***
                              (0.072)

Constant                     0.321***
                              (0.100)

-----------------------------------------------
Observations                    526
R2                             0.461
Adjusted R2                    0.453
Residual Std. Error      0.393 (df = 517)
F Statistic           55.246*** (df = 8; 517)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

a) Everything else equal, how does the wage of a single male compare to the wage of a married female?

单身+男性 vs 已婚+女性

分析married、female、married：female三项系数
对于单身男性（married=0 female=0）：y = constant
对于已婚女性（married=1 female=1）：y = constant + 0.213 - 0.110 - 0.301 = constant - 0.198
即单身男性比已婚女性的工资高0.198个单位

b) For a worker with 10 years of work experience, what is the expected effect of an additional year of work experience on her wage?

(Hint: the answer is slightly different if you use an exact calculation or if you take the derivative. Both answers will be correct.)

与工作经验相关的系数：exper、I（exper2）

法1：代入作差
exper=10,y1 = constant + 10*0.027 + 100*（-0.001）= constant+0.17
exper=11,y2 = constant + 11*0.027 + 121*(-0.001) = constant +0.176
y2 - y1 = 0.006
即对于有10年工作经验的工人来说，额外的一年工作经验对她的工资水平的期望影响是0.006个单位

法2：求导
dy/dx = 0.027 - 0.002 * exper
此时取exper=10 求出dy/dx =0.007
即对于有10年工作经验的工人来说，额外的一年工作经验对她的工资水平的期望影响是0.007个单位

9. (10 points, 2 points each) True or False (Binary choice models)

a) After estimating the linear probability model, you should compute the marginal effect.

b) After estimating the logit model or the probit model, you should compute the marginal effect.

c) The marginal effect of an x variable measures the average increase in Prob(Y=1) when that x variable increases by one.

d) After estimating a logit or a probit model, you should check the predicted value of the regression to see whether the predicted values are mostly between 0 and 1.

e) The logit model and the probit model are better than the linear probability model because the effect on Prob(Y=1) when the value of a dummy variable increases from 0 to 1 is more stable under the logit model and the probit model.

Binary choice models 二元选择模型 week4

a) FALSE 线性概率模型算出来的就是边际，所以不需要再计算边际值了

b) TRUE
模型带有非线性时，不能直接解释系数，常常通过计算marginal effect
比如logit模型中估计的系数不能直接解读，通过计算系数对应的概率变化，即marginal effect

c) TRUE 系数解读 自变量x增加一单位会导致因变量取1的平均变化情况

d) FALSE LPM会在（0，1）之外，logit和probit都不会

e) TRUE

10. (6 points) Measurement errors in the x variable: Suppose the true model that generates y and x is as follows:

$$\mathrm{y_i} = \alpha + \beta x_i + e_i$$

However, you do not observe the true value x_i. Instead, you observe $x_i^* = x_i + v_i$, where v_i is the measurement error in the x variable.　自变量x存在测量误差

Thus, when you estimate the regression model, you are in fact estimating the following model with measurement error in the x variable:

$$\mathrm{y_i} = b_0 + b_1 x_i^* + u_i$$

Question: What is the relationship between $\hat{b}_1$ (the estimated regression coefficient of y on x* when there are measurement errors in the x variable) and β (the true regression coefficient)?

Describe the steps you took to get to the answer.

b1^hat的绝对值比β更小，即b1^hat更接近0

首先根据题目给出的条件，我们可以知道：yi = b0+b1xi+ui+b1vi(代入xi* = xi+vi)

根据β的定义，β=cov(x，y)/var(x)

此时b1=cov(yi,xi)/var(xi)=[cov(b0,xi)+cov(b1xi,xi)+cov(ui,xi)+cov(b1v1,xi*-v1)/var(xi)

真实的β=[cov(b0,xi)+cov(b1xi,xi)+cov(ui,xi)]/var(xi)，且分子只有cov(b1xi,xi)非0

所以b1=β+ cov(b1v1，xi*-v1)/var(xi) ==> β和cov(b1v1，xi*-v1)/var(xi) 反向变动

老师在录音里面说为什么β和后面那一项会反向变动不用解释，可以记结论

(这里好像和老师讲的例子有个地方正好相反 测量误差那里老师上课讲的应该是xi= xi*+vi，所以代入以后是负号)

b1^hat = β · var(xi) /[var(xi)+var(vi)]--课件里的公式

tobit heckman
考适用场景、系数解读

11. (8 points, 2 points each) True or False (Tobit and Heckman selection)

a) The OLS is not valid when the data is MCAR.

b) The Tobit model is designed to handle situations where you observe a censored version of the true dependent variable. For example, the true dependent variable is y and you observe only y* = max(0,y). If you would like to know the effect of x on y, you should compute the marginal effect after estimating the Tobit model.

c) The Heckman selection model is designed to handle situations where selection into the sample may depend on some unobserved benefit, so that an estimation based on the OLS will suffer from an omitted variable bias. The Heckman selection model includes the Inverse Mills ratio from the second stage estimation as an additional control variable in the first stage regression to address this omitted variable bias concern.

d) You can use exactly the same set of x variables in the first stage and the second stage of the Heckman selection model.

a) FALSE MCAR=missing completely at random week5 5.1
当数据是 “完全随机缺失” (MCAR) 时，缺失数据不会引发统计问题（只需舍弃那些存在缺失值的观测数据即可）

b) TRUE censored version、tobit model week5 5.3
托宾模型常用于处理因变量存在截断（censored）的情况，当实际观测到的因变量是经过某种截断处理后的版本时，通过估计托宾模型并进一步计算边际效应，能够分析自变量对真实因变量的影响

c) FALSE heckman 两阶段回归 week5 5.4
第一阶段：利用最大似然估计法（MLE）对 Probit 模型进行估计，计算inverse mills ratio
逆米尔斯比率反映了样本选择过程中未观测到的因素对个体进入样本概率的影响
第二阶段：建立一个描述我们真正感兴趣的变量（如收入、消费等）与解释变量之间关系的结果方程，将第一阶段计算得到的逆米尔斯比率作为一个额外的解释变量加入到结果方程中,如果逆米尔斯比例的系数显著不为零，则说明存在样本选择偏差，通过加入逆米尔斯比率有效地纠正了这种偏差。
ZY老师上课基本没介绍这一部分，只提了适用情况有选择性披露问题、政策的选择性参与问题

d) FALSE 简单来说第一阶段是选择方程，强调影响样本选择；第二阶段是结果方程，关注结果与解释变量间的关系，所以两者不能完全一样

12. (9 points, 3 points each) The instrumental variable approach can be described as follows. You are studying the relationship between y and x: 工具变量

$$y_i = b_0 + b_1 x_i + u_i$$

but x_i and u_i are not independent, which violates the OLS assumptions and biases the estimated coefficient $\hat{b}_1$.

For example, y_i is a firm's innovation activities (e.g. number of new products developed), and x_i is the amount of VC investment the firm receives; in this case, u_i can include many factors that affects the firm's y_i beyond VC investment that it received.

a) Suppose whether the firm's founder graduated from a top 2 university positively attracts VC investment and positively affects innovation activities. How will this affect the correlation between x_i and u_i, and how will this affect the estimated coefficient $\hat{b}_1$?

xi ui正相关 b1^hat会被高估 因为cov(yi,xi)增大了 多了一项cov（ui,xi）

b) The instrumental variable approach uses an instrumental variable z_i, and involves estimating a first stage regression, computing the predicted value $\hat{x}_i$, and use the predicted value in the original regression:

$$x_i = a_0 + a_1 z_i + u_i$$
$$y_i = b_0 + b_1 \hat{x}_i + u_i$$

What are the names and the contents of the two assumptions required in order for the instrumental variable z_i to be valid?

相关性：工具变量z必须和x相关，cov（xi，zi）不等于0
外生性：工具变量z必须与ui无关，即z只通过x影响y

c) Based on these two assumptions, evaluate whether the number of VC funds in the firm's headquarter city is a valid instrument for the above example. Explain why.

相关性：企业总部所在城市风投基金数量与企业
外生性：不一定满足，该公司所在城市的 VC 数量可能反映该城市的等级水平（与之相关），而不同城市的等级水平对该城市是否存在其他鼓励创新的政策相关，因此该变量可能通过其他路径来影响因变量

RDD

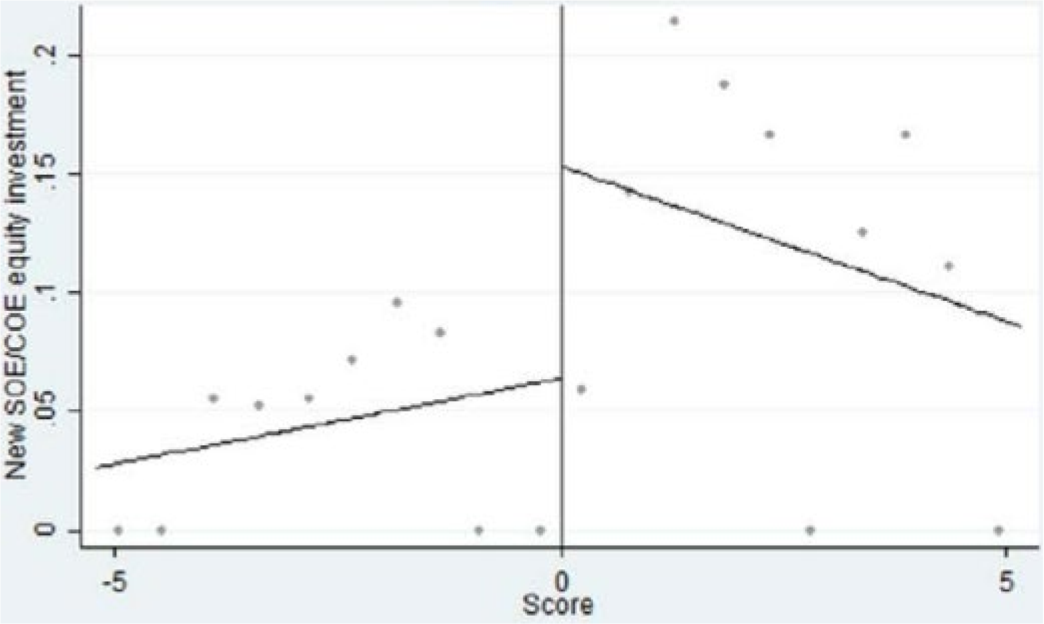
13. (7 points) Write down the regression equation for the following regression discontinuity plot.

The name of the y variable is SOEInvestment (vertical axis), i.e. equity investment the firm received from state-owned investors.

The horizontal axis is a firm's score in the Innofund (科技部创新基金) grant evaluation process.

When Score>=0, the firm successfully gets the grant and certification from the Innofund.

Specify which regression coefficient is the coefficient of interest in the regression equation you write down.



时间序列

14. (14 points, 2 points each) True or False (Time series and panel data)

a) You can control for seasonality in a time series regression with monthly data by including month of the year dummies (Jan, Feb, …) as control variables.

b) The unit root test uses a regression based on the AR(1) model:

$$\mathrm{y_t} = \alpha + \rho y_{t-1} + e_t$$

If you cannot reject the null hypothesis $\rho = 0$, then $\mathrm{y_t}$ has a unit root.

c) When $\mathrm{y_t}$ has a unit root, $\Delta\mathrm{y_t} = \mathrm{y_t} - \mathrm{y_{t-1}}$ will NOT have a unit root.

d) HAC standard errors can deal with regression errors that are correlated over time.

e) Clustered standard errors can deal with regression errors that are correlated over time, but can only be used with panel data.

f) With panel data, you can control for unobservable characteristics that are time-invariant using individual fixed effects. Then, the regression coefficient measures the effect of the x variable, holding fixed the same individual.

g) With panel data, you can control for the effect of macroeconomic and aggregate factors using time fixed effects.

15. (6 points, 2 points each) True or False (Event studies)

a) The for-loop version of the event study R code we studied in class explicitly addresses the issue that events are overlapping (so that abnormal returns may be correlated across events) by using bootstrap standard errors, so that the t-stats produced by that R code is correct.

b) The for-loop version of the event study R code we studied in class did not address the issue that events are overlapping (so that abnormal returns may be correlated across events.) Therefore, the average CAR produced by that R code is incorrect.

c) After defining the event and the window parameters, the event study estimates the normal return and the abnormal returns using data from the event window.

16. (6 points) You are given with the following regression equation:

$$R_{it} = \alpha_i + \beta_i R_{mt} + \sum_{s=\tau_1}^{s=\tau_2} \gamma_s 1\{t = s\} + \epsilon_{it}$$

a) Please explain the meaning and role of each of the variables and the coefficients in this regression equation. (2 points)

b) Please give an expression for CAR(-1,1) under this regression model. (2 points)

c) Provide one advantage and one disadvantage of this regression model, compared to the for-loop version of the event study R code we studied in class. (2 points)