Applied Quantitative Method (II) Final Project

經濟碩一 阮昱眾

1. Introduction: research questions

本研究以1975年美國已婚女性勞動資料，女性的勞動薪資及女性在勞動市場之供給需求一直是勞動研究極受關注的部分，本研究將分析1975年影響已婚女性工作薪資之因素，分別以參數模型與非參數模型分析並透過模型檢定來比較兩者結果差異，最後再以結果較佳的模型做更深入的分析。

2. Econometric Models and Methods:

Parametric Model (OLS) : $y = X'\beta + e$
Nonparametric Model : $y = m(X) + e$, using Local Linear and LSCV

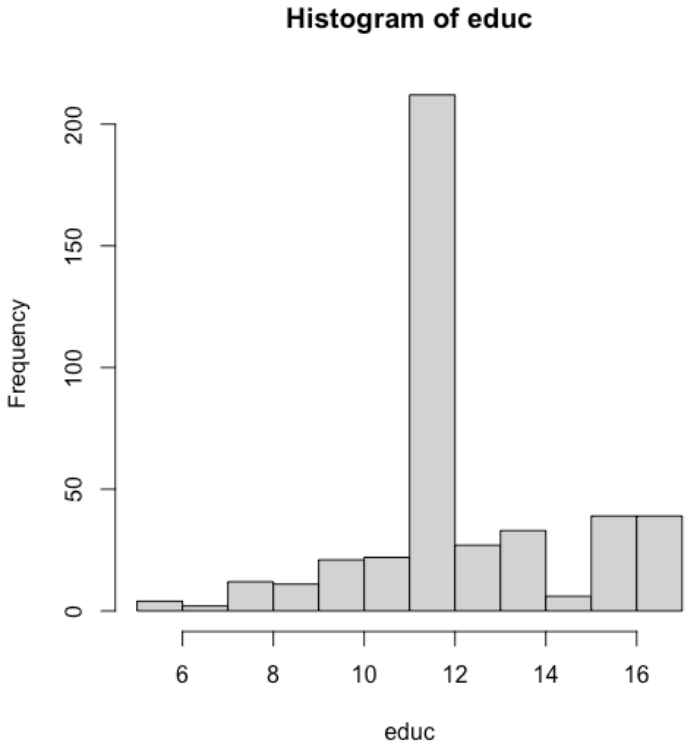
3. Data:

Data sources：Principles of Econometrics, 4th Edition / data files / mroz.dta

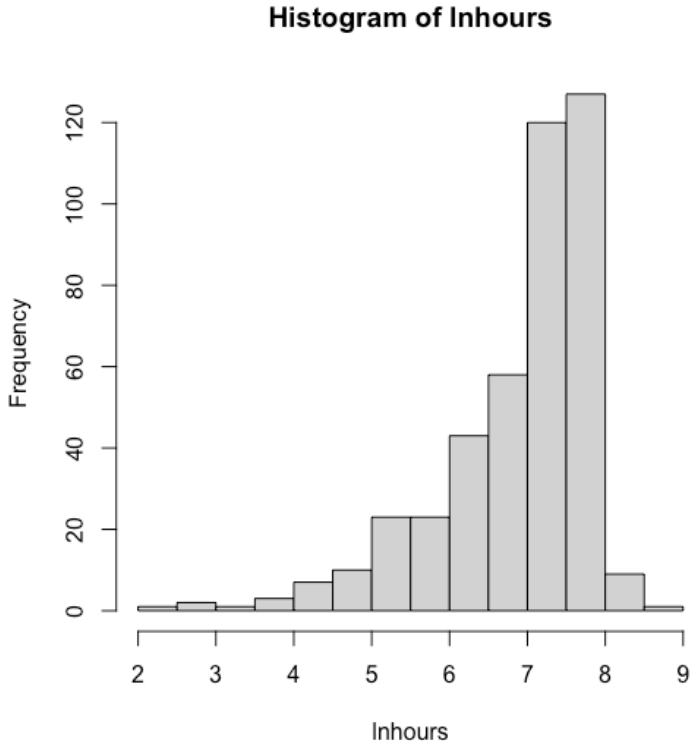
Variable definitions : wage = Married woman's 1975 average hourly earning, in dollars
hours = Married woman's hours worked in 1975
educ = Married woman's education attainment, in years
largecity = dummy variable equal 1 if live in large city, else equal 0

Independent variable : lnwage = log(wage)

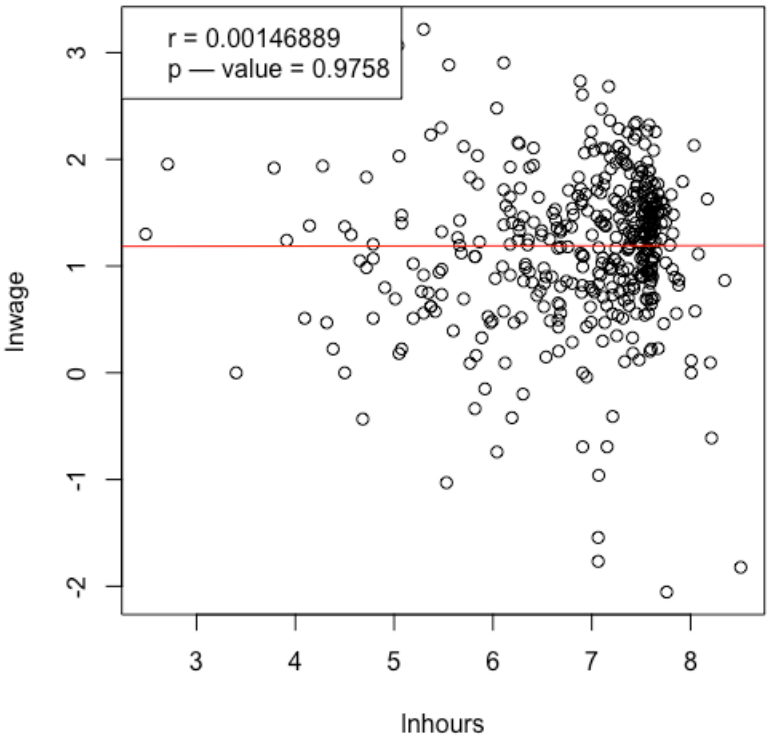
Explanatory variable : 1. lnhours = log(hours)
2. educ
3. largecity



3–1. 分析前透過直方圖可以發現在1975年美國已婚女性的教育年數呈現集中於平均數的分配，相當於高中畢業程度，次高為大學畢業程度，因此為探討教育程度與薪資程度之關係，後續分析會特別著重於高中畢業(12年)與大學(16年)之比較。



3–2. 由直方圖觀察到1975年女性工時呈現左偏分配，多數已婚女性工時集中於高工時部分，因此將著重於第三四分位數的工時。



3-3.由於工作時數與總薪資勢必是呈現正的線性關係，因此本研究著重於探討總工作時數與平均時薪的關係且其是否存在線性關係，於實際分析前，先以簡單的相關係數檢定來觀察時薪與工時的關係，透過相關係數檢定結果兩者不存在線性關係，剛好適合以非參數的框架來分析，並且將搭配其他變數一同加入模型。

4. Empirical Results:

4-1參數模型結果：

```
> summary(par.model)

Call:
lm(formula = lnwage ~ lnhours + educ + largecity, x = TRUE, y = TRUE)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1308 -0.3031  0.0662  0.4004  2.1358

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.41363    0.31495  -1.313    0.190
lnhours      0.02878    0.03417   0.842    0.400
educ         0.10774    0.01467   7.343 1.07e-12 ***
largecity    0.06604    0.06940   0.952    0.342
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6803 on 424 degrees of freedom
Multiple R-squared:  0.1212,    Adjusted R-squared:  0.115
F-statistic: 19.49 on 3 and 424 DF,  p-value: 7.387e-12
```

上圖為以OLS作為估計方法所得之結果，個別係數之不顯著以及較低的R-sqaured都顯示了在這三個解釋變數下所構建的線性參數模型的解釋力不足以及估計結果不準確，在OLS的框架下，僅能得知教育水準可能是一個重要的解釋變數。

4-2非參數模型結果：

```
> hlr.test

Consistent Model Specification Test
Parametric null model: lm(formula = lnwage ~ lnhours + educ + largecity, x =
                          TRUE, y = TRUE)
Number of regressors: 3
Wild Bootstrap (199 replications)

Test Statistic 'Jn': 4.597214   P Value: 0.0050251 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Null of correct specification is rejected at the 1% level
```

4-2.1 Correct Parametric Specification Test (Hsiao, Li and Racine)

透過檢定結果拒絕虛無假設，得到在以這三個解釋變數來解釋已婚女性薪資的框架下以非參數模型的設置優於參數模型。

4-2.2 Irrelevant Regressors Test (Racine, Hart and Li)

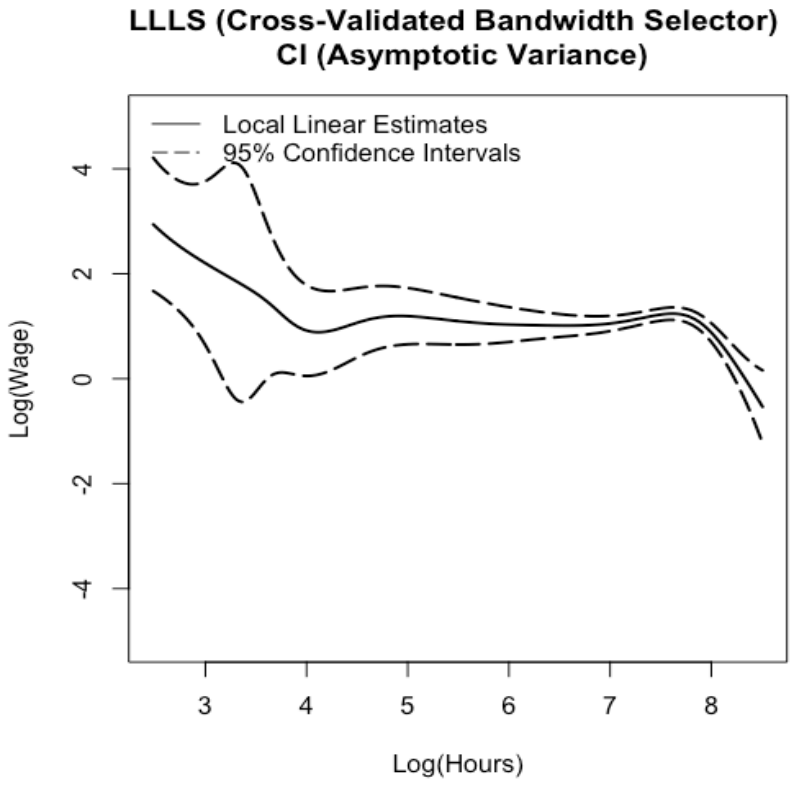
```
> rhl.test

Kernel Regression Significance Test
Type I Test with Wild Bootstrap (199 replications, Pivot = TRUE, joint = FALSE)
Explanatory variables tested for significance:
lnhours (1), ordered(educ) (2), factor(largecity) (3)

                lnhours ordered(educ) factor(largecity)
Bandwidth(s): 0.4302564      0.275633         0.4802569

Individual Significance Tests
P Value:
lnhours           0.015075 *
ordered(educ)     0.010050 *
factor(largecity) 0.055276 .
---
```

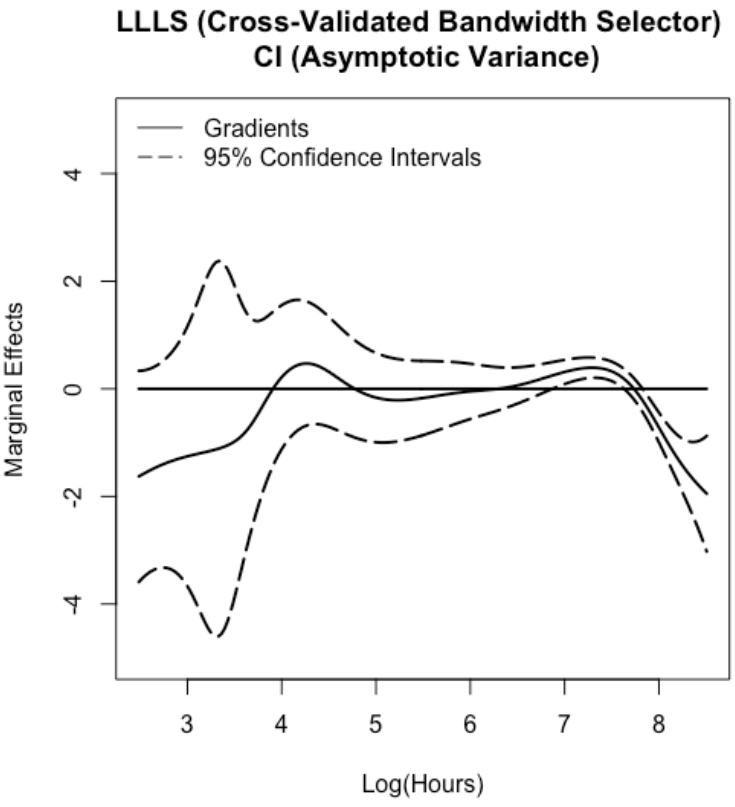
```
> rhl.test_joint

Kernel Regression Significance Test
Type I Test with Wild Bootstrap (199 replications, Pivot = TRUE, joint = TRUE)
Explanatory variables tested for significance:
lnhours (1), ordered(educ) (2), factor(largecity) (3)

                lnhours ordered(educ) factor(largecity)
Bandwidth(s): 0.4302564      0.275633         0.4802569

Joint Significance Test
P Value:  0.01005 *
```

透過解釋變數之重要性檢定，在90%信心水準下，無論是個別檢定還是聯合檢定皆拒絕虛無假設，充分證據顯示以上三個變數都是重要的解釋變數。

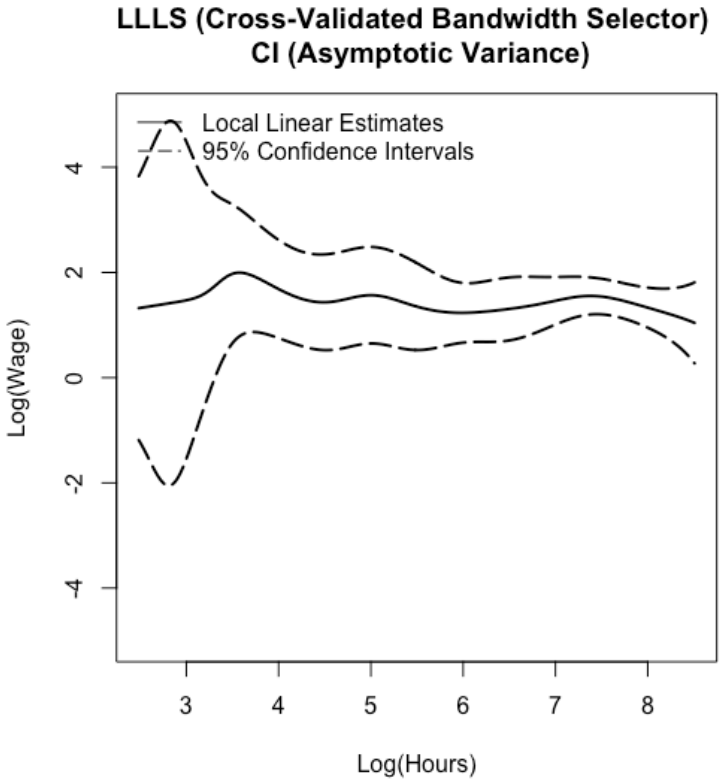
根據上述兩個檢定可得知在此以非參數模型分析優於參數模型，因此選定已非參數模型作為分析模型，以下將深入探討非參數迴歸中個別解釋變數如何影響已婚女性的時薪。

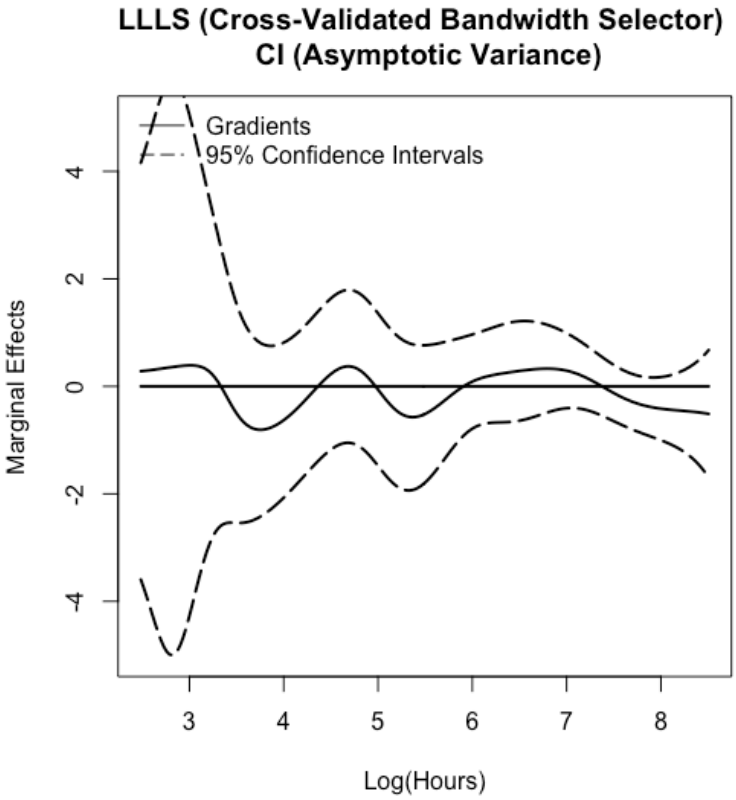
4-2.3.1 工時 (lnhours) 對於 時薪 (lnwage) 之估計結果 (教育年數為12年 大城市)





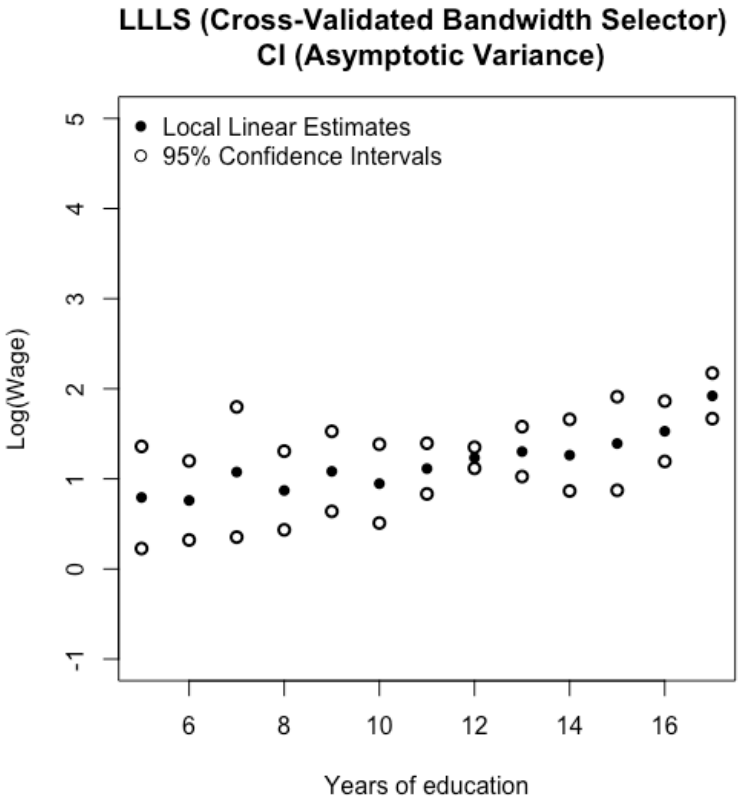
4–2.3.2 工時 (lnhours) 對於 時薪 (lnwage) 之估計結果 (教育年數為16年 大城市)

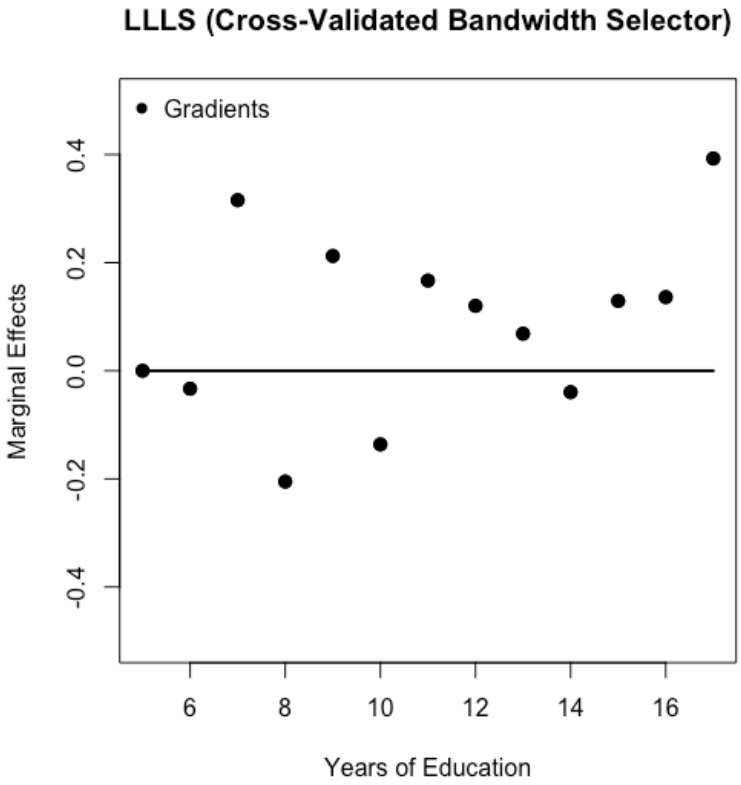




比較4–2.3.1與4–2.3.2結果可得知在大城市，大學畢業的已婚女性有著較穩定的工時、時薪關係，因此受教育的好處也在此體現出來，學歷較高者可以有著更穩定的工時與時薪關係，但無論學歷高低與否工時與時薪的關係並沒有一個固定趨勢或關係該結果也符合現實生活。

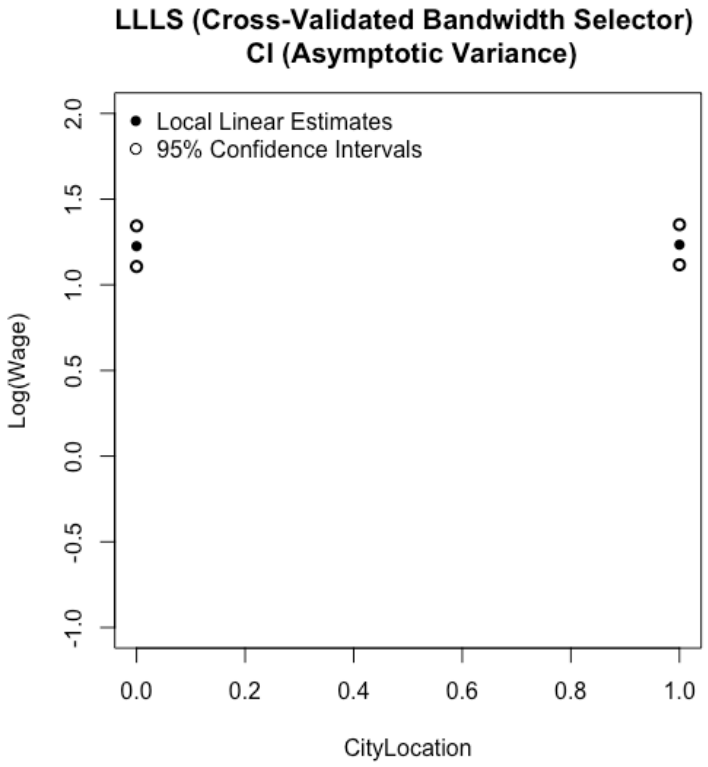
4–2.4.1 教育程度(educ) 對於 時薪 (lnwage) 之估計結果 (第三四分位數工時 大城市)

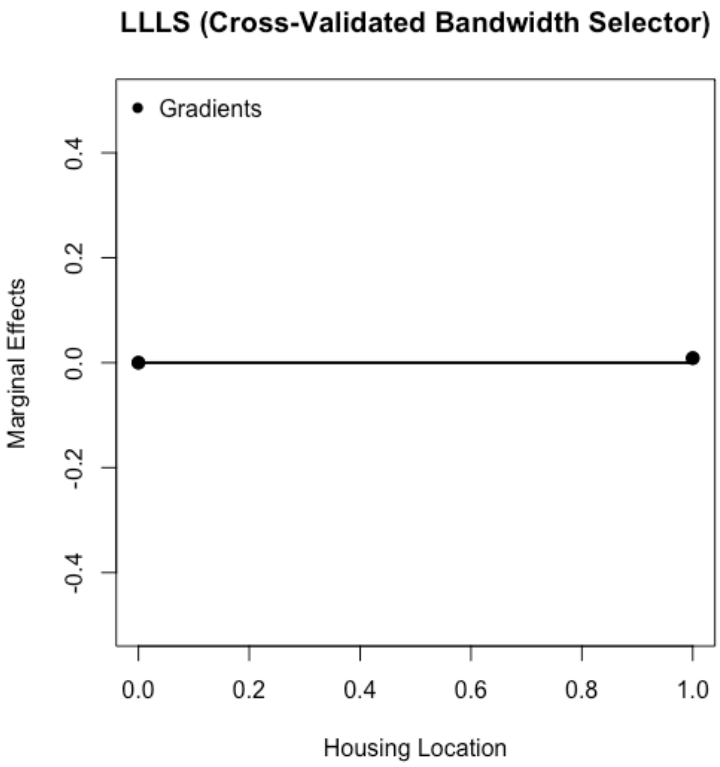




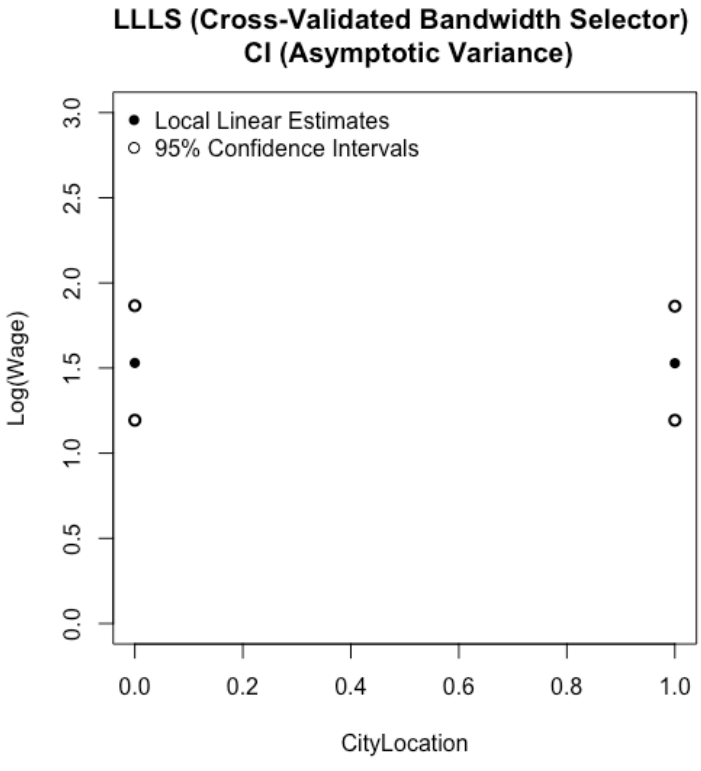
由上述結果可得知在大城市中且固定工時於第三四分位數水準下，教育水準在高中後有較明顯上升的趨勢，高中以下則不一定，可推斷在當時的大城市中，就讀大學可以帶來更多的時薪。

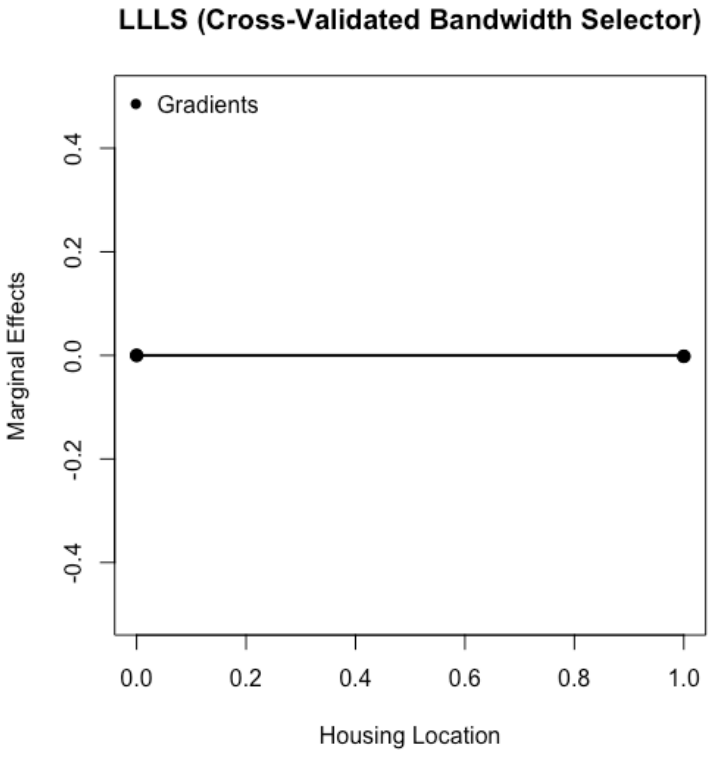
4–2.5.1 城市(largecity) 對於 時薪 (lnwage) 之估計結果 (第三四分位數工時 高中畢業)





4–2.5.2 城市(largecity) 對於 時薪 (lnwage) 之估計結果 (第三四分位數工時 大學畢業)





比較上述兩者結果，可以發現在1975年，無論學歷高低與否，大小城市之間的時薪並沒有顯著的差異，也應對了4-2.2的 Irrelevant Regressors Test結果，城市這個變數的P-value較高，相對較不重要。

5. Conclusion

由2個檢定結果顯示在解釋變數為工時、教育水準以及大城市與否的框架下，以非參數的模型來解釋女性時薪水準是較為適合的。

透過非參數實證分析結果顯示：

- 教育水準在高中以下區間，提升教育水準對於時薪並沒有顯著的影響關係。
- 教育水準在高中以上區間，提升教育水準對於時薪有顯著的正向影響關係。
- 總工時與時薪之關係，在學歷為大學以上有較穩定的關係。
- 總工時與時薪之關係，在學歷為大學以下並沒有穩定的正向或反向關係。
- 在1975年，大城市與小城市之時薪水準相差無異，無論教育水準高低皆得到此結果。

6. References

- 上課筆記與code
- Nonparametric Econometrics : The np package (Hayfield and Racine 2008)
- Principle of econometrics 4th edition (R.Carter Hill, William E. Griffiths, Guay C. Lim)

7. Appendix: codes

```
library("np")
load(file = "/Users/ellisruan/Desktop/npdata/mroz.rda")

mroz <- subset(mroz, lfp==1)
attach(mroz)

lnwage <- log(wage)

lnhours <- log(hours)
educ <- educ
largecity <- largecity
n <- length(lnwage)

# parametric model

par.model <- lm(lnwage ~ lnhours + educ + largecity, x=TRUE, y=TRUE)
```

```
summary(par.model)

# Correct Specification Test (only for LC)

hlr.test <- npcmstest(formula = lnwage ~ lnhours + ordered(educ) + factor(largecity),
                      okertype="wangvanryzin",
                      model=par.model,
                      boot.method="wild",
                      boot.num=199)
hlr.test


# Irrevelant Regressor Test

llls.bw <- npregbw(formula = lnwage ~ lnhours + ordered(educ) + factor(largecity),
                   okertype="wangvanryzin",
                   regtype="ll",
                   bwmethod="cv.ls")
llls.bw

rhl.test <- npsigtest(bws=llls.bw,
                      index=c(1,2,3),
                      boot.method="wild",
                      boot.num=199)

rhl.test_joint <- npsigtest(bws=llls.bw,
                            index=c(1,2,3),
                            joint=TRUE,
                            boot.method="wild",
                            boot.num=199)
rhl.test
rhl.test_joint

# Interval of fitted curve and gradient (Asymptotic)

 ## how do lnhours effect lnwage (educ level at 12 years and locate in large city)
  ### fitted curve

data.train <- data.frame(lnhours,
                         ordered(educ),
                         factor(largecity))

data.eval <- data.frame(lnhours=seq(from=min(lnhours),to=max(lnhours),length=200),
                        educ=ordered(rep(x=12,times=200)),
                        largecity=factor(rep(x=1,times=200)))
```

```
llls <- npreg(bws=llls.bw,
              residuals=TRUE,
              gradients=TRUE,
              data=data.train,
              newdata=data.eval)

fit <- lllls$mean
fit.se <- lllls$merr
fit.lower <- fit-1.96*fit.se
fit.upper <- fit+1.96*fit.se
lnhours.star <- seq(from=min(lnhours),to=max(lnhours),length=200)

plot(lnhours.star,
     fit,type="l",
     col="black",
     xlab="Log(Hours)",
     ylab="Log(Wage)",
     main="LLLS (Cross-Validated Bandwidth Selector) \n CI (Asymptotic Variance)",
     ylim=c(-5,5),
     lwd=2)
lines(lnhours.star,fit.lower,lty=5,col="black",lwd=2)
lines(lnhours.star,fit.upper,lty=5,col="black",lwd=2)
legend("topleft",
       c("Local Linear Estimates","95% Confidence Intervals"),
       col=c("black","black"),
       lty=c(1,5),
       bty="n")

   ### gradient

grad <- lllls$grad[,1]
grad.se <- lllls$gerr[,1]
grad.lower <- grad-1.96*grad.se
grad.upper <- grad+1.96*grad.se

zero <- array(0,dim=c(200,1))

plot(lnhours.star,
     grad,type="l",
     col="black",
     xlab="Log(Hours)",
     ylab="Marginal Effects",
     main="LLLS (Cross-Validated Bandwidth Selector) \n CI (Asymptotic Variance)",
     ylim=c(-5,5),
```

```
    lwd=2)
lines(lnhours.star,grad.lower,lty=5,col="black",lwd=2)
lines(lnhours.star,grad.upper,lty=5,col="black",lwd=2)
lines(lnhours.star,zero,lty=1,col="black",lwd=2)
legend("topleft",
       c("Gradients","95% Confidence Intervals"),
       col=c("black","black"),
       lty=c(1,5),
       bty="n")


  ## How do lnhours effect lnwage (educ level at 16 years and locate in large city)
   ### fitted curve

data.train <- data.frame(lnhours,
                         ordered(educ),
                         factor(largecity))

data.eval <- data.frame(lnhours=seq(from=min(lnhours),to=max(lnhours),length=200),
                        educ=ordered(rep(x=16,times=200)),
                        largecity=factor(rep(x=1,times=200)))

llls <- npreg(bws=llls.bw,
              residuals=TRUE,
              gradients=TRUE,
              data=data.train,
              newdata=data.eval)

fit <- llls$mean
fit.se <- llls$merr
fit.lower <- fit-1.96*fit.se
fit.upper <- fit+1.96*fit.se
lnhours.star <- seq(from=min(lnhours),to=max(lnhours),length=200)

plot(lnhours.star,
     fit,type="l",
     col="black",
     xlab="Log(Hours)",
     ylab="Log(Wage)",
     main="LLLS (Cross-Validated Bandwidth Selector) \n CI (Asymptotic Variance)",
     ylim=c(-5,5),
     lwd=2)
lines(lnhours.star,fit.lower,lty=5,col="black",lwd=2)
lines(lnhours.star,fit.upper,lty=5,col="black",lwd=2)
legend("topleft",
```

```
       c("Local Linear Estimates","95% Confidence Intervals"),
       col=c("black","black"),
       lty=c(1,5),
       bty="n")

### gradient

grad <- llls$grad[,1]
grad.se <- llls$gerr[,1]
grad.lower <- grad-1.96*grad.se
grad.upper <- grad+1.96*grad.se

zero <- array(0,dim=c(200,1))

plot(lnhours.star,
     grad,type="l",
     col="black",
     xlab="Log(Hours)",
     ylab="Marginal Effects",
     main="LLLS (Cross-Validated Bandwidth Selector) \n CI (Asymptotic Variance)",
     ylim=c(-5,5),
     lwd=2)
lines(lnhours.star,grad.lower,lty=5,col="black",lwd=2)
lines(lnhours.star,grad.upper,lty=5,col="black",lwd=2)
lines(lnhours.star,zero,lty=1,col="black",lwd=2)
legend("topleft",
       c("Gradients","95% Confidence Intervals"),
       col=c("black","black"),
       lty=c(1,5),
       bty="n")


  ## how do educ effect lnwage (lnhours level at 0.75 and locate in large city)
   ### fitted curve

data.eval <- data.frame(lnhours=rep(x=uocquantile(x=lnhours,prob=0.75),times=13),
                        educ=ordered(sort(unique(educ))),
                        largecity=factor(rep(x=1,times=13)))

llls <- npreg(bws=llls.bw,
              residuals=TRUE,
              gradients=TRUE,
              data=data.train,
              newdata=data.eval)
```

```
fit <- llls$mean
fit.se <- llls$merr
fit.lower <- fit-1.96*fit.se
fit.upper <- fit+1.96*fit.se

educ.star <- sort(unique(educ))

plot(educ.star,
     fit,type="p",
     pch=16,col="black",
     xlab="Years of education",
     ylab="Log(Wage)",
     main="LLLS (Cross-Validated Bandwidth Selector) \n CI (Asymptotic Variance)",
     ylim=c(-1,5),
     lwd=2)
points(educ.star,fit.lower,col="black",lwd=2)
points(educ.star,fit.upper,col="black",lwd=2)
legend("topleft",
       c("Local Linear Estimates","95% Confidence Intervals"),
       col=c("black","black"),
       pch=c(16,1),
       bty="n")

   ### gradient

grad <- llls$grad[,2]
grad.se <- llls$gerr[,2]
grad.lower <- grad-1.96*grad.se
grad.upper <- grad+1.96*grad.se

grad <- c(0,grad[2:length(grad)])
grad.lower <- c(0,grad.lower[2:length(grad.lower)])
grad.upper <- c(0,grad.upper[2:length(grad.upper)])

zero <- array(0,dim=c(13,1))

plot(educ.star,
     grad,
     type="p",
     pch=16,
     col="black",
     xlab="Years of Education",
     ylab="Marginal Effects",
     main="LLLS (Cross-Validated Bandwidth Selector)",
     ylim=c(-0.5,0.5),
```

```
     lwd=2)
points(educ.star,grad.lower,col="black",lwd=2)
points(educ.star,grad.upper,col="black",lwd=2)
lines(educ.star,zero,lty=1,col="black",lwd=2)
legend("topleft",c("Gradients"),col=c("black"),pch=c(16),bty="n")




  ## how do city effect lnwage (lnhours level at 0.75 and educ level at 12)
   ### fitted curve
data.eval <- data.frame(lnhours=rep(x=uocquantile(x=lnhours,prob=0.75),times=2),
                        educ=ordered(rep(x=12,times=2)),
                        largecity=factor(sort(unique(largecity))))

llls <- npreg(bws=llls.bw,
              residuals=TRUE,
              gradients=TRUE,
              data=data.train,
              newdata=data.eval)

fit <- llls$mean
fit.se <- llls$merr
fit.lower <- fit-1.96*fit.se
fit.upper <- fit+1.96*fit.se

largecity.star <- sort(unique(largecity))

plot(largecity.star,
     fit,
     type="p",
     pch=16,
     col="black",
     xlab="CityLocation",
     ylab="Log(Wage)",
     main="LLLS (Cross-Validated Bandwidth Selector) \n CI (Asymptotic Variance)",
     ylim=c(-1,2),
     lwd=2)
points(largecity.star,fit.lower,col="black",lwd=2)
points(largecity.star,fit.upper,col="black",lwd=2)
legend("topleft",
       c("Local Linear Estimates","95% Confidence Intervals"),
       col=c("black","black"),
       pch=c(16,1),
       bty="n")
```

```
   ### gradient

grad <- llls$grad[,3]
grad.se <- llls$gerr[,3]
grad.lower <- grad-1.96*grad.se
grad.upper <- grad+1.96*grad.se

zero <- array(0,dim=c(2,1))

plot(largecity.star,
     grad,
     type="p",
     pch=16,
     col="black",
     xlab="Housing Location",
     ylab="Marginal Effects",
     main="LLLS (Cross-Validated Bandwidth Selector)",
     ylim=c(-0.5,0.5),lwd=2)
points(largecity.star,grad.lower,col="black",lwd=2)
points(largecity.star,grad.upper,col="black",lwd=2)
lines(largecity.star,zero,lty=1,col="black",lwd=2)
legend("topleft",c("Gradients"),col=c("black"),pch=c(16),bty="n")

  ## how do city effect lnwage (lnhours level at mean and educ level at 16)
   ### fitted curve

data.eval <- data.frame(lnhours=rep(x=uocquantile(x=lnhours,prob=0.75),times=2),
                        educ=ordered(rep(x=16,times=2)),
                        largecity=factor(sort(unique(largecity))))

llls <- npreg(bws=llls.bw,
              residuals=TRUE,
              gradients=TRUE,
              data=data.train,
              newdata=data.eval)

fit <- llls$mean
fit.se <- llls$merr
fit.lower <- fit-1.96*fit.se
fit.upper <- fit+1.96*fit.se

largecity.star <- sort(unique(largecity))

plot(largecity.star,
```

```
    fit,
    type="p",
    pch=16,
    col="black",
    xlab="CityLocation",
    ylab="Log(Wage)",
    main="LLLS (Cross-Validated Bandwidth Selector) \n CI (Asymptotic Variance)",
    ylim=c(0,3),
    lwd=2)
points(largecity.star,fit.lower,col="black",lwd=2)
points(largecity.star,fit.upper,col="black",lwd=2)
legend("topleft",
       c("Local Linear Estimates","95% Confidence Intervals"),
       col=c("black","black"),
       pch=c(16,1),
       bty="n")

  ### gradient
grad <- llls$grad[,3]
grad.se <- llls$gerr[,3]
grad.lower <- grad-1.96*grad.se
grad.upper <- grad+1.96*grad.se

zero <- array(0,dim=c(2,1))

plot(largecity.star,
    grad,
    type="p",
    pch=16,
    col="black",
    xlab="Housing Location",
    ylab="Marginal Effects",
    main="LLLS (Cross-Validated Bandwidth Selector)",
    ylim=c(-0.5,0.5),lwd=2)
points(largecity.star,grad.lower,col="black",lwd=2)
points(largecity.star,grad.upper,col="black",lwd=2)
lines(largecity.star,zero,lty=1,col="black",lwd=2)
legend("topleft",c("Gradients"),col=c("black"),pch=c(16),bty="n")
```