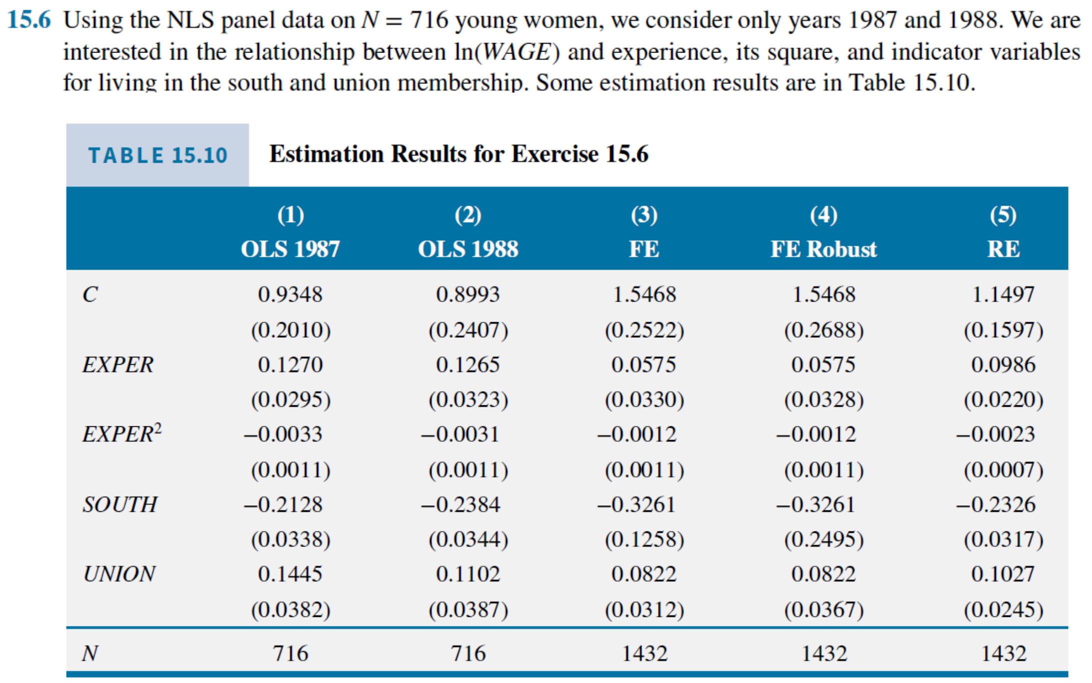
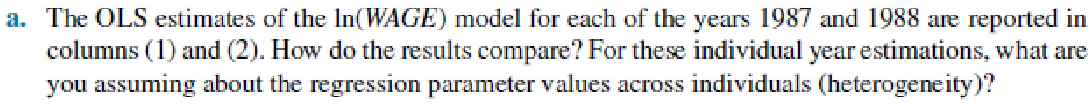


15.6 Using the NLS panel data on $N = 716$ young women, we consider only years 1987 and 1988. We are interested in the relationship between $\ln(WAGE)$ and experience, its square, and indicator variables for living in the south and union membership. Some estimation results are in Table 15.10.

TABLE 15.10 **Estimation Results for Exercise 15.6**

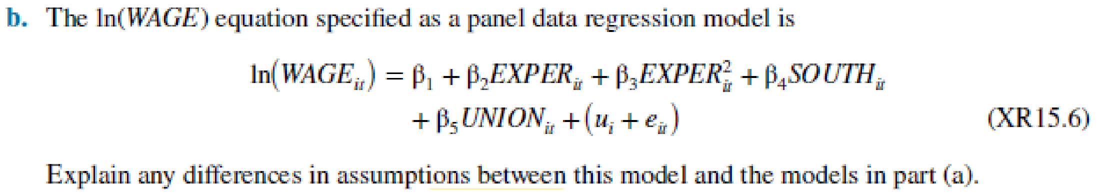
	(1) OLS 1987	(2) OLS 1988	(3) FE	(4) FE Robust	(5) RE
C	0.9348	0.8993	1.5468	1.5468	1.1497
	(0.2010)	(0.2407)	(0.2522)	(0.2688)	(0.1597)
EXPER	0.1270	0.1265	0.0575	0.0575	0.0986
	(0.0295)	(0.0323)	(0.0330)	(0.0328)	(0.0220)
$EXPER^2$	−0.0033	−0.0031	−0.0012	−0.0012	−0.0023
	(0.0011)	(0.0011)	(0.0011)	(0.0011)	(0.0007)
SOUTH	−0.2128	−0.2384	−0.3261	−0.3261	−0.2326
	(0.0338)	(0.0344)	(0.1258)	(0.2495)	(0.0317)
UNION	0.1445	0.1102	0.0822	0.0822	0.1027
	(0.0382)	(0.0387)	(0.0312)	(0.0367)	(0.0245)
N	716	716	1432	1432	1432



a. The OLS estimates of the $\ln(WAGE)$ model for each of the years 1987 and 1988 are reported in columns (1) and (2). How do the results compare? For these individual year estimations, what are you assuming about the regression parameter values across individuals (heterogeneity)?

(a)

1987 年與 1988 年 OLS 估計結果差異非常小，差異約在一個標準誤以內。OLS 模型假設全體個體的參數值（包括截距）都是相同的，背後假設是沒有個體之間的異質性。



b. The $\ln(WAGE)$ equation specified as a panel data regression model is

$$\ln(WAGE_{it}) = \beta_1 + \beta_2 EXPER_{it} + \beta_3 EXPER_{it}^2 + \beta_4 SOUTH_{it} + \beta_5 UNION_{it} + (u_i + e_{it}) \qquad \text{(XR15.6)}$$

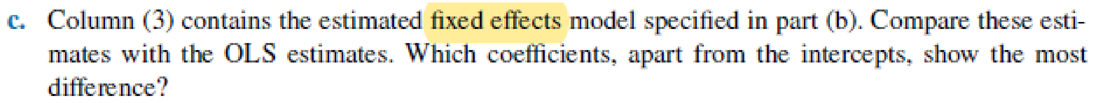
Explain any differences in assumptions between this model and the models in part (a).

(b)

上述方程式加入個體和時間的指標 i,t，以標示可觀察變數。

也加入誤差 ei，是隨個體和時間變化的隨機誤差，稱為「個體特異性誤差成分」。

另一誤差 ui 是 time-invariant，代表個體間的未觀察到異質性，並假設在樣本期間不會改變。



c. Column (3) contains the estimated fixed effects model specified in part (b). Compare these estimates with the OLS estimates. Which coefficients, apart from the intercepts, show the most difference?

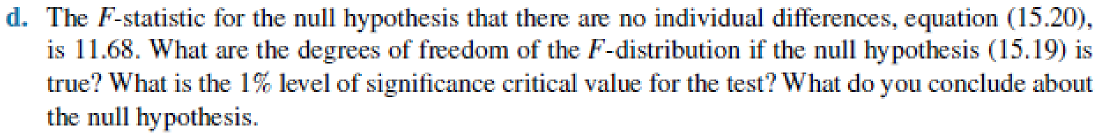
(c)

根據固定效果模型，以下為 95%的信賴區間：

- EXPER：(-0.0085, 0.1235)
- $EXPER^2$：(-0.0034, 0.001)

- SOUTH：(-0.5777, -0.0745)
- UNION：(0.0198, 0.1446)

其中，OLS 估計中 EXPER 的係數不在固定效果的信賴區間內，這表示兩者估計在這變數上有顯著差異。其他變數的 OLS 估計係數都落在固定效果模型的信賴區間內，表示它們之間差異**並不顯著**。



d. The *F*-statistic for the null hypothesis that there are no individual differences, equation (15.20), is 11.68. What are the degrees of freedom of the *F*-distribution if the null hypothesis (15.19) is true? What is the 1% level of significance critical value for the test? What do you conclude about the null hypothesis.

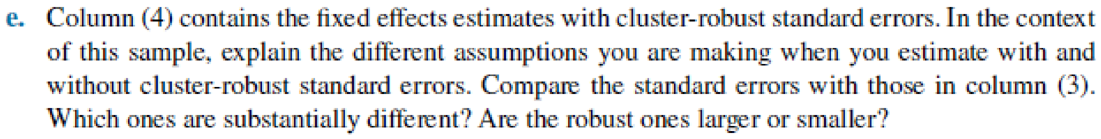
(d)

F 統計量虛無假設為「不存在個體差異」。

F 分布的**分子自由度**為 N−1=716−1=715，因為存在多組假設，如 β1i=β1(i+1)。

分母自由度為 NT−N−(K−1)=1432−716−4=712。

其中 N=716 是個體數量，T=2 是時間點數量，K=5 是模型中參數個數。分母自由度要扣除 N 是因為模型中隱含了許多個體指標變數以控制個體差異。1%顯著水準下的臨界值約為 1.19，因為 F=11.68 大於臨界值，故拒絕虛無假設，表示存在顯著的個體差異。故**固定效果模型**更適合。



e. Column (4) contains the fixed effects estimates with cluster-robust standard errors. In the context of this sample, explain the different assumptions you are making when you estimate with and without cluster-robust standard errors. Compare the standard errors with those in column (3). Which ones are substantially different? Are the robust ones larger or smaller?

(e)

在進行「within」轉換後，隨機誤差變成 $\tilde{e}_{it}=e_{it}-\bar{e}_i$。當個別特有誤差 ei 彼此不相關時，這些經轉換後的隨機誤差會呈現序列相關。

使用 cluster 穩健標準誤，等於允許 ei 在個體和時間上存在異質性（heteroskedasticity），代表誤差的變異大小不一定相同，可能隨著個體或時間改變，不是固定的「同方差」；並且 ei 在時間序列（serial correlation）上可以存在自我相關，代表同一個體隨著時間變化的誤差項可能彼此相關（有「記憶」），不一定相互獨立。

檢視上表，可以發現傳統標準誤與穩健標準誤：

- 對於 EXPER 和 EXPER^2，穩健標準誤幾乎與傳統標準誤相同（比率約為 1 和 1.006）。
- 對於 SOUTH 變數，穩健標準誤約為傳統標準誤的兩倍。
- 對於 UNION 變數，穩健標準誤約為傳統標準誤的 1.18 倍。

15.20 This exercise uses data from the STAR experiment introduced to illustrate fixed and random effects for grouped data. In the STAR experiment, children were randomly assigned within schools into three types of classes: small classes with 13–17 students, regular-sized classes with 22–25 students, and regular-sized classes with a full-time teacher aide to assist the teacher. Student scores on achievement tests were recorded as well as some information about the students, teachers, and schools. Data for the kindergarten classes are contained in the data file *star*.

a. Estimate a regression equation (with no fixed or random effects) where *READSCORE* is related to *SMALL*, *AIDE*, *TCHEXPER*, *BOY*, *WHITE_ASIAN*, and *FREELUNCH*. Discuss the results. Do students perform better in reading when they are in small classes? Does a teacher's aide improve scores? Do the students of more experienced teachers score higher on reading tests? Does the student's sex or race make a difference?

b. Reestimate the model in part (a) with school fixed effects. Compare the results with those in part (a). Have any of your conclusions changed? [*Hint*: specify *SCHID* as the cross-section identifier and *ID* as the "time" identifier.]

c. Test for the significance of the school fixed effects. Under what conditions would we expect the inclusion of significant fixed effects to have little influence on the coefficient estimates of the remaining variables?

d. Reestimate the model in part (a) with school random effects. Compare the results with those from parts (a) and (b). Are there any variables in the equation that might be correlated with the school effects? Use the LM test for the presence of random effects.

e. Using the *t*-test statistic in equation (15.36) and a 5% significance level, test whether there are any significant differences between the fixed effects and random effects estimates of the coefficients on *SMALL*, *AIDE*, *TCHEXPER*, *WHITE_ASIAN*, and *FREELUNCH*. What are the implications of the test outcomes? What happens if we apply the test to the fixed and random effects estimates of the coefficient on *BOY*?

f. Create school-averages of the variables and carry out the Mundlak test for correlation between them and the unobserved heterogeneity.

(a)

$$READSCORE = \beta_0 + \beta_1 SMALL + \beta_2 AIDE + \beta_3 TCHEXPER + \beta_4 BOY + \beta_5 WHITE\_ASIAN + \beta_6 FREELUNCH$$

(i) 若學生在小班而非大班學習，其平均閱讀成績將高出 5.8 分。該係數在 1%的顯著水準下與零有顯著差異。

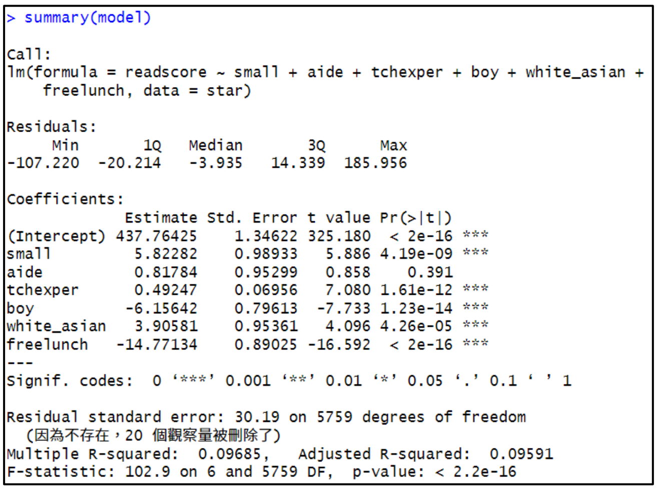
(ii) 是否有助理老師對平均閱讀成績沒有顯著影響。

(iii) 每增加一年教學經驗，平均閱讀成績預估提升 0.49 分，且係數與零顯著不同。

(iv) 男生的平均閱讀成績比女生低 6 分，白人或亞裔學生的平均閱讀成績則比黑人學生高 3.9 分。

(v) 領取免費午餐的學生，其平均閱讀成績預估比未領取者低 14.8 分。

除了 AIDE 變數外，其他所有變數的係數皆達顯著水準。



```
> summary(model)

Call:
lm(formula = readscore ~ small + aide + tchexper + boy + white_asian + 
    freelunch, data = star)

Residuals:
     Min       1Q   Median       3Q      Max 
-107.220  -20.214   -3.935   14.339  185.956 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 437.76425    1.34622 325.180  < 2e-16 ***
small         5.82282    0.98933   5.886 4.19e-09 ***
aide          0.81784    0.95299   0.858    0.391    
tchexper      0.49247    0.06956   7.080 1.61e-12 ***
boy          -6.15642    0.79613  -7.733 1.23e-14 ***
white_asian   3.90581    0.95361   4.096 4.26e-05 ***
freelunch   -14.77134    0.89025 -16.592  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.19 on 5759 degrees of freedom
  (因為不存在，20 個觀察量被刪除了)
Multiple R-squared:  0.09685,	Adjusted R-squared:  0.09591 
F-statistic: 102.9 on 6 and 5759 DF,  p-value: < 2.2e-16
```

(b)

重新估計(a)題的模型，但加入「學校固定效果（school fixed effects）」，控制個體之間不變的特質。

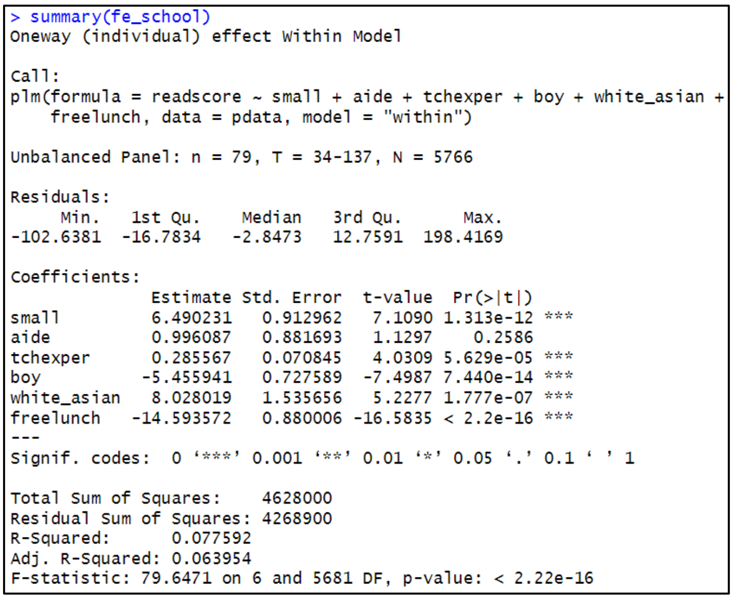
(i) 小班教學使平均閱讀分數提高 6.49 分，比 OLS 模型所估計的略高。

(ii) 教學經驗對平均閱讀分數的估計影響降為每增加一年經驗，分數提高 0.29 分。

(iii) 男生與女生平均閱讀分數的差異估計比 OLS 結果略小，為-5.456。

(iv) 白人或亞裔學生與黑人學生的平均分數差異大約翻倍，達到 8 分左右。

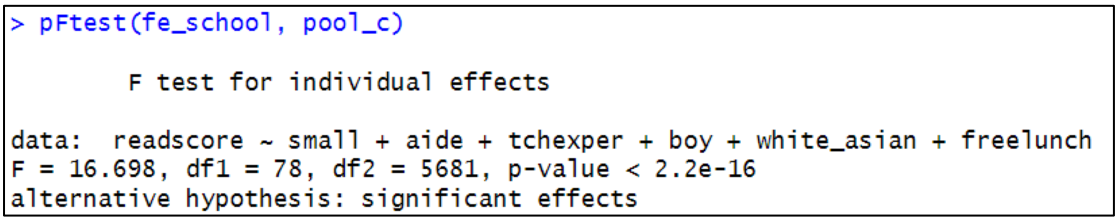
(v) 領取免費午餐的學生對平均閱讀分數，在固定效果模型和 OLS 模型估計出的值差不多。



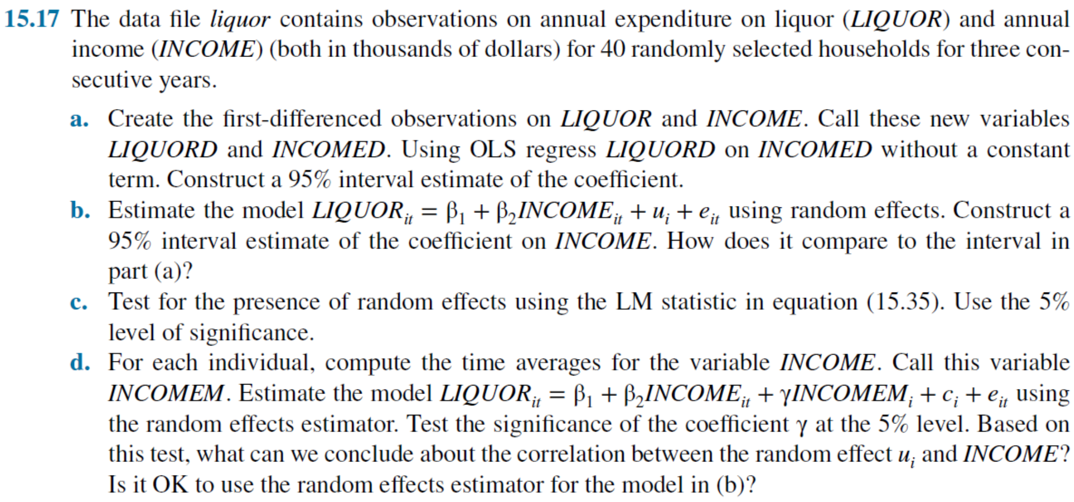
```
> summary(fe_school)
Oneway (individual) effect Within Model

Call:
plm(formula = readscore ~ small + aide + tchexper + boy + white_asian + 
    freelunch, data = pdata, model = "within")

Unbalanced Panel: n = 79, T = 34-137, N = 5766

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max. 
-102.6381  -16.7834   -2.8473   12.7591  198.4169 

Coefficients:
              Estimate Std. Error  t-value  Pr(>|t|)    
small         6.490231   0.912962   7.1090 1.313e-12 ***
aide          0.996087   0.881693   1.1297    0.2586    
tchexper      0.285567   0.070845   4.0309 5.629e-05 ***
boy          -5.455941   0.727589  -7.4987 7.440e-14 ***
white_asian   8.028019   1.535656   5.2277 1.777e-07 ***
freelunch   -14.593572   0.880006 -16.5835 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    4628000
Residual Sum of Squares: 4268900
R-Squared:      0.077592
Adj. R-Squared: 0.063954
F-statistic: 79.6471 on 6 and 5681 DF, p-value: < 2.22e-16
```

(c)

用 **F 檢定**比較含有學校固定效果模型與不含學校固定效果的模型。如果學校虛擬變數與其他解釋變數之間不相關，那麼將其納入或排除對迴歸係數估計的影響將很小。F 檢定統計量為 16.70，而臨界值 $F_{(0.95,78,5681)}$=1.2798。因此拒絕虛無假設（即學校之間沒有顯著差異）。



```
> pFtest(fe_school, pool_c)

        F test for individual effects

data:  readscore ~ small + aide + tchexper + boy + white_asian + freelunch
F = 16.698, df1 = 78, df2 = 5681, p-value < 2.2e-16
alternative hypothesis: significant effects
```



15.17 The data file *liquor* contains observations on annual expenditure on liquor (*LIQUOR*) and annual income (*INCOME*) (both in thousands of dollars) for 40 randomly selected households for three consecutive years.

a. Create the first-differenced observations on *LIQUOR* and *INCOME*. Call these new variables *LIQUORD* and *INCOMED*. Using OLS regress *LIQUORD* on *INCOMED* without a constant term. Construct a 95% interval estimate of the coefficient.

b. Estimate the model $LIQUOR_{it} = \beta_1 + \beta_2 INCOME_{it} + u_i + e_{it}$ using random effects. Construct a 95% interval estimate of the coefficient on *INCOME*. How does it compare to the interval in part (a)?

c. Test for the presence of random effects using the LM statistic in equation (15.35). Use the 5% level of significance.

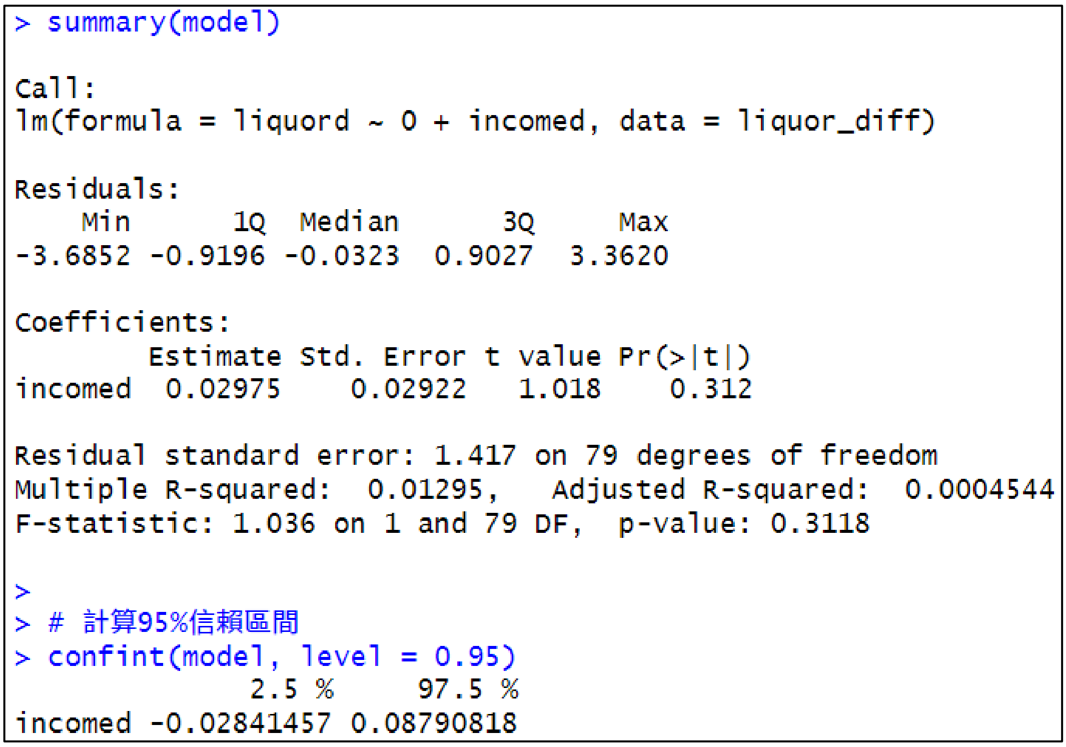
d. For each individual, compute the time averages for the variable *INCOME*. Call this variable *INCOMEM*. Estimate the model $LIQUOR_{it} = \beta_1 + \beta_2 INCOME_{it} + \gamma INCOMEM_i + c_i + e_{it}$ using the random effects estimator. Test the significance of the coefficient γ at the 5% level. Based on this test, what can we conclude about the correlation between the random effect u_i and *INCOME*? Is it OK to use the random effects estimator for the model in (b)?

(a)

$$\widehat{LIQUORD} = 0.02975 \cdot \mathrm{INCOMED}$$

INCOMED在 95%信賴區間估計為[-0.028416,0.08791]。

此區間包含零，因此沒有足夠證據反駁「收入不影響酒類支出」的假設。



```
> summary(model)

Call:
lm(formula = liquord ~ 0 + incomed, data = liquor_diff)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.6852 -0.9196 -0.0323  0.9027  3.3620 

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
incomed  0.02975    0.02922   1.018    0.312

Residual standard error: 1.417 on 79 degrees of freedom
Multiple R-squared:  0.01295,	Adjusted R-squared:  0.0004544 
F-statistic: 1.036 on 1 and 79 DF,  p-value: 0.3118

> 
> # 計算95%信賴區間
> confint(model, level = 0.95)
              2.5 %     97.5 %
incomed -0.02841457 0.08790818
```