Data Modelling & Analytics
Individual Assignment 3

Load and prepare the data

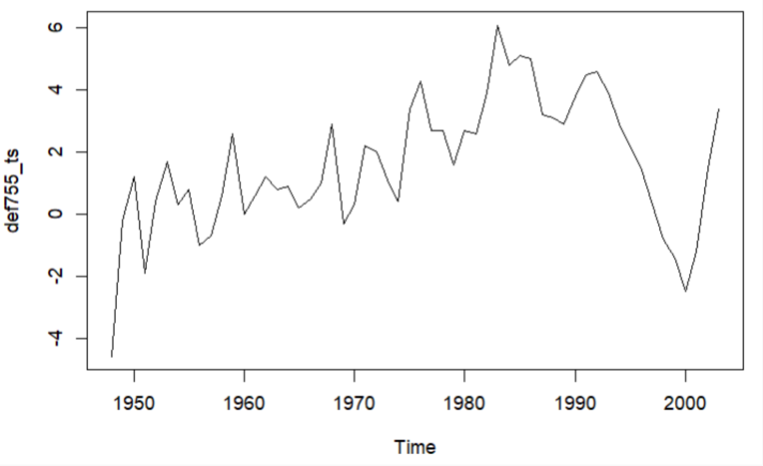
```
> # Load the data set 'intdef'
> data(intdef, package='wooldridge')
> intdef755 <- intdef
> # Rename variables
> intdef755 <- rename(intdef755, year755 = year, i3755 = i3, inf755 = inf,
def755 = def)
>
> # Load the data set 'Gasoline'
> data(Gasoline, package = 'plm')
> Gasoline755 <- Gasoline
> # Rename variables
> Gasoline755 <- rename(Gasoline755, country755 = country, year755 = year,
lgaspcar755 = lgaspcar, lincomep755 = lincomep, lrpmg755 = lrpmg, lcarpcap7
55 = lcarpcap)
```

Part A: Time series

For part A, please use the data "intdefxxx".

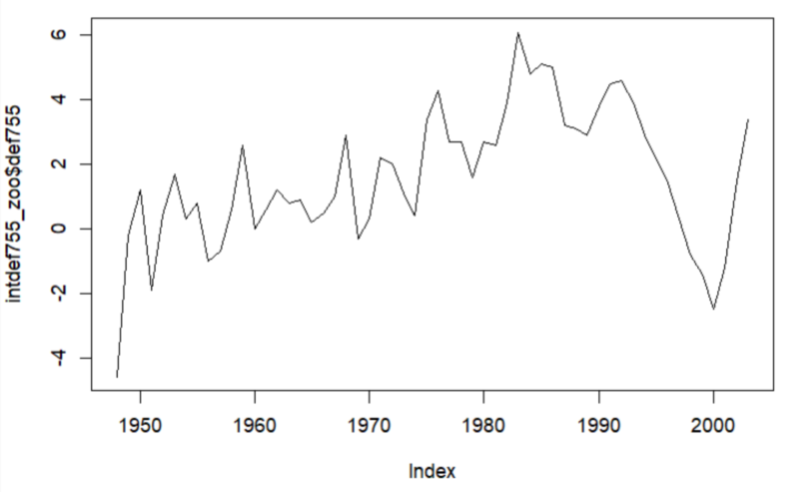
1. [0.8 points] Define variable defxxx from the data "intdefxxx" as a ts object and plot it.

```
> # Define variable def755 from the data "intdef755" as a ts object
> def755_ts <- ts(intdef755$def755, start = 1948)
>
> # Plot time series
> plot(def755_ts)
```



2. [1 point] Define “intdefxxx” as a zoo object containing all data. Make a time series plot of variable defxxx. Compare this plot with the one that you make for question 1. Are the two plots the same?

```
> # Define a "zoo" object containing all data
> intdef755_zoo <- zoo(intdef755, order.by = intdef755$year755)
>
> # Time series plot of variable def755
> plot(intdef755_zoo$def755)
```



The two plots are the same. Only the name of the two axes are different.

3. [2.5 points] Use the command “lm” to fit a finite distributed lag (FDL) model of order 3:

- Dependent variable: i3xxx
- Independent variables: infxxx, defxxx, defxxx lagged by one time unit, defxxx lagged by two time units, defxxx lagged by three time units

Compare this FDL model of order 3 with the static time series model in slide 5 of video 5.1. Which model should you choose, this FDL model of order 3 or the model in slide 5? Please provide an explanation about how you make the decision.

```
> # Generate the lags of variable def755 manually
> # Create a variable for def755-1: lagged by one time unit
> intdef755['Ldef755'] <- Lag(intdef755$def755, +1)
> # Create a variable for price_t-2: lagged by two time units
> intdef755['Ldef7552'] <- Lag(intdef755$def755, +2)
> # Create a variable for price_t-3: lagged by three time units
> intdef755['Ldef7553'] <- Lag(intdef755$def755, +3)
>
> # A FDL model: Run a linear regression using lm command
> fdl_lm <- lm(i3755 ~ inf755 + def755 + Ldef755 + Ldef7552 + Ldef7553, dat
a=intdef755)
> summary(fdl_lm)

Call:
lm(formula = i3755 ~ inf755 + def755 + Ldef755 + Ldef7552 + Ldef7553,
    data = intdef755)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9696 -0.9565 -0.2407  0.7665  4.4336

Coefficients:
```

```
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.700041   0.409620   4.150 0.000138 ***
inf755       0.601733   0.078526   7.663 8.15e-10 ***
def755       0.088537   0.177261   0.499 0.619778    
Ldef755      0.094998   0.220365   0.431 0.668371    
Ldef7552    -0.001538   0.219013  -0.007 0.994427    
Ldef7553     0.449741   0.164359   2.736 0.008741 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.63 on 47 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.6932,	Adjusted R-squared:  0.6606 
F-statistic: 21.24 on 5 and 47 DF,  p-value: 4.787e-11
```

We should choose this FDL model of order 3 instead of the static time series model in slide 5. Although the IV 'inf' and 'def' are significant in the static time series model, if we want to do calculations in time series we will have to specify in R that the data type we are using is time series. The DV is i3(3 month T-bill rate). So an FDL model of order 3 would help us find the long-run propensity (LRP), which measures the cumulative effect of a change 'inf' and 'def' on i3 over time.

4. [2 points] Use the command "dynlm" to fit a FDL model of order 3 with the same dependent variable and independent variables as those in question 3:

- Dependent variable: i3xxx
- Independent variables: infxxx, defxxx, defxxx lagged by one time unit, defxxx lagged by two time units, defxxx lagged by three time units

Use the command "stargazer" to make a table of regression results in questions 3 and 4:

- column (1) shows the result that you get in question 3 using the command "lm"
- column (2) shows the result that you get in this question using the command "dynlm"

Does command "dynlm" give you the same result as command "lm"?

```
> # Define yearly time series as a ts object
> intdef755_ts <- ts(intdef755, start=1948)
> 
> # A FDL model: Run a linear regression using dynlm command
> fdl_dyn <- dynlm(i3755 ~ inf755 + def755 + L(def755) + L(def755, 2) + L(d
ef755, 3),
+                  data=intdef755_ts)
> summary(fdl_dyn)

Time series regression with "ts" data:
Start = 1951, End = 2003

Call:
dynlm(formula = i3755 ~ inf755 + def755 + L(def755) + L(def755, 
    2) + L(def755, 3), data = intdef755_ts)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9696 -0.9565 -0.2407  0.7665  4.4336 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.700041   0.409620   4.150 0.000138 ***
```

```
inf755        0.601733   0.078526   7.663 8.15e-10 ***
def755        0.088537   0.177261   0.499 0.619778
L(def755)     0.094998   0.220365   0.431 0.668371
L(def755, 2) -0.001538   0.219013  -0.007 0.994427
L(def755, 3)  0.449741   0.164359   2.736 0.008741 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.63 on 47 degrees of freedom
Multiple R-squared:  0.6932,   Adjusted R-squared:  0.6606
F-statistic: 21.24 on 5 and 47 DF,  p-value: 4.787e-11

> #Table of regression result
> stargazer(fdl_lm, fdl_dyn, type="text")
```

```
===========================================================
                                 Dependent variable:
                            ----------------------------
                                       i3755
                                 OLS          dynamic
                                              linear
                                 (1)            (2)
-----------------------------------------------------------
inf755                         0.602***      0.602***
                               (0.079)        (0.079)

def755                          0.089          0.089
                               (0.177)        (0.177)

Ldef755                         0.095
                               (0.220)

Ldef7552                        -0.002
                               (0.219)

Ldef7553                       0.450***
                               (0.164)

L(def755)                                      0.095
                                              (0.220)

L(def755, 2)                                  -0.002
                                              (0.219)

L(def755, 3)                                  0.450***
                                              (0.164)

Constant                       1.700***      1.700***
                               (0.410)        (0.410)

-----------------------------------------------------------
```

```
Observations                          53            53
R2                                  0.693         0.693
Adjusted R2                         0.661         0.661
Residual Std. Error (df = 47)       1.630         1.630
F Statistic (df = 5; 47)          21.241***     21.241***
==========================================================
Note:                            *p<0.1; **p<0.05; ***p<0.01
```

The estimations from command "dynlm" give the same result as command "lm"

5. [1.5 points] Test whether you should add a time trend in the above FDL model of order 3. Specifically, you compare the model in question 4 with the following FDL model:

- dependent variable: i3xxx
- independent variables: infxxx, defxxx, defxxx lagged by one time unit, defxxx lagged by two time units, defxxx lagged by three time units, and time trend

Which model should you choose, the FDL model with a time trend or without a time trend? Please provide an explanation about how you make the decision.

```
> #FDL model with time trend
> fdl_dyn_t <- dynlm(i3755 ~ inf755 + def755 + L(def755) + L(def755, 2) + L
(def755, 3) + trend(intdef755_ts),
+                    data=intdef755_ts)
> summary(fdl_dyn_t)

Time series regression with "ts" data:
Start = 1951, End = 2003

Call:
dynlm(formula = i3755 ~ inf755 + def755 + L(def755) + L(def755,
    2) + L(def755, 3) + trend(intdef755_ts), data = intdef755_ts)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1098 -0.9643 -0.0765  0.7953  4.4534

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.32679    0.56468   2.350   0.0231 *
inf755               0.59936    0.07863   7.623 1.07e-09 ***
def755               0.07619    0.17787   0.428   0.6704
L(def755)            0.09116    0.22058   0.413   0.6813
L(def755, 2)         0.00911    0.21947   0.042   0.9671
L(def755, 3)         0.40083    0.17218   2.328   0.0244 *
trend(intdef755_ts)  0.01574    0.01638   0.961   0.3415
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.631 on 46 degrees of freedom
Multiple R-squared:  0.6993,   Adjusted R-squared:   0.66
F-statistic: 17.83 on 6 and 46 DF,  p-value: 1.515e-10

>
> #Compare the two models using anova
> anova(fdl_dyn, fdl_dyn_t)
Analysis of Variance Table
```

```
Model 1: i3755 ~ inf755 + def755 + L(def755) + L(def755, 2) + L(def755,
    3)
Model 2: i3755 ~ inf755 + def755 + L(def755) + L(def755, 2) + L(def755,
    3) + trend(intdef755_ts)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     47 124.85
2     46 122.39  1    2.4579 0.9238 0.3415
```

p-value=0.3415 > 0.05, the null hypothesis cannot be rejected. So we should choose the restricted model which does not include time trend.

6. [2 points] Based on your chosen model in question 5, calculate the estimated value of long-run propensity (LRP) of variable defxxx. Test whether this LRP is significant. Interpret LRP.

```
> # Calculate the estimated value of long run propensity (LRP) of variable
def755
> b <- coef(fdl_dyn)
> b["def755"]+b["L(def755)"]+b["L(def755, 2)"]+b["L(def755, 3)"]
   def755
0.6317375
>
> # Test whether LRP is significant
> # F test H0: LRP=0
> linearHypothesis(fdl_dyn,"def755 + L(def755) + L(def755, 2) + L(def755, 3
) = 0")
Linear hypothesis test

Hypothesis:
def755  + L(def755)  + L(def755, 2)  + L(def755, 3)  = 0

Model 1: restricted model
Model 2: i3755 ~ inf755 + def755 + L(def755) + L(def755, 2) + L(def755,
    3)

  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1     48 182.20
2     47 124.85  1    57.351 21.59 2.749e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The estimated long run propensity (LRP) is 0.6317375
- LRP is significant (F test, p value = 2.749e-05 < 0.05)
- Interpretation: As for the effect of def755 (*deficit*) on i3755 (*3 month T-bill rate*): The estimated long-run propensity (LRP) is 0.6317 and it is significant (F test, p value = 2.749e-05 < 0.05). This means, if the (*deficit*) increases one amount unit, then, after three time units, the (*3 month T-bill rate*) will eventually increase by 0.6317.

Part B: Panel data

For part B, please use the data "Gasolinexxx".

1. [0.2 points] Which variables are the entity index and time index of the panel data "Gasolinexxx"?

```
> head(Gasoline755)
  country755 year755 lgaspcar755 lincomep755   lrpmg755 lcarpcap755
1    AUSTRIA    1960    4.173244   -6.474277 -0.3345476   -9.766840
2    AUSTRIA    1961    4.100989   -6.426006 -0.3513276   -9.608622
3    AUSTRIA    1962    4.073177   -6.407308 -0.3795177   -9.457257
4    AUSTRIA    1963    4.059509   -6.370679 -0.4142514   -9.343155
5    AUSTRIA    1964    4.037689   -6.322247 -0.4453354   -9.237739
6    AUSTRIA    1965    4.033983   -6.294668 -0.4970607   -9.123903
```

In the panel data Gasoline755, entity index is country755 and time index is year755

2. [1 point] Create a new variable (a new column) called "m_lincomepxxx" in the data "Gasolinexxx" such that, for every entity, the value of the variable "m_lincomepxxx" is the mean of lincomepxxx across different years.

```
> # For every country755, calculate the mean of gross lincomep755 across di
fferent years
> d1 <- aggregate(Gasoline755$lincomep755, list(Gasoline755$country755), FU
N = mean)
>
> # Rename the columns of the data frame d1
> colnames(d1) <- c("country755", "m_lincomep755")
>
> # Combine the data frame d1 with the original data set Gasoline755
> Gasoline755 <- left_join(d1, Gasoline755)
Joining, by = "country755"
```

3. [1 point] Define the data "Gasolinexxx" as a panel data frame in R. What are panel dimensions? What are the meanings of the numbers n, T, and N in RStudio console output? What are the time-invariant and individual-invariant variables of this panel?

```
> # Define panel data frame
> Gasoline755_pdata <- pdata.frame(Gasoline755, index = c("country755", "ye
ar755"))
> # What are panel dimensions?
> pdim(Gasoline755_pdata)
Balanced Panel: n = 18, T = 19, N = 342
> # What time-invariant and individual-invariant variables?
> pvar(Gasoline755_pdata)
no time variation:       country755 m_lincomep755
no individual variation: year755
```

- n = 18 => There are 18 countries
- T = 19 => Every individual has 19 rows (19 years data) in the panel
- N = 342 => There are 342 observations
- time-invariant variables: 'country755' and 'm_lincomep755'
- individual-invariant variables: year755

4. [2.5 points] Make the following two plots:
  - Plot 1: A plot of dependent variable lgaspcarxxx and yearxxx for every entity

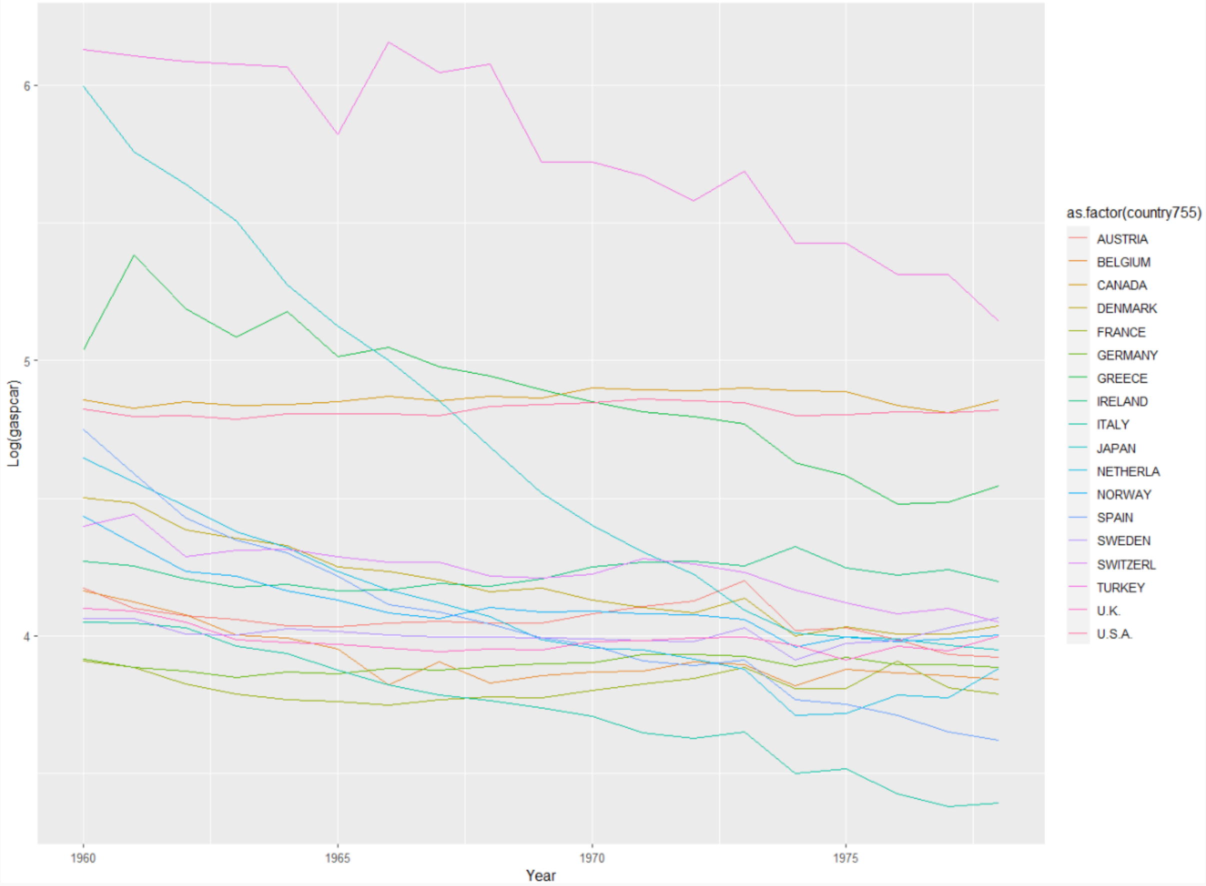
Hints: The format of the plot should be similar to the example plot presented in the videos. In the plot, there should be a line for every entity. Make sure that the labels of your plot are clear to see.

– Plot 2: A plot for fixed effects: Heterogeneity across entities

  Hints: The dependent variable is still lgaspcarxxx. The format of the plot should be similar to the example plot presented in the videos. In the plot, there should be black points and a red line. Make sure that the labels of your plot are clear to see.

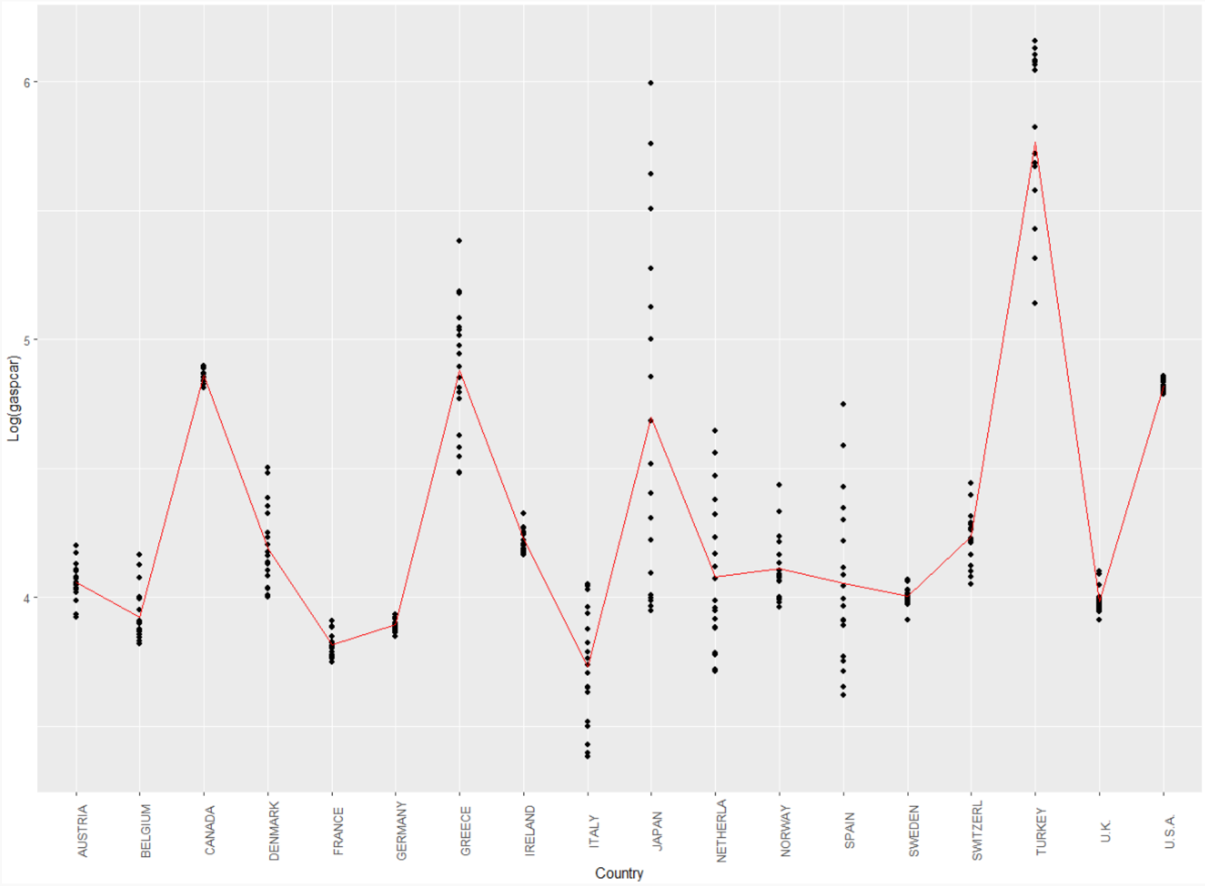
What do you observe from the two plots? Please describe the two plots. Based on the two plots, do you think whether the individual fixed effects should be taken into consideration or not? Please explain why you think the individual fixed effects should or shouldn't be taken into consideration.

```
> #Plot 1: A plot of dependent variable lgaspcar755 and year755 for every c
ountry
> ggplot(data = Gasoline755, aes(x = year755, y = lgaspcar755)) +
+   geom_line(aes(colour = as.factor(country755))) +
+   labs(x = "Year",  y = "Log(gaspcar)")
```




```
> #Plot 2: A plot for fixed effects: Heterogeneity across entities
> #For every country, calculate the mean of gross lgaspcar755 across differ
ent country
> d1 <- aggregate(Gasoline755$lgaspcar755, list(Gasoline755$country755), FU
N = mean)
>
> #Rename the columns of the data frame d1
> colnames(d1) <- c("country755", "m_lgaspcar755")
>
> #Combine the data frame d1 with the original data set Gasoline755
```

```
> d2 <- left_join(d1, Gasoline755)
Joining, by = "country755"
>
> #Use ggplot to make a plot
> ggplot(data = d2, aes(x = as.character(country755), y = lgaspcar755)) +
+   scale_x_discrete(labels = as.character(Gasoline755$country755),
+                    breaks = Gasoline755$country755) +
+   theme(axis.text.x = element_text(angle = 90)) +
+   geom_point() +
+   geom_line(aes(x = as.numeric(country755), y = m_lgaspcar755), col = "re
d") +
+   labs(x = "Country", y = "Log(gaspcar)")
```



Plot 1: The x axis is *year*, the y axis is *logarithm of gasoline use per car*. And each line represents one country. In general, the gasoline use per car decreases over time. However, Turkey, Greece, Canada, and the USA have a relatively higher gasoline user per car than other countries. Also, the line for Turkey, Japan, Spain, and Netherland has a steeper downward slope than others'. Therefore, we may need to consider the heterogeneity across countries when we fit a panel data model (1).

Plot 2: We plot the *logarithm of gasoline use per car* across different *countries.* We can see that Canada, Greece, Japan, Turkey and the USA have a higher mean of *logarithm of gasoline use per car* than the other countries. The entity fixed effects aim to account for characteristics of countries that do not change over time. Thus, we probably need to include individual fixed effects in our model to account for the heterogeneity across countries (2).

From (1) and (2), I believe individual fixed effects should be taken into consideration

5. [2.5 points] Use the command "lm" to fit a least squares dummy variable (LSDV) model that considers individual fixed effects:

- Dependent variable: lgaspcarxxx
- Independent variables: lincomepxxx, lrpmgxxx, lcarpcapxxx, etc.

Use the command "plm" to estimate a FE estimator (or within estimator) that considers individual fixed effects:

- Dependent variable: lgaspcarxxx
- Independent variables: lincomepxxx, lrpmgxxx and lcarpcapxxx

Use the command "stargazer" to make a table of the results in this question:

- column (1) shows the result of the LSDV model
- column (2) shows the result of the FE estimator
- only include variables lincomepxxx, lrpmgxxx, and lcarpcapxxx in the table

Compare the result of the LSDV model and the result of the FE estimator. Do you get the same estimated coefficients and standard errors of variables lincomepxxx, lrpmgxxx and lcarpcapxxx?

```
# Least squares dummy variable (LSDV) model
> fe_lsdv <- lm(lgaspcar755 ~ lincomep755 + lrpmg755 + lcarpcap755 + factor
(country755), data = Gasoline755)
> summary(fe_lsdv)

Call:
lm(formula = lgaspcar755 ~ lincomep755 + lrpmg755 + lcarpcap755 +
    factor(country755), data = Gasoline755)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37877 -0.03976  0.00465  0.04541  0.36286

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 2.28586    0.22832  10.011  < 2e-16 ***
lincomep755                 0.66225    0.07339   9.024  < 2e-16 ***
lrpmg755                   -0.32170    0.04410  -7.295 2.35e-12 ***
lcarpcap755                -0.64048    0.02968 -21.580  < 2e-16 ***
factor(country755)BELGIUM  -0.12030    0.03415  -3.523 0.000489 ***
factor(country755)CANADA    0.75598    0.04075  18.554  < 2e-16 ***
factor(country755)DENMARK   0.10360    0.03660   2.830 0.004944 **
factor(country755)FRANCE   -0.08108    0.03356  -2.416 0.016256 *
factor(country755)GERMANY  -0.13599    0.03188  -4.266 2.63e-05 ***
factor(country755)GREECE    0.05125    0.04153   1.234 0.218049
factor(country755)IRELAND   0.30647    0.03529   8.683  < 2e-16 ***
factor(country755)ITALY    -0.05331    0.03711  -1.436 0.151868
factor(country755)JAPAN     0.09007    0.03861   2.333 0.020262 *
factor(country755)NETHERLA -0.05106    0.03358  -1.521 0.129280
factor(country755)NORWAY   -0.06916    0.04041  -1.711 0.087967 .
factor(country755)SPAIN    -0.60408    0.09122  -6.622 1.49e-10 ***
factor(country755)SWEDEN    0.74049    0.18008   4.112 4.99e-05 ***
factor(country755)SWITZERL  0.11665    0.03471   3.360 0.000872 ***
factor(country755)TURKEY    0.22413    0.04764   4.704 3.79e-06 ***
factor(country755)U.K.      0.05959    0.03019   1.974 0.049237 *
factor(country755)U.S.A.    0.76940    0.04458  17.260  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.09233 on 321 degrees of freedom
Multiple R-squared:  0.9734,   Adjusted R-squared:  0.9717
F-statistic: 586.6 on 20 and 321 DF,  p-value: < 2.2e-16

> # Fixed effects (FE) estimator (or within estimator)
> fe_plm <- plm(lgaspcar755 ~ lincomep755 + lrpmg755 + lcarpcap755, data =
Gasoline755,
+                index = c("country755", "year755"), effect = "individual",
+                model = "within")
> summary(fe_plm)
Oneway (individual) effect Within Model

Call:
plm(formula = lgaspcar755 ~ lincomep755 + lrpmg755 + lcarpcap755,
    data = Gasoline755, effect = "individual", model = "within",
    index = c("country755", "year755"))

Balanced Panel: n = 18, T = 19, N = 342

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-0.378774 -0.039758  0.004650  0.045412  0.362856

Coefficients:
             Estimate Std. Error  t-value  Pr(>|t|)
lincomep755  0.662250   0.073386   9.0242 < 2.2e-16 ***
lrpmg755    -0.321702   0.044099  -7.2950 2.355e-12 ***
lcarpcap755 -0.640483   0.029679 -21.5804 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    17.061
Residual Sum of Squares: 2.7365
R-Squared:      0.8396
Adj. R-Squared: 0.82961
F-statistic: 560.093 on 3 and 321 DF, p-value: < 2.22e-16

> #stargazer
> stargazer(fe_lsdv,fe_plm,
+           type="text",
+           column.labels=c("LSDV", "FE"),
+           keep=c("lincomep755", "lrpmg755", "lcarpcap755"))
```

```
======================================================================
                              Dependent variable:
                    --------------------------------------------------
                                   lgaspcar755
                           OLS                      panel
                                                    linear
                          LSDV                       FE
                           (1)                       (2)
----------------------------------------------------------------------
lincomep755             0.662***                  0.662***
                        (0.073)                   (0.073)


lrpmg755               -0.322***                 -0.322***
                        (0.044)                   (0.044)
```

```
lcarpcap755                  -0.640***              -0.640***
                              (0.030)                (0.030)

---------------------------------------------------------------------
Observations                   342                    342
R2                            0.973                  0.840
Adjusted R2                   0.972                  0.830
Residual Std. Error     0.092 (df = 321)
F Statistic           586.556*** (df = 20; 321) 560.093*** (df = 3; 321)
=====================================================================
Note:                                    *p<0.1; **p<0.05; ***p<0.01
```

Compare the result: We get the same estimated coefficients and standard errors of variables "lincomep755", "lrpmg755", and "lcarpcap755" . In fact, the LSDV and FE models are equivalent to each other

6. [0.5 points] Instead of using variable lincomepxxx as an independent variable, Lucy wants to use m_lincomepxxx as an independent variable. Can she get an estimated coefficient on the variable m_lincomepxxx if she uses a FE model, yes or no? Please provide an explanation for your answer.

```
> pvar(Gasoline755_pdata)
no time variation:       country755 m_lincomep755
no individual variation: year755
```

From Question 3, command 'pvar(Gasoline755_pdata)' has shown us that variables country755 and m_lincomep755 are time-invariant independent variables. Meanwhile, an FE model cannot estimate the coefficients on time-invariant independent. Therefore Lucy cannot use m_lincomep755 as an independent variable

7. [1.5 points] Should you add time fixed effects to the FE model in question 5? In other words, should you choose a FE estimator with both individual and time fixed effects or only with individual fixed effects? Please provide an explanation about how you make the decision.

```
> #Fixed effects(FE) estimator (or within estimator) with time fixed effect
> fe_plm_time <- plm(lgaspcar755 ~ lincomep755 + lrpmg755 + lcarpcap755, da
ta = Gasoline755,
+                     index = c("country755", "year755"), effect = "twoways"
,
+                     model = "within")
> summary(fe_plm_time)
Twoways effects Within Model

Call:
plm(formula = lgaspcar755 ~ lincomep755 + lrpmg755 + lcarpcap755,
    data = Gasoline755, effect = "twoways", model = "within",
    index = c("country755", "year755"))

Balanced Panel: n = 18, T = 19, N = 342

Residuals:
```

```
       Min.     1st Qu.      Median     3rd Qu.        Max.
-0.41920085 -0.03886111  0.00018502  0.04199566  0.23067839

Coefficients:
             Estimate Std. Error  t-value  Pr(>|t|)
lincomep755  0.051369   0.091386   0.5621    0.5745
lrpmg755    -0.192850   0.042860  -4.4995 9.718e-06 ***
lcarpcap755 -0.593448   0.027669 -21.4479 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    10.644
Residual Sum of Squares: 1.997
R-Squared:      0.81239
Adj. R-Squared: 0.78886
F-statistic: 437.338 on 3 and 303 DF, p-value: < 2.22e-16
> fe_plm_time <- plm(lgaspcar755 ~ lincomep755 + lrpmg755 + lcarpcap755, da
ta = Gasoline755,
+                    index = c("country755", "year755"), effect = "twoways"
,
+                    model = "within")
> summary(fe_plm_time)
Twoways effects Within Model

Call:
plm(formula = lgaspcar755 ~ lincomep755 + lrpmg755 + lcarpcap755,
    data = Gasoline755, effect = "twoways", model = "within",
    index = c("country755", "year755"))

Balanced Panel: n = 18, T = 19, N = 342

Residuals:
       Min.     1st Qu.      Median     3rd Qu.        Max.
-0.41920085 -0.03886111  0.00018502  0.04199566  0.23067839

Coefficients:
             Estimate Std. Error  t-value  Pr(>|t|)
lincomep755  0.051369   0.091386   0.5621    0.5745
lrpmg755    -0.192850   0.042860  -4.4995 9.718e-06 ***
lcarpcap755 -0.593448   0.027669 -21.4479 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    10.644
Residual Sum of Squares: 1.997
R-Squared:      0.81239
Adj. R-Squared: 0.78886
F-statistic: 437.338 on 3 and 303 DF, p-value: < 2.22e-16
> # Test whether we should add time fixed effects for FE estimator
> pFtest(fe_plm_time, fe_plm)

        F test for twoways effects

data:  lgaspcar755 ~ lincomep755 + lrpmg755 + lcarpcap755
F = 6.2338, df1 = 18, df2 = 303, p-value = 5.36e-13
alternative hypothesis: significant effects
```

I performed a Ftest (pFtest) between FE estimator with time fixed effect and FE estimator without time-fixed effect. F statistic = 6.2338, p-value = 5.36e-13< 0.05, the null hypothesis that there is no time fixed effects is rejected. We should use a FE estimator that also considers time fixed effects.

8. [4 points] In the following analysis, the dependent variable is still lgaspcarxxx. The independent variables are based on the model chosen by you in question 7:

- If in question 7 you choose a model that doesn't include time fixed effects, then in the following analysis, your independent variables are lincomepxxx, lrpmgxxx, and lcarpcapxxx.
- If in question 7 you choose a model that includes time fixed effects, then in the following analysis, please take the time fixed effects into consideration by including year dummies as your independent variables in your regressions. So your independent variables are lincomepxxx, lrpmgxxx, lcarpcapxxx, and year dummies.

Estimate a pooled OLS model, a FE model, and a RE model using the above dependent variable and independent variables. Use the command "stargazer" to make a table of the results:

- column (1) shows the result of pooled OLS
- column (2) shows the result of the FE model
- column (3) shows the result of the RE model
- only include variables lincomepxxx, lrpmgxxx and lcarpcapxxx in the table

Which model should you choose among pooled OLS, FE model, and RE model? Please provide an explanation about how you make the decision. Based on your final chosen model, is the coefficient on lrpmgxxx significant or not?

```
> #Includes time fixed effect (adding year dummies as IV)
> #Pooled OLS model
> pooled_plm_time <- plm(lgaspcar755 ~ lincomep755 + lrpmg755 + lcarpcap755
+                        + factor(year755), data = Gasoline755, index = c("
country755", "year755"),
+                        model = "pooling")
> summary(pooled_plm_time)
Pooling Model

Call:
plm(formula = lgaspcar755 ~ lincomep755 + lrpmg755 + lcarpcap755 +
    factor(year755), data = Gasoline755, model = "pooling",
    index = c("country755", "year755"))

Balanced Panel: n = 18, T = 19, N = 342

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-0.391090 -0.153023 -0.060439  0.167940  0.576696

Coefficients:
                      Estimate Std. Error  t-value Pr(>|t|)
(Intercept)          2.4821273  0.1475186  16.8259   <2e-16 ***
lincomep755          0.8998965  0.0370783  24.2701   <2e-16 ***
lrpmg755            -0.8991473  0.0311874 -28.8304   <2e-16 ***
lcarpcap755         -0.7642396  0.0191903 -39.8242   <2e-16 ***
factor(year755)1961 -0.0067867  0.0714537  -0.0950   0.9244
factor(year755)1962 -0.0274129  0.0714912  -0.3834   0.7016
factor(year755)1963 -0.0477171  0.0715585  -0.6668   0.5054
factor(year755)1964 -0.0541619  0.0716596  -0.7558   0.4503
factor(year755)1965 -0.0479861  0.0717556  -0.6687   0.5041
factor(year755)1966 -0.0302447  0.0718620  -0.4209   0.6741
factor(year755)1967 -0.0202598  0.0719790  -0.2815   0.7785
```

```
factor(year755)1968 -0.0143913  0.0721091  -0.1996   0.8419
factor(year755)1969 -0.0542155  0.0723071  -0.7498   0.4539
factor(year755)1970 -0.0693975  0.0724975  -0.9572   0.3392
factor(year755)1971 -0.0601962  0.0726382  -0.8287   0.4079
factor(year755)1972 -0.0738927  0.0728597  -1.0142   0.3113
factor(year755)1973 -0.0959981  0.0731650  -1.3121   0.1904
factor(year755)1974 -0.0547243  0.0730354  -0.7493   0.4542
factor(year755)1975 -0.0061533  0.0731047  -0.0842   0.9330
factor(year755)1976 -0.0275159  0.0732676  -0.3756   0.7075
factor(year755)1977 -0.0350204  0.0734096  -0.4771   0.6336
factor(year755)1978 -0.0622416  0.0735833  -0.8459   0.3983
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    102.74
Residual Sum of Squares: 14.697
R-Squared:      0.85695
Adj. R-Squared: 0.84757
F-statistic: 91.2881 on 21 and 320 DF, p-value: < 2.22e-16
> #FE model
> fe_plm_time_1 <- plm(lgaspcar755 ~ lincomep755 + lrpmg755 + lcarpcap755
+                       + factor(year755), data = Gasoline755,
+                     index = c("country755", "year755"), effect = "individu
al",
+                     model = "within")
> summary(fe_plm_time_1)
Oneway (individual) effect Within Model

Call:
plm(formula = lgaspcar755 ~ lincomep755 + lrpmg755 + lcarpcap755 +
    factor(year755), data = Gasoline755, effect = "individual",
    model = "within", index = c("country755", "year755"))

Balanced Panel: n = 18, T = 19, N = 342

Residuals:
       Min.     1st Qu.      Median     3rd Qu.        Max.
-0.41920085 -0.03886111  0.00018502  0.04199566  0.23067839

Coefficients:
                     Estimate Std. Error  t-value  Pr(>|t|)
lincomep755          0.051369   0.091386   0.5621 0.5744611
lrpmg755            -0.192850   0.042860  -4.4995 9.718e-06 ***
lcarpcap755         -0.593448   0.027669 -21.4479 < 2.2e-16 ***
factor(year755)1961  0.040970   0.027248   1.5036 0.1337236
factor(year755)1962  0.044249   0.027595   1.6035 0.1098635
factor(year755)1963  0.064744   0.028277   2.2897 0.0227268 *
factor(year755)1964  0.105995   0.029297   3.6179 0.0003479 ***
factor(year755)1965  0.124134   0.030049   4.1310 4.677e-05 ***
factor(year755)1966  0.167830   0.031046   5.4058 1.310e-07 ***
factor(year755)1967  0.198832   0.032048   6.2042 1.801e-09 ***
factor(year755)1968  0.230077   0.033201   6.9299 2.537e-11 ***
factor(year755)1969  0.242999   0.035304   6.8831 3.374e-11 ***
factor(year755)1970  0.275080   0.037182   7.3982 1.365e-12 ***
factor(year755)1971  0.304198   0.038516   7.8980 5.262e-14 ***
factor(year755)1972  0.332136   0.040532   8.1944 7.160e-15 ***
factor(year755)1973  0.369707   0.043209   8.5562 5.913e-16 ***
factor(year755)1974  0.327938   0.042232   7.7652 1.266e-13 ***
factor(year755)1975  0.362392   0.041877   8.6538 2.984e-16 ***
factor(year755)1976  0.370891   0.043626   8.5016 8.651e-16 ***
factor(year755)1977  0.385702   0.044559   8.6559 2.939e-16 ***
```

```
factor(year755)1978  0.400956   0.046409   8.6397 3.295e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    17.061
Residual Sum of Squares: 1.997
R-Squared:      0.88295
Adj. R-Squared: 0.86827
F-statistic: 108.839 on 21 and 303 DF, p-value: < 2.22e-16
> #RE model
> re_plm_time <- plm(lgaspcar755 ~ lincomep755 + lrpmg755 + lcarpcap755
+                       + factor(year755), data = Gasoline755,
+                       index = c("country755", "year755"), effect = "indivi
dual",
+                       model = "random")
> summary(re_plm_time)
Oneway (individual) effect Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = lgaspcar755 ~ lincomep755 + lrpmg755 + lcarpcap755 +
    factor(year755), data = Gasoline755, effect = "individual",
    model = "random", index = c("country755", "year755"))

Balanced Panel: n = 18, T = 19, N = 342

Effects:
                   var  std.dev share
idiosyncratic 0.006591 0.081183 0.147
individual    0.038340 0.195805 0.853
theta: 0.9053

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-0.438574 -0.042035  0.001770  0.049205  0.262524

Coefficients:
                     Estimate Std. Error  z-value  Pr(>|z|)
(Intercept)         -0.253081   0.332126  -0.7620 0.4460574
lincomep755          0.203648   0.072781   2.7981 0.0051402 **
lrpmg755            -0.287121   0.041747  -6.8776 6.086e-12 ***
lcarpcap755         -0.606100   0.024710 -24.5287 < 2.2e-16 ***
factor(year755)1961  0.031380   0.029320   1.0703 0.2845054
factor(year755)1962  0.028921   0.029627   0.9762 0.3289854
factor(year755)1963  0.040812   0.030218   1.3506 0.1768284
factor(year755)1964  0.072743   0.031108   2.3384 0.0193670 *
factor(year755)1965  0.086578   0.031780   2.7243 0.0064439 **
factor(year755)1966  0.124155   0.032629   3.8051 0.0001417 ***
factor(year755)1967  0.149924   0.033493   4.4763 7.594e-06 ***
factor(year755)1968  0.175379   0.034483   5.0859 3.659e-07 ***
factor(year755)1969  0.177925   0.036224   4.9119 9.022e-07 ***
factor(year755)1970  0.201222   0.037813   5.3215 1.029e-07 ***
factor(year755)1971  0.225522   0.038892   5.7987 6.683e-09 ***
factor(year755)1972  0.245353   0.040590   6.0447 1.497e-09 ***
factor(year755)1973  0.272253   0.042853   6.3532 2.109e-10 ***
factor(year755)1974  0.240432   0.041668   5.7702 7.917e-09 ***
factor(year755)1975  0.276336   0.041564   6.6484 2.963e-11 ***
factor(year755)1976  0.278795   0.042940   6.4926 8.434e-11 ***
factor(year755)1977  0.289351   0.043843   6.5998 4.118e-11 ***
factor(year755)1978  0.297254   0.045368   6.5521 5.673e-11 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    17.829
Residual Sum of Squares: 2.4487
R-Squared:      0.86265
Adj. R-Squared: 0.85364
Chisq: 2009.89 on 21 DF, p-value: < 2.22e-16
> # Pooled OLS vs. FE model (pFtest)
> pFtest(fe_plm_time_1, pooled_plm_time)

        F test for individual effects

data:  lgaspcar755 ~ lincomep755 + lrpmg755 + lcarpcap755 + factor(year755)
F = 113.35, df1 = 17, df2 = 303, p-value < 2.2e-16
alternative hypothesis: significant effects

> # FE model vs. RE model (Hausman Test)
> phtest(fe_plm_time_1, re_plm_time)

        Hausman Test

data:  lgaspcar755 ~ lincomep755 + lrpmg755 + lcarpcap755 + factor(year755)
chisq = 141.85, df = 21, p-value < 2.2e-16
alternative hypothesis: one model is inconsistent

> # Stargazer
> stargazer(pooled_plm_time, re_plm_time, fe_plm_time_1,
+           type="text",
+           column.labels=c("Pooled OLS", "RE", "FE"),
+           keep=c("lincomep755", "lrpmg755", "lcarpcap755"))
```

```
============================================================================
                                  Dependent variable:
             ---------------------------------------------------------------
                                      lgaspcar755
                   Pooled OLS            RE                  FE
                      (1)               (2)                 (3)
----------------------------------------------------------------------------
lincomep755         0.900***          0.204***             0.051
                    (0.037)           (0.073)             (0.091)

lrpmg755           -0.899***         -0.287***           -0.193***
                    (0.031)           (0.042)             (0.043)

lcarpcap755        -0.764***         -0.606***           -0.593***
                    (0.019)           (0.025)             (0.028)

----------------------------------------------------------------------------
Observations          342               342                 342
R2                   0.857             0.863               0.883
Adjusted R2          0.848             0.854               0.868
F Statistic  91.288*** (df = 21; 320) 2,009.887*** 108.839*** (df = 21; 303)
============================================================================
Note:                                            *p<0.1; **p<0.05; ***p<0.01
```

- Pooled OLS vs. FE model (pFtest): F statistic = 113.35, p-value < 2.2e-16 < 0.05, the null hypothesis that there is no individual fixed effects is rejected. We should use the FE estimator rather than the pooled OLS.
- FE model vs. RE model (Hausman Test: phtest): p-value < 2.2e-16 < 0.05, the null hypothesis that the preferred model is RE model is rejected. We should choose the FE model rather than the RE model.
- To conclude, the above two tests revealed that FE model is preferred than both pooled OLS and RE model. We should choose **FE model**
- From the chosen FE model we can observe that the lrpmg755 coefficient is significant (p-value = 9.718e-06 <0.05).

9. [2 points] Test whether there is considerable serial correlation in your chosen model of question 8. Based on the test, should you use the standard errors that you get in question 8 or the robust standard errors? Please provide an explanation about how you make the decision. Based on the decision, will you change your conclusion about whether lrpmgxxx is significant or not?

If you decide to use robust standard errors, please calculate the robust standard errors and use the command "stargazer" to make a table of your results:

- column (1) shows the result of your chosen model with standard errors in question 8
- column (2) shows the result of your chosen model with robust standard errors

```
> # Test for serial correlation:
> pbgtest(fe_plm_time_1)

	Breusch-Godfrey/Wooldridge test for serial correlation in panel mod
els

data:  lwage ~ married + union + factor(year)
chisq = 305.04, df = 8, p-value < 2.2e-16
alternative hypothesis: serial correlation in idiosyncratic errors

>
> #Robust standard errors
> coeftest(fe_plm_time_1, vcovHC)

t test of coefficients:

                 Estimate Std. Error t value  Pr(>|t|)
married          0.058337   0.021296  2.7394 0.0061843 **
union            0.083370   0.023015  3.6223 0.0002958 ***
factor(year)1981 0.113549   0.024571  4.6212 3.941e-06 ***
factor(year)1982 0.167669   0.024228  6.9205 5.257e-12 ***
factor(year)1983 0.210939   0.024912  8.4673 < 2.2e-16 ***
factor(year)1984 0.278407   0.027618 10.0805 < 2.2e-16 ***
factor(year)1985 0.327462   0.026994 12.1307 < 2.2e-16 ***
factor(year)1986 0.386807   0.028244 13.6954 < 2.2e-16 ***
factor(year)1987 0.447037   0.027328 16.3583 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

>
> # Calculate robust standard errors for FE model
> fe_plm_time_cov <- vcovHC(fe_plm_time_1)
```

```
> fe_robust_se <- sqrt(diag(fe_plm_time_cov))
>
> #Stargazer
> stargazer(fe_plm_time_1, fe_plm_time_1,
+           type="text",
+           column.labels=c("FE", "FE r.se."),
+           se = list(NULL, fe_robust_se),
+           keep=c("lincomep755", "lrpmg755", "lcarpcap755"))
```

```
=======================================================
                            Dependent variable:
                        ----------------------------
                               lgaspcar755
                             FE         FE r.se.
                            (1)           (2)
-------------------------------------------------------
lincomep755                0.051         0.051
                          (0.091)       (0.231)

lrpmg755                 -0.193***      -0.193
                          (0.043)       (0.124)

lcarpcap755              -0.593***     -0.593***
                          (0.028)       (0.080)

-------------------------------------------------------
Observations                342           342
R2                         0.883         0.883
Adjusted R2                0.868         0.868
F Statistic (df = 21; 303) 108.839***  108.839***
=======================================================
Note:                      *p<0.1; **p<0.05; ***p<0.01
```

- Breusch-Godfrey/Wooldridge test (pbgtest): P-value < 2.2e-16 < 0.05: the null hypothesis that there is no serial correlation is rejected. There is strong evidence for serial correlation. We should use **robust standard errors** instead of standard errors from question 8.
- After we use Robust standard errors technique to correct the potential heteroskedasticity and autocorrelation consistent standard errors, the **coefficient of independent variable lrpmg755 becomes <u>insignificant</u>**.