# Econometrics III
# Assignment Part IV

Thao Le (523716)
David Gyaraki (582340)

## Contents

```
# load packages
if(!require(pacman)){install.packages("pacman")}

p_load(devtools,tidyverse,dplyr,ggplot2,latex2exp,stargazer, fixest,
 ↪  modelsummary, knitr, readr, tseries, lmtest, forecast, dynlm, vars,
 ↪  xtable, ecm)

dfAssign_p4 <- as.data.frame(read_csv("data/data_assign_p4.csv"))

# Encode quarters
dfAssign_p4 <-
 ↪  cbind(dfAssign_p4,c(seq(1,nrow(dfAssign_p4),length.out=nrow(dfAssign_p4))))
colnames(dfAssign_p4) <- c("obs", "CONS", "INC", "TIME")
```

We are always more interested in estimating ADL(1,1) instead of 1st diff ADL(1,1) because we are not able to say anything about the long run equilibrium in case the 1st diff are used

# 1 Question 1

```
# Set the parameters
gamma <- 0.5
sigma_u <- 1
sigma_v <- 1
phi <- 0.9
init_step = 200
max_T = 2000

# Function to simulate all data
sim_data <- function(T) {
  vt <- rnorm(T, mean = 0, sd = sqrt(sigma_v))
  ut <- rnorm(T, mean = 0, sd = sqrt(sigma_u))

  #X_t and Y_t in the cointergrated model
  xt_cointergrated <- cumsum(vt)
  yt_cointergrated <- gamma * xt_cointergrated + ut

  #X_t and Y_t in the stationary model
  xt_staionary <- numeric(T)
  xt_staionary[1] <- vt[1]
```

```r
  for (t in 2:T) {
    xt_staionary[t] <- phi * xt_staionary[t-1] + vt[t]
  }
  yt_staionary <- gamma * xt_staionary + ut

  return(list(xt_cointergrated = xt_cointergrated, yt_cointergrated =
   ↪  yt_cointergrated,
              xt_staionary = xt_staionary, yt_staionary = yt_staionary))
}

# Initialize dataframe to store results
sim_results <- data.frame(
  T = numeric(),
  beta_hat_1 = numeric(),
  t_stat_1 = numeric(),
  R_squared_1 = numeric(),
  beta_hat_2 = numeric(),
  t_stat_2 = numeric(),
  R_squared_2 = numeric(),
  stringsAsFactors = FALSE
)

# For each sample size T from T = 200,400,..., 1000
for (T in seq(init_step, max_T, 200)) {
  lData = sim_data(T)

  # Cointegrated series
  X1_t = lData$xt_cointergrated
  Y1_t = lData$yt_cointergrated

  # Stationary series
  X2_t = lData$xt_staionary
  Y2_t = lData$yt_staionary

  # Perform linear regression and get summary stats on both models
  model1 <- lm(Y1_t ~ X1_t)
  summary1 <- tidy(model1)
  beta_hat1 <- summary1$estimate[2]
  t_stat1 <- summary1$statistic[2]
  R_squared1 <- summary(model1)$r.squared
```

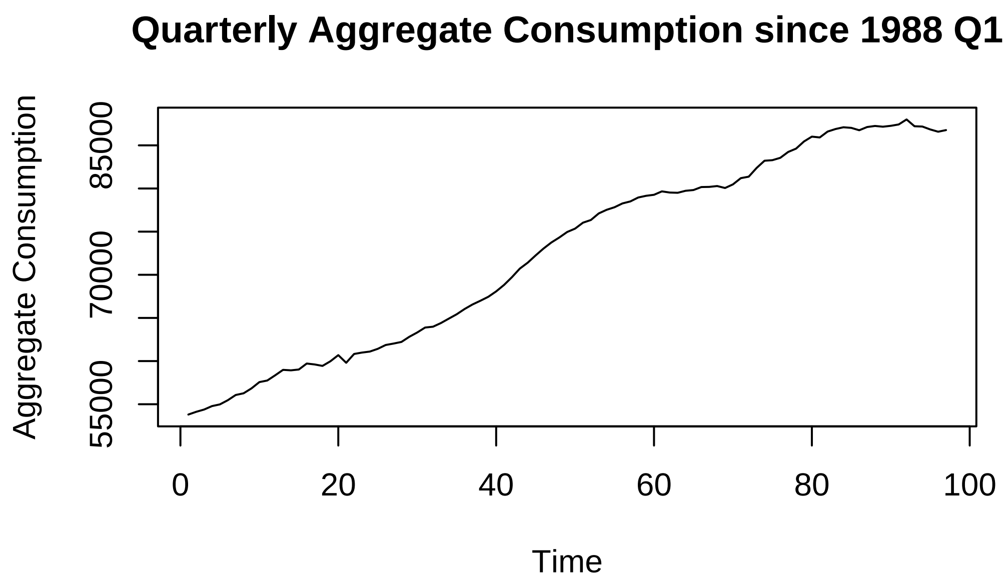
```r
  model2 <- lm(Y2_t ~ X2_t)
  summary2 <- tidy(model2)
  beta_hat2 <- summary2$estimate[2]
  t_stat2 <- summary2$statistic[2]
  R_squared2 <- summary(model2)$r.squared

  # Store results
  sim_results <- sim_results %>%
    add_row(T = T,
            beta_hat_1 = beta_hat1,
            t_stat_1 = t_stat1,
            R_squared_1 = R_squared1,
            beta_hat_2 = beta_hat2,
            t_stat_2 = t_stat2,
            R_squared_2 = R_squared2,
            )
}
```

| | T | beta_hat_1 | t_stat_1 | R_squared_1 | beta_hat_2 | t_stat_2 | R_squared_2 |
|---|---|---|---|---|---|---|---|
| 1 | 200 | 0.54 | 24.62 | 0.75 | 0.51 | 13.08 | 0.46 |
| 2 | 400 | 0.50 | 75.90 | 0.94 | 0.49 | 21.06 | 0.53 |
| 3 | 600 | 0.50 | 85.59 | 0.92 | 0.49 | 25.48 | 0.52 |
| 4 | 800 | 0.50 | 124.18 | 0.95 | 0.51 | 33.60 | 0.59 |
| 5 | 1000 | 0.50 | 137.93 | 0.95 | 0.49 | 32.81 | 0.52 |
| 6 | 1200 | 0.50 | 211.06 | 0.97 | 0.50 | 40.04 | 0.57 |
| 7 | 1400 | 0.50 | 553.08 | 1.00 | 0.50 | 39.90 | 0.53 |
| 8 | 1600 | 0.50 | 244.21 | 0.97 | 0.49 | 43.78 | 0.55 |
| 9 | 1800 | 0.50 | 255.31 | 0.97 | 0.51 | 48.00 | 0.56 |
| 10 | 2000 | 0.50 | 415.59 | 0.99 | 0.51 | 51.27 | 0.57 |

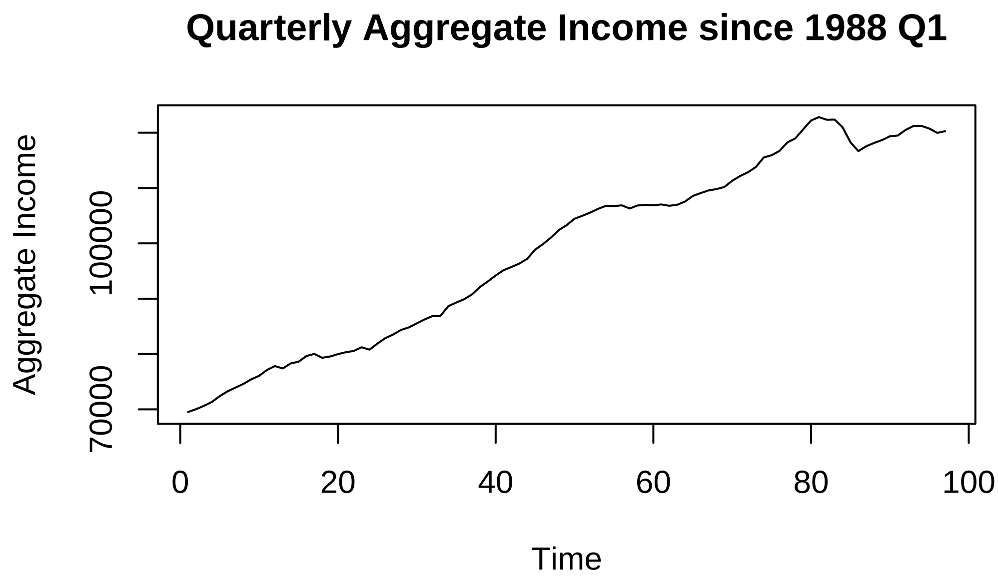
Looking at the results of model 1 (using non-stationary cointegrated variables), we can see that the $\hat{\beta}$ estimates are quite reliable and consistent, with higher T-statistics and higher R-squared scores. This indicates that we did not run into the spurious regression problems. On the other hand, the result of model 2 suggests that the model perform poorly, with lower T-statistics and low R-squared scores suggest that the model perform poorly and the results are not reliable. This suggests that we run into the spurious regression problems.

## 2 Question 2

```
plot.ts(dfAssign_p4$CONS, main="Quarterly Aggregate Consumption since
↪ 1988 Q1", ylab="Aggregate Consumption")
```

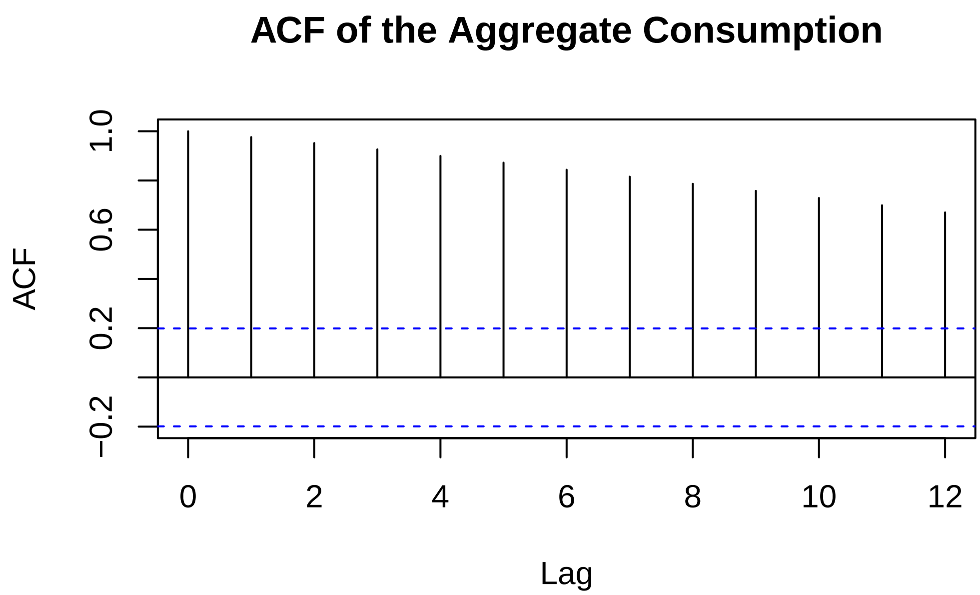



```
plot.ts(dfAssign_p4$INC, main="Quarterly Aggregate Income since 1988
↪ Q1", ylab="Aggregate Income")
```

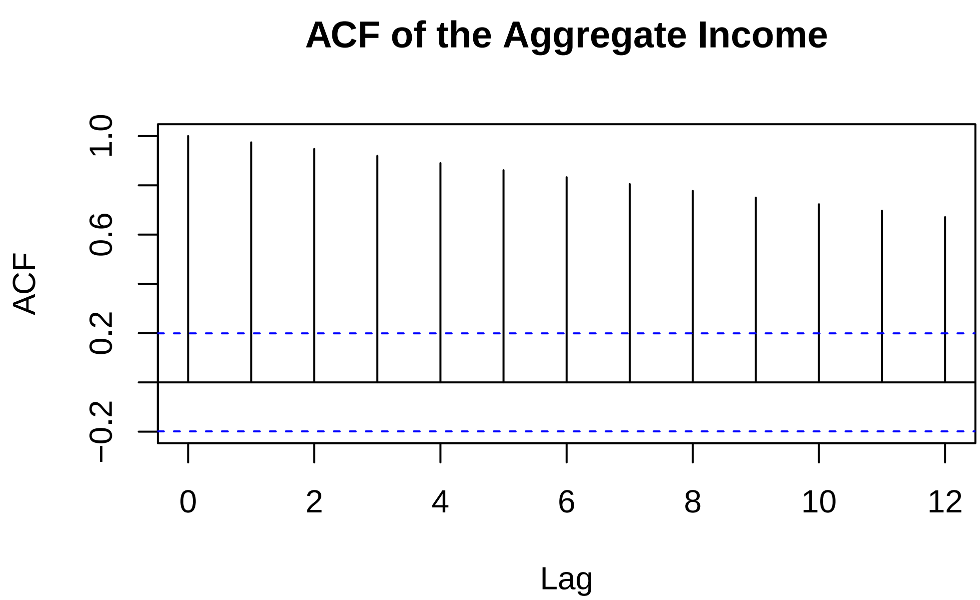



The two plots above show the shape of the quarterly aggregate consumption and quarterly aggregate income in the Netherlands. From this we can see that due to the aggregate nature of the two series, the series are not likely to be stationary at first glance.

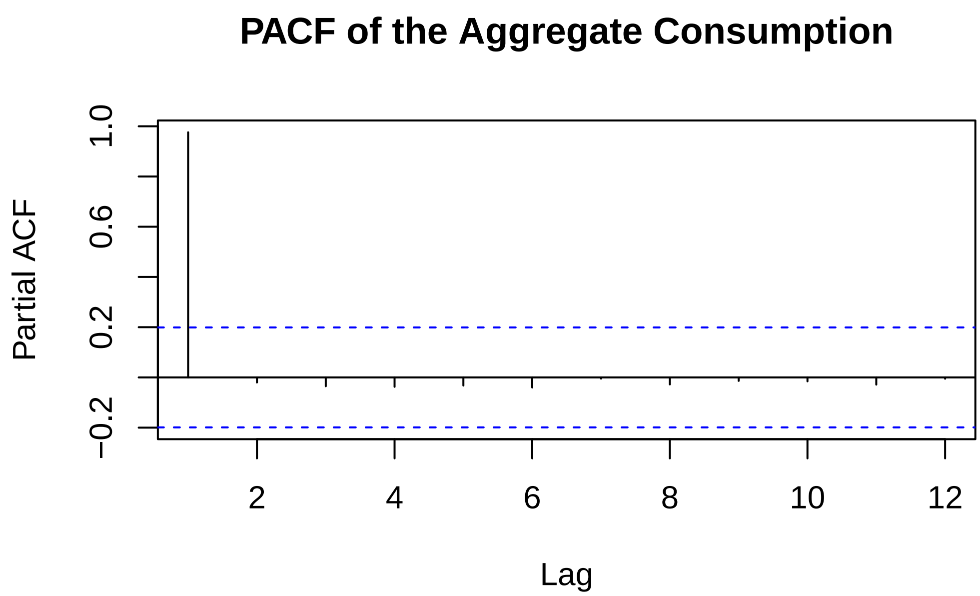
```
acf(dfAssign_p4$CONS,12,pl=T, main="ACF of the Aggregate Consumption")
```



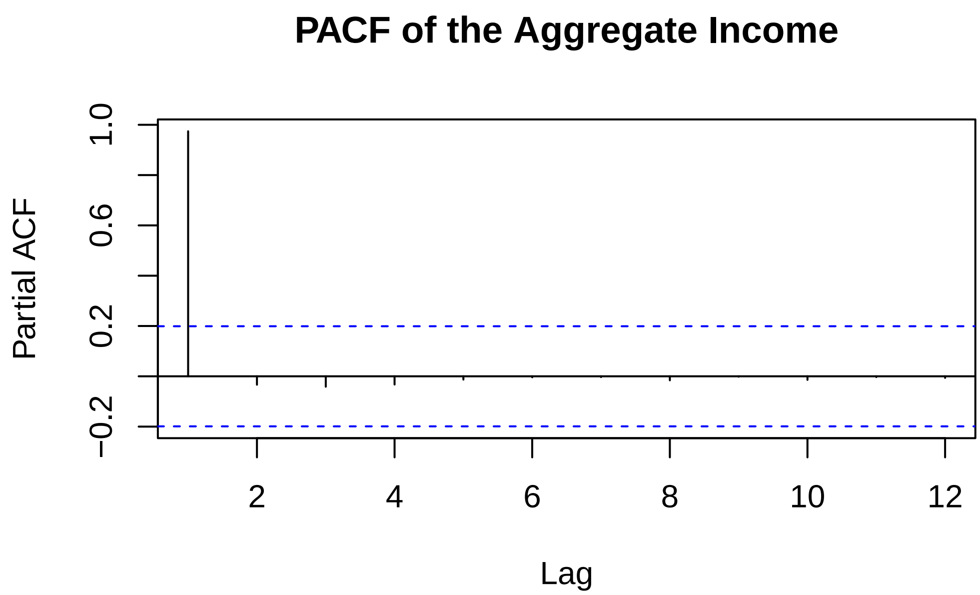

```r
acf(dfAssign_p4$INC,12,pl=T, main="ACF of the Aggregate Income")
```



```r
pacf(dfAssign_p4$CONS,12,pl=T, main="PACF of the Aggregate Consumption")
```




```r
pacf(dfAssign_p4$INC,12,pl=T, main="PACF of the Aggregate Income")
```




From the ACF and PACF plots for the consumption and income components of the GDP, we can see that due to the aggregate nature of the series, both the consumption and income autocorrelations show significant correlation (above the white-noise threshold) across all lags all the way to the 12th period lag. However, when we investigate the partial autocorrelation function results, both variables have insignificant correlations for all lags (1st to 12th). While some correlations are more pronounced for consumption than for income, no values reach the white-noise threshold and therefore are not very significant.

## 3 Question 3

```
adf.test(dfAssign_p4$CONS)
```

```
	Augmented Dickey-Fuller Test

data:  dfAssign_p4$CONS
Dickey-Fuller = -1.214, Lag order = 4, p-value = 0.9006
alternative hypothesis: stationary
```

```r
adf.test(dfAssign_p4$INC)
```

```
    Augmented Dickey-Fuller Test

data:  dfAssign_p4$INC
Dickey-Fuller = -1.4134, Lag order = 4, p-value = 0.819
alternative hypothesis: stationary
```

Based on the test results, we cannot reject in either case that the series would be non-stationary. Since we do not find sufficient evidence for the stationarity, we can stay at the null-hypothesis of the series being non-stationary.

# 4 Question 4

```r
d1_CONS <- diff(dfAssign_p4$CONS, differences = 1)

d1_INC <- diff(dfAssign_p4$INC, differences = 1)

adf.test(d1_CONS)
```

```
    Augmented Dickey-Fuller Test

data:  d1_CONS
Dickey-Fuller = -2.1092, Lag order = 4, p-value = 0.5313
alternative hypothesis: stationary
```

```r
adf.test(d1_INC)
```

```
    Augmented Dickey-Fuller Test

data:  d1_INC
Dickey-Fuller = -3.7139, Lag order = 4, p-value = 0.02705
alternative hypothesis: stationary
```

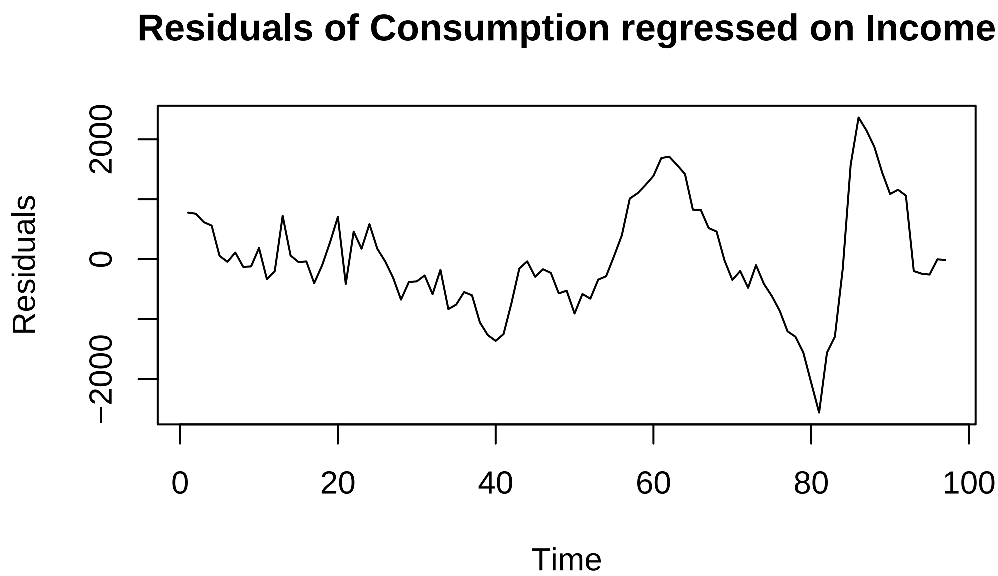
In the case of the first-difference series, we still cannot reject the null hypothesis on the first case (consumption), however the second Augmented Dickey-Fuller test has a p-value of $p = 2.71\%$ and hence can be rejected at a 5% significance level. This means that the first difference of the income component of the GDP might be stationary. In terms of the order of integration, these results imply that the income series is most likely an $I(1)$ series (integrated of order 1), while for consumption, this may be $I(2)$ or higher order, but the ADF test implies that even the first difference of the series is not stationary.

## 5 Question 5

```r
model5 <- lm(CONS ~ INC, data=dfAssign_p4)

summary(model5)
```

```
Call:
lm(formula = CONS ~ INC, data = dfAssign_p4)

Residuals:
    Min      1Q  Median      3Q     Max 
-2559.4  -524.3  -128.0   561.2  2364.8 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 6.783e+03  5.550e+02   12.22   <2e-16 ***
INC         6.651e-01  5.534e-03  120.19   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 911.2 on 95 degrees of freedom
Multiple R-squared:  0.9935,	Adjusted R-squared:  0.9934 
F-statistic: 1.445e+04 on 1 and 95 DF,  p-value: < 2.2e-16
```

```r
residuals <- model5$residuals

dfAssign_p4 <- cbind(dfAssign_p4, residuals)

plot.ts(dfAssign_p4$residuals, main="Residuals of Consumption regressed
 ↪  on Income", ylab="Residuals")
```




From the residuals plot, one can already see that we would suspect stationarity in the residuals, which would imply cointegration between the two variables.

```
adf.test(residuals)
```

```
	Augmented Dickey-Fuller Test

data:  residuals
Dickey-Fuller = -3.5848, Lag order = 4, p-value = 0.03839
alternative hypothesis: stationary
```

```
adf_res = ur.df(dfAssign_p4$residuals, type =  "none", lags = 10,
      selectlags = "BIC")

summary(adf_res)
```

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################
```

```
Test regression none


Call:
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:
     Min       1Q   Median       3Q      Max 
-1179.90  -270.52    -7.59   274.20  1411.39 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
z.lag.1     -0.17739    0.05162  -3.436 0.000924 ***
z.diff.lag1  0.19325    0.10272   1.881 0.063428 .  
z.diff.lag2  0.26739    0.10461   2.556 0.012410 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 421.2 on 83 degrees of freedom
Multiple R-squared:  0.1657,	Adjusted R-squared:  0.1356 
F-statistic: 5.497 on 3 and 83 DF,  p-value: 0.001709


Value of test-statistic is: -3.4363 

Critical values for test statistics: 
     1pct  5pct 10pct
tau1 -2.6 -1.95 -1.61
```

The ADF test is conducted by starting with 10 number of lags and a general-to-specific approach, where we remove lags based on the Schwartz Information Criteria at each step (without considering drift or trend). Based on the ADF test on the residuals, we can observe that the non-stationarity hypothesis is rejected. This implies that the residuals series is probably stationary, which implies existing cointegration between consumption and income. Then we can observe, that the model with the best SIC is the one:

$$\Delta Z_t = \gamma Z_{t-1} + \delta_1 \Delta Z_{t-1} + \delta_2 \Delta Z_{t-2} \tag{1}$$

Then the test statistic for the ADF test is $-3.436$, which is below the 5% significance level of $-1.95$, hence the consumption and income variables are likely cointegrated with order

$CI(1,1)$ which comes from the starting assumption of this question that both variables separately are considered $I(1)$. This cointegration is also confirmed by the coefficient test of $\gamma = -0.177$ in the model, which shows that the coefficient is not statistically equal to 0.

## 6 Question 6

```
ecm_model <- ecmback(y=dfAssign_p4['CONS'], xtr = dfAssign_p4['INC'],
↪  xeq=dfAssign_p4['residuals'], lags = 1, criterion = "BIC",
↪  includeIntercept = F)

summary(ecm_model)
```

```
Call:
lm(formula = as.formula(formula), data = x, weights = weights)

Residuals:
     Min       1Q   Median       3Q      Max 
-1045.09  -160.57    54.12   196.27   733.58 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
deltaINC       0.2294926  0.0433945   5.289 8.14e-07 ***
residualsLag1 -0.1540825  0.0337558  -4.565 1.53e-05 ***
yLag1          0.0029312  0.0005107   5.739 1.18e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 298.3 on 93 degrees of freedom
Multiple R-squared:  0.6465,    Adjusted R-squared:  0.6351 
F-statistic: 56.69 on 3 and 93 DF,  p-value: < 2.2e-16
```

```
mean(diff(dfAssign_p4$CONS, differences = 1))/mean(diff(dfAssign_p4$INC,
↪  differences = 1))
```

```
[1] 0.6495517
```

In this case, we can use the same series of estimated residuals as used in Section 5 and the $CI(1,1)$ property to estimate the error correction model. Above we can see the results of an error correction model, where the model is estimated in the following form:

$$\Delta Y_t = \gamma Z_{t-1} + \delta_0 Y_{t-1} + \beta_0 \Delta X_t + \epsilon_t, \tag{2}$$

where the Y marks the consumption series, X marks the income series and the Z is the residual from the simple regression of consumption on income. We can infer from the results that the short run multiplier is $\beta_0 = 0.23$ and for the long run multiplier we need to first specify the model $\bar{Y} = \alpha + \delta_1 \bar{X}$, where $\delta_1$ is the long run multiplier. We can calculate this by restricting the $\alpha$ to 0, and then we get that $\Delta\bar{Y} = \delta_1 \Delta\bar{X}$, resulting in $\delta_1 = 0.65$. This means that in the short run, increase in income will slightly increase consumption, but in the long term, this effect will become more pronounced. To interpret the error correction coefficient of $\gamma = -0.15$, we can say that consumption is corrected with a slight error correction of $-0.15 \times \epsilon_t$ if the consumption deviates from the long term equilibrium.

## 7 Question 7

Based on the results of Section 6, we can observe an error correction coefficient of $\gamma = -0.15$, which is a relatively low error correction coefficient. There is no overshooting, because $|\gamma| < 1$, which means that the model applies a slight error correction when the consumption is far from its long term equilibrium. Next, we can address the question of Granger causality by estimating an ADL(1,1) model and checking the significance of the first exogenous lag coefficient ($X_{t-1}$) or directly looking at the Granger causality test. Following the latter, we can observe the results for the test as below, where the p-value of $0.005$ shows that the income Granger causes the consumption at 5% level. If we reverse the order however, there is not enough evidence of the reverse at 5% level, only at 10%. This implies that income Granger causes consumption, but probably consumption does not Granger cause income.

```r
grangertest(CONS ~ INC, order = 3, data = dfAssign_p4)
```

```
Granger causality test

Model 1: CONS ~ Lags(CONS, 1:3) + Lags(INC, 1:3)
Model 2: CONS ~ Lags(CONS, 1:3)
  Res.Df Df      F    Pr(>F)
1     87
2     90 -3 4.6045 0.004873 **
---
```

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

```r
grangertest(INC ~ CONS, order = 3, data = dfAssign_p4)
```

```
Granger causality test

Model 1: INC ~ Lags(INC, 1:3) + Lags(CONS, 1:3)
Model 2: INC ~ Lags(INC, 1:3)
  Res.Df Df      F  Pr(>F)
1     87
2     90 -3 2.3272 0.08018 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# 8 Question 8

In order to predict the consumption in 2009 Q3, we need to re-estimate the model using data only up to that point in order to avoid predicting the data we used for estimation.

```r
dfQ8 <- dfAssign_p4[1:85,]

#modelres <- lm(CONS ~ INC, data=dfQ8)
#
#residuals <- modelres$residuals
#
#dfQ8['residuals'] <- residuals

ecm_model_q8 <- ecmback(y=dfQ8['CONS'], xtr = dfQ8['INC'],
↪  xeq=dfQ8['residuals'], lags = 1, criterion = "BIC", includeIntercept
↪  = F)

summary(ecm_model_q8)
```

```
Call:
lm(formula = as.formula(formula), data = x, weights = weights)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-1070.62  -160.17    30.22   183.09   702.40

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
deltaINC       0.1940756  0.0451806   4.296 4.81e-05 ***
residualsLag1 -0.1491552  0.0379356  -3.932 0.000177 ***
yLag1          0.0036199  0.0005607   6.456 7.40e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 279.5 on 81 degrees of freedom
Multiple R-squared:  0.7117,	Adjusted R-squared:  0.701
F-statistic: 66.66 on 3 and 81 DF,  p-value: < 2.2e-16
```

```r
ecmpredict(ecm_model_q8, dfAssign_p4[86:88,], init = dfAssign_p4[86,2])
```

```
               87       88
86748.80 86881.61 86991.78
```

From the prediction, we can see that the predicted aggregate consumption in period 87 (2009 Q3) is 86881.6, while the actual value is 87121.6. Therefore, we predict an increase of $86881.6 - 86748.8 = 132.8$ in aggregate consumption, expecting a very slight raise. Since the lagged residual value in 2009 Q2 is $2364.76$ and our coefficient in the model is $-0.149$, the error correction actually takes away $2364.76 * 0.149 = 352.35$ from the predicted increase. This sounds like a huge impact but it can mostly be attributed to the fact that the residuals around this period (namely in Q2 2009 which we can treat as observed) are quite large in absolute terms due to the volatile nature of the economy in those periods. If we did not consider the error correction, we would actually overshoot the actual consumption value in the period, while in this case we undershoot it.