Exam Econometrics (MSB104)

SOLUTION PROPOSAL

Subject code: MSB104

Date of exam: 15.06.2021

Language: English (you may submit your answer in English or any scandinavian tongue)

Course coordinator: Henrik Lindegaard Andersen (hlan@hvl.no (mailto:hlan@hvl.no))

General information

- State any references clearly (as your assignment will be cross-checked in text analysis software).
- Remember that the exam is *INDIVIDUAL*. It is NOT allowed to collaborate with others during the exam. Otherwise, all aids are allowed.
- You may write by hand and/or use any text editor; your answer must be uploaded to WISEflow as ONE final PDF-document.
- Do NOT write any personal identifiers on your hand-in (e.g. name or student id).
- You are NOT supposed to gather data OR to run any regression in this assignment.
- NB! Dot (.) is used as a decimal separator (e.g. Pi is 3.14159). Comma (,) is used as a thousand separator for readability (e.g. one thousand is 1,000).
- Your answer to each sub-question within Part I (210 minutes) will be given an equal weight in the evaluation, and equally for Part II (90 minutes).
- Do you have questions to the exam text? Part I/II: 92345700/41611857 (Henrik/Jørn).

Part I: Regression analysis with cross-sectional data and OLS (210 minutes)

Part I consists of 14 questions labeled a) to n). Each question is given equal weight.

Imagine yourself in some years, when one of your future kids are about to begin secondary education (“videregående skole”). You want to give him, or her, some good career advice, and because you have access to a random sample (n=2,786) of Norwegians aged 25–64 years, with a positive earned income, you run a linear regression of annual earned income (in NOK 1,000) on some explanatory variables and an intercept.

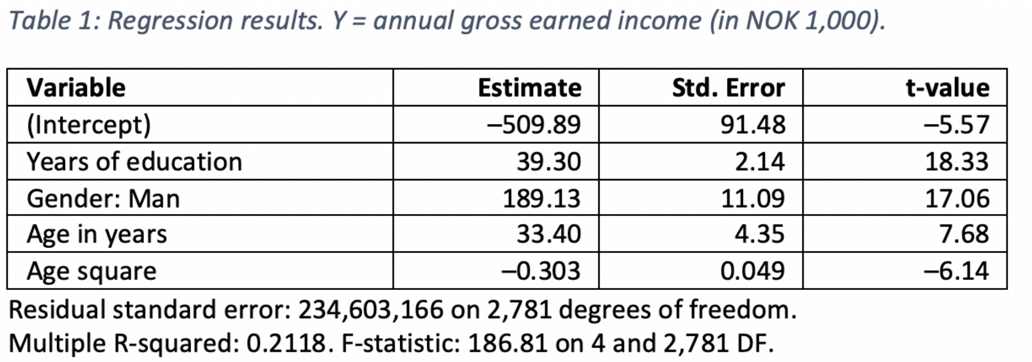


Table 1: Regression results. Y = annual gross earned income (in NOK 1,000).

Variable	Estimate	Std. Error	t-value
(Intercept)	–509.89	91.48	–5.57
Years of education	39.30	2.14	18.33
Gender: Man	189.13	11.09	17.06
Age in years	33.40	4.35	7.68
Age square	–0.303	0.049	–6.14

Residual standard error: 234,603,166 on 2,781 degrees of freedom.
Multiple R-squared: 0.2118. F-statistic: 186.81 on 4 and 2,781 DF.

The explanatory variables are: years of education after compulsory primary school (e.g. equal to 3 if a person only has ‘videregående skole’), a gender dummy, and age measured in years, as well as age squared. The results are summarized in Table 1.

Now, answer the following questions:

a. Interpret the estimates: e.g. what are the consequences of a one unit change in years of education, man, and age?
b. Briefly, comment on the statistical significance of the estimates.
c. Explain what the R-squared and the F-statistic tells us. For the F-statistic, show how it is calculated. Note: You may want to know that the sum of squared residuals from the restricted model is 297,640,088.
d. Explain why you would want to include age squared – and why is it OK to do so, when you are running a “linear” regression?
e. Calculate [1] the expected annual earned income at the age of 35, if your kid is a boy, and obtains 4 years of education after primary school. If your kid is a girl, [2] how many years of education would she need, to obtain – roughly – the same expected earned income as the boy at the same age?
f. Calculate a 90% confidence interval (CI) for the education estimate.
g. You worry that there might be a bias in your estimate for the years of education. Imagine that people with higher ability also obtain more education, but ability is unobserved. Explain the direction of the bias in the education coefficient and state your assumptions.
h. You have included age in your regression because you think is may be a good measure (a proxy-variable, that is) for labour market experience, which is not available in your data. It seems reasonable to assume that age and experience are highly correlated. Carefully explain what would happen to the coefficient and the standard error on age, if you were to include both age and experience - at the same time - in the regression.

Now you run the same model, but you include some extra explanatory variables – all are binary indicator variables. The results are listed in Table 2.

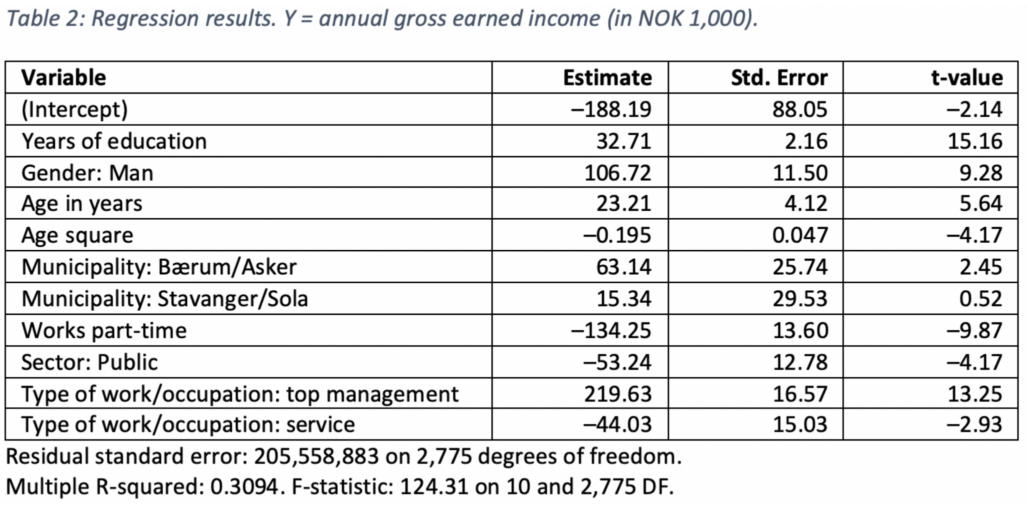


Table 2: Regression results. Y = annual gross earned income (in NOK 1,000).

Variable	Estimate	Std. Error	t-value
(Intercept)	–188.19	88.05	–2.14
Years of education	32.71	2.16	15.16
Gender: Man	106.72	11.50	9.28
Age in years	23.21	4.12	5.64
Age square	–0.195	0.047	–4.17
Municipality: Bærum/Asker	63.14	25.74	2.45
Municipality: Stavanger/Sola	15.34	29.53	0.52
Works part-time	–134.25	13.60	–9.87
Sector: Public	–53.24	12.78	–4.17
Type of work/occupation: top management	219.63	16.57	13.25
Type of work/occupation: service	–44.03	15.03	–2.93

Residual standard error: 205,558,883 on 2,775 degrees of freedom.
Multiple R-squared: 0.3094. F-statistic: 124.31 on 10 and 2,775 DF.

The added variables indicate whether an individual lives in Bærum/Asker or Stavanger/Sola municipalities, works part-time, is employed in the public sector, or has a top management job (e.g. CEOs or top officials) or a service occupation (e.g. bartenders, waiters, kitchen staff etc.).

i. Write up the relevant equation and carefully explain why the residual standard error drops by approximately 29 million in Table 2 compared to Table 1.

j. Find the p-value for the Bærum/Asker-municipality estimate for a one-tailed test. Carefully explain what it means. You may either calculate the exact p-value using software or approximate it using the tables in the appendix of the book.

k. You live in Haugesund, but after seeing these estimates, you consider moving to Stavanger in order to obtain a higher earned income. Statistically and econometrically speaking, this might not be a wise idea for two reasons. State these reasons.

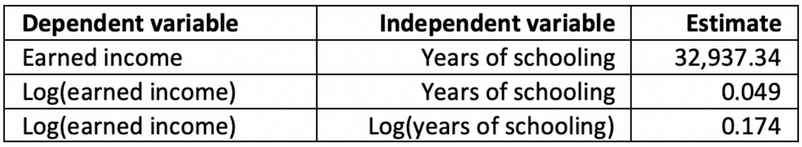
l. The point estimate for service-jobs is –44. Explain how this is to be interpreted, i.e. what is the group of reference?

You suspect that the variance of the unobserved determinants of earned income increases with the level of education. To test this, you calculate the squared residuals using the estimates in Table 2. You then regress these squared residuals on all the explanatory variables listed in Table 2. This auxiliary regression produces an F-statistic of 8.48 with 10 and 2,775 degrees of freedom.

m. Briefly, explain what this test tells you (i.e. what's the null hypothesis, the approximate p-value, and what is the conclusion). Also explain what the underlying problem is, and what may be done.

Instead of the models above, where you used earned income in NOK 1,000, you could have run models with other scaling's or functional forms. Now imagine that you focus on the earned income in kroner; the first row in Table 3 shows the estimate on years of education in a simple linear regression using this dependent variable. The second and third row, respectively, shows the estimate from a log-level and log-log model.

Table 3: Simple linear regression estimates involving logarithms



Dependent variable	Independent variable	Estimate
Earned income	Years of schooling	32,937.34
Log(earned income)	Years of schooling	0.049
Log(earned income)	Log(years of schooling)	0.174

Answer the following question:

n. Explain the interpretation of the coefficient estimates in Table 3.

Part II: Regression analysis with time series data and simultanous equations models (90 minutes)

Part II consists of three subsections. Each subsection is given equal weight. It is sufficient to provide short and punctuated answers to all of the questions.

Short questions

1. In time series econometrics, explain the conceptual differences for the data generating process (DGP) before and after its realization.

Formally, a sequence of random variables indexed by time is called a stochastic process or a time series process. ("Stochastic" is a synonym for random.) When we collect a time series data set, we obtain one possible outcome, or realization, of the stochastic process.

2. Explain why a contemporanous exogenous process differs from a strictly exogenous process.

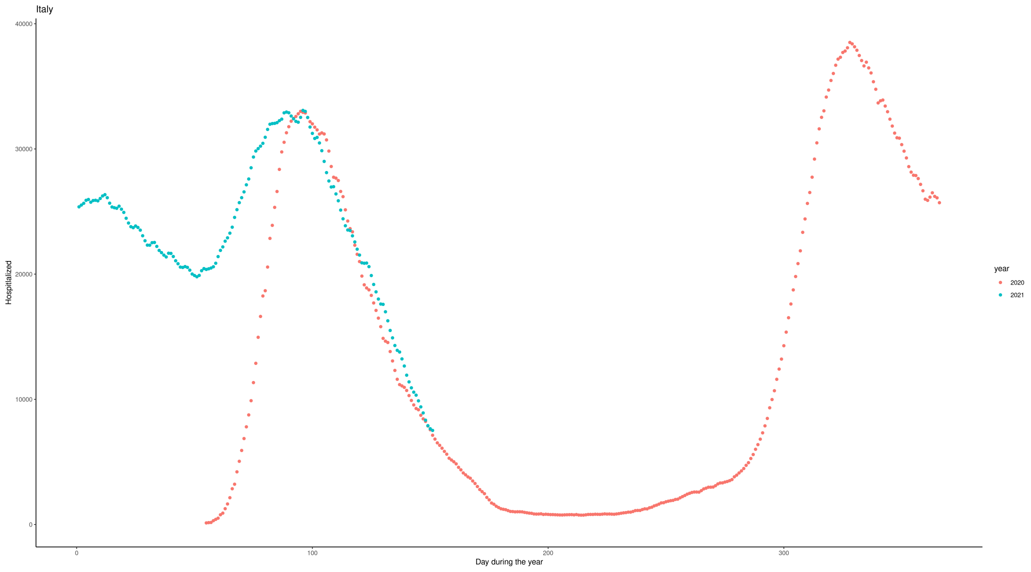
The TS.3' assumption that $E(u_t|x_t) = 0$ implies that u_t and the explanatory variables (X) are contemporaneously uncorrelated: $Corr(x_{t,j}, u_t) = 0$ for all js. $E(u_t|X) = 0$ requires more than contemporaneous exogeneity: $u_{s,j}$ must be uncorrelated with $x_{s,j}$, even when $j \neq$i In this case we say that the explanatory variables are strictly exogenous.

3. Assume that you detect that TS.3 is violaed but not TS.3'. Given that the other time series assumptions are satisfied, how would this impact your results?

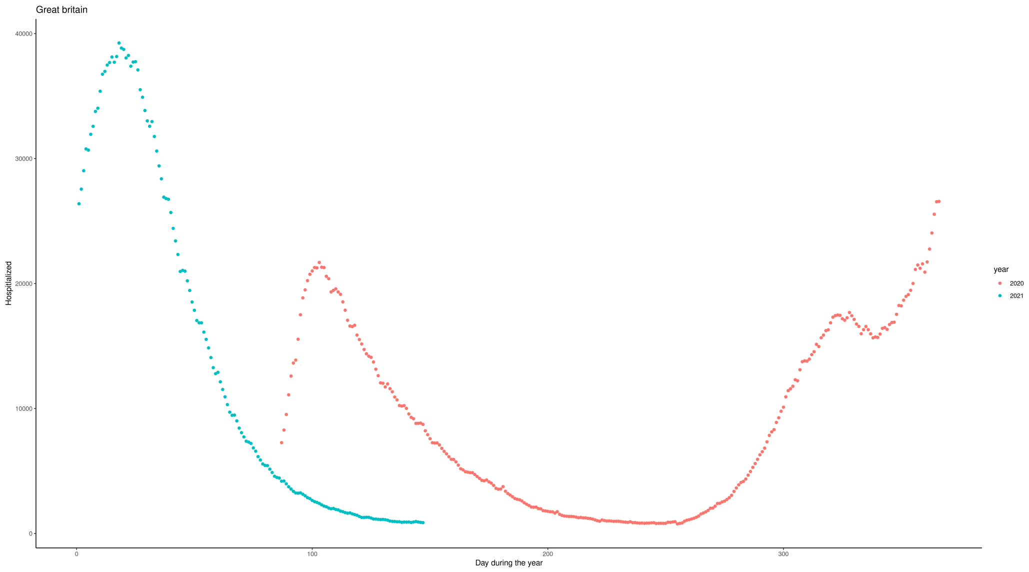
TS.3' is sufficient for proving consistency, while violation of TS.3 implies that the OLS estimator is unabiased. As such OLS properties are only valid in large samples.

4. Assume that you want to estimate in a time series model the impact of a lockdown on the number of hospitalized Covid patients in Italy. Given the plot below, which variable do you think should be included in such a model in order to eliminate the problem of spurious regression? (red dot = 2020, blue dot = 2021, x="Day during the year", y="Hospitalized")

The year-to-year plot clearly indicates that the number of Covid-patients follows a seasonal pattern. Variables that captures this should be included in order to avoid the problem of spurious regression.



5. Why do you think the year-to-year pattern of the plot from Great Britain differ such from Italy, and how would you account for this in a time series regression model? (red dot = 2020, blue dot = 2021, x="Day during the year", y="Hospitalized")



The seasonal pattern seems here to be broken. This is probably due to that UK is the leading country when it comes vaciation for a large fration of its population. A variable that captures the degree of vaciation should be included in order to capture this.

6. Why are demand and supply curves (with structural interpretation) difficult to estimate when using real life data?

The equilibrium condition combined with the equation for the supply and demand curve constiute a simultanous equation (SEM) model in whichthe variable price and quantiy commonly determined. This imply endogeneity problem in the form of simulteneity problem, which implies that the SEM model can not be given a structural interpretation.

7. Is there a approach to solve this problem?

The use of IV-variabel approach: Either the demand or the supply curve need an observed variable z_i in which changes in the value of the variable would reveal the occurnse of the other curve.

Stochastical regression models

We have the following AR(1) process.

$$y_t = \mu + \rho y_{t-1} + u_t \text{ where } u_t \sim N(0, \sigma_u^2) \tag{1}$$

and moving average MA(1) process:

$$y_t = \gamma + e_t + \theta e_{t-1} \text{ where } e_t \sim N(0, \sigma_e^2) \tag{2}$$

1. Find for the two model specifications above its mean and and one-period ahead forecast value

AR(1)

$$E(y_t) = E(\mu + \rho(\rho y_{t-2} + e_{t-1}) + e_t) = \mu + E(\sum_{j=0}^{\infty} \alpha^j (\mu + e_{t-j})) = \sum_{j=0}^{\infty} \alpha^j \mu \tag{3}$$

$$y_{t+1} = \mu + \rho y_t + u_{t+1} \Rightarrow E(y_{t+1}|y_t) = E(\mu + \rho y_t + e_{t+1}|y_t) = \mu + \rho y_t \tag{4}$$

MA(1)

$$E(y_t) = E(\gamma + e_t + \theta e_{t-1}) = \gamma \tag{5}$$

$$y_{t+1} = \gamma + e_{t+1} + \theta e_t \Rightarrow E(y_{t+1}|y_t) = E(\gamma + e_{t+1} + \theta e_t) = \gamma + \theta e_t \tag{6}$$

2. Why can’t OLS be applied to the estimation to both of the model specifications?

Fitting the MA estimates is more complicated than it is in autoregressive models (AR models), because the lagged error terms are not observable. This means that iterative non-linear fitting procedures need to be used in place of linear least squares.

3. Do you think its advisable to use the model specification of the MA(1) process to model the stock market?

No, the stock market is commonly viewed to follow a Random-Walk process. If a MA(1) was the case, the estimated regression could be used as a “money-making machine” since profit would be ensured in the long run if buying under the trend and selling over the trend.

Application: The Phillips-Curve

Cf. appendix part II for output information.

a. Explain the main difference between the two model specifications

The first moel specifies the staic relationship between the impact unemployment has on the level of inflation. In the second model, the level of inflation has been replaced by the change in its inflation rate.

b. In the diagnostic part of a time series regression model, does the output provide any clues of which specification that should be used?

Yes, π variabel seems to be integrated of order one, while differencing it provides an I(0) variable.

c. Interpret the results from the estimation of the coefficients and explain why the standard deviations differ in the two cases?

Solution: - Model 2 seems to best among the two: More significant variables, higher R-square. - Differ: Standard differ since HAC seeks to correct for the presence of serial correlation, usually be giving higher standard deviation values.

Appendix

Part II

Data sample (realized DGP)

```
#knitr::purl("Applications_Phillips_static.Rmd")
# Loading data
rm(list=ls())
library(sandwich)
library(lmtest)
library(wooldridge)
library(ggplot2)
library(plotly)
phillips_rdgp <- phillips # Realized DGP
#write.csv(phillips_rdgp, file='/home/joernih/tmp/R/tmp.csv')
#From `r phillips_rdgp$year[1]` to `r rev(phillips_rdgp$year)[1]`
```

Data generating process (DGP) and its regression model

Model specification 1:

$$inf_t = \beta_0 + \beta_1 unem_t + u_t, t = 1, 2, \ldots, n \tag{7}$$

```
model_ph <- lm(phillips_rdgp$inf ~ phillips_rdgp$unem)
```

Model specification 2:

$$\Delta inf_t = \beta_0 + \beta_1 unem_t + u_t, t = 1, 2, \ldots, n \tag{8}$$

```
model_phc <- lm(phillips_rdgp$cinf ~ phillips_rdgp$unem)
```

Model estimation

```
ols_ph <- summary(model_ph)
ols_phc <-summary(model_phc)
```

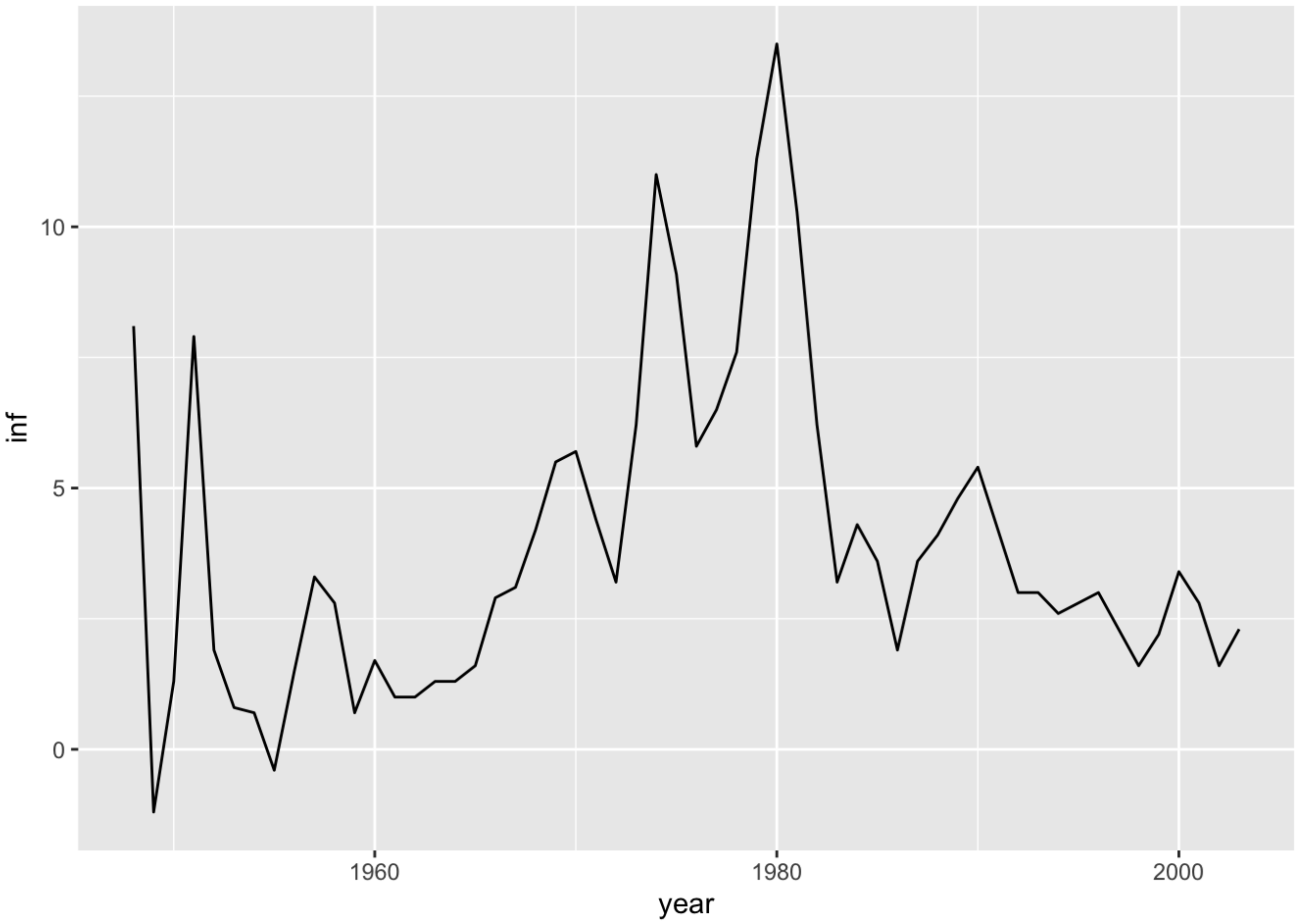
Diagnostics

Detecting

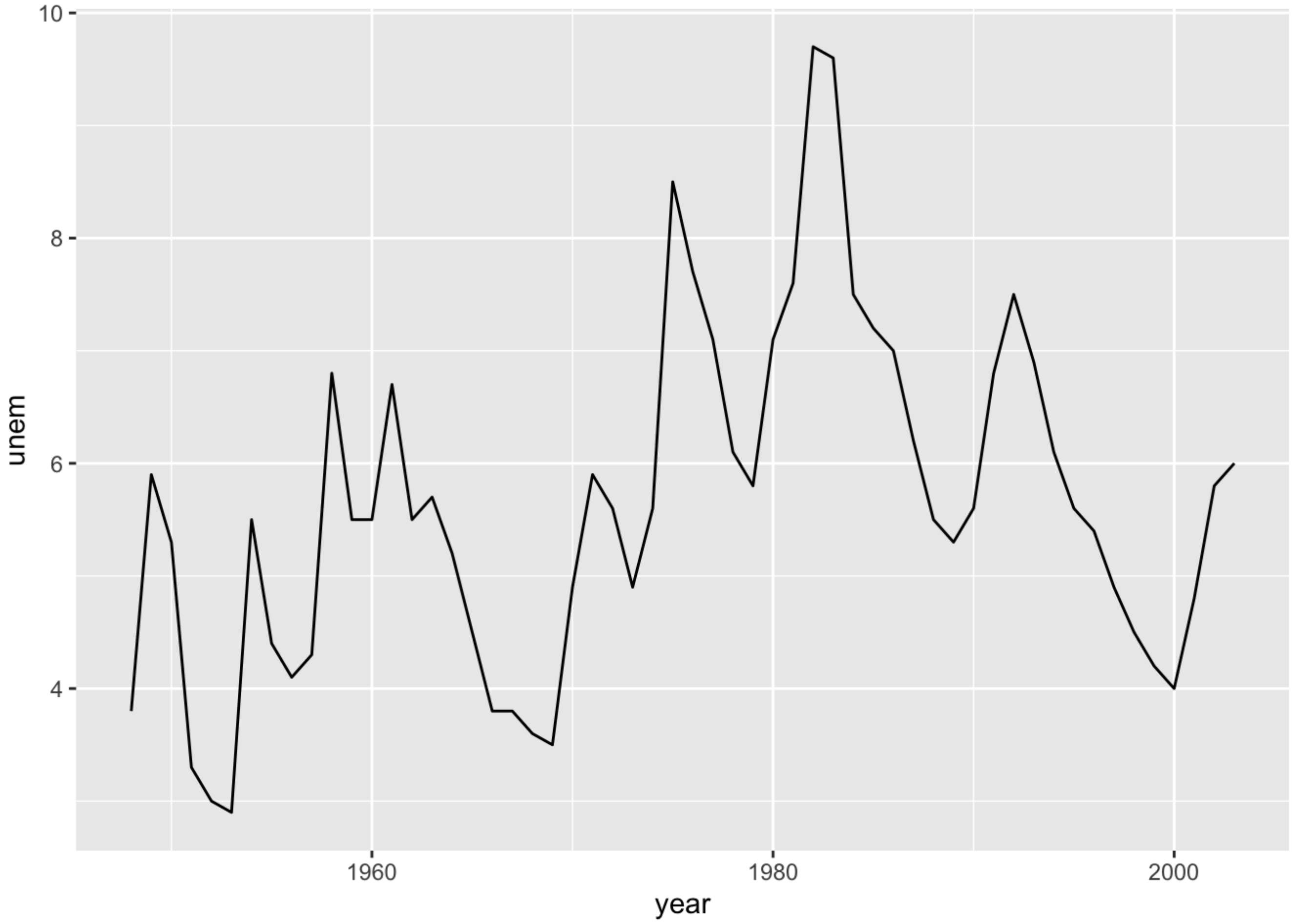
Time trends and/or weakly dependency (I(0) or I(1))

Graphical

```
ggplot2::ggplot(data=phillips_rdgp) + ggplot2::geom_line(aes(x=year,y=inf))
```

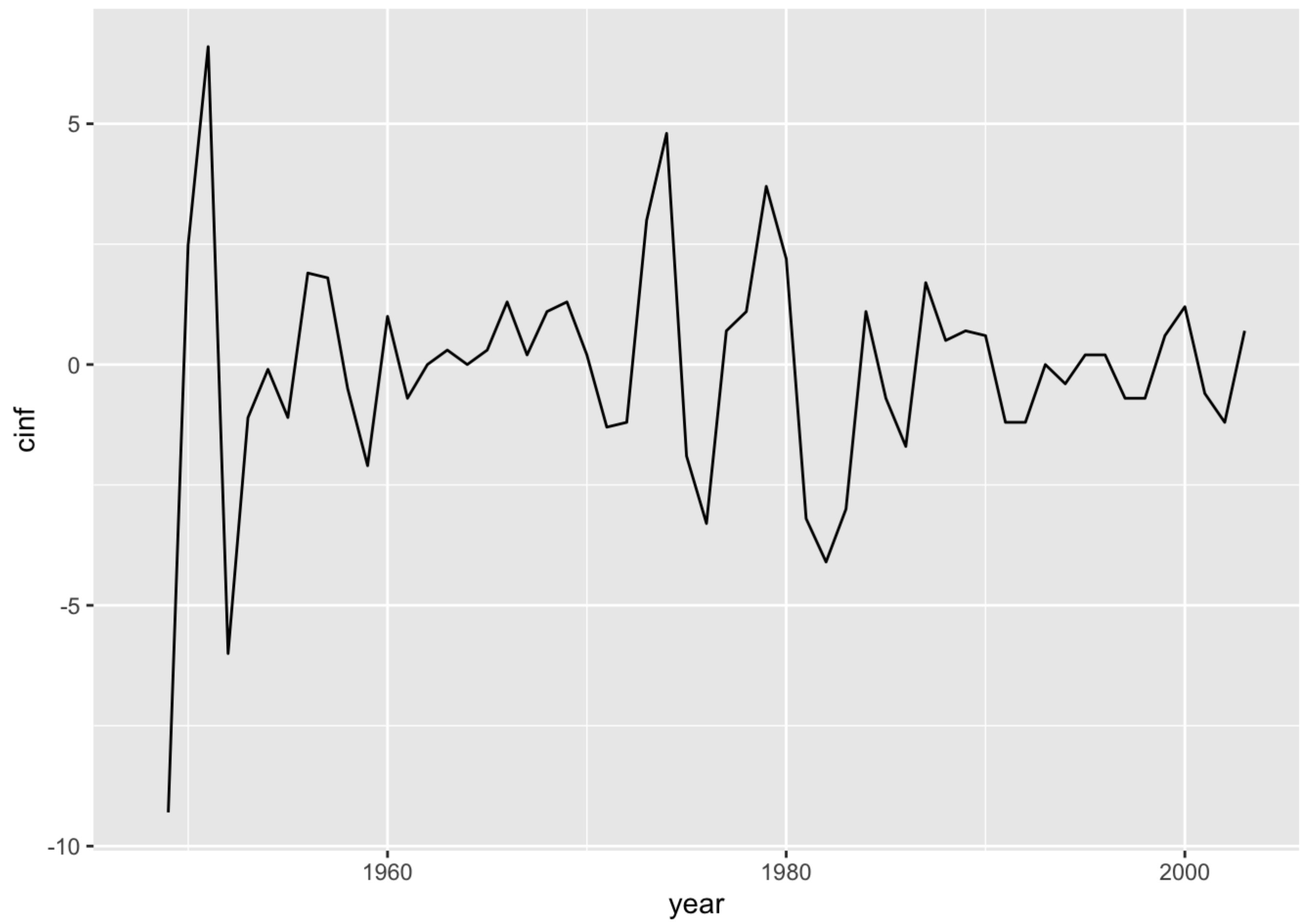


```
ggplot2::ggplot(data=phillips_rdgp) + ggplot2::geom_line(aes(x=year,y=unem))
```



```
ggplot2::ggplot(data=phillips_rdgp) + ggplot2::geom_line(aes(x=year,y=cinf))
```

```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```



Formally

Estimation and output results for model 1

Non-robust

```
summary(model_ph, robust=FALSE)
```

```
##
## Call:
## lm(formula = phillips_rdgp$inf ~ phillips_rdgp$unem)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2176 -1.7812 -0.6659  1.1473  8.8795
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.0536     1.5480   0.681   0.4990
## phillips_rdgp$unem   0.5024     0.2656   1.892   0.0639 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.972 on 54 degrees of freedom
## Multiple R-squared:  0.06215,    Adjusted R-squared:  0.04479
## F-statistic: 3.579 on 1 and 54 DF,  p-value: 0.06389
```

Robust (HAC) (https://www.r-econometrics.com/methods/hcrobusterrors/) t-testing

```
# Robust t test
coeftest(model_ph, vcov = vcovHC(model_ph, type = "HC0"))
```

```
##
## t test of coefficients:
##
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.05357    1.35491  0.7776  0.44020
## phillips_rdgp$unem  0.50238    0.24346  2.0635  0.04388 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Estimation and output results for model 2

Non-robust

```
summary(model_phc, robust=FALSE)
```

```
##
## Call:
## lm(formula = phillips_rdgp$cinf ~ phillips_rdgp$unem)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0741 -0.9241  0.0189  0.8606  5.4800
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.8282     1.2249   2.309   0.0249 *
## phillips_rdgp$unem  -0.5176     0.2090  -2.476   0.0165 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.307 on 53 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.1037, Adjusted R-squared:  0.08679
## F-statistic: 6.132 on 1 and 53 DF,  p-value: 0.0165
```

Robust (HAC) (https://www.r-econometrics.com/methods/hcrobusterrors/) t-testing

```
# Robust t test
coeftest(model_phc, vcov = vcovHC(model_phc, type = "HC0"))
```

```
##
## t test of coefficients:
##
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.82820    1.42601  1.9833  0.05253 .
## phillips_rdgp$unem -0.51765    0.22758 -2.2746  0.02700 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
knitr::knit_exit()
```