# Exam Econometrics (MSB104)

# SOLUTION PROPOSAL

## **Subject code**: MSB104

## **Date of exam**: 30.11.2020

## **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.
- 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).

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

You have been hired by “Tromb AS” –a local landlord– who rents out apartments in Haugesund, Stavanger and Bergen. The rental market is highly competitive and therefore Tromb is eager to price his apartments just right. Your task is to carry out a statistical analysis of the *rental* market for Tromb.

The Tromb-business is family run, and no-one have any formal education in economics or econometrics, so you must be careful to explain your results.

# A: Build you own model

i. Carefully specify a good *econometric model* of the actual monthly price of an apartment in Haugesund today. The model must be linear in parameters and it must include exactly five $x$-variables. Assume that any cross-sectional data, that you want, is available to you, but for Haugesund only. Remember to explain the unit of measurement for each of your variables (e.g. `size’ measures the size of the apartment in square metres).

ii. Briefly discuss each element of your equation and explain what sign ($\pm$) you expect on each $\hat{\beta}_j$, if you could run the regression. Also explain if you expect the actual partial effect of $\beta_j$ to be linear, although you perhaps did not include any higher order polynomial functions for the particular $x_j$ to capture non-linearity.

iii. Now, first explain generally what multicollinearity is, and what its consequences are for the variance and bias (4-5 typed lines will be sufficient). Next, in the context of the model you have written, do you think that you have problems with multicollinearity (you only have to give one example)?

iv. First, explain generally what heteroskedasticity is, and what the consequences are for variance and bias in the OLS estimators of $\beta_j$ (6-7 typed lines). Second, in the context of your model, do you think that you have problems with heteroskedasticity (one example will suffice) and, if you do, how would you fix it?

# B: Interpretation and inference in a basic model

Table 1 in the Appendix shows the regression output from **R** using a random cross-section that “Tromp AS” has provided. The sample was collected from finn.no and contains 70 rental apartments located in Haugesund, Stavanger and Bergen. The multiple regression model was the following:

$$\texttt{price}_i = \beta_0 + \beta_1 \cdot \texttt{kvm}_i + \beta_2 \cdot \texttt{HGSD}_i + \beta_3 \cdot \texttt{etasje}_i + u_i \tag{1}$$

The $x$-variables have the following units: `kvm` is size measured in square metres, `HGSD` is a dummy variable equal to 1 if the city is Haugesund and zero otherwise, and `etasje` is the location of the apartment (0 is basement, 1 is ground floor, and so on). $\beta_0$ is a constant and $u_i$ is an error term. `price` is measured in 1000 kr. and it is the asking price.

Now, complete the following assignments:

a. Give a careful interpretation of the estimates and their units in Table 1.

b. Is the price really lower in Haugesund? State your hypotheses and give at careful interpretation, as well as a non-technical conclusion that Tromb will understand.

c. What is the approximate p-value of your test above and what does it it mean (use Table G.2 in the book to give your answer)?

d. Calculate a 95% confidence interval for $\beta_2$. Carefully interpret your findings, so that a layman will understand them.

e. Now, assume the existence of an unobserved variable $\mathtt{km}$, which measures the distance to the city centre for an apartment, and assume that the correlation between $\mathtt{km}$ and $\mathtt{HGSD}$ is strictly negative, i.e. $\mathtt{Corr}(\mathtt{km}, \mathtt{HGSD}) < 0$. First, explain what you think will happen to $\hat{\beta}_2$ if you were to include this new variable $\mathtt{km}$ in the regression. Second, explain why it is unreasonable to assume that $\beta_2$ has a *causal* interpretation in equation (1).

f. Finally, discuss the goodness-of-fit as given in Table 1 for the purpose of prediction, i.e. $\mathbf{E}\left(\mathtt{price} \mid \mathbf{x}\right)$, as well as inference for any particular $\beta_j$, i.e. $\partial\,\mathtt{price}/\partial\beta_j$.

# C: Interpretation and inference in a log-level model

Now you run a new econometric model based on the same data as in Question B. The model is given below in equation (2), and the results are given in Table 2 in the Appendix.

$$\log(\mathtt{price}_i) = \beta_0 + \beta_1 \cdot \mathtt{kvm}_i + \beta_2 \cdot \mathtt{HGSD}_i + \beta_3 \cdot \mathtt{etasje}_i + \beta_4 \cdot \mathtt{rom}_i + u_i \tag{2}$$

All the $x$-variables are the same as stated above except that $\mathtt{rom}$ is added. This variables measures the number of bedrooms in the apartment.

i. Interpret the estimates for $\beta_1$ and $\beta_4$. Compare $R^2$ of this model with the $R^2$ of 0,35 from the model in equation (1).

ii. Test the following joint null hypothesis $H_0$: $\beta_1 = 0$ and $\beta_4 = 0$. You may find it usefull to know that the mean of the squared residuals (i.e. $\frac{1}{70} \cdot \sum \hat{u}_i^2$) from the regression in equation (2) is 0,0509, while the mean of the squared residuals from a regression of $\log(\mathtt{price})$ on $\mathtt{HGSD}$ and $\mathtt{etasje}$ (plus a constant) is 0,0684. Do you reject the null hypothesis at the 1% level?

iii. Now, compare the econometric model in equation (2) with an otherwise identical model, execpt you do not include $\mathtt{rom}$. Explain what happens to the sampling variance of the OLS slope estimator $\hat{\beta}_1$ when you add $\mathtt{rom}$ to the model. Note that the correlation between $\mathtt{rom}$ and $\mathtt{kvm}$ is 0,72. You may relate the answer to the components of the variance formula

$$\texttt{Var}\left(\hat{\beta}_j\right) = \frac{\sigma^2}{\texttt{SST}_j \cdot \left(1 - R_j^2\right)}$$

iv. Assume that the correlation between `rom` and `HGSD` is zero. What is the implication for $\texttt{Var}\left(\hat{\beta}_2\right)$ if we add `rom` to the econometric model?

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

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

# Small questions

1. In a realized set of time series observations, would an arbitrary reordering of the observations impact on the estimated results?
2. Imagine a situation in which you need to capture in a time series regression model the economic impact of a lockdown to the society from Covid-19. What type of time series variable would you include to capture this effect?
3. In many time series model applications, it seems reasonable to assume that the TS'.3 assumption that $E(u_t|X_t) = 0$ holds, but not $E(u_t|X) = 0$. Explain what is the difference between these two assumptions.
4. Compared table 1 to to 2, In addition to $E(u_t|X_t) = 0$, what additional requriments are needed for the estimators to be consistently estimated?
5. Assume that you have estimated a time series model for the future interest rate setting with an uncertainty band. What impact do serial correlation have on these results
6. Why do find economic researchers often find it convenient to not adjust for serial correlation in the error term, but rather present the results based on robst HAD standard error
7. What is meant by simultaneity and in what way violated OLS assumption. Name a couple of examples how this occur in a time series context.

# Stochastical regression models

We have the following finite distributed lagged model:

$$y_t = \alpha + \delta_0 z_t + \delta_1 z_{t-1} + u_t \tag{3}$$

1. Show that

a. Temporary change by 1 in period t can be written as:

- period t : $\delta_0$
- period t+1 : $\delta_1$
- period t+2: $0$

b. Permanent change by 1 in period t

- period t : $\delta_0$
- period t+1 : $\delta_0 + \delta_1$
- period t+2: $\delta_0 + \delta_1$

3. A random walk model without and with drift is given as

$$y_t = y_{t-1} + e_t \tag{6}$$

$$y_t = \alpha_0 + y_{t-1} + e_t \tag{7}$$

Where $e_t \sim i.i.d. N(0, \sigma_e^2)$

4. For (1), (2) find its (i) expected value, (ii) variance and (iii) forecast 2-periods ahead

5. If you were given the task of forecasting stock market return for (a) 20 days ahead and (b) 20 year head. Which of the two model would you select for the two tasks and why?

# Application: Effects of personal exemption on fertility rates (cf. appendix for part II)

a. Explain what is the difference between the two model specifications

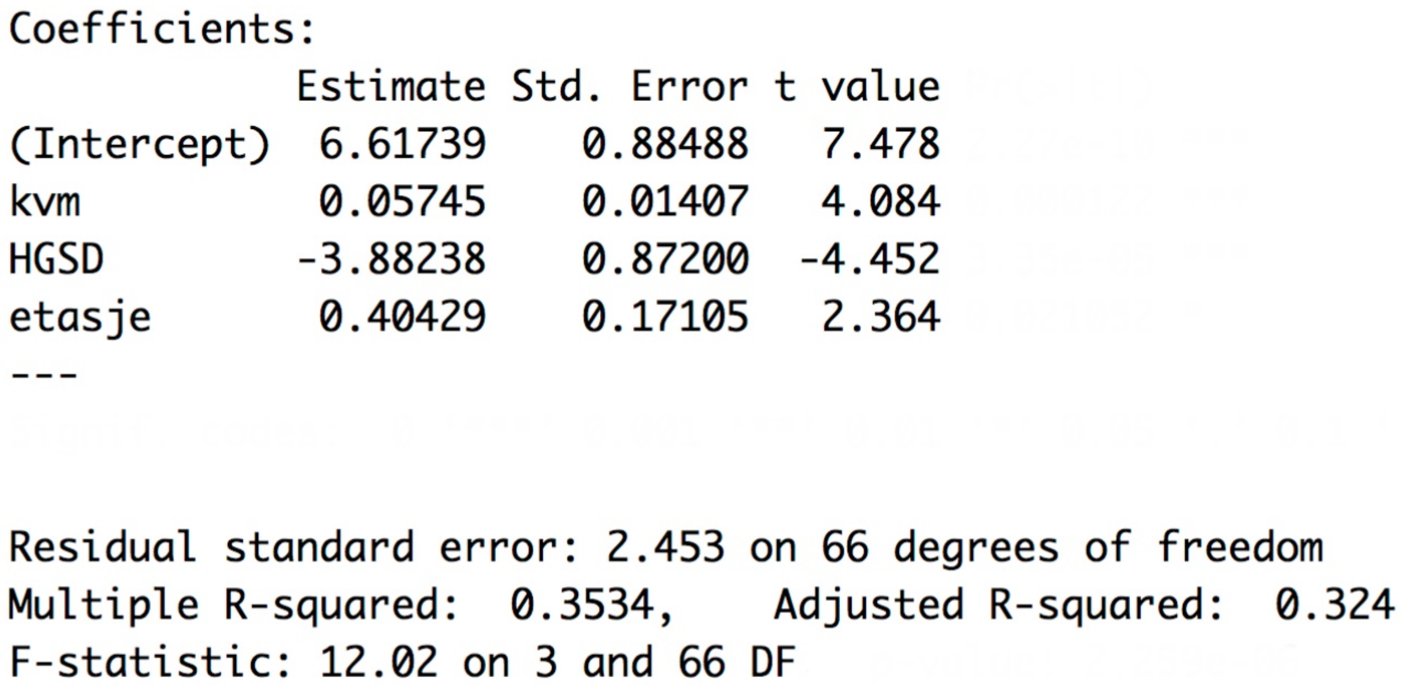
b. In the diagnostic part of checking the time series model, provide a list of of the different issues (no explanations) that could violate either the TS.1 or TS'2 assumptions.

c. Under the diagnostic part, the graphical analysis suggests that the problem of … solved under the model specificaiton 2. However, the model specification can be improved upon by doing one thing. Any suggestions

d. Interpret the result from the coefficients estimation and explain why the standard deviation differ in the two cases?

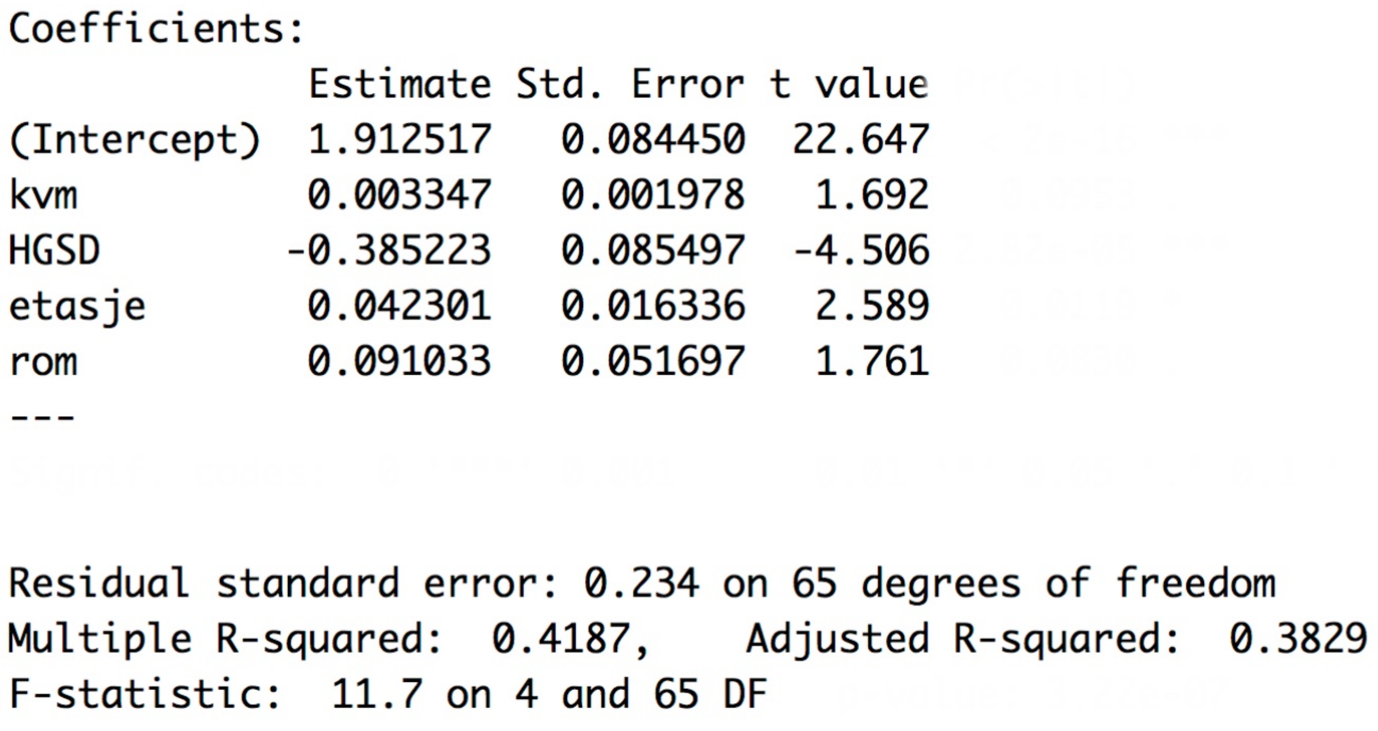
# Appendix

## Part I



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.61739    0.88488   7.478 2.27e-10 ***
kvm          0.05745    0.01407   4.084 0.000122 ***
HGSD        -3.88238    0.87200  -4.452 3.35e-05 ***
etasje       0.40429    0.17105   2.364 0.021052 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.453 on 66 degrees of freedom
Multiple R-squared:  0.3534,	Adjusted R-squared:  0.324 
F-statistic: 12.02 on 3 and 66 DF,  p-value: 2.258e-06
```



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.912517   0.084450  22.647  < 2e-16 ***
kvm          0.003347   0.001978   1.692   0.0953 .  
HGSD        -0.385223   0.085497  -4.506 2.82e-05 ***
etasje       0.042301   0.016336   2.589   0.0119 *  
rom          0.091033   0.051697   1.761   0.0830 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.234 on 65 degrees of freedom
Multiple R-squared:  0.4187,	Adjusted R-squared:  0.3829 
F-statistic:  11.7 on 4 and 65 DF,  p-value: 3.22e-07
```

# Part II

## 1. Data sample (realized DGP)

```
# Loading data
rm(list=ls())
library(sandwich)
library(lmtest)
library(wooldridge)
library(ggplot2)
library(plotly)
fertility_rdgp <- fertil3 # Realized DGP
```

From 1913 to 1984

## 2. Data generating process (DGP) and its regression model

Model specification 1:

$$gfr_t = \alpha_0 + \delta_0 pe_t + \delta_1 pe_{t-1} + \delta_2 pe_{t-1} + u_t, t = 1, 2, \ldots, n \tag{8}$$

```
model_ferd <- lm(fertility_rdgp$gfr ~ fertility_rdgp$pe+ fertility_rdgp$pe_1+fertility_r
dgp$pe_2)
```

Model specification 2:

$$\Delta gfr_t = \alpha_0 + \delta_0 \Delta pe_t + \delta_1 \Delta pe_{t-1} + \delta_2 \Delta pe_{t-1} + u_t, t = 1, 2, \ldots, n \tag{9}$$

```
model_ferdc <- lm(fertility_rdgp$cgfr ~ fertility_rdgp$cpe +fertility_rdgp$cpe_1+fertili
ty_rdgp$cpe_2)
```

## 3. Model estimation

```
ols_ferd <- summary(model_ferd)
ols_ferdc <- summary(model_ferdc)
```

## 4. Diagnostics

### Detecting

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

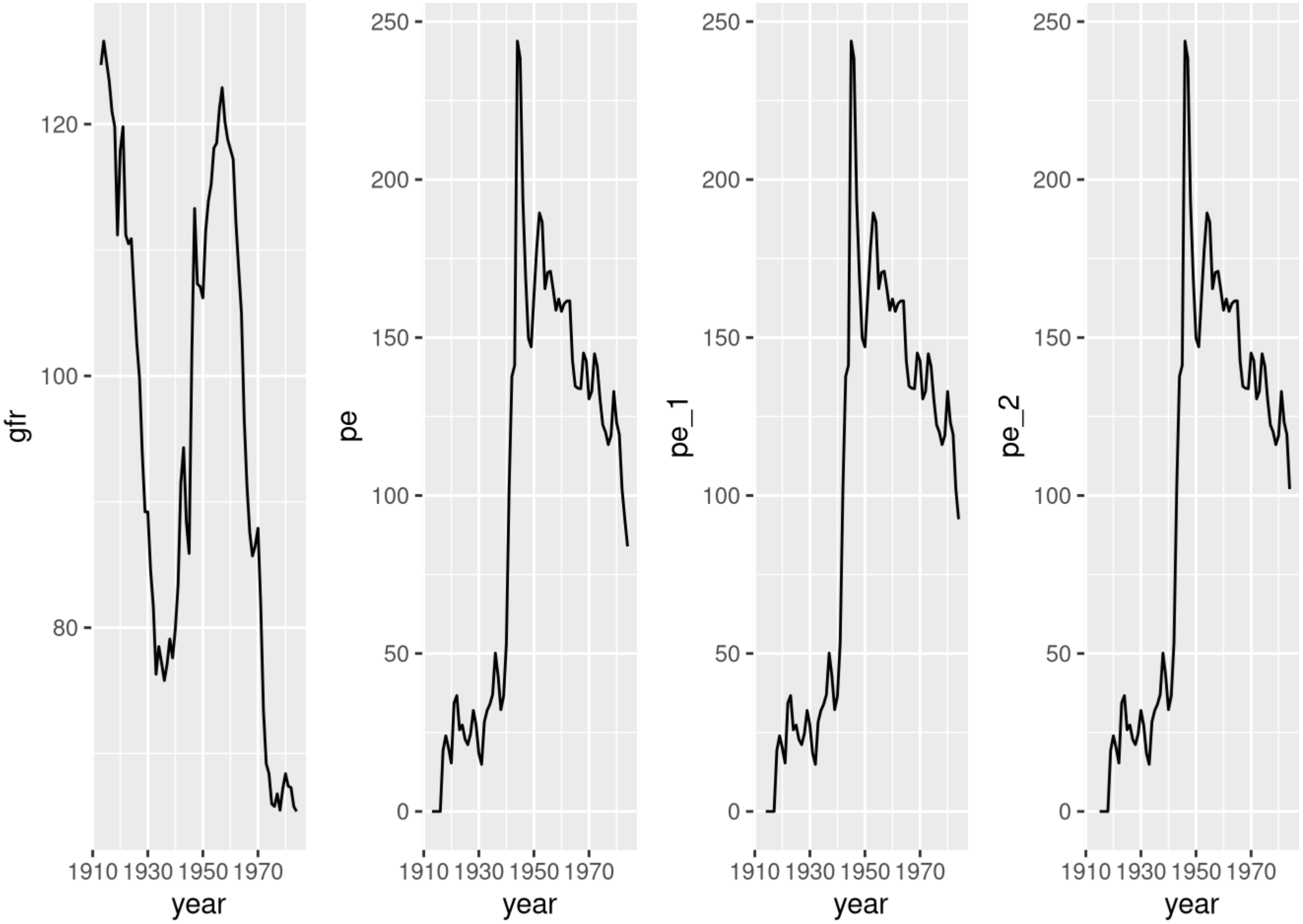
##### Graphical

Model specification 1:

```
p1 <- ggplot2::ggplot(data=fertility_rdgp) + ggplot2::geom_line(aes(x=year,y=gfr))
p2 <- ggplot2::ggplot(data=fertility_rdgp) + ggplot2::geom_line(aes(x=year,y=pe))
p3 <- ggplot2::ggplot(data=fertility_rdgp) + ggplot2::geom_line(aes(x=year,y=pe_1))
p4 <- ggplot2::ggplot(data=fertility_rdgp) + ggplot2::geom_line(aes(x=year,y=pe_2))
gridExtra::grid.arrange(p1, p2, p3, p4,nrow = 1)
```

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

```
## Warning: Removed 2 row(s) containing missing values (geom_path).
```



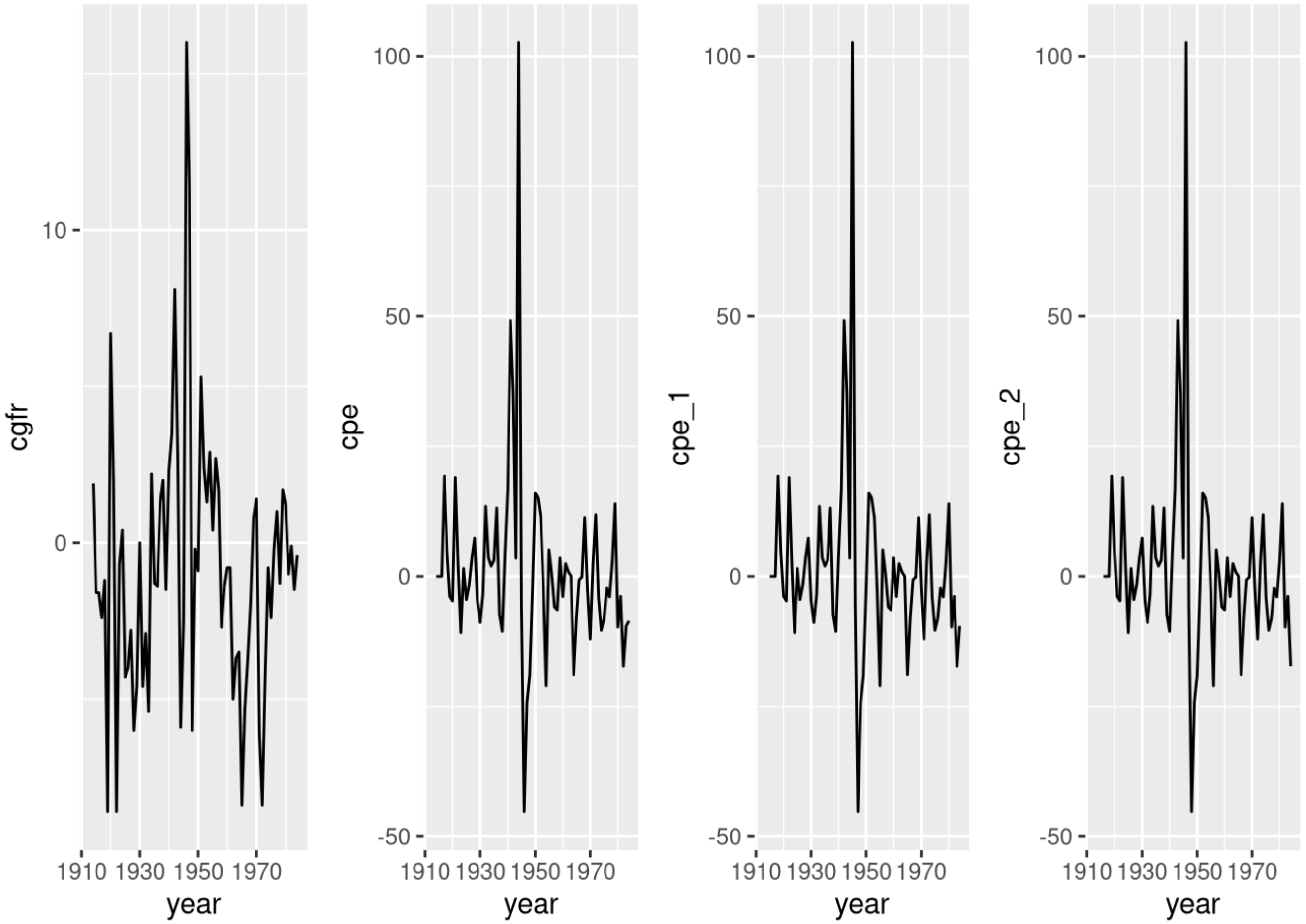
Model specification 2:

```
cp1 <- ggplot2::ggplot(data=fertility_rdgp) + ggplot2::geom_line(aes(x=year,y=cgfr))
cp2 <- ggplot2::ggplot(data=fertility_rdgp) + ggplot2::geom_line(aes(x=year,y=cpe))
cp3 <- ggplot2::ggplot(data=fertility_rdgp) + ggplot2::geom_line(aes(x=year,y=cpe_1))
cp4 <- ggplot2::ggplot(data=fertility_rdgp) + ggplot2::geom_line(aes(x=year,y=cpe_2))
gridExtra::grid.arrange(cp1, cp2, cp3, cp4,nrow = 1)
```

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

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

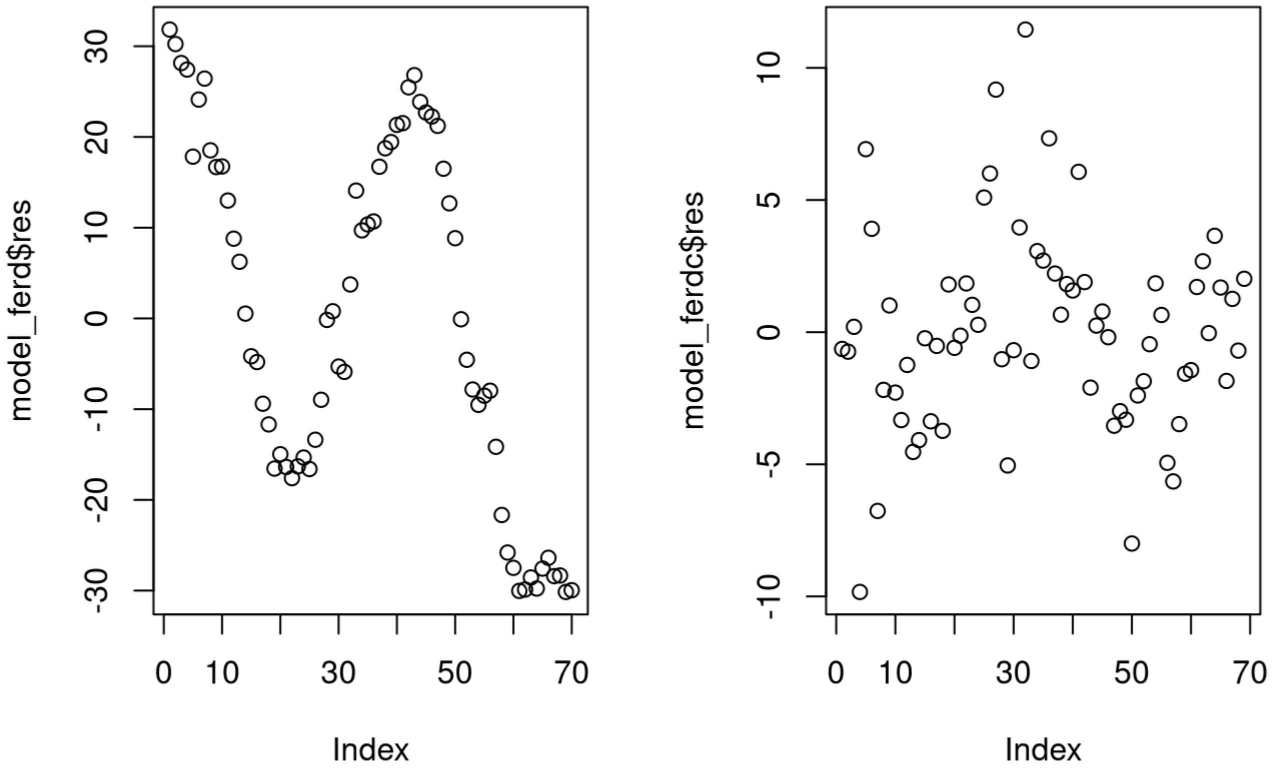
```
## Warning: Removed 2 row(s) containing missing values (geom_path).
## Warning: Removed 3 row(s) containing missing values (geom_path).
```



Formally

### Serial correlation

```
par(mfrow=c(1,2))
plot(model_ferd$res)
plot(model_ferdc$res)
```



Option 1: t-test AR(1) on model 1

```
## Serial correlation with strictly exogenous regressors
#### Step 1: Find the estimated residual
resval <- model_ferd$res
#### Step 2: Estimat the estimated residual on itself
model_ar <- lm(resval~lag(resval))
#### Step 3: Show the results from the estimation
summary(model_ar)
```

```
## Warning in summary.lm(model_ar): essentially perfect fit: summary may be
## unreliable
```

```
## 
## Call:
## lm(formula = resval ~ lag(resval))
## 
## Residuals:
##        Min         1Q     Median         3Q        Max 
## -6.285e-14 -6.800e-17  9.130e-16  1.805e-15  9.743e-15 
## 
## Coefficients:
##              Estimate Std. Error   t value Pr(>|t|)    
## (Intercept) 1.509e-30  9.483e-16 0.000e+00        1    
## lag(resval) 1.000e+00  4.936e-17 2.026e+16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 7.934e-15 on 68 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1 
## F-statistic: 4.104e+32 on 1 and 68 DF,  p-value: < 2.2e-16
```

Option 2: Testing for serial correlation with general regressors on model 1

```
## Perform Breusch-Godfrey test for first-order serial correlation:
bgtest(fertility_rdgp$gfr ~ fertility_rdgp$pe_1+fertility_rdgp$pe_2)
```

```
## 
##  Breusch-Godfrey test for serial correlation of order up to 1
## 
## data:  fertility_rdgp$gfr ~ fertility_rdgp$pe_1 + fertility_rdgp$pe_2
## LM test = 64.287, df = 1, p-value = 1.076e-15
```

# 7. Estimation and output results

## Non-robust

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

```
##
## Call:
## lm(formula = fertility_rdgp$cgfr ~ fertility_rdgp$cpe + fertility_rdgp$cpe_1 +
##     fertility_rdgp$cpe_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8307 -2.1842 -0.1912  1.8442 11.4506
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -0.96368    0.46776  -2.060  0.04339 *
## fertility_rdgp$cpe   -0.03620    0.02677  -1.352  0.18101
## fertility_rdgp$cpe_1 -0.01397    0.02755  -0.507  0.61385
## fertility_rdgp$cpe_2  0.10999    0.02688   4.092  0.00012 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.859 on 65 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.2325, Adjusted R-squared:  0.1971
## F-statistic: 6.563 on 3 and 65 DF,  p-value: 0.0006054
```

## Robust t-testing (HAC)

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

```
##
## t test of coefficients:
##
##                       Estimate Std. Error t value  Pr(>|t|)
## (Intercept)          -0.963679   0.450703 -2.1382   0.03627 *
## fertility_rdgp$cpe   -0.036202   0.033268 -1.0882   0.28053
## fertility_rdgp$cpe_1 -0.013971   0.032601 -0.4285   0.66968
## fertility_rdgp$cpe_2  0.109990   0.026082  4.2171 7.816e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```