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).
- Do you have questions to the exam text? Part I/II: 92345700/41611857 (Henrik/Jørn).

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).

Solution: One example may be

$$\begin{aligned}\texttt{price}_i = \beta_0 + \beta_1 \cdot \texttt{antall rom}_i + \beta_2 \cdot \texttt{bygget år}_i + \beta_3 \cdot \texttt{nylig oppusset}_i \\ + \beta_4 \cdot \texttt{etasje}_i + \beta_5 \cdot \texttt{rejsetid til sentrum}_i + u_i\end{aligned}$$

- **`antall rom` is equal to one if the apartment has only one room (i.e., a studio apartment), two if there is one bedroom and so on.**
- **`bygget år` is the calendar year that the house where the apartment sits was build.**
- **`nylig oppusset` is a dummy variable equal to one if the apartment is newly renovated, and zero otherwise.**
- **`etasje` is equal to zero if the apartment is located in the basement, one if it is on the ground floor, and so on.**
- **`rejsetid til sentrum` is the traveling time to the city centre (town hall) as measured by the fastest means of transportation (walk, bicycle or public bus).**

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 β_j to be linear, although you perhaps did not include any higher order polynomial functions for the particular x_j to capture non-linearity.

Solution: I expect the coefficient on `antall rom` to be positive, because more rooms makes the apartment more desirable (e.g. to students who want to split the rent), but also because apartments with more rooms are generally larger and therefor more costly. Maybe one ought to allow for diminishing effect when `antall rom` increases by including $\texttt{rom}^2$.

The estimated coefficient on `bygget år` ought to be negative because newer apartments are more desirable, so when this x-variable increase by one unit, the price is expected to increase.

But since old apartments may be modernized, I want to allow for a shift in the association between `bygget år` and `price` if the apartment is renovated. I expect that a renovation has a positive association with the price and therefore should the coefficient on `nylig oppusset` be positive.

`etasje`: I expect the coefficient to be positive. In reality the price increase may not be linear, but this is not captured in this model. This is so because an basement located in the basement may be disproportionally unattractive and, likewise, it does not make a huge difference if the apartment is located on the 6th or 7th floor, for example.

`rejsetid til sentrum`: I expect the coefficient to be negative, because a further distance makes the apartment less attractive. Here the true effect may also be diminishing.

My dependent variable `price` is the monthly price of an apartment—the actual price that the renter pays, and u is an unobserved error term that captures everything that I have not controlled for in my econometric model. Finally, β_0 is an intercept which captures the hypothetical price of an apartment where all my x-variables equals zero.

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)?

Solution: Multicollinearity occurs when you have a variable x_1 in your regression, and you add another variable x_2 that is strongly correlated with x_1. Depending on the degree of correlation the variance of the $\hat{\beta}_1$ may increase dramatically, but there are not consequences insofar $E(\hat{\beta}_1) = \beta_1$.

I do not think that I have any major problems with multicollinearity in my econometric model. If I should mention one case, then it would be `bygget år` and `rejsetid til sentrum` because it may be so, that newer apartments are located further away from the city centre (a positive correlation). Perhaps related, you could have a correlation between `rejsetid til sentrum` and `etasje`.

iv. First, explain generally what heteroskedasticity is, and what the consequences are for variance and bias in the OLS estimators of β_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?

Heteroskedasticity occurs when the conditional variance of the error term is not constant

$$\mathtt{Var}(u \mid x_1, x_2, \ldots, x_k) \neq \sigma^2$$

Heteroskedasticity does not cause bias in the OLS estimators, because we only used MLR.1–MLR4 to prove unbiasedness. MLR.5 (homoskedasticity) was used to ensure unbiasedness of $\mathtt{Var}(\hat{\beta}_j)$. Therefore, in the presence of heteroskedasticity, the variance of the OLS estimator is biased, and it is no longer valid for constructing CIs or t-statistics.

Yes, I may have problems with heteroskedasticity in my model, as the variation in `price` may increase as, say, the apartment increases in size, which I measure by the number of bedrooms. I could remedy this by using a heteroskedasticity robust variance estimator.

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:

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

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). β_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.

Solution: The asking price of an apartment (in 1000 kr.) increase by 0,057 if an apartment increase by one square metre. Since the unit of the y-variable is in 1000 kr. this means 57 kr. for each extra square meter or 570 kr. for 10 square metres (which, I believe, is about the size of a room).

If the apartment is located in Haugesund, then the price decrease by 3882 kr. (3,882 times 1000).

If the apartment is located on the first floor, as opposed to the ground floor, the price increase by 404 kr. (0,404 times 1000).

The constant is equal to 6,6 which means that a hypothetical apartment located in the basement, that has a size of 0 square metres, and is not in Haugesund costs 6600 kr.

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.

Solution: My null hypothesis is $H_0: \beta_{\texttt{Haugesund}} = 0$ versus $H_1: \beta_{\texttt{Haugesund}} < 0$

The t-statistic for $\hat{\beta}_{\texttt{Haugesund}}$ is -4,45, and the one-tailed 5% critical value with 66 degrees of freedom (70-3-1) is approximately -1,67 (I use the t-distribution table to find this value with df=60, which is the closest I come to 66). Therefore, I reject the null hypothesis.

This means that I am very certain that the price is lower in Haugesund—the uncertainty is only 5% in my test.

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)?

Solution: I can approximate the p-value using Table G.2. The critical value with 66 degrees of freedom associated with a 0,5% significance level (the lowest in the table) is -2,66. That means that the p-values is less than 0,5%. Compared to the outcome we observe (-3,88), the p-value is the probability of observing an outcome as extreme (or more), given H_0 is true, which is highly unlikely.

d. Calculate a 95% confidence interval for β_2. Carefully interpret your findings, so that a layman will understand them.

Solution: There are 66 degrees of freedom, and the formula for the confidence interval is

$$\hat{\beta}_2 \pm c \cdot \texttt{se}(\hat{\beta}_2)$$

The only thing I need to find in Table G.2. is c, which is the critical value associated with a 95% confidence interval. I use df=60 and a 2-tailed probability of 0,05, which gives me a critical value of 2,000. Therefore I get $-3,88 \pm 2,00 \cdot 0,872 = (-5,524; -2,136)$.

This means that in 95% of the cases (if I were to repeat the whole exercise), I would get an estimate of $\hat{\beta}_2$ contained in this interval. I could also state that I am 95% certain that the price of an apartment in Haugesund is between 2136 kr. and 5524 kr. lower than in other cities in the population.

e. Now, assume the existence of an unobserved variable `km`, which measures the distance to the city centre for an apartment, and assume that the correlation between `km` and `HGSD` is strictly negative, i.e. $\texttt{Corr}(\texttt{km}, \texttt{HGSD}) < 0$. First, explain what you think will happen to $\hat{\beta}_2$ if you were to include this new variable `km` in the regression. Second, explain why it is unreasonable to assume that β_2 has a *causal* interpretation in equation (1).

Solution: The slope parameter on `km` would be negative, because an apartment located further away from the city centre would be less attractive. Hence, I expect the bias in β_2 to be positive, and therefore I would expect an even lower price for an apartment in Haugesund if I were to include the omitted variable `km`.

The critical assumption is MLR.4. In equation (1) the error term contains the unobserved `km`, which is both correlated with `price` and `HGSD`. Therefore MLR.4 is violated and we have a case of omitted variable bias that inflates β_2.

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

Solution: The goodness-of-fit is measured by R^2 which is equal to 0,35. This means that 35% of the variation in prices are explained by the model. If I were to predict the price of an apartment, I think this number is too low—I would need to include more relevant x-variables, that help me make a more accurat prediction of the price. If the purpose is to make inference about β_j then the R^2 is not important. . . .

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.

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

All the x-variables are the same as stated above except that `rom` is added. This variables measures the number of bedrooms in the apartment.

i. Interpret the estimates for β_1 and β_4. Compare R^2 of this model with the R^2 of 0,35 from the model in equation (1).

Solution: The coefficient on the `kvm` variable tells us that the price increase by 3,3% if the apartment increase 10 square metres in size, ceteris paribus. The coefficient `rom` tells us that the price increase by 9% if the apartment increases 1 room in size, holding everything else fixed. The R^2 of this equation is not comparable to the R^2 of the former model, because the dependent variable has changes from level to the natural log.

ii. Test the following joint null hypothesis $H_0 \colon \beta_1 = 0$ 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 $\log(\texttt{price})$ on `HGSD` and `etasje` (plus a constant) is 0,0684. Do you reject the null hypothesis at the 1% level?

Solution: The F-statistic is given by

$$\frac{\left(\texttt{SSR}_r - \texttt{SSR}_{ur}\right)/q}{\texttt{SSR}_{ur}/(n-k-1)}$$

In this case the ur denotes the model that includes the variables in the null hypothesis and $\texttt{SSR}_{ur}$ is calculated as n times the mean of the squared residuals: $0.0509 \cdot 70 = 3.56$. Likewise $\texttt{SSR}_r = 0.0684 \cdot 70 = 4.79$. $q = 2$ is the number of exclusion restrictions, which is given by our null hypothesis. So we get:

$$\frac{(4.79 - 3.56)/2}{3.56/(70 - 4 - 1)} = 11,1$$

As there are two degrees of freedom in the numerator and 65 degrees of freedom in the denominator, we get a critical value of about 4.98 using Table G.3c (the 1% level) for the F distribution. The test-statistic clearly falls in the reject region, which is why we reject the null hypothesis and conclude that at least one of the parameters listen in our null hypothesis is different from zero.

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

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

Solution: This is clearly a case of multicollinearity between `rom` and `kvm`. The R_j^2 comes from a regression of `kvm` on `rom` and the other x-variables. As you add this highly correlated variable `rom` to the auxiliary regression, the R_j^2 increase (probably dramatically) which is why term in the denominator of the variance $(1 - R_j^2)$ decrease (i.e. gets closer to zero), and hence the variance of $\hat{\beta}_{\texttt{kvm}}$ increase.

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?

Solution: Nothing should happen to the variance then, as the covariance is zero . . .

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 a realized set of time series observations, would an arbitrary reordering of the observations have any impact on the estimated results?

Solution: Yes, the nature of time series is based upon the notion of temporal ordering of observations.

2. Facing a situation in which you for a time series regression model need to capture the economic impact of a lockdown due to Corona-virus. What type of time series variable would you use to capture such an effect?

Solution: Dummy variable can be used to capture the state of a lockdown. But since an index number can aggregate vast amount of information in a single quantity (e.g., school-closing, shop-closing, work-at-home etc.), this type of variable is if accessiblethe preferred one.

3. For many time series model applications, it is reasonable to assume that TS'.3 assumption $E(u_t|X_t) = 0$ holds, but not $E(u_t|X) = 0$. Explain the main difference between these two assumptions.

Solution: In the first the error term only conditioned on that u_t is uncorrelated with the explanatory variables in our specified model. For the later, it must be uncorrelated with explanatory variables in all previous and future periods

4. It is more common for OLS-estimators in a regression model to be consistent than unbiased, since we can replace the assumption of TS.3 with the TS'.3. However, there are some additional requirements needed for this to be the case. Can you name them?

Solution: (1) Need a large sample. The time series in the regression mode need to be (1) stationary and (2) weakly dependent.

5. Let's say you have estimated a macroeconomic model for the purpose of forecasting (i.e., provided point estimates with error bands) the future path of the development of short term interest rates. What impact would the detection of serial correlation have on these results?

Solution: (1) Point estimate correct, (2) Error bands wrong (most likely too narrow)

6. Why do economic researchers often find it more convenient not to adjust for serial correlation in the error term in a regression model, but rather present the results based on robust HAC standard errors?

Solution: Correcting for serial correlation could in practice be difficult since it is based on the assumption of (1) strictly exogenous regressors and (2) that the error term it follows an AR(1) process.

7. What is meant by simultaneity and how does it violate the OLS classical assumption? Give an example how simultaneity can occur in the context of a time series regression model.

Solution: Simultaneity occurs if the dependent variable jointly determine with at least one of the exogenous variable in the regression model. Violates TS.3/TS'.3. - Example: (1) Police force and murder rate in a city. (2) Exchange rate and interest rate setting by the central bank.

Stochastical regression models

We have the following finite distributed lagged model:

$$y_t = \alpha + \theta_0 z_t + \theta_1 z_{t-1} + u_t \tag{3}$$

1. Show that

a. Temporary change by 1 in period t would imply changes equal to:

- period t : θ_0
- period t+1 : θ_1
- period t+2: 0

Temporary change: (c+1) in period t, otherwise c

$$y_{t-1} = \alpha + \theta_0 c + \theta_1 c + u_t$$
$$y_t = \alpha + \theta_0 (c+1) + \theta_1 c + u_t$$
$$y_{t+1} = \alpha + \theta_0 c + \theta_1 (c+1) + u_t$$
$$y_{t+2} = \alpha + \theta_0 c + \theta_1 c + u_t$$

Which implies:

$$y_t - y_{t-1} = \theta_0$$
$$y_{t+1} - y_{t-1} = \theta_1$$
$$y_{t+2} - y_{t-1} = 0$$

b. Permanent change by 1 in period t would imply change equal to

- period t : θ_0
- period t+1 : $\theta_0 + \theta_1$
- period t+2: $\theta_0 + \theta_1$

Permanent change: change + 1 in period t and onwards, otherwise c:

$$y_{t-1} = \alpha + \theta_0 c + \theta_1 c + u_t$$
$$y_t = \alpha + \theta_0 (c+1) + \theta_1 c + u_t$$
$$y_{t+1} = \alpha + \theta_0 (c+1) + \theta_1 (c+1) + u_t$$
$$y_{t+2} = \alpha + \theta_0 (c+1) + \theta_1 (c+1) + (c+1) + u_t$$

Which implies:

$$y_t - y_{t-1} = \theta_0$$
$$y_{t+1} - y_{t-1} = \theta_0 + \theta_1$$
$$y_{t+2} - y_{t-1} = \theta_0 + \theta_1 \text{long-run multiplier}$$

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

$$y_t = y_{t-1} + e_t \tag{4}$$

$$y_t = \beta_0 + y_{t-1} + e_t \tag{5}$$

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

For (4), (5) find its (i) expected value, (ii) variance and (iii) forecast h-periods ahead

Random walk (without drift):

$$y_t = y_{t-1} + e_t$$

$$(i)$$

$$E(y_t) = E(y_{t-2} + e_{t-1} + e_t) = E(y_0 + e_1 + e_2 \ldots + e_t) = E(y_0) = y_0$$

$$(ii)$$

$$Var(y_t) = Var(y_0 + e_1 + e_2 \ldots + e_t) = t\sigma_e^2$$

$$(iii)$$

$$E(y_{t+h}|y_t) = E(y_{t+h} + e_{t+h}|y_t) = E(y_t + e_{t+h} + e_{t+h-1} + \ldots + e_t|y_t) = y_t$$

Random walk with drift:

$$y_t = \beta_0 + y_{t-1} + e_t$$

$$(i)$$

$$E(y_t) = E(\beta_0 + \beta_0 + y_{t-2} + e_{t-1} + e_t) = E(t\beta_0 + y_0 + e_1 + e_2 \ldots + e_t) = E(t\beta_0 + y_0) = t\beta_0 + y_0$$

$$(ii)$$

$$Var(y_t) = Var(ty_0 + e_1 + e_2 \ldots + e_t) = t\sigma_e^2$$

$$(iii)$$

$$E(y_{t+h}|y_t) = E(\beta_0 + y_{t+h} + e_{t+h}|y_t) = E(h\beta_0 + y_t + e_{t+h} + e_{t+h-1} + \ldots + e_t|y_t) = h\beta_0 + y_t$$

3. Given the task of modeling the stock market return for (i) 20 days ahead and (ii) 20 years ahead. Which version of the random walk models above would you employ for case (i) and case (ii)?

20 days ahead $\Rightarrow$ Random walk without drift, due to zero growth rate prediction in the short run. 20 years ahead $\Rightarrow$ Random walk with drift, due to expectation of long run positive growth in the economy.

Application: Effects of personal exemption on fertility rates

Cf. appendix part II for output information.

a. Explain the main difference between the two model specifications

Solution: Model 2 provides the variables in model 1 on differentiated form, representing changes in variables instead of levels.

b. In the diagnostic part of a time series regression model, which of the issues that are commonly looked for in a time series model are analyzed in the output and which one are left out?

Solution: 1. Omitted variables: (1) trend: left out (formal).
1. Whether the time series variable are persistent: included. 1. Serial correlation: included. 1. Heteroskedasticity: left out. 1. Simultaneity: left out.

c. In the diagnostic part from the output results, differentiating the variables (model specification 2) provides variables that are much less persistent. Still, there is clearly a spike for the variables during the second world war. What are the consequences of such a spike and how should it be taken care off?

Solution: The spike introduces heteroskedasticity (variance of y_t dependendt on time). Problem can probably be solved by introducing a dummy variable that captures the effect of the second world war.

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

Solution: - Significant: Only lag 2. Interpretation: Takes some time for changes in exemption to be transmitted to changes in fertility - Differ: Standard differ since HAC seeks to correct for the presence of serial correlation, usually be giving higher standard deviation values.

Appendix

Part I

Table 1

```
Coefficients:
            Estimate Std. Error t value
(Intercept)  6.61739    0.88488   7.478
kvm          0.05745    0.01407   4.084
HGSD        -3.88238    0.87200  -4.452
etasje       0.40429    0.17105   2.364
---


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
```

Table 2

```
Coefficients:
             Estimate Std. Error t value
(Intercept)  1.912517   0.084450  22.647
kvm          0.003347   0.001978   1.692
HGSD        -0.385223   0.085497  -4.506
etasje       0.042301   0.016336   2.589
rom          0.091033   0.051697   1.761
---


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
```

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:

$$gfr_t = \alpha_0 + \delta_0 pe_t + \delta_1 pe_{t-1} + \delta_2 pe_{t-1} + u_t, t = 1, 2, \ldots, n \tag{6}$$

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

Model specification 2:

$$\Delta gfr_t = \alpha_0 + \delta_0 \Delta pe_t + \delta_1 \Delta pe_{t-1} + \delta_2 \Delta pe_{t-1} + u_t, t = 1, 2, \ldots, n \tag{7}$$

```
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))

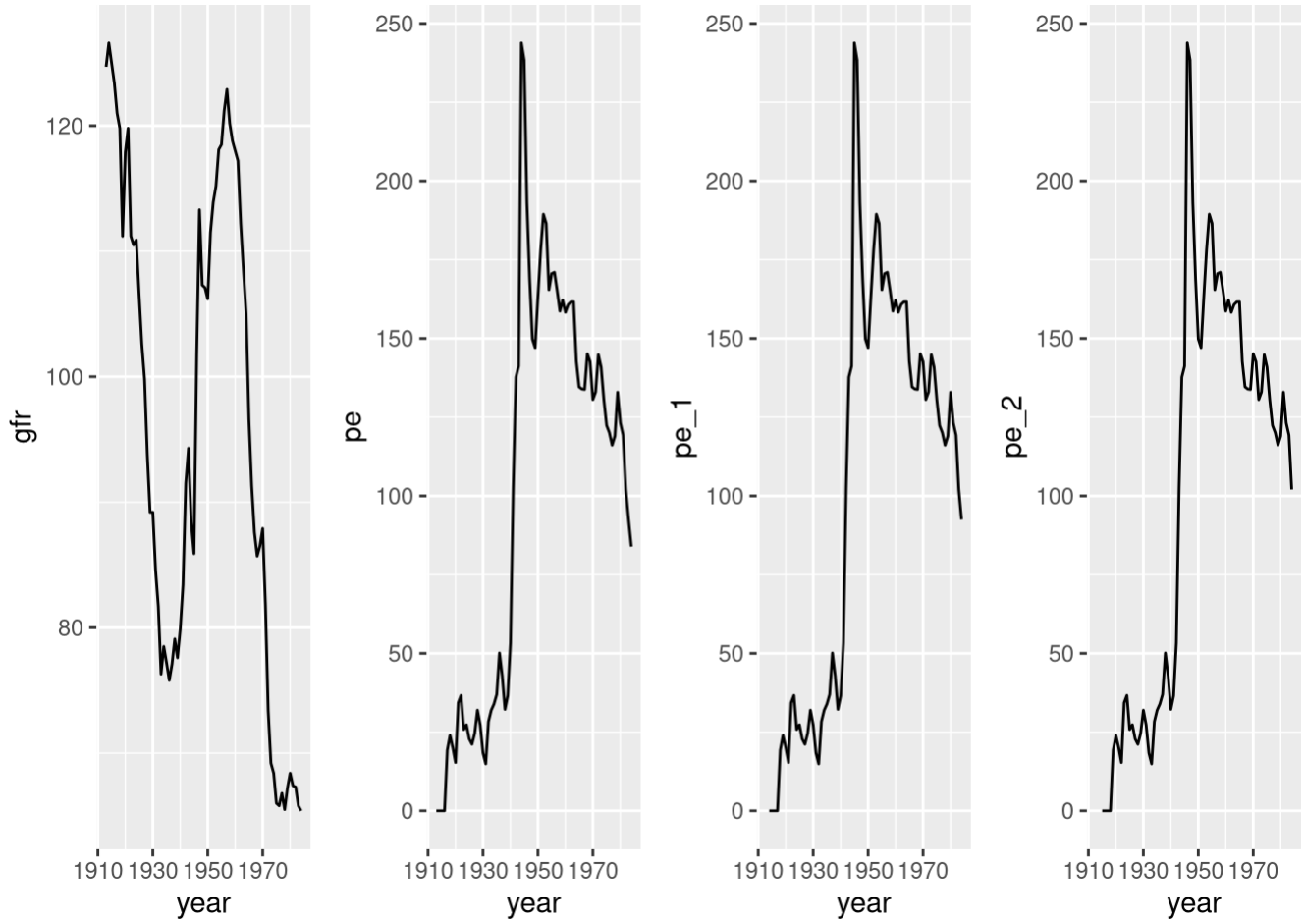
Graphical

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:

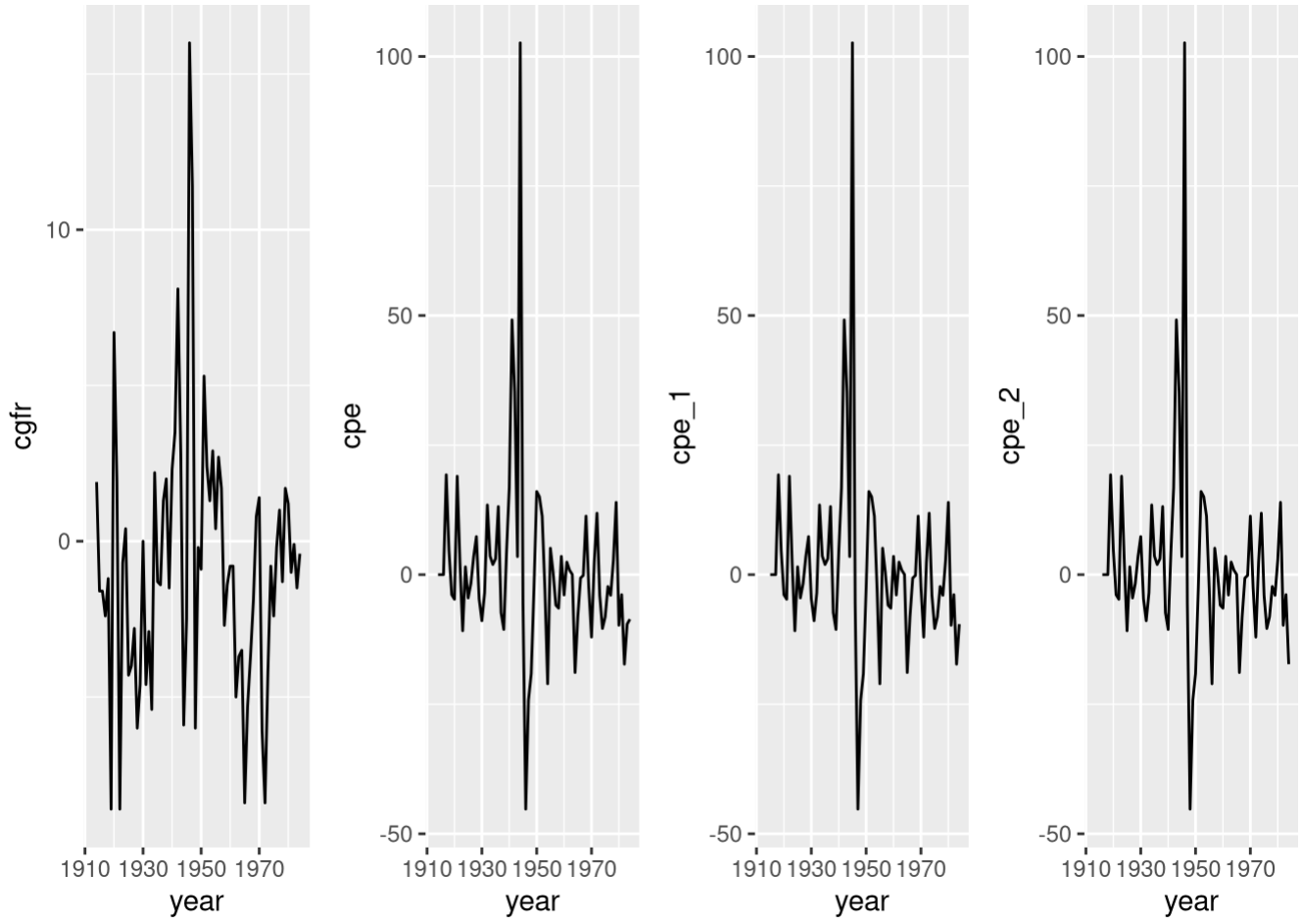
```
cp1 <- ggplot2::ggplot(data=fertility_rdgp) + ggplot2::geom_line(aes(x=year,y=cgfr))
cp2 <- ggplot2::ggplot(data=fertility_rdgp) + ggplot2::geom_line(aes(x=year,y=cpe))
cp3 <- ggplot2::ggplot(data=fertility_rdgp) + ggplot2::geom_line(aes(x=year,y=cpe_1))
cp4 <- ggplot2::ggplot(data=fertility_rdgp)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

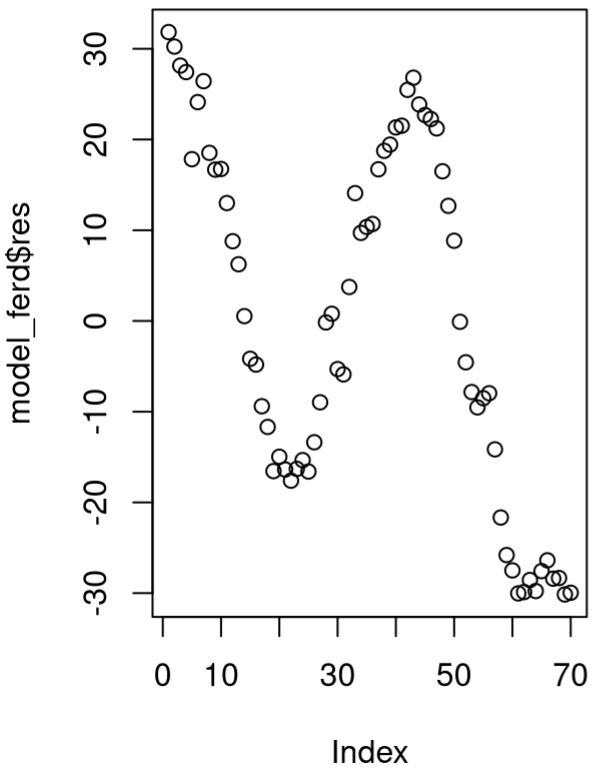
```
## Warning: Removed 3 row(s) containing missing values (geom_path).
```

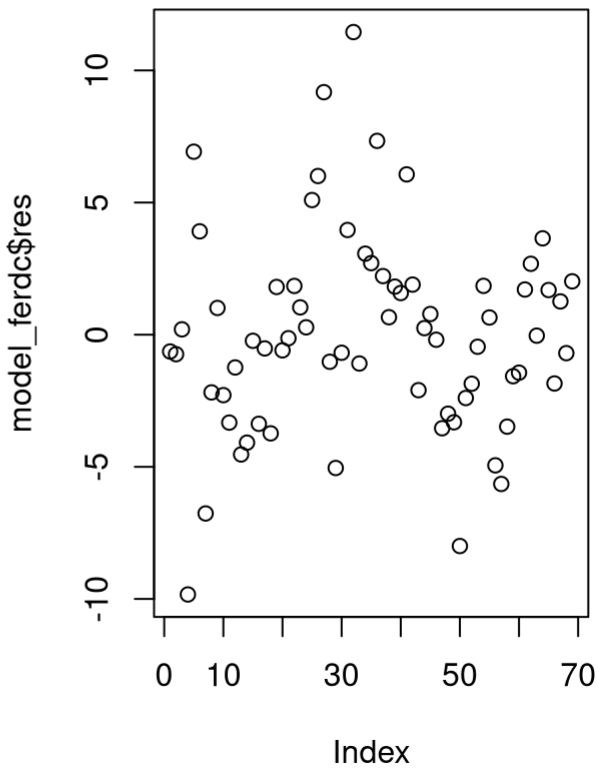


Formally

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
```