Technical University of Denmark

Written examination, May 16, 2019, 15:00-17:00.

Page 1 of 5 pages.

Course name: Introductory Econometrics.

Course number: 42008.

Aids allowed: All aids, except internet and any kind of communication.

Exam duration: 2 hours.

Weighting: Exercise 1: 40%, Exercise 2: 60%.

Read/skim through all questions first. Answer short and precisely and remember to move on if you are stuck at a question. Below, HN refers to our main textbook by David Hendry and Bent Nielsen (2007).

You can answer digitally and/or handwritten. Hence, a combination is allowed, but we recommend that you write your answers by hand, as there will be symbols and calculations.

Only answers in English are allowed.

Exercise 1: The AR(1) model (40%)

1. Consider the AR(1) model, $\mathcal{M}$, which, in regression equation form, is:

$$Y_t = \gamma_1 + \gamma_2 Y_{t-1} + u_t \text{ for } t = 1, 2, .., T \tag{1}$$

   where, for the conditional distribution of $u_1, u_2, ..., u_T$ given Y_0, we assume,

   i) **Independence:** $u_1, u_2, ..., u_T$ are mutually independent,

   ii) **Normality:** $u_t \sim N(0, \sigma^2)$,

   and,

   iii) **Parameter space:** $(\gamma_1, \gamma_2, \sigma^2)' \equiv \boldsymbol{\omega} \in \mathbb{R}^2 \times \mathbb{R}_+ \equiv \Omega$ ($'$denotes the transpose).

   As usual, $\mathcal{M}$ can also be written as a set of hypothetical joint data densities (conditional on Y_0), i.e. $\{\mathrm{f}_{\boldsymbol{\omega}}(y_1, y_2, ..., y_T \mid y_0), \boldsymbol{\omega} \in \Omega\}$. Suppose that the Data Generation Process (DGP) is described by the joint data density, $D(y_1, y_2, ..., y_T \mid y_0)$. What does it mean that $\mathcal{M}$ is correctly specified?

2. Explain the steps in deriving the log-likelihood function associated with $\mathcal{M}$ and obtaining the Maximum Likelihood Estimator (MLE), $\widehat{\boldsymbol{\omega}}$.

3. For each of the assumptions, i), ii) and iii), state relevant mis-specification analyses (tests and possibly graphical tests).

4. The MLEs $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$, are given by,

$$\widehat{\gamma}_1 = \overline{Y} - \widehat{\gamma}_2 \overline{Y}_{(-)} \text{ and } \widehat{\gamma}_2 = \frac{\sum_{t=1}^{T} Y_t (Y_{t-1} - \overline{Y}_{(-)})}{\sum_{t=1}^{T} (Y_{t-1} - \overline{Y}_{(-)})^2},$$

   where $\overline{Y} = T^{-1} \sum_{t=1}^{T} Y_t$ and $\overline{Y}_{(-)} = T^{-1} \sum_{t=1}^{T} Y_{t-1}$. Derive $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ based on an orthogonal reparameterization (see HN, §5.2.3) of $\mathcal{M}$. [Hint: Using the approach in HN, §5.2.3 or the slides from Lecture 5, you can start from rewriting eq. (1) as $Y_t = \gamma_1 X_{1,t} + \gamma_2 X_{2,t} + u_t$, where $X_{1,t} \equiv 1$, $X_{2,t} \equiv Y_{t-1}$.]

Exercise 2: Regression analysis of the RECS data (60%)

We consider the RECS data again. However, suppose we have a smaller subsample of the data set, consisting of 234 observations. This sample was randomly chosen from the original data set that we have studied during the course. As previously, we are interested in what demographic and economic variables matter for household electricity consumption. In Table 1, the variables we consider are described.

TABLE 1		
	R nomenclature	**Description**
Y	`LKWH.pers`	Logarithm of `KWH/NHSLDMEM` where `KWH` is total electricity usage in kwhs. For `NHSLDMEM` see below.
X_2	`NET`	Indicator: `net` $= 1$ if internet access at home, otherwise `net` $= 0$.
X_3	`SMART`	Indicator: `smart` $= 1$ if respondent has viewed smart meter interval data, otherwise `smart` $= 0$.
X_4	`NHSLDMEM`	Number of household members.
X_5	`MONEYPY`	Annual gross household income for the last year (1-8).
X_6	`HHSEX`	Indicator: `HHSEX` $= 1$ if respondent is female.
X_7	`HHAGE`	Respondent age, 18-110.
X_8	`ATHOME`	Number of weekdays someone is at home (1-5).

1. Consider the following R-script and explain briefly what it does.

```
dat.RECS<-read.csv("RECS.csv",sep = ',', header = T)
dat.RECS$LKWH.pers<-log(dat.RECS$KWH/dat.RECS$NHSLDMEM)
set.seed(22);
sample.size<-234;
dat.RECS.s<-dat.RECS[sample(nrow(dat.RECS), sample.size), ]
dat.RECS.s$NET<-ifelse(dat.RECS.s$INTERNET==1,1,0)
dat.RECS.s$SMART<-ifelse(dat.RECS.s$INTDATAACC==1,1,0)
```

2. Consider the following R output from running a regression of the logarithm of kwh consumption per person (Y), on a constant and the first six regressors, $X_2, .., X_7$, in Table 1. That is, a regression model like that in HN, §7.1, but with $k = 7$ instead of $k = 3$ (i.e. a 7-variable model in the language of HN). Denote this model M1, which has the maximized value of the

log-likelihood at -200.85. The R output from M1 is:

```
## Call:
## lm(formula = LKWH.pers ~ NET + SMART + NHSLDMEM + MONEYPY + HHSEX +
##     HHAGE, data = dat.RECS.s)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5641 -0.3636 -0.0073  0.3872  1.2852
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.00808    0.23184   34.54  < 2e-16 ***
## NET          0.06170    0.12794    0.48    0.630
## SMART       -0.49861    0.24265   -2.05    0.041 *
## NHSLDMEM    -0.19726    0.03060   -6.45  6.8e-10 ***
## MONEYPY      0.04446    0.01799    2.47    0.014 *
## HHSEX       -0.01052    0.07711   -0.14    0.892
## HHAGE        0.01119    0.00249    4.50  1.1e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.58 on 227 degrees of freedom
## Multiple R-squared:  0.308,  Adjusted R-squared:  0.29
## F-statistic: 16.9 on 6 and 227 DF,  p-value: 4.51e-16
```

Assuming that exogeneity and independence assumptions hold, we focus here on the mis-specification tests for normality and constant error variance. I.e. the Jarque-Bera test for normality (HN, §9.2.1) and an augmented version of the White's test in HN §9.3.1, respectively: The computed value of the Jarque-Bera statistic was 1.27. The White's statistic is here denoted W, and it holds that $W \sim \chi^2(9)$. The computed value of W was 12.02. What do you conclude from these two mis-specification tests, when testing at a 5% level of significance? Justify your answer.

3. Comment on the signs of the estimated coefficients corresponding to regressors `SMART`, `NHSLDMEM`, `MONEYPY` and `HHAGE`. Do the signs seem reasonable?

4. Compute a 95% confidence interval for the parameter on the `SMART` regressor.

5. We see that in M1, the variables `NET` and `HHSEX` are individually highly insignificant. Imposing the two corresponding zero restrictions gives a restricted version of M1, call it M2. The R

output from M2 is:

```
## Call:
## lm(formula = LKWH.pers ~ SMART + NHSLDMEM + MONEYPY + HHAGE,
##     data = dat.RECS.s)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5554 -0.3626 -0.0102  0.3954  1.2897
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.04311    0.17809   45.16  < 2e-16 ***
## SMART       -0.49484    0.24150   -2.05   0.0416 *
## NHSLDMEM    -0.19584    0.03034   -6.45  6.4e-10 ***
## MONEYPY      0.04630    0.01740    2.66   0.0084 **
## HHAGE        0.01108    0.00247    4.49  1.1e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.577 on 229 degrees of freedom
## Multiple R-squared:  0.307,  Adjusted R-squared:  0.295
## F-statistic: 25.4 on 4 and 229 DF,  p-value: <2e-16
```

M2 has the maximized value of the log-likelihood at -200.98. Perform the LR test of excluding `NET` and `HHSEX` jointly. Use a 5% level of significance. What do you conclude?

6. We now augment M2 with the last regressor in Table 1, `ATHOME`. This gives the model M3. The conclusions from the mis-specification analysis in M3 are the same as for M1. The R output from M3 is:

```
## Call:
## lm(formula = LKWH.pers ~ SMART + NHSLDMEM + MONEYPY + HHAGE +
##     ATHOME, data = dat.RECS.s)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.43449 -0.36070 -0.01793  0.37381  1.22831
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.963238   0.178263  44.671  < 2e-16 ***
## SMART       -0.549801   0.239195  -2.299 0.022436 *
## NHSLDMEM    -0.207489   0.030256  -6.858  6.5e-11 ***
## MONEYPY      0.052252   0.017318   3.017 0.002841 **
## HHAGE        0.009518   0.002503   3.803 0.000184 ***
## ATHOME       0.050639   0.018924   2.676 0.007992 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5698 on 228 degrees of freedom
## Multiple R-squared:  0.3285, Adjusted R-squared:  0.3138
## F-statistic: 22.31 on 5 and 228 DF,  p-value: < 2.2e-16
```

It appears that there is a (moderate) drop in the estimated regression coefficient on the age variable, `HHAGE`, compared to M2. Provide an interpretation of this. Discuss also to what extent the coefficient on `HHAGE` in M3 can be regarded as being structural or not and suggest ways to assess this.