Test exam in Introductory Econometrics (42008) April 30, 2020

This is an individual 2-hours all-aid exam. You are allowed to use your laptop with Rstudio installed. You can have access to the internet but may not communicate with anyone. Read/skim through all questions first. Answer short and precisely and remember to move on if you are stuck at a question.

Exercise 1: The multiple regression model (50%)

1. Consider the regression model, $\mathcal{M}$: **Fit the model with sample data, and perform mis-specification tests. If all tests are passed, then the model is correctly specified.**

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + u_i$$

where $X_{2,i}, X_{3,i}$ are exogenous, $(u_i, X_{2,i}, X_{3,i})$ are independent $(u_i \mid X_{2,i}, X_{3,i}) \sim N(0, \sigma^2)$ and $(\beta_1, \beta_2, \beta_3, \sigma^2) \in \mathbb{R}^3 \times \mathbb{R}_+$. Explain what it means that $\mathcal{M}$ is correctly specified?

2. Explain the main implication for estimation and inference of the assumption that $X_{2,i}, X_{3,i}$ are exogenous. **It suffices to look at the conditional likelihood if the parameters of the conditional and the marginal likelihood are unrelated.**
3. What is the interpretation of the regression parameter β_2?
4. Derive the maximum likelihood estimators (MLEs) of β_1 and β_2 under the restriction $\beta_3 = 1$
5. Explain how this restriction may be tested by LR testing.

Exercise 2: Regression analysis for the RECS data (50%)

We consider now the RECS data again. However, suppose we have a smaller subsample of the data set, consisting of 305 observations. This was randomly chosen from the original data set. As previously, we are interested in what demographic and economic variables matter for household electricity consumption. In Table 1, the variables we consider from the RECS data 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 next line.
X_2	`NHSLDMEM`	Number of household members
X_3	`EDUCATION`	Highest education completed by respondent (1-5)
X_4	`MONEYPY`	Annual gross household income for the last year (1-8)
X_5	`HHSEX`	Respondent gender. 1 if female.
X_6	`HHAGE`	Respondent age, 18-110
X_7	`ATHOME`	Number of weekdays someone is at home (1-5)

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

```
library(stats);
dat.RECS<-read.csv("RECS.csv",sep = ',', header = T)
dat.RECS$KWH.pers<-dat.RECS$KWH/dat.RECS$NHSLDMEM
dat.RECS$LKWH.pers<-log(dat.RECS$KWH.pers)
seed<-8;
set.seed(seed)
sample.size<-305;
dat.RECS.s<-dat.RECS[sample(nrow(dat.RECS), sample.size), ]
M1.s<-lm(LKWH.pers ~  NHSLDMEM + EDUCATION + MONEYPY + HHSEX + HHAGE + ATHOME, data = dat.RECS.s)
(sum.m1<-summary(M1.s))
```

2. Consider the following R output from running a regression of the logarithm of kwh consumption per person (Y) on a constant and the six regressors, $X_2, .., X_7$ in Table 1. That is a regression model like in Exercise 1 but with more regressors. Denote this model M1, which has the maximized value of the log-likelihood at -264.3511.

```
## Call:
## lm(formula = LKWH.pers ~ NHSLDMEM + EDUCATION + MONEYPY + HHSEX +
##     HHAGE + ATHOME, data = dat.RECS.s)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.43875 -0.40704  0.05903  0.40347  1.44541
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.027986   0.207997  38.597  < 2e-16 ***
## NHSLDMEM    -0.183410   0.026686  -6.873 3.69e-11 ***
## EDUCATION   -0.020649   0.035185  -0.587   0.5577
## MONEYPY      0.042966   0.017966   2.391   0.0174 *
## HHSEX        0.102454   0.068896   1.487   0.1380
## HHAGE        0.010032   0.002269   4.421 1.38e-05 ***
## ATHOME      -0.015398   0.017900  -0.860   0.3904
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5824 on 298 degrees of freedom
## Multiple R-squared:  0.289,  Adjusted R-squared:  0.2747
## F-statistic: 20.18 on 6 and 298 DF,  p-value: < 2.2e-16
```

   Assuming that exogeneity and independence assumptions hold we may here focus on the mis-specification test for normality and constant error variance: The corresponding mis-specification test statistic for normality (Jarque-Bera statistic, §9.2.1) was 3.3202. An augmented version of White's test for heteroskedasticity (§9.3.1) was performed. The computed value of this White's statistic was 8.028658. This version of Whites statistic is distributed as $\chi^2(11)$. What do you conclude from these mis-specification tests when testing at a 5% level of significance? Justify your answer.

3. Compute a 99% confidence interval for the parameter on the `NHSLDMEM` regressor. How do you interpret this?

4. Denote the true (or population-) value of the parameter corresponding to `HHSEX`, by β_5^0. Conduct a (one-sided) t-test, at 10% level of significance , of the null hypothesis $\beta_5^0 = 0$ vs. the positive alternative, $\beta_5^0 > 0$. Interpret the result.

5. M1 implies that the regressors `EDUCATION` and `MONEYPY` are treated as quantitative variables. An argument for this could be to save on parameters. However, a more reasonable model would include indicator variables instead. Output from such a model (call it M2) is the following:

```
## Call:
## lm(formula = LKWH.pers ~ NHSLDMEM + EDU + MONEY + HHSEX + HHAGE +
##     ATHOME, data = dat.RECS.s)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.47099 -0.40424  0.06518  0.40452  1.44550
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.058316   0.257160  31.336  < 2e-16 ***
## NHSLDMEM    -0.184959   0.027878  -6.635 1.60e-10 ***
## EDU2        -0.024923   0.185267  -0.135   0.8931
## EDU3        -0.022653   0.188783  -0.120   0.9046
## EDU4        -0.084617   0.201348  -0.420   0.6746
## EDU5        -0.077480   0.203826  -0.380   0.7041
## MONEY2       0.039470   0.116517   0.339   0.7350
## MONEY3       0.102494   0.130119   0.788   0.4315
## MONEY4       0.122496   0.140537   0.872   0.3841
## MONEY5       0.206732   0.158590   1.304   0.1934
## MONEY6       0.213951   0.157031   1.362   0.1741
## MONEY7       0.225613   0.203021   1.111   0.2674
## MONEY8       0.314990   0.161907   1.945   0.0527 .
## HHSEX        0.099646   0.071014   1.403   0.1616
## HHAGE        0.009952   0.002358   4.220 3.28e-05 ***
## ATHOME      -0.015722   0.018266  -0.861   0.3901
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.591 on 289 degrees of freedom
## Multiple R-squared:  0.2898, Adjusted R-squared:  0.2529
## F-statistic: 7.862 on 15 and 289 DF,  p-value: 8.254e-15
```

beta_edu2 = 2 * beta_edu1
beta_edu3 = 3 * beta_edu1
beta_edu4 = 4 * beta_edu1
beta_edu5 = 5 * beta_edu1

where `EDU2-EDU5` are the indicator/binary where `EDU2` =1 if `EDUCATION` =2, 0 otherwise, and `EDU3`=1 if `EDUCATION` =3, 0 otherwise etc. (same notation for the `MONEYPY` variable). Conclusions from the mis-specification from Q2 are unchanged. Compare the output of M1 and M2. Show that M1 corresponds to M2 with a set of parameter restrictions (i.e. state these restrictions).

6. Suppose these parameter restrictions are accepted so we can return to M1 above. We see that in M1, the variables `EDUCATION`, `HHSEX` and `ATHOME` are individually insignificant at a 5% level of significance. Jointly excluding these three regressors from M1, we get the restricted version of M1, call it M3, which has the maximized value of the log-likelihood at -265.8895.

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.102192   0.156780  51.679  < 2e-16 ***
## NHSLDMEM    -0.186145   0.025506  -7.298 2.61e-12 ***
## MONEYPY      0.039393   0.015321   2.571   0.0106 *
## HHAGE        0.009462   0.002071   4.568 7.18e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5824 on 301 degrees of freedom
## Multiple R-squared:  0.2818, Adjusted R-squared:  0.2746
## F-statistic: 39.36 on 3 and 301 DF,  p-value: < 2.2e-16
```

Perform the LR test corresponding to excluding these variables jointly. Use a 5% level of significance. What do you conclude?