STOCKHOLM UNIVERSITY
Department of Statistics
Johan Koskinen

WRITTEN RE-EXAMINATION, ECONOMETRICS I
2023-06-08

Time for examination: 14.00-19.00

Allowed tools: Pocket calculator, own formula sheet (1 double-sided A4 page), Course text-book: Wooldridge, J.M. *Introductory Econometrics - a Modern Approach (any edition)*
Note that no formula sheet will be provided.

The exam consists of 4 independent problems. Well motivated and clear solutions are required for full scoring on a problem. Don't forget to state any necessary assumptions or conditions where needed.
Passing rate: 50% of overall total, which is 100 points. For detailed grading criteria, see the course description. Answers may be given in English or Swedish.

Good luck!

Problem 1. (35 points)
The academic publishing business has a huge financial turnover, with worldwide sales amounting to more than USD 19 billion. This positions the business somewhere between the music industry and the film industry. Their business model is coming under increasing criticism by Universities seeing as the publishers make huge profits from University subscriptions while academics are expected to volunteer their time to do all the work, i.e. write papers and review them for free. Journals, on the other hand, could argue that they add value by providing academics with a platform for promoting their research and that a journal should be judge by how much publicity, and thereby status, that they generate. Let us look at how popular journals are and how this is a function of how many citations they provide for a given price. Table 1 describes data on economics journals.

Table 1: Summary of variables

	mean	sd	Description
PricePerCitation	2.548	3.466	Total number of citations of papers in the journal divide by the subscription price
Age	33.094	25.711	How many years have the journal existed (in year 2000)
Characters	2.673	1.600	A scaled measure of how many characters an issue contains
Subscriptions	196.867	204.529	Number of libraries that subscribe to the journal

The models for the following four population equations are estimated

$$
\begin{aligned}
(I)\ \log(Subscriptions) &= \beta_0 + \beta_1 \log(PricePerCitation) + u \\
(II)\ \log(Subscriptions) &= \beta_0 + \beta_1 \log(PricePerCitation) + \beta_4 \log(Age) + \beta_6 \log(Characters) + u \\
(III)\ \log(Subscriptions) &= \beta_0 + \beta_1 \log(PricePerCitation) + \beta_2 \log(PricePerCitation)^2 \\
&+ \beta_3 \log(PricePerCitation)^3 + \beta_4 \log(Age) \\
&+ \beta_5 [\log(Age) \times \log(PricePerCitation)] + \beta_6 \log(Characters) + u \\
(IV)\ \log(Subscriptions) &= \beta_0 + \beta_1 \log(PricePerCitation) + \beta_4 \log(Age) \\
&+ \beta_5 [\log(Age) \times \log(PricePerCitation)] + \beta_6 \log(Characters) + u
\end{aligned}
$$

The key interest is in the effect of price per citation on the number of subscriptions and model (II) includes *Age* and the length (*Characters*) of the journal as controls.

The estimation results for the four models are provided in Table 2

(a) Test if the effect of (log) price per citation in (I) is statistically significantly different from 0 using a confidence interval. Is there an assumption about the error term that requires an approximation be made?

(b) What is the percentage change in the number of subscriptions if a journal increases its citations by 10% with the price unchanged according to Model II?

(c) What is the predicted *number* of citations under models (I) and (IV) for the 'average' journal (i.e. a journal that has the mean values of the predictors)? Are your predictions

Table 2: Estimation results for models I, II, III, and IV

	Dependent variable:			
	log(Subscriptions)			
	(I)	(II)	(III)	(IV)
	(1)	(2)	(3)	(4)
log(PricePerCitation)	−0.533	−0.408	−0.961	−0.899
	(0.034)	(0.044)	(0.160)	(0.145)
I(log(PricePerCitation)ˆ2)			0.017	
			(0.025)	
I(log(PricePerCitation)ˆ3)			0.004	
			(0.006)	
log(Age)		0.424	0.373	0.374
		(0.119)	(0.118)	(0.118)
log(Characters)		0.206	0.235	0.229
		(0.098)	(0.098)	(0.096)
log(PricePerCitation):log(Age)			0.156	0.141
			(0.052)	(0.040)
Constant	4.766	3.207	3.408	3.434
	(0.055)	(0.380)	(0.374)	(0.367)
Observations	180	180	180	180
R^2	0.557	0.613	0.635	0.634
Adjusted R^2	0.555	0.607	0.622	0.626
Residual Std. Error	0.750	0.705	0.691	0.688
F Statistic	224.037	93.009	50.149	75.749

likely to be biased?

(d) Test whether the squared and cubed transformation of log price per citation are needed

(e) A librarian called Ken, who is responsible for his University's journal subscriptions, says about *American Economic Review*: 'We only get it because everyone else has it'. Discuss briefly what assumptions such purchasing behaviour might break.

Problem 2. (25 points)
A researchers is assuming a simple model (model I) for how how much you spend on food relates to your income

$$(I)\ foodexp = \beta_0 + \beta_1 income + u,$$

where *income* is weekly income in \$100, and *foodexp* is the weekly food expenditure in \$. Estimating the model $\widehat{foodexp}_i = \hat{\beta}_0 + \hat{\beta}_1 income_i$ using OLS he gets the following results

```
Call:
lm(formula = foodexp ~ income, data = food)

Coefficients:
            Estimate Std. Error
(Intercept)   83.416     43.410
income        10.210      2.093
---


Residual standard error: 89.52 on 38 degrees of freedom
Multiple R-squared:  0.385,Adjusted R-squared:  0.3688
F-statistic: 23.79 on 1 and 38 DF,  p-value: 1.946e-05
```

He then estimates the coefficients of the model

$$(II)\ \hat{u}_i^2 = \hat{\delta}_0 + \hat{\delta}_1 income_i + t_i,$$

where $\hat{u}_i$ are the residuals (`uhatsq`) from the first regression. The results are

```
Call:
lm(formula = uhatsq ~ income, data = food)

Coefficients:
            Estimate Std. Error
(Intercept)  -5762.4     4823.5
income         682.2      232.6
---
Residual standard error: 9947 on 38 degrees of freedom
Multiple R-squared:  0.1846,Adjusted R-squared:  0.1632
F-statistic: 8.604 on 1 and 38 DF,  p-value: 0.005659
```

Defining $g_i = \hat{u}_i^2$ he also estimates the coefficients of the model

$$(III)\ \log(g_i) = \alpha_0 + \alpha_1 \log(income_i) + e_i,$$

and defines $\hat{\sigma}_i^2 = e^{\hat{g}_i}$ (`vari`), where the predictions are

$$\hat{g}_i = \hat{\alpha}_0 + \hat{\alpha}_1 \log(income_i)$$

Finally, model IV is estimated as the regression

$$foodexp_i^* = \gamma_0 \hat{\sigma}_i^{-1} + \gamma_1 income_i^* + \nu_i,$$

where $foodexp_i^* = foodexp_i/\hat{\sigma}_i$ (`foodexp.w`) and $income_i^* = income_i/\hat{\sigma}_i$ (`income.w`). The OLS estimates of γ_0 and γ_1 are

```
lm(formula = foodexp.w ~ 0 + sqrt(1/vari) + income.w)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4222 -0.9811 -0.0789  1.3996  2.6088

Coefficients:
             Estimate Std. Error
sqrt(1/vari)  76.0538     9.7135
income.w      10.6335     0.9715
---
Residual standard error: 1.547 on 38 degrees of freedom
Multiple R-squared:  0.9523,Adjusted R-squared:  0.9498
F-statistic: 379.7 on 2 and 38 DF,  p-value: < 2.2e-16
```

(a) Is there evidence for any of the standard assumptions of SLR being violated in model I. If so, provide two consequences for inference.

(b) What sign will $\hat{\alpha}_1$ have?

(c) Is either of $\hat{\gamma}_1$ and $\hat{\beta}_1$ unbiased?

(d) What is $\sum_{i=1}^{n}(foodexp_i - \hat{\gamma}_0 - \hat{\gamma}_1 income_i)^2/\hat{\sigma}_i^2$?

Problem 3. (20 points)
For a survey of high school graduates we have following information

	mean	sd	Description
wage	9.501	1.343	hourly wage
education	13.808	1.789	number of years of education
afric	0.166		Dummy for African-American
hisp	0.191		Dummy for Hispanic
female	0.549		Dummy for Female
unemp	7.597	2.764	county unemployment rate
urban	0.233		Is the school in an urban area?
distance	1.803	2.297	distance from 4-year college (in 10 miles)

Consider the structural model

$$\log(wage) = \beta_0 + \beta_1 education + \beta_2 unemp + \beta_3 afric + \beta_4 hisp + \beta_5 female + \beta_6 urban + u$$

and assume that we suspect that *education* is endogenous but the rest of the predictors are exogenous. For *education* and the variable *distance*, we have the following two equations

```
lm(formula = education ~ distance, data = ColDat)
Coefficients:
            Estimate Std. Error
(Intercept) 13.93861    0.03290
distance    -0.07258    0.01127
---

Residual standard error: 1.782 on 4737 degrees of freedom
```

and

```
Call:
lm(formula = education ~ afric + hisp + female + unemp + urban +
    distance, data = ColDat)

Coefficients:
             Estimate Std. Error
(Intercept) 14.060680   0.083075
afric       -0.524317   0.072444
hisp        -0.274761   0.067879
female      -0.024645   0.051731
unemp        0.010267   0.009768
urban       -0.092308   0.065039
distance    -0.086846   0.012244
---
```

```
Residual standard error: 1.77 on 4732 degrees of freedom
Multiple R-squared:  0.02298,Adjusted R-squared:  0.02174
F-statistic: 18.55 on 6 and 4732 DF,  p-value: < 2.2e-16
```

The predicted values `yhat` from this regression are used as predictors in the regression

```
Call:
lm(formula = log(wage) ~ yhat + afric + hisp + female + unemp +
    urban, data = ColDat)

Coefficients:
              Estimate Std. Error
(Intercept)  1.2171787  0.1515969
yhat         0.0673242  0.0107992
afric       -0.0277621  0.0078353
hisp        -0.0335043  0.0061216
female      -0.0076101  0.0039698
unemp        0.0142234  0.0007245
urban        0.0064494  0.0047979
---
Residual standard error: 0.1355 on 4732 degrees of freedom
Multiple R-squared:  0.1099,Adjusted R-squared:  0.1087
F-statistic: 97.35 on 6 and 4732 DF,  p-value: < 2.2e-16
```

(a) For *distance* to be a valid instrument for the endogenous variable, what are the exclusion criteria for the exogenous variables? Discuss briefly.

(b) If the exclusion criteria are met, what other condition needs to be satisfied for *distance* to be a valid instrument for the endogenous variable? Perform a formal test

(c) Does education have a causal effect on (log) earnings (wage)? Can you determine if the instrument is weak or not?

Problem 4. (20 points)
A health economist is interested in resource management among hospitals and collects data on 85 hospitals in the region Lazio in Italy, giving the variables

	mean	sd	Description
LHU	0.529		Dummy for Local Health Units
RES	0.071		Dummy for research or University hospital
PRIVATE	0.529		Dummy for private hospital
SIZE	406.412	493.658	Number of employees at Hospital

Furthermore, for each (ordered) pair of hospitals $i = 1, \dots, m$, she counts how many patients were transfered from one hospital to the other in a year. In addition she records the geographical location of the sending hospital and receiving hospital. For each pair of hospitals, this gives the following variables

	mean	sd	Description
patientTransfer	1.383	8.776	Number of patients transfered from sending hospital to receiving hospital
DIST	50.460	39.139	Distance in kilometres between sending and receiving hospital

The initial idea was to estimate a gravity model

$$patientTransfer = SIZEsender^{\alpha_1} SIZEreceiv^{\alpha_2} lnDIST^{-\gamma} e^u,$$

where, for each pair, $SIZEsender$ is the size of the hospital that sends patients, and $SIZEreceiv$ is the size of the hospital receiving patients. However, 87% of pairs of hospitals do not transfer any patients between them. She decides instead to model if a hospital sends any patient to another hospital instead and defines the variable $trans_i$ to be 1 if $patientTransfer_i > 0$, and 0 otherwise, for pair i. The following linear probability model (LPM) is assumed

$$\begin{aligned} E(trans \mid \mathbf{x}) &= \beta_0 + \beta_1 LHUsender + \beta_2 LHUreceiv + \beta_3 RESsender + \beta_4 RESreceiv \\ &+ \beta_5 PRIVATEsender + \beta_6 PRIVATEreceiv + \beta_7 lnSIZEsender + \beta_8 lnSIZEreceiv \\ &+ \beta_9 LHUsender \times LHUreceiv + \beta_{10} RESsender \times RESreceiv \\ &+ \beta_{11} PRIVATEsender \times PRIVATEreceiv + \beta_{12} lnDIST \end{aligned}$$

In this equation ln means natural logarithm, for example $SIZEsender = \log(lnSIZEsender)$, in other words the logarithm of the number of staff of the hospital that sends patients in a sender-receiver pair. For the hospital types, the suffix *sender* refers to the hospital that potentially sends and *receiv* the hospital that potentially receives patients.

The LPM, using robust standard errors, is estimated as

```
                         Estimate Std. Error
(Intercept)            -0.2456964  0.0487173
```

```
LHUsender                    -0.0283724  0.0201472
LHUreceiv                     0.0345335  0.0207747
RESsender                     0.0200296  0.0250857
RESreceiv                    -0.1909323  0.0213702
PRIVATEsender                -0.1300690  0.0198389
PRIVATEreceiv                -0.1521700  0.0188334
lnSIZEsender                  0.0773063  0.0044251
lnSIZEreceiv                  0.0716800  0.0043848
lnDIST                       -0.0820730  0.0054242
LHUsender:LHUreceiv          -0.0440559  0.0202876
RESsender:RESreceiv           0.0818306  0.0941448
PRIVATEsender:PRIVATEreceiv   0.1274284  0.0206091
```

To better predict outcomes, a logistic regression is also estimated with the same covariates:

```
Call:
glm(formula = trans ~ LHUsender + LHUreceiv + LHUsender * LHUreceiv +
    RESsender + RESreceiv + RESsender * RESreceiv + PRIVATEsender +
    PRIVATEreceiv + PRIVATEsender * PRIVATEreceiv + lnSIZEsender +
    lnSIZEreceiv + lnDIST, family = binomial(link = "logit"))

Coefficients:
                            Estimate Std. Error
(Intercept)                 -7.88070    0.46063
LHUsender                   -0.07705    0.14978
LHUreceiv                    0.41784    0.14367
RESsender                    0.01078    0.14081
RESreceiv                   -1.47862    0.17102
PRIVATEsender               -0.50845    0.14661
PRIVATEreceiv               -1.21937    0.16025
lnSIZEsender                 0.85164    0.04757
lnSIZEreceiv                 0.72578    0.04437
lnDIST                      -0.72588    0.03806
LHUsender:LHUreceiv          0.07982    0.15682
RESsender:RESreceiv          0.03998    0.44050
PRIVATEsender:PRIVATEreceiv  0.83753    0.25401
---
    Null deviance: 6790.4  on 7137  degrees of freedom
Residual deviance: 4894.9  on 7125  degrees of freedom
AIC: 4920.9
```

The researchers also wants a simplified model so she estimates a model without the hospital types as predictors:

```
Call:
```

```
glm(formula = trans ~ lnSIZEsender + lnSIZEreceiv + lnDIST,
family = binomial(link = "logit"))

Coefficients:
             Estimate Std. Error
(Intercept)  -8.89698    0.33012
lnSIZEsender  0.87988    0.03606
lnSIZEreceiv  0.70058    0.03418
lnDIST       -0.48299    0.02944
---
    Null deviance: 6790.4  on 7137  degrees of freedom
Residual deviance: 5241.1  on 7134  degrees of freedom
AIC: 5249.1
```

(a) How much less likely is a private hospital to send patients to another hospital compare to an LHU, everything else equal, according to the linear probability model?

(b) Using the LPM, a hospital A has a predicted probability of 0.1 of sending patient to hospital B. How many times further apart would they have to be for the predicted probability to be 0? (Hint: by what factor would you have to multiply the current distance?)

(c) On the 5%-level, do you draw different conclusions about the receiving LHU hospitals based on LPM and logistic?.

(d) Consider the results for the *simplified* logistic regression model. For a small hospital with 96 employees, what is the difference in the predicted probability of sending patients to a hospital 10 kilometres away that has 200 staff versus sending to a hospital with 1000 staff 200 kilometres away?