STOCKHOLM UNIVERSITY
Department of Statistics
Johan Koskinen

WRITTEN EXAMINATION, ECONOMETRICS I
2023-04-27

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 partially independent problems using the same dataset. 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 US General Social Survey, is a nationally representative survey in the US conducted every year or every other year. In the 2021 GSS (GSS21) a suite of demographics were recorded for respondents. The respondents were also asked a number of questions on opinions, attitudes, and behaviours. A researcher wants to know what the effect of education is on the beliefs about climate change. Table 1 describes the variables a researcher wants to use to investigate this.

Table 1: Summary of variables

Variable Name	Description
clmtwrld	What do you think the impact of climate change will be on the world as a whole, on a scale from 0 (extremely bad) to 10 (extremely good)
trrerch	How much do you trust in University research centers on a scale of 0 (not at all) to 10 (completely)
trmedia	How much do you trust in the news media on a scale of 0 (not at all) to 10 (completely)
trlegis	How much do you trust in the U.S. Congress on a scale of 0 (not at all) to 10 (completely)
educ	Number of years of education
sex	Male (0) or Female (1)
$INCOM16$	Socio-economic background: At the age of 16, how was your family's income on a scale from 1 (far below average) to 5 (far above average). The middle point, 3, is average.

The following population model is assumed

$$clmtwrld = \beta_0+\beta_1 trresrch+\beta_1 trmedia+\beta_3 trbusind+\beta_4 trlegis+\beta_5 educ+\beta_6 sex+\beta_7 INCOM16+u$$

The researcher estimates the following model (Model A)

```
Call:
lm(formula = clmtwrld ~ trresrch + trmedia + trbusind + trlegis +
    educ + sex + INCOM16, data = myDat)

Coefficients:
            Estimate Std. Error
(Intercept)  5.10258    0.40025
trresrch    -0.28764    0.03008
trmedia     -0.26864    0.02703
trbusind     0.33432    0.03068
trlegis      0.10585    0.03000
educ        -0.09409    0.02241
sex         -0.13684    0.11564
INCOM16      0.07898    0.06103
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 1.952 on 1182 degrees of freedom
Multiple R-squared:   0.31,Adjusted R-squared:  0.3059
F-statistic: 75.88 on 7 and 1182 DF,  p-value: < 2.2e-16
```

(a) Formally test whether education has an effect on belief in climate change. Interpret and specify what assumptions you need.

(b) What is the difference in believed climate impact between someone who has complete trust in media and someone who has no trust at all, all else equal?

Another researcher is concerned that with the US becoming so polarised and with mainstream media and science being increasingly challenged by disinformation, the model may not apply equally to all Americans. She decides to run the regression separately for those $n_1 = 422$ respondents that voted for Trump in the 2016 (Model B.1), and the $n_0 = 768$ that did not vote for Trump (Model B.0). The estimations yielded $SSR_1 = 1444.622$ for Trump voters, and $SSR_0 = 2749.379$ for non-Trump voters.

(c) Test whether there is any difference in the population equation between the two populations. Set up the tests in terms of linear restrictions. (Hint: it might be useful to write out the sum of square residuals for the unrestricted model. Further, in R output, the 'residual standard error' is $\hat{\sigma}$)

(d) The researcher also estimates the model for two separate populations using interaction effects with the binary variable $Trump16$ (that is 1, if respondent voted Trump, and 0 otherwise). In this model (Model C) the slope for *trmedia* is -0.198 (standard error: 0.034), and the slope for the interaction of *trmedia* with $Trump16$ is 0.172 (standard error: 0.070). How do you interpret this? (Note: the t-statistic for *trmedia* in Model B.1 is -0.432)

(e) Yet another researcher is looking for policy-relevant conclusions. Arguing that we cannot manipulate peoples's level of education, what is really important is how knowledgeable people are about climate change. For a variable *know*, that ranges from 0 (no knowledge) to 100 (perfect knowledge), she says that if you regressed *know* on all of the predictors in model A, then an extra year of education would lead to an increase of 2.5 in *know*, everything else equal. If *know* were added to Model A, how large would the slope $\hat{\beta}_{k+1}$ for *know* have to be for the correct estimate $\hat{\beta}_5$ for *educ* to be 0 (call the OLS in Model A for example $\tilde{\beta}_5$, to distinguish it from $\hat{\beta}_5$).

Problem 2. (25 points)
Figure 1 provides a plot of the squared residuals from Model C against *trmedia*.

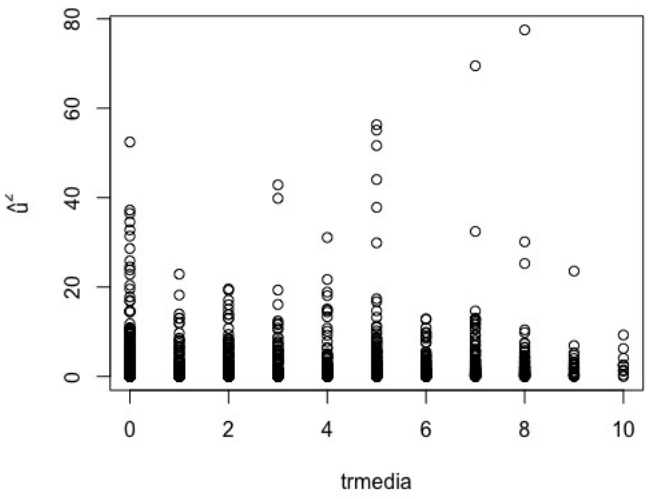

Figure 1: Residuals

(a) Does the plot indicate that any of the assumptions might be violated? If so, can you suggest a way of transforming the equation of Model C to an equation for which MLR.1 through MLR.5 are satisfied (making assumptions only based on your assessment of the Figure)?

Let $\hat{u}_i^2$ be the residuals from Model C and $\hat{y}_i$ the predicted values, for $i = 1, \ldots, 1190$. A regression

$$\hat{u}_i^2 = \underset{(0.70229)}{0.67362} + \underset{(0.52964)}{2.13303}\hat{y}_i + \underset{(0.08967)}{-0.31862}\hat{y}_i^2$$

is estimated, which yields an $SSR = 52887.53$ and $SST = 53689.56$.

(b) What conclusions can you draw about the distribution of the residuals in Model C?

Each of the predictors that are included in Model C but that are *not* included in Model A are regressed on the predictors in Model A in 8 separate regressions, one at a time. The residual from the j'th regression for the i'th respondent is called $\tilde{r}_{ij}$. For each observation i, 8 new predictors $w_{ij} = \tilde{r}_{ij}\hat{u}_i$ are constructed. Finally, OLS estimates are estimated for 1 regressed on $w_{i1}, \ldots, w_{i8}$. The $SSR = 1123.639$ and $R^2 = 0.05584$ for this equation.

(c) Does this affect your conclusions in Question 1 c?

Problem 3. (20 points)
A researcher argues that one of the problems with Models A and C is that the variable *trmedia* is endogenous. Let us assume that the other predictors are exogenous but that there is another exogenous variable *oppgunlaw* that we can use to instrument *trmedia* in the structural equation of Model A. The variable *oppgunlaw* is 1 if the respondent would oppose any law which would require a person to obtain a police permit before he or she could buy a gun, and zero otherwise. The correlation between *trmedia* and *oppgunlaw* is -0.381.

(a) Explain what the conditions are for *oppgunlaw* being a valid instrumental variable for *trmedia*. Comment on their plausibility.

The following reduced form equation for *trmedia* is estimated

```
             Estimate Std. Error
(Intercept) -0.311617   0.432686
trresrch     0.431063   0.029082
trbusind    -0.002924   0.032220
trlegis      0.445095   0.028582
educ         0.035193   0.023488
sex         -0.275934   0.121467
INCOM16     -0.038383   0.063861
oppgunlaw   -1.171043   0.139709
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.041 on 1182 degrees of freedom
Multiple R-squared:  0.4696,Adjusted R-squared:  0.4664
F-statistic: 149.5 on 7 and 1182 DF,  p-value: < 2.2e-16
```

(b) Is the second condition (relevance) fullfilled?

The predicted values $\widehat{trmedia}$ from the estimated reduced form equation are called `yhat`. Model A is re-estimated using $\widehat{trmedia}$ (`yhat` in the output) in lieu of *trmedia*, yielding the following results

```
            Estimate Std. Error
(Intercept)  4.45351    0.43407
trresrch    -0.04024    0.06331
yhat        -0.77508    0.11662
trbusind     0.31969    0.03153
trlegis      0.34836    0.06231
educ        -0.06687    0.02370
sex         -0.22338    0.11977
INCOM16      0.04604    0.06281
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 1.995 on 1182 degrees of freedom
Multiple R-squared:  0.2793,Adjusted R-squared:  0.275
F-statistic: 65.44 on 7 and 1182 DF,  p-value: < 2.2e-16
```

(c) Does the instrumented variable *trmedia* have a stronger or weaker effect that the incorrect (biased) OLS estimate? Has the removing the endogeneity cause the effect to disappear (do the appropriate test)

Problem 4. (20 points)
A lobbyist for the National Rifle Association (NRA) is interested in the effect of legislation on gun sales. He uses the 2021 GSS and *owngun* as his outcome variable. The variable *owngun* indicates if a respondent's household owns a gun (1 if YES, 0 if NO). In addition to the variables *educ*, $female$, $INCOM16$, and *oppgunlaw* used above, he divides respondents into four geographical regions, indicated by the dummy variables $Northeast$, $Midwest$, $South$, and $West$. He also includes the variable *polviews*, that ranks the respondent on a 7-point scale, ranging from very liberal (1) to very conservative (7).

He estimates the following linear probability model

```
              Estimate Std. Error
(Intercept)  0.1725235  0.0710107
educ        -0.0005541  0.0035762
female      -0.1053824  0.0185263
INCOM16     -0.0101917  0.0097796
MidwestTRUE  0.1029413  0.0293348
SouthTRUE    0.1429267  0.0278237
WestTRUE     0.0410997  0.0299783
polviews     0.0459640  0.0062035
oppgunlaw    0.1940222  0.0214149
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.463 on 2581 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.1128,Adjusted R-squared:  0.1101
F-statistic: 41.02 on 8 and 2581 DF,  p-value: < 2.2e-16
```

(a) What is the estimated difference in $E(owngun \mid \mathbf{x}^{(1)}) - E(owngun \mid \mathbf{x}^{(0)})$, where $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(0)}$ are the covariates of a person who opposes legislation and a person who *does not*, respectively, but that are otherwise equal?

(b) As he does not know how to calculate robust standard errors, he also estimates a model without any controls to look at the effect only of *oppgunlaw*. This yields the intercept $\hat{\beta}_0 = 0.32133$ and slope $\hat{\beta}_1 = 0.27230$ for *oppgunlaw*. The number of people that oppose legislation is $n_1 = \sum_i oppgunlaw_i = 1046$, and the number of people that do *not* oppose legislation is $n_0 = 1807$. Test whether $\beta_1 = 0$ (Hint: What is the average/expected value and variance of a binary variable? Also, note that for binary outcome y, $\hat{\beta}_1 = \bar{y}_1 - \bar{y}_0$. Motivate why you pool proportions or not).

Another person, a law-maker and activist, realises that estimating a linear regression for binary outcomes is not a great idea, so she estimates the following probit model

```
             Estimate Std. Error
(Intercept) -0.910034   0.201856
educ        -0.001526   0.010131
female      -0.293633   0.052349
INCOM16     -0.030101   0.027795
MidwestTRUE  0.305628   0.084883
SouthTRUE    0.412709   0.080527
WestTRUE     0.127080   0.087621
polviews     0.128635   0.017597
oppgunlaw    0.511083   0.059218
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3494.1  on 2589  degrees of freedom
Residual deviance: 3192.2  on 2581  degrees of freedom
  (8 observations deleted due to missingness)
AIC: 3210.2
```

(c) Test if there is an effect of opposing gun laws on gun ownership, everything else equal.

(d) If you encounter a man, in the Midwest, who has 12 years of school, came from the poorest conditions ($INCOME16$=1), is as conservative as you can be (a score of 7), and opposes gun laws, what is the predicted probability according to the model that he (or his household) has a gun?[1]

[1]There is one such person in this subset and he owns a gun. For the same type of person, but from average social background, 8 out of 10 own guns.