Final Exam SØK1005 – Spring 2021

1 Do-file

```
4   use HTVexam
5
6   browse
7
8   summarize lwage educ motheduc fatheduc urban
9
10  regress lwage educ motheduc fatheduc urban
11
12  regress lwage educ urban
13
14  correlate motheduc fatheduc
15
```

2 Understanding Abstracts

A) From this abstract I expect that a version of the following models has been estimated.

$$STEM_i = \beta_0 + \beta_1 MathRank_i + u_i \quad (1)$$

$$ArtsAndSocialSciences_i = \beta_0 + \beta_1 MathRank_i + u_i \quad (2)$$

$$STEM_i = \beta_0 + \beta_1 EnglishRank_i + u_i \quad (3)$$

$$ArtsAndSocialSciences_i = \beta_0 + \beta_1 EnglishRank_i + u_i \quad (4)$$

Here I have first reported the simple linear regression models that I expect has been estimated. In the abstract they state that they find that a higher rank in math increases the likelihood of choosing STEM and decrease the likelihood of choosing Arts and Social Sciences. From this we can tell that they have estimated the effect of math rank on the students' preferences rankings for both STEM and Arts and Social Sciences. This I have reported in regression functions (1) and (2). In regression function (1) we expect the estimation of β_1 to be positive since they found that higher math rank increases the likelihood of choosing STEM, and in regression function (2) we expect the estimation of β_1 to be negative since they found that higher math rank decreased the likelihood of choosing Arts and Social Sciences.

Next, they state that a higher rank in English leads to an increase in the probability of choosing Arts and Social Sciences and decrease the probability of choosing STEM. From the we can tell that they have estimated the effect of English rank on the students' preferences rankings for both STEM and Arts and Social Sciences. This I have reported in regression functions (3) and (4). We expect the estimation of β_1 in

regression function (3) to be negative since they found that a higher rank in English decreased the probability of choosing STEM. In regression function (4) we expect the estimation of β_1 to be positive sine they found that a higher English rank increased the probability of choosing Arts and Social Sciences.

I expect this dataset to be a cross-section. This is because there is no mention of a time frame, and in the first sentence they state that they use unique data on preference rankings for all high school students who apply for college students in Ireland. Given that this is only done at one point in time, the data would take the form of a cross-sectional dataset.

An alternative or addition to the simple linear regression models I have reported above is the following multiple linear regressions.

$$STEM_i = \beta_0 + \beta_1 MathRank_i + \beta_2 EnglishRank_i + \beta_3 boy_i + u_i \quad (5)$$

$$ArtsAndSocialSciences_i = \beta_0 + \beta_1 MathRank_i + \beta_2 EnglishRank_i + \beta_3 boy_i + u_i \quad (6)$$

In these models' boy is a dummy variable to capture average differences between boys and girls.

B) From this abstract I would expect the following simple linear regression models to be estimated.

$$FirstYearGrades_{it} = \beta_0 + \beta_1 ShiftingToSemester_{it} + u_{it} \quad (7)$$

$$ProbFullCourseLoad_{it} = \beta_0 + \beta_1 ShiftingToSemester_{it} + u_{it} \quad (8)$$

$$TimingMajorChoice_{it} = \beta_0 + \beta_1 ShiftingToSemester_{it} + u_{it} \quad (9)$$

The reason why I expect the independent variable to be shifting to a semester (ShiftingToSemester) is because they in the last sentence report the effect of shifting to a semester on different variables. The variables are first-year grades (FirstYearGrades), probability of enrolling in full course load (ProbFullCourseLoad), and timing of major choice (TimingMajorChoice). Since they are estimating the effect of shifting to a semester on these variables, these would be the dependent variables. They report that the estimated effect shifting to a semester is lower first year grade, and thus the β_1 in regression function (7) would be negative. The effect of shifting to a semester is also estimated to be negative on both probability of enrolling in full course load and the timing of major choice, and thus we expect β_1 to be negative in the sample regression function of regressions (8) and (9).

In the abstracts second sentence they say that they have used panel data. This is also what we would expect as they are estimating the effect of an event on different variables. It would therefore be appropriate to use a panel dataset since this would give observation both before and after the event.

3 Estimation in Stata

A) The task asks us to estimate the effect of education on log-wages, and accounting for the effect of mother's and father's education and residence in an urban area. The population regression function will be as follows.

$$lwage = \beta_0 + \beta_1 educ + \beta_2 motheduc + \beta_3 fatheduc + \beta_4 urban + u$$

Doing the regression in Stata we get the following result.

```
. regress lwage educ motheduc fatheduc urban
```

Source	SS	df	MS	Number of obs	=	496
				F(4, 491)	=	24.01
Model	25.9875467	4	6.49688669	Prob > F	=	0.0000
Residual	132.853143	491	.270576666	R-squared	=	0.1636
				Adj R-squared	=	0.1568
Total	158.84069	495	.320890282	Root MSE	=	.52017

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0693345	.0115045	6.03	0.000	.0467303	.0919386
motheduc	.0202912	.0124639	1.63	0.104	-.0041979	.0447803
fatheduc	.0148721	.0090647	1.64	0.102	-.0029384	.0326825
urban	.1010183	.0638069	1.58	0.114	-.0243499	.2263865
_cons	1.023523	.1522257	6.72	0.000	.7244286	1.322617

The sample regression function is thus:

$$\widehat{lwage} = 1.023523 + 0.0693345 educ + 0.0202912 motheduc + 0.0148721 fatheduc + 0.1010183 urban$$

Next, we interpret the results. The everything else equal effects are:

- It is estimated that increasing education by one year (seems to be years looking at the data) will increase hourly wage by 6.93%. It should also be mentioned here that education is reported as highest completed by 1991. Same goes for mother's and father's education.
- A one-year increase in mother's education is estimated to increase wage by 2.03%.
- Similarly, are a one-year increase in father's education estimated to increase wages by 1.49%.
- It is estimated that living in an urban area will on average increase wages by 10.10%. This is the case that the urban-dummy equals 1.
- Lastly, we have the constant. This is the estimated log-wage in the case that education, mother's education, and father's education all equal 0, and we are

outside an urban area. The log-wage is 1.023523, giving and expected hourly wage of \$2.78 ($e^{1.023523} = 2.782981$).

B) We are going to test if there is evidence of a wage differential in urban areas and non-urban areas. To do this we will choose the null hypothesis that there is no difference and the alternative hypothesis that there is difference. This can be formulated as following.

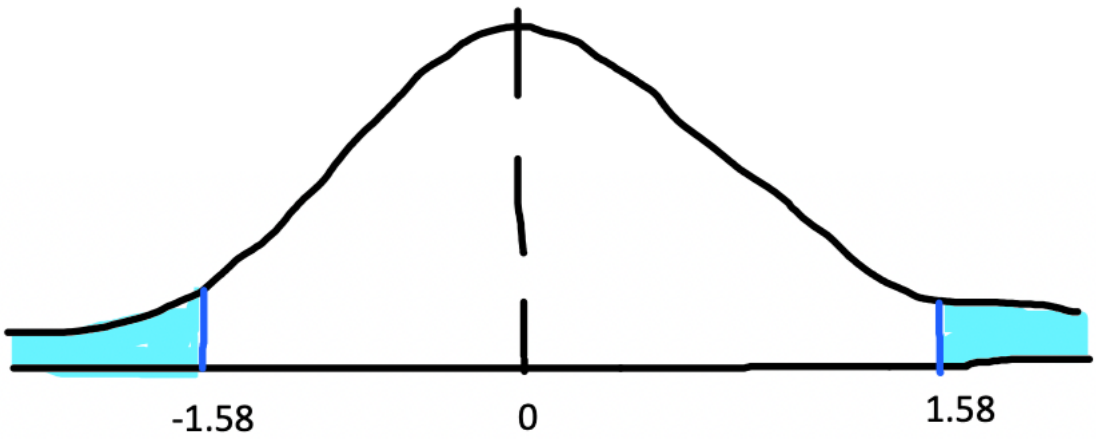
$$H_0: \beta_4 = 0$$

$$H_1: \beta_4 \neq 0$$

Since this is a single hypothesis test, we are using a t-test. The test statistic is as follows under the null hypothesis.

$$TS = \frac{\widehat{\beta_4} - \beta_4}{se(\widehat{\beta_4})} = \frac{\widehat{\beta_4} - 0}{se(\widehat{\beta_4})} = 1.58$$

The number of observations is 496 and we have 4 variables, and thus the test statistic is distributed as $t - stat \sim t_{491}$ under the null hypotheses. The p-value Stata reports for the urban-dummy is 0.114. This means that the probability of observing such a test statistic under the null hypotheses is 11.4%. We can show this graphically.



Under the null hypothesis the probability of landing in the white area is 88.6% and the probability of landing in the blue area is 11.4%. This is what the p-value is, it reports the probability of observing a test statistic as extreme under the null hypothesis.

The usual confidence intervals we use are 10%, 5% and 1%. These would all mean observing a test statistic that is more unlikely than the one we have observed, since we had the p-value 0.114. In other words, are we not able to reject the null hypothesis at any of these significance levels, but we would have been able to reject

the null hypothesis at an 11.4% significance level. However, we can conclude that there is not sufficient evidence to say that there are different wages in urban areas, even though there is some evidence for it.

C) We are now going to test if mothers and father education has a jointly significant effect on wages. To do this we choose the null hypothesis that they are not jointly significant, and the alternative hypothesis that they are not. This can be formulated as follows.

$$H_0: \beta_2 = \beta_3 = 0$$

$$H_1: not\ H_0$$

The test statistic for the joint hypothesis test, called F-statistic can be defined in two ways:

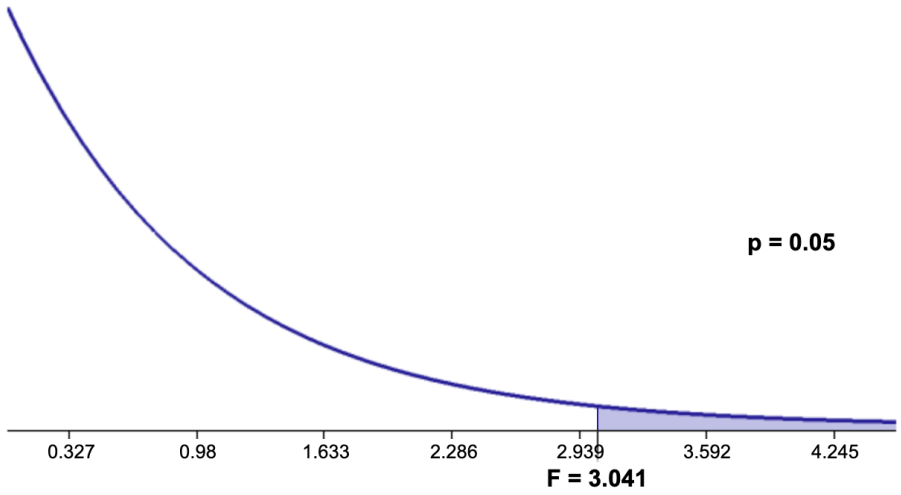
$$F - stat = \frac{SSR_r - SSR_u}{SSR_u} \cdot \frac{n-k-1}{q}$$

$$F - stat = \frac{R_u^2 - R_r^2}{1 - R_u^2} \cdot \frac{n-k-1}{q}$$

The distribution of this F-statistic is $F - stat \sim F_{2;491}$. We have the nominator degrees of freedom equal to 2 because we have 2 restrictions in the restricted model, and the denominator degrees of freedom equal to 491 since we have 496 observations and 4 variables in the unrestricted model. We can now check the tables to find the critical value and define a rejection region for this test.

$$F - stat > 2.9957$$

Graphically we can see this as the F-statistic having to exceed approximately 3, and thereby being in the blue are.



We will perform both the tests, however they should both give the same result. The result from Stata when estimating the restricted model is as following.

```
. regress lwage educ urban
```

Source	SS	df	MS	Number of obs	=	496
				F(2, 493)	=	42.32
Model	23.2756302	2	11.6378151	Prob > F	=	0.0000
Residual	135.565059	493	.274979836	R-squared	=	0.1465
				Adj R-squared	=	0.1431
Total	158.84069	495	.320890282	Root MSE	=	.52439

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0865828	.010203	8.49	0.000	.0665362	.1066295
urban	.1264772	.0638019	1.98	0.048	.00112	.2518344
_cons	1.210796	.1361355	8.89	0.000	.9433185	1.478273

$$\widehat{lwage} = 1.210796 + 0.0865828educ + 0.1264772urban$$

Next, we calculate the F-statistic.

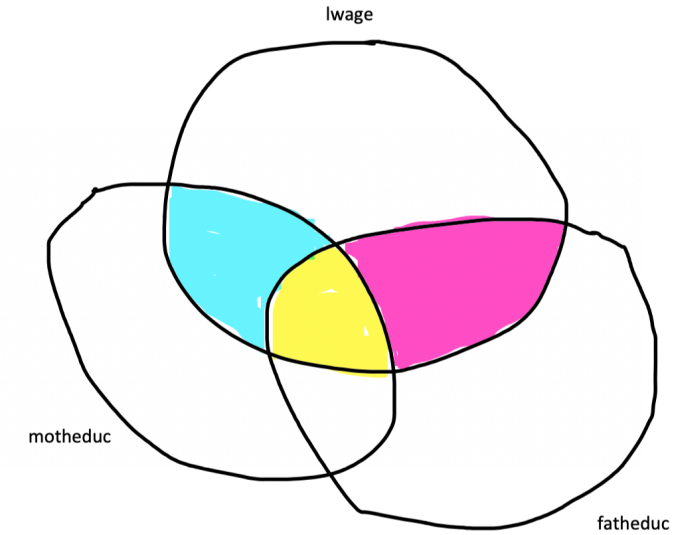
$$F - stat = \frac{SSR_r - SSR_u}{SSR_u} \cdot \frac{n-k-1}{q} = \frac{135.565059 - 132.853143}{132.853143} \cdot \frac{491}{2} = 5.01$$

$$F - stat = \frac{R_u^2 - R_r^2}{1 - R_u^2} \cdot \frac{n-k-1}{q} = \frac{0.1636 - 0.1464}{1 - 0.1636} \cdot \frac{491}{2} = 5.04$$

The difference in the results is due to rounding (or typing) errors. However, the result from both tests exceeds the critical value of 2.9957. Thus, we can reject the hull hypothesis that mother's and father's education does not influence wages at a 5% significance level. In other words, do we have sufficient evidence to suggest that mother's and father's educations have a jointly statistically significant effect on wages.

D) By looking at the Stata output window we see that mother's education has a t-statistic of 1.63 and father's education has a t-statistic of 1.64. This means that if we were to perform a test on either of them being individually statistically significant that would be the result from the t-test. The p-values are respectively 0.104 and 0.102 for mother's and father's education. This means that the probability of observing such t-statistics for mother's and father's educations under the null hypothesis is just above 10% for both. We would thus not be able to reject either if we had performed a test at a 10%, 5% or 1% significance level. Based on this we can conclude that neither mother's nor father's education has an individually statistically significant effect on wages.

E) The results we have contained in task C) and D) are compatible. The reason for this can intuitively be explained by using a Ballentine-diagram.



Here we see that the partialled out effect of mothers of mother's education on lwage is the blue area, and the partialled out effect of father's education on lwage is the pink are. When we are performing the individual tests, we are only testing these two areas individually, but when we are doing the joint test, we are including the yellow area as well as both the blue and pink. This we can do because we do not care if the effect is caused by mother's or father's education, only if it is caused by them jointly or at least of them. We are testing the total effect of the two variables. We can see from this that doing the joint test has a bigger explanatory momentum on lwage, and especially of the correlation is big. That is if the overlapping effect of mother's and father's education (the yellow area) is large.

Correlation is a key thing here. Because having high correlation between explanatory variable makes the variance inflation factor (VIF) big, and this makes the standard error of the coefficient big. The reason for this is that when the variables are highly correlated it becomes difficult to precisely estimate the effect of a single variable. The variance of a coefficient is given by the following expression.

$$V(\widehat{\beta}_j) = \frac{\sigma^2}{SST_j} \cdot \frac{1}{1 - R_j^2} = \frac{\sigma^2}{SST_j} \cdot VIF$$

When the correlation increases R-squared between the explanatory variables increases and the VIF increases. Oppositely, when the correlation decreases the R-squared between the explanatory variables decreases and the VIF decreases.

The problem with having a high VIF, and thus a high variance and standard error is that it becomes difficult to reject the null hypothesis when doing individual testing. It is important to note here that this does not make the variance biased, only imprecise. When we do joint tests, we can walk around this problem as we are not interested in the individual effect but the joint effect. This means as I explained to

begin with that, we can also include the area that represents the joint effect in the test, making it easier to reject the null hypothesis.

To sum up, we have that if mothers and father education has a high correlation the answers from task C) and D) are compatible. We can expect mother's and father's education to have some correlation as many meet the people they get together with in an age where one might be studying or working, and thus meeting a lot of people in similar fields. We can check the correlation between the two variables in Stata as well.

	motheduc	fatheduc
motheduc	**1.0000**	
fatheduc	**0.5729**	**1.0000**

We see that the variables have quite a high correlation of 0.5729.

4 Interpretation of an empirical analysis

A) Column to reports the effects of log health expenditures and log unemployment on log circulatory disease. The table reports that the expected effect of a 1% increase in health expenditures is a 0.436% increase in circulatory diseases, everything else equal. Next, we have that a 1% increase in the unemployment rate is estimated to increase circulatory diseases with 0.265%.

B) The R-squared is defined as follows.

$$R^2 = \frac{SSE}{SST}, \qquad 0 \leq R^2 \leq 1$$

The R-squared is the explained sum of squares over the total sum of squares. They are defined as following.

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - \hat{\bar{y}})^2$$

The total sum of squares is the total variance in y and the explained sum of squares is the variance in y that is explained by x or x's. This means that in the model with a higher R-squared, more of the variance in the log circular diseases is explained by the explanatory variables. However, this is not what The Department of Health should think about when they are deciding which model to use.

What they should really care about is which model better predicts the effect of health expenditures on circular diseases. For the model to unbiasedly (and precisely) predict the effect of health expenditures they need MLR1-MLR6 to hold. MLR1-MLR3 is that the model needs to be linear in parameters, the sampling of observations needs to be random and there needs to be enough variation in the x's ($V(x_i) \neq 0$). Since we cannot observe the dataset, there is not much we can say about these assumptions.

The more interesting to discuss in this case is MLR4, the zero conditional mean assumption ($E(u_i|x_i) = 0$). This assumption states that the expected value of the error term should be zero and that the independent variable should be uncorrelated with the error term. The expected value of the error term, or the expected value of the residuals which is our closest estimate of the error term, is zero. This we know because if it is not zero, we can fix this simply by moving the intersect up and down. The second part of the assumption is usually the more challenging. This is that the independent variable should be uncorrelated with the error term. Normally, since we normally don't know the population regression function, this is an assumption that is difficult not to break. In our case we have that even though this assumption is probably broken in both the models, the model in column (2) is doing a better job at trying to minimize this. That is because unemployment is likely to be correlated with both health expenditures and circulatory diseases, meaning that excluding it in model (1) will give a biased result. Having a lot of circulatory diseases in a community will probably increase the number of unemployed and increase health expenditures. Another variable that might be related to both are age of population, if we assume that older people might have higher health expenditures and are more likely to experience circulatory diseases.

Next, we have MLR5 and MLR6. MLR5 is homoskedasticity assumption ($V(u_i|x_i) = \sigma^2$). This is the assumption that variance in the error term should not vary across different values of the independent variable. Since breaking this assumption does not really make a bias in the estimation it is not that important here. Finally, we have MLR6 the assumption of normality in the error term. This assumption encompasses MLR4 and MLR5 and will be broken if either of those are broken.

To sum up, the important thing when choosing a model is choosing the best model that best predicts the parameter of interest. This means that one should choose the model that unbiasedly predicts the effect of health expenditures, and to do that the assumptions mentioned above need to hold. The R-squared is not what one should be fixated on, because it only tells how much of the total variance in the dependent variable is explained by the independent variables. This is different from unbiasedly predicting an effect, because in addition to having several explanatory variables we need to include the right explanatory variables not to break assumption 4. However, I would agree that model (2) is better because it includes one more explanatory variable, and thus is a little better than model (1). Still both models are probably biased as there are several other important variables that are missing from both, for example age. This means that The Department of Health should be careful with making policy based on either model.

C) Including a set of dummy-variables, one for each municipality would maybe strengthen the explanatory power of the model. This would be because one would be able to capture the effect that different municipalities might have different average per capita circulatory diseases. The reason for doing this would essentially be to reduce the risk of breaking assumption 4 as explained above.

However, I do not really believe that there is a strong correlation between which municipality you live in and circulatory diseases and health expenditures. This is because the differences in municipalities that might affect both circular diseases and health expenditures not really are because of the specific municipality, but rather other factors that might better be captured in their own variable. Such variables might be average age of the population, wealth of population, and diet. The age has an effect because higher age is increasing the probability of circulatory disease and it is more likely that you need medical help. Diet also effect the risk of diseases and the amount of money spent on health expenditures.

5 Theory question

To address the problem, we will start with fixating the population regression model the young econometricians are trying to estimate.

$$CovidCases = \beta_0 + \beta_1 PrecenceOf5G + u$$

The problem they are facing is that they only have observations of the covid cases when the number of cases exceeds 10 000. Even though this model presents several OLS estimator issues we will start by looking at the issue at hand. Only having observations when y > 10 000 breaks assumption 2 of random sampling and will cause the estimator to become biased and less precise. The reason why we are breaking the random sampling assumption is that the probability of each country being selected is not equal when one only can include countries with more than 10 000 covid cases. In addition, the number of observations will decrease.

The expected result of such a bias if the real β_1 is positive is a downward bias, and the expected result if the real β_1 is negative is an upward bias. This is because when we exclude the smaller observations when β_1 is positive we will get a less steep curve with a higher intersect. When β_1 is negative we will get a less steep curve as well, but now with a lower intersect. In addition, the standard error will increase, because the sample will get smaller.

Either way, the estimated effect of β_1 will be biased. This can be shown by using Monte Carlo simulations, as we did in assignment 3, and this is a good aid in ascertain the properties of the OLS estimators. In a Monte Carlo simulation one decides the true population regression function and based on this draw many samples with many observations. Doing a regression on each sample one can calculate the average effect of the parameters, and thus showing that they do in fact deliver the unbiased results on average. In addition, one can on purpose break different assumptions to show how they alter the results. This way, one can visually get an idea of the importance of the assumptions.

In addition to breaking assumption 2, it is very clear that they are also breaking assumption 4. That is because there are many variables that might be related both to the number of covid cases and the presence of 5G network. One thing is population density. If the population density is high the number of covid cases is likely to increase both relatively and numerally. That is because many people together in a smaller place has a better potential for exponential spread, and it is more cost efficient to build out the 5G network in place where many people can use it per square kilo meter.