1 Do-file

```
clear all

use HTVexam.dta

// 3a

regress lwage educ motheduc fatheduc urban

// 3b
regress lwage educ motheduc fatheduc urban

// 3c

regress lwage educ motheduc fatheduc urban
**unrestrcited model

regress lwage educ urban
**restricted model


// 3d

regress lwage educ motheduc fatheduc urban
```

2 Understading Abstracts

a)

$$STEM_i = \beta_0 + \beta_1 math_i - \beta_2 english_i + \beta_3 boy_i + u_i$$

or

$$ArtsSciences_i = \beta_0 - \beta_1 maths_i + \beta_3 english_i + u_i$$

We see that the dependent variables are STEM or Arts and Social Sciences. It depends on what we want to estimate either STEM or Arts and Social Sciences. The independent

variables are the ranks the students get in math and English and if the student is a boy or not. In the text, it says that the effects of subject ranks on STEM are larger for boys, that is why I put in variable boy as a dummy variable. Math has a positive estimated sign in the top model and minus in the second. Vise versus for English. Boy has a positive sign for top model. The dataset is likely to be a cross-section. It seems that the investigating happened at a specific time, at the end of high school. They have checked all high school students who apply for college at a specific time. But it doesn't say anything specific about it in the abstracts.

b)

$$on - time\ gradution\ rates_{it} = \beta_0 - \beta_1 switchtosemester_{it} + u_{it}$$
$$first - year\ grades_{it} = \beta_0 - \beta_1 switchtosmester_{it} + u_{it}$$
$$enrollingfullcourse_{it} = \beta_0 - switchtosemester_{it} + u_{it}$$
$$timingofmajorchoice_{it} = \beta_0 - switchtosemester_{it} + u_{it}$$

As clearly seen, the independent variable is switching from quarters to semesters. It is found switching to semesters negatively impacts on-time graduation rates. It is also found that shifting to a semester lowers first-year grades, decreases the probability of enrolling in a full course load and delays the timing of major choice. Therefore, the sign in front of switch to semester (independent variable) is minus. It has a negative impact on the dependent variables, on-time graduation rates, first-year grades, probability of enrolling in a full course load and timing of major choice. In the abstracts, they say that they use panel data. They check people´s dependent variable when there are quarters and when there are semesters. That is why I am using it as subscripts.

3 Estimation in Stata

a)

```
. 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
------------------------------------------------------------------------------
```

$$\log(\widehat{wage}) = 1.0235 + 0.069educ + 0.0202motheduc + 0.0148fatheduc + 0.101urban$$

The results suggest that, all else equal:

$\widehat{\beta_0}, \mathbf{1.0235}, \boldsymbol{Intercept}$: If educ, motheduc and fatheduc is the lowest grade and doesn't live in urban area then we expect wage to be approximately to be 1.02 dollars per hour.

$\widehat{\beta_1}, \mathbf{0.069}, \boldsymbol{educ}$: An increase of a higher grade, educ are expected to increase hourly wage by 6,9 percent (0,069*100%).

$\widehat{\beta_2}, \mathbf{0.0202}, \boldsymbol{motheduc}$: An increase of a higher grade of highest grade by the mother is expected to increase hourly wage by 2,02 percent (0,0202*100%).

$\widehat{\beta_3}, \mathbf{0.0148}, \boldsymbol{fatheduc}$: An increase of a higher grade of highest grade by the father is expected to increase hourly wage by 1,48 percent (0,0148*100%).

$\widehat{\beta_4}, \mathbf{0.101}, \boldsymbol{urban}$: If the working man has residence in an urban area, then we expect hourly wage to increase by 10,1 percent (0,101*100%).

b)

```
. 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

Testing procedure

1. Specify H_0 and H_A

$\boldsymbol{H_0: \beta_4 = 0}$, Living in an urban are doesn't impact hourly wage

$\boldsymbol{H_A: \beta_4 \neq 0}$, Living in an urban area has a impact on hourly wage.

P-value is the smallest level where we can reject the null hypothesis. In Stata, they calculate the t-value by checking against the null hypothesis stated. They then calculate the p-value for the t-value. It then calculates for the smallest significance level for which we can reject H_0.

When using the p-value reported in Stata, we can compare the significance level and the p-value reported in Stata. When:

$$\alpha\ (significance\ level) > p - value$$

Then we reject the null hypothesis.

If we have the opposite:

$$\alpha\ (significance\ level) < p - value$$

Then we fail to reject the null hypothesis.

In Stata, p-value for urban is 0.114. This means that we can reject the null hypothesis at the lowest significance level of 0,114 or 11,4%. We can set a very high significance level, for example 20% and we will reject the null hypothesis. This means that urban is statistically significant.

$$0.2 > 0.114$$

But setting such a high significance level isn't of much help. Checking the alternative hypothesis with a lower significance level is much better. Therefor we set a significance level of 10%, 0,01. Remember that we are checking against a two-sided test. We check the significance level against the p-value.

$$0.1 < 0.114$$

The p-value is larger than the significance level. This means the we fail to reject the null hypothesis with 10% significance level. We don't have enough evidence that living in an urban area has an impact on hourly wages.

c)

Since I am testing mother and father education have a joint effect I need to use F-test.

Testing procedure for F-statistic

1. Specify H_0 and H_A

$\boldsymbol{H_0: \beta_2, \beta_3 = 0,}$ Mother and father education doesn't have any effect on individual wages.

$\boldsymbol{H_A: not\ H_0,}$ Mother and father has a jointly effect on individual wages in some form.

```
. 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
------------------------------------------------------------------------------

. **unrestrcited model
.
. 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
------------------------------------------------------------------------------
```

2. Define the F-statistics (F-stat)

$$F-stat = \frac{\frac{SSR_r - SSR_{ur}}{q}}{\frac{SSR_{ur}}{n-k-1}} = \frac{SSR_r - SSR_{ur}}{SSR_{ur}} * \frac{n-k-1}{q}$$

or

$$F-stat = \frac{R^2_{ur} - R^2_r}{1-R^2_{ur}} * \frac{n-k-1}{q}$$

$R^2_u = 0.1636$

$\boldsymbol{R_r^2 = 0.1465}$

$\boldsymbol{n = 496}$ **observations**

$\boldsymbol{k = 4}$ **parameters in the unrestricted regression**

$\boldsymbol{q = 2}$ **restrictions in the restricted regression**

$$\boldsymbol{F - stat = \frac{R_{ur}^2 - R_r^2}{1 - R_{ur}^2} * \frac{n - k - 1}{q} = \frac{0.1636 - 0.1465}{1 - 0.1636} * \frac{496 - 4 - 1}{2} = 5.019}$$

3. Define the distribution $\boldsymbol{F - stat \sim F_{q,\ n-k-1}}$

We have a F-distribution. Our model has n=496 observations, k = 4 variables and q = 2 restrictions.

$$\boldsymbol{FS \sim F_{2,\ 496-4-1} = FS \sim F_{2,\ 491}}$$

4. Define the rejection region based on a significance level

I want to test with a 5% significance level. This is the usual significance level to use.

The rejection region is:

$$F - stat > c$$

$$F - stat > 2.9957$$

$$5.019 > 2.9957$$

5. Conclude

Since F-stat > c, we can reject the null hypothesis with a significance level of 5%. We can conclude that we have enough evidence to say that mother and father education are jointly statistically significant related to individual wages in some form. It seems that parents' education effects in one form or another individual wage.

d)

```
. 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
```

There are several ways to answer if motheduc and fatheduc are individually significant. I can look at the p-value for the variables and compare it to a significance level or check with the 95% confidence level.

Motheduc:

It has a p-value of 0.104. It means we can reject the null hypothesis at a lowest significance level of 10.4% for a two-tailed test. Since State check these hypothesis:

$$\boldsymbol{H_0: \beta_2 = 0}$$

$$\boldsymbol{H_A: \beta_2 \neq 0}$$

That is why it is a two-tailed test.

As stated earlier, we can compare significance level with the p-value to see if motheduc or fatheduc are individually significant.

I want to check for a significance level of 10%.

$$0.104 > 0.1$$

This means that motheduc is individually insignificant, therefor we fail to reject the null hypothesis. It seems that motheduc doesn't have an effect on individual wage.

Fatheduc:

It has a p-value of 0.102. So here again, we compare the significance level with the p-value. The p-value is larger than the significance level.

$$0.102 > 0.1$$

Fatheduc is individually insignificant. We fail to reject the null hypothesis. It seems that fatheduc don´t have an effect on individual wage.

Another way to test individually significant, is to check if the null hypothesis is within the confidence interval that is stated in Stata.

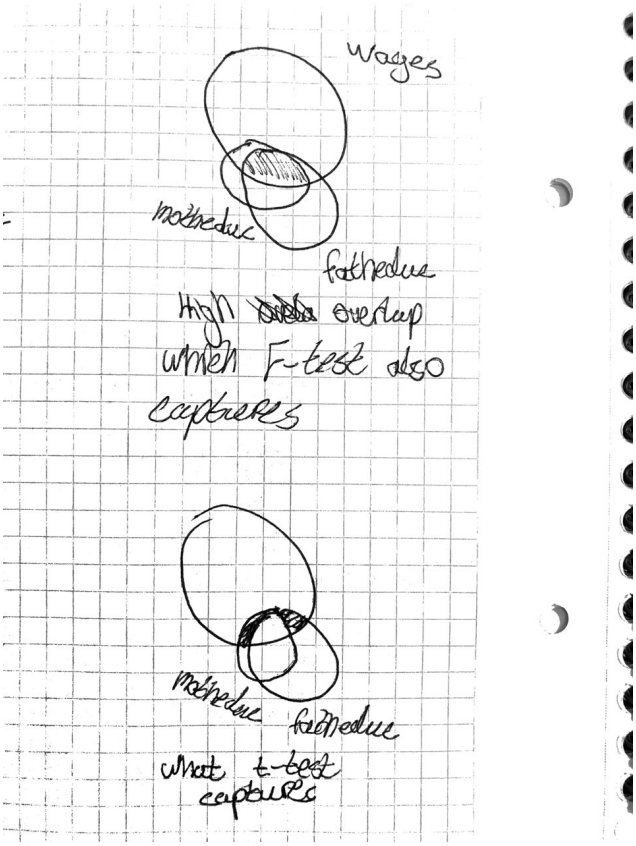
$$\boldsymbol{H_0: \beta_2 = 0}$$

If zero is within the confidence interval then we fail to reject the null hypothesis. If it is not in the confidence level, then we can reject the null hypothesis. And seen in the State table, zero is within the confidence interval for both variables. We therefore fail to reject the null hypotheses.

e)

I would say they are compatible. When doing a regression/OLS, we want to find the specific effect x_i has on y. When doing a regression, the procedure will net out/partialled out the effect different x´s have in common on y. We found out that motheduc and fatheduc wasn't individually significant, but jointly significant. It seems that motherduc and fatheduc correlate/overlapa bit. There is high collinearity between the two variables. Their specific effect on individual wages isn't much and therefore isn't individually significant. Their joint effect (overlap part) gets partialled out in a student-test. When doing a F-test, the test will capture their joint effect on individual (overlapped part). That is why they are jointly significant but not individual significant.

In the figure, I didn't draw the in the other variables, because it will ruin the intuition and make it messier.



4 Interpretation of an empirical analysis.

a)

The results suggest that, all else equal:

$\widehat{\boldsymbol{\beta_1}}, \mathbf{0.436}, \boldsymbol{log\ health\ expenditures}$: This a log-log-model. Which means:

$$\beta_1 = \frac{\frac{dy}{y}}{\frac{dx}{x}} = \frac{relative\ change\ in\ y}{relative\ change\ in\ x}$$

A change by one percent in x equals a $\beta_1\%$ change in y.

An increase in health expenditures by 1% is expected to increase circulatory diseases by 0.436% (0.436%).

$\widehat{\boldsymbol{\beta_2}}, \mathbf{0.265}, \boldsymbol{log\ unemployment}$: An increase in unemployment by 1% is expected to increase circulatory diseases by 0.265% (0.265%).

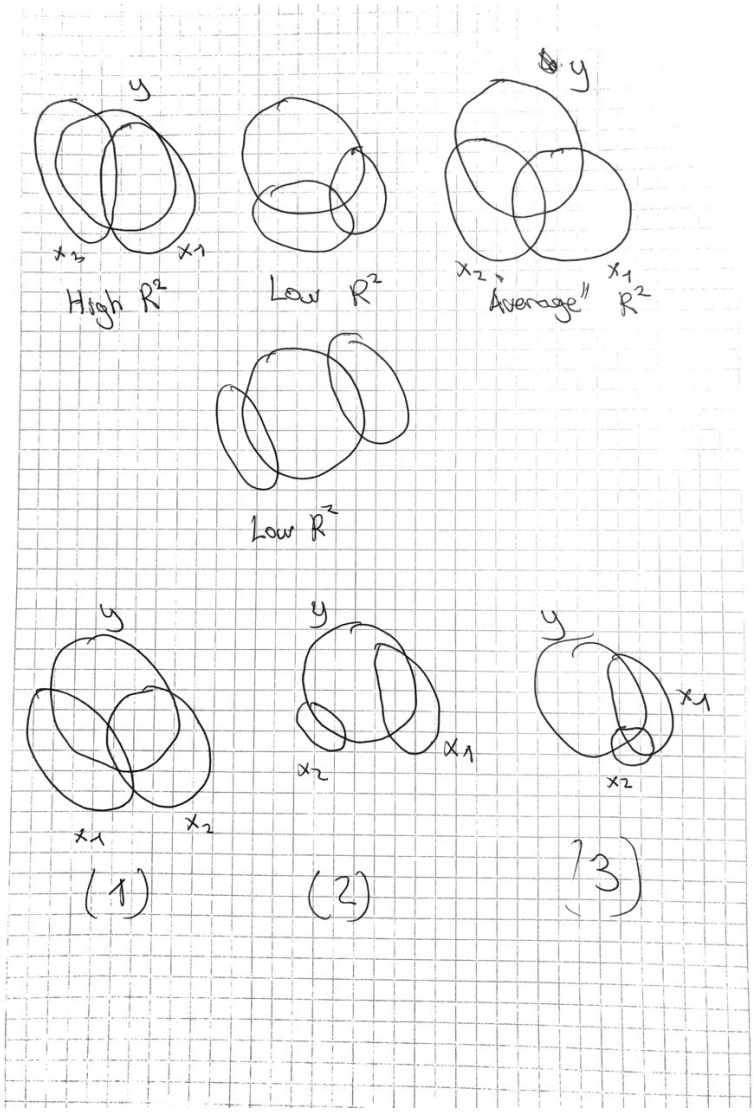
b)

R^2 is a measure of how much our model explain the dependent variable by using total variations in y that is explained by x. We can state the formula for R^2as:

$$R^2 = \frac{SSE}{SST} = \frac{explained\ sum\ of\ squared}{total\ sum\ of\ squares}$$

When thinking of Ballantine diagrams, we what to capture the specific effect x has on y. The more specific overlap between a x variable and y variable the higher the R^2 gets.

I wouldn't necessary agree with the Department of health. Adding another variable will most likely increase R-squared since more of the y diagram is overlapped with the new variable. Just comparing the two R-squared, it goes from 11% to 27%. Model 2 captures 27% on the effect on y. It isn't much, but better than 11%. The question we need to answer is how important is the new variable? In this case, it is log of the unemployment rate. Adding this variable, does it have an impact of circulatory diseases? Is it a right variable to include to use for their policy conclusions? Adding a variable will increase R-squared, but that doesn't mean with should use the model with the highest R-squared, since in the real world the variable might not be realistic significant. Unemployment might increase circulatory diseases, since people won´t work and get the "brain-exercise" by working. Or being employed might stress people or get people to eat fast food since they are exhausted from work. Therefore, it might increase peoples chance of getting circulatory diseases. I therefor disagree with the views of the Department of Health. The model of 2 isn't an appropriate model to use. I would argue that model 2 is close to the model 3. See figure under. Under I have drawed some Ballentine figures around this problem. Figure 2 has non correlation between x_1 and x_2. While figure 1 has some correlation between x´s and high overlap with y. If unemployment would have high effect on diseases, it would look something like figure 1.



c)

These dummy variables capture each municipality circulatory diseases, health expenditures and unemployment. We will therefore get more observations. They capture average differences in circulatory diseases across municipality. A reason for including them is to see how health expenditures and unemployment effect circulatory diseases at a municipality level. It allows circulatory diseases to vary on average across municipalities. By including them the Department of health can look at the intercept for each municipality, but the effect the independent variables on the dependent variables are the same. Rather than deciding a policy for the whole of Norway, they can set the new policy conclusions after what they find from the new model with dummy variables for each municipality.

5 Theory question

I would make simulations (Monte Carlo) in Stata with their dataset. I would like to take different random samples from their dataset. I would then collect the different β´s and then take the average of all the different β´s to see the effect total number of 5G operators has on Covid cases. By doing this, I use the central limit theorem. I take a sample of the dataset. Find the β and collect it. Then I take another sample of the dataset. Remember it must be random. Find the β and collect it. Then after doing this x amount of times. I would recommend doing it many times. 200, 1000 maybe. I can use State to create a loop to collect the different $\beta´s$.

Before doing the process of collecting the data, we need some assumptions to hold for this to work.

I assume it is a single linear regression

$$covidcases_i = \beta_0 + \beta_1 number\ of\ 5G\ operators + u_i$$

$$\widehat{covidcases_\iota} = \widehat{\beta_0} + \widehat{\beta_1} number\ of\ 5G\ operators$$

SLR.1

The model is to be linear in the parameters. As seen over it is.

SLR.2

The dataset must be randomly sampled. This implies that each i has the same probability of being selected. By getting rid of the countries that doesn't report Covid cases since data aren't collected, breaks this assumption. The estimation of the β will be biased, since we have removed a section of our observations. The bias is likely to be in the downward direction and the intercept will increase.

SLR.3

There needs to be enough variability in the x_i.

$$V(x_i) \neq 0$$

This is because

$$\widehat{\beta_1} = \frac{covariance_{xy}}{variance\_x}$$

The estimated beta will be zero. I assume that the dataset has variability in covid cases. Highly unlikely that each country has the same amount of covid cases.

SLR.4

$$E(u|x_i) = 0$$

What this means is that the average value of the error term is zero across different x-values of the population. And it means that error term and x doesn't correlate. This assumption is most likely broken. There might be other variables that are in the error term that effect the amount covid cases. Other variables outside the model, is correlated with other variables and the dependent variable. Example, the population in the country might effect number of 5 operators and the amount of covid cases. Not all countries has 5G yet.

SLR.5

Homoskedacitiy, which means same variability:

$$V(u|x_i) = \sigma^2$$

The variability of error across x needs to be constant. This is most likely broken. Some countries have a high population, but not a lot of covid cases. And vise versus.

SLR.6

Normality and it states that the error term has a normal distribution as:

$$u \sim N(0, \sigma^2)$$

If SLR.4 or SLR.5 is broken then SLR.6 is broken.

Under these assumptions, the OLS estimator is unbiased. The procedure is described over would work. It is the procedure that is unbiased under these assumptions, it is not the estimation from different samples are not unbiased. Therefore, we create a loop as mentioned over to get the estimated unbiased estimator, but it would be biased since some observations are gone. If none of the assumptions were broken and had data on all countries, then we could find the best linear unbiased estimator (BLUE) and find the effect of 5G operators on covid cases.