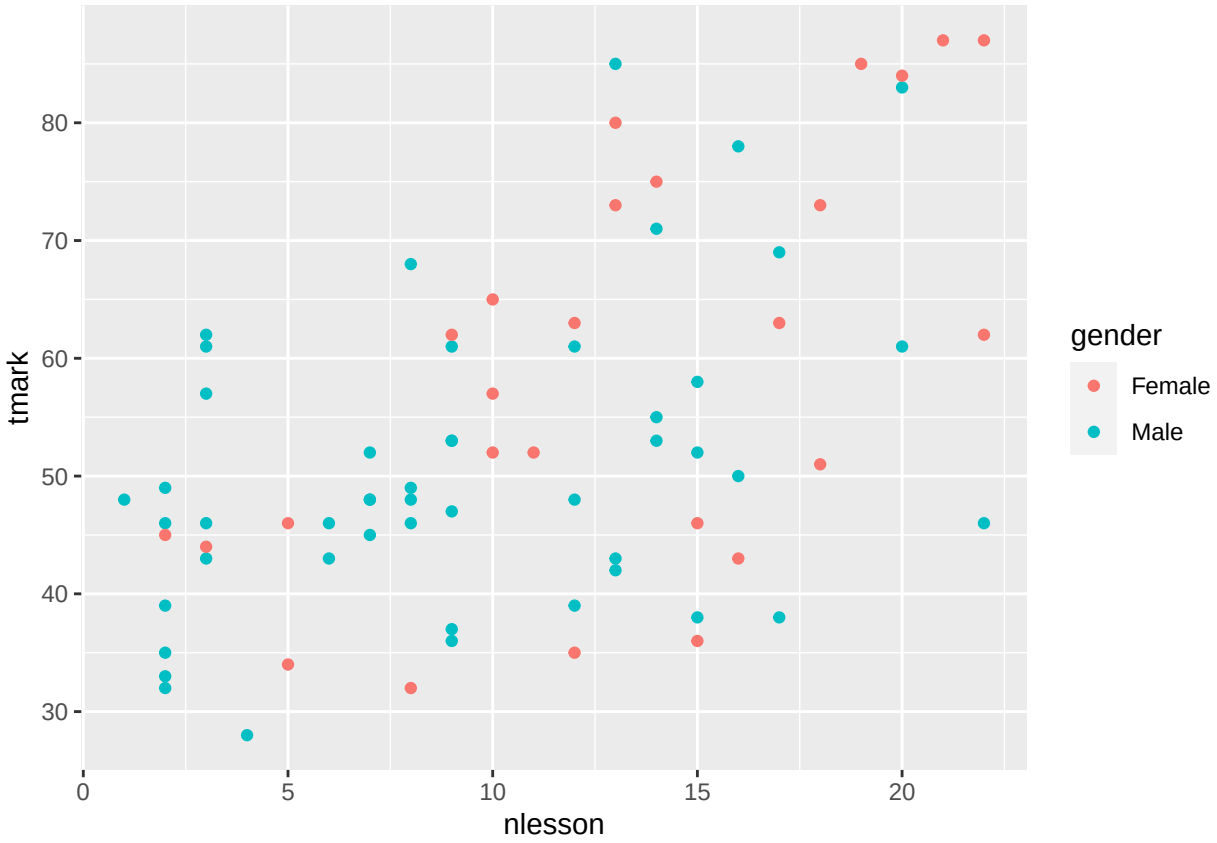
Introduction

In this report, we are going to investigate how seminar attendance (and various other factors) affects test performance. The dataset we are going to use is in Excel format. This data is cross-sectional, with each observation corresponding to one student, and contains information on the total number of seminars attended, the total grade obtained for the subject, as well as a host of other potential explanatory factors for a sample of 100 students in Uzbekistan studying Fundamentals of Statistics course in 2018. Please, see the attached document titled "Description of the variables" for the full information on the all available variables.

First, we viewed, cleaned, examined, graphed the data and started with an ordinary least squares model of the form

$$tmark_i = \beta_0 + \beta_1 nlesson_i + u_i$$

, where *tmark* is the final exam score and *nlesson* is the total number of seminars attended during the semester. The resuslt are displayed below



The interpretation of the constant here is the total mark that would be achieved without any seminars attended. The R-squared is about 28%. This means that a linear model with seminar attendance explains a substantial part of variation in total mark away from its mean. The results of a Breusch-Pagan are also shown. It appears that there is heteroscedasticity across the nlesson variable as we reject the null of no heteroscedasticity at any reasonable significance level. Logging the dependent variable has had some effect of reducing the dispersion, however not sufficiently so as to remove the heteroscedasticity at the 5% significance. Eventough the adjusted R-squares is slightly lower in the logged model, the substantial reduction in the standard errors lead us to conclude this may be a better specification. The coefficient on *nlesson* suggests that an additional seminar causes on average a 2.4% increase in total mark.

Table 1:

	Dependent variable:	
	mark_t_FoS	log(mark_t_FoS)
	(1)	(2)
nlesson	1.362*** (0.253)	0.024*** (0.005)
Constant	39.021*** (3.072)	3.687*** (0.057)
Observations	74	74
R^2	0.288	0.267
Adjusted R^2	0.278	0.257
Residual Std. Error (df = 72)	12.748	0.236
F Statistic (df = 1; 72)	29.061***	26.253***
Note:		*p<0.1; **p<0.05; ***p<0.01

Instrumental variables estimation

There are likely a number of omitted factors which are correlated with nlesson. This means that nlesson is likely endogenous, causing the OLS estimates of parameters be both biased and inconsistent.

An example of an ommited factor might be an individuals interest in the class, where we would expect a positive correlation between interest and seminar attendance. Some measures of ability, effort or preparation are also likely correlated with seminar attendance and with the total mark. These characteristics encourage students to attend seminars and to increase their marks on exams at the same time. Hence they are positively linked to both independent and dependent variables, which means that the the estimate of seminar attendance effect over-estimated.

For simplicity, let us consider the structural equation of the form

$$tmark_i = \beta_0 + \beta_1 nlesson_i + u_i$$

where we expect β_1 to be positive as attending Statistics seminars should benefit the students.

Endogeneity from an omitted variables bias can be eliminated (or at least mitigated) when a suitable proxy variable is given for an unobserved explanatory variable. Thus, we considered entrance test scores as proxies for ability and the number of files which a student downloaded together with the the number of prescribed chapters read as proxies for interest.

$$tmark_i = \beta_0 + \beta_1 nlesson_i + \beta_2 entmath_i + \beta_3 entIELTS_i + \beta_4 nfile + \beta_5 chapters + u_i$$

with $\beta_i, i = 2, \ldots 5$ also expected to be positive.

Another way to get around the issue of endogenous regressors is to use an appropriate instrument. A good IV for *nlesson* has no direct effect on score and is not correlated with student ability and motivation. Let us evaluate whether *working* (whether a student of the Statistics course was working part time in 2018) is likely to be a reasonable instrument to use for seminar attendance).

Part- time work should affect *nlesson*, because having having a job will make an individual less likely to attend seminars. It also is likely uncorrelated with the omitted factors (important in determining *tmark*) in our regression. One could argue that it influences ability and interest, but in our view this is likely a quite weak effect. Hovewer, full-time work is more likely to be correlated with interest and ability. It has been fairly well established that socioeconomic status affects student performance. The error term u contains, among other things, family income, which has positive effect on test scores. In that sense, *working* is certainly not exogenous as it is also very likely correlated with family income.

The requirements for an instrument z to be consistent are that it satisfies $Cov(z, u) = 0$; (instrument exogeneity) and that z must also be related, either positively or negatively, to the endogenous explanatory variable x, i.e. relevant. We are going to try using the *working* variable, expecting negative correlation with *nlesson* and verifing it by regressing *nlesson* on *ptwork* and similarly defined $ftwork$, in turn performing a 2SLS first stage estimation. The estimated sign and even magnitude of the parameters corresponds to the predictions outlined above. We would argue our instruments are sufficiently strong since they explain about 7% of the variance in *nlesson*, likely outweighing the correlation with omitted variables.

Table 2:

	Dependent variable:
	nlesson
ptwork	-3.135^{**} (1.391)
ftwork	-5.147^{**} (2.369)
Constant	12.576^{***} (0.991)
Observations	74
R^2	0.096
Adjusted R^2	0.071
Residual Std. Error	5.692 (df = 71)
F Statistic	3.777^{**} (df = 2; 71)
Note:	*p<0.1; **p<0.05; ***p<0.01

Results

```
##
##  2SLS Estimates
##
## Model Formula: log(mark_t_FoS) ~ nlesson
##
## Instruments: ~ptwork + ftwork
##
## Residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.55969 -0.18663  0.05605  0.00000  0.16696  0.48454
##
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 3.52484478 0.17482859 20.16172  < 2e-16 ***
## nlesson     0.03924948 0.01618363  2.42526 0.017806 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2531627 on 72 degrees of freedom
```

We find that our instrument does not work as expected. The return to 1 incremental seminar from OLS estimates, about a 2.4% increase in total mark, is smaller than that suggested by instrumental variables estimators, at around 4%!

At this point we decide to control for the unobserved factors. Considering the proxy model as our new structural equation, we arrive to a model which should be consistent as long as the 2SLS assumptions of Ínearity in parameters, random sampling, rank condition and exogenous instrumental variables:

```
##
##  2SLS Estimates
##
## Model Formula: log(mark_t_FoS) ~ nlesson + ent_math + ent_ielts + nfile + chapters
##
## Instruments: ~ptwork + ftwork + ent_math + ent_ielts + nfile + chapters
##
## Residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.44870 -0.10887  0.01861  0.00000  0.10979  0.40736
##
##                 Estimate   Std. Error  t value   Pr(>|t|)
## (Intercept)  2.293734638  0.359580565  6.37892 1.8297e-08 ***
## nlesson      0.022599289  0.020197199  1.11893  0.2671063
## ent_math     0.002917758  0.002320305  1.25749  0.2128782
## ent_ielts    0.191263011  0.046559375  4.10794  0.0001095 ***
## nfile        0.001934344  0.001277433  1.51424  0.1345988
## chapters    -0.014563417  0.008294110 -1.75587  0.0836128 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1900992 on 68 degrees of freedom

##
## z test of coefficients:
##
##               Estimate Std. Error z value  Pr(>|z|)
## (Intercept)  2.2937346  0.3595806  6.3789 1.783e-10 ***
## nlesson      0.0225993  0.0201972  1.1189   0.26317
## ent_math     0.0029178  0.0023203  1.2575   0.20858
## ent_ielts    0.1912630  0.0465594  4.1079 3.992e-05 ***
## nfile        0.0019343  0.0012774  1.5142   0.12996
## chapters    -0.0145634  0.0082941 -1.7559   0.07911 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis under the Hausman test is that both models produce consistent estimates. With a p value close to 1, we cannot reject the null hypothesis. This result corresponds to the fact that that the return to 1 incremental seminar from OLS estimates, at about a 1.9% increase in total mark, is now closer to the IV estimation at around 2.3% In the regression-based test for endogeneity, the coefficient of "res" is insignificant. Therefore, neither of these tests proves endogeneity in the structural model. We can test whether the IVs are exogenous as we have more instruments. With a very small p-value, we reject that all the instruments are exogenous. The errors on the estimated slope coefficient from instrumental variables estimates are roughly five times those of OLS! This is typical of instrumental variables estimators.As the BP test still indicates the presence of heteroscedasticity, fGLS is needed in order to achieve robust inference. Please view the attached .Rmd and consult Wooldridge (Introductory econometrics: a modern approach, 7th ed., 531-533) if necessary. The paper from Card (1995), http://davidcard.berkeley.edu/papers/geo_var_schooling.pdf inspired our methodology.

Table 3: Structural model

	Dependent variable:
	log(mark_t_FoS)
nlesson	0.019^{***} (0.004)
ent_math	0.003 (0.002)
ent_ielts	0.191^{***} (0.046)
nfile	0.002^{***} (0.001)
chapters	-0.015^{*} (0.008)
Constant	2.331^{***} (0.287)
Observations	74
R^2	0.557
Adjusted R^2	0.524
Residual Std. Error	0.189 (df = 68)
F Statistic	17.093^{***} (df = 5; 68)
Note:	*p<0.1; **p<0.05; ***p<0.01