1. Theory

Question 1

a) i)

Given by lecture:

$$\text{TSS} = \text{ESS} + \text{RSS}$$

$$\text{TSS} = \Sigma_{i=1}^{n} y_i^2$$

$$\text{ESS} = \Sigma_{i=1}^{n} \hat{y}_i^2$$

$$\text{RSS} = \Sigma_{i=1}^{n} e_i^2$$

Formal proof of TSS = ESS + RSS:

$$\text{TSS} = \Sigma_{i=1}^{n} y_i^2 = \Sigma_{i=1}^{n} (y_i - \bar{y})^2 = \Sigma_{i=1}^{n} (y_i - \bar{y} + \hat{y}_i - \hat{y}_i)^2 = \Sigma_{i=1}^{n} \big((\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)\big)^2 =$$

$$\Sigma_{i=1}^{n} \big((\hat{y}_i - \bar{y}) + \hat{e}_i\big)^2 = \Sigma_{i=1}^{n} ((\hat{y}_i - \bar{y})^2 + 2\hat{e}_i(\hat{y}_i - \bar{y}) + \hat{e}_i^{\,2}) = \Sigma_{i=1}^{n} (\hat{y}_i - \bar{y})^2 +$$

$$2\Sigma_{i=1}^{n} \hat{e}_i(\hat{y}_i - \bar{y}) + \Sigma_{i=1}^{n} \hat{e}_i^{\,2} = \Sigma_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \Sigma_{i=1}^{n} \hat{e}_i^{\,2} + 2\Sigma_{i=1}^{n} \hat{e}_i(\widehat{\beta_0} + \widehat{\beta_1 x_{i1}} + (\dots) +$$

$$\widehat{\beta_k x_{ik}} - \bar{y}) = \Sigma_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \Sigma_{i=1}^{n} \hat{e}_i^{\,2} + 2\big(\widehat{\beta_0} - \bar{y}\big)\Sigma_{i=1}^{n} \hat{e}_i^{\,2} + 2\widehat{\beta_1}\Sigma_{i=1}^{n} \hat{e}_i x_{i1} + (\dots) +$$

$$2\widehat{\beta_k}\Sigma_{i=1}^{n} \hat{e}_i x_{i1} = \Sigma_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \Sigma_{i=1}^{n} \hat{e}_i^{\,2} = \Sigma_{i=1}^{n} \hat{y}_i^2 + \Sigma_{i=1}^{n} \hat{e}_i^{\,2} = ESS + RSS$$

ii)

Formal proof of $R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{e\prime e}{\tilde{y}\prime\tilde{y}}$

$$R^2 = \frac{ESS}{TSS} = \frac{\Sigma_{i=1}^{n} \hat{y}_i^2}{\Sigma_{i=1}^{n} y_i^2} = \frac{ESS}{ESS + RSS} = \frac{\Sigma_{i=1}^{n} \hat{y}_i^2}{\Sigma_{i=1}^{n} \hat{y}_i^2 + \Sigma_{i=1}^{n} e_i^2} = \frac{\Sigma_{i=1}^{n} \hat{y}_i^2 + \Sigma_{i=1}^{n} e_i^2 - \Sigma_{i=1}^{n} e_i^2}{\Sigma_{i=1}^{n} \hat{y}_i^2 + \Sigma_{i=1}^{n} e_i^2}$$

$$= \frac{\Sigma_{i=1}^{n} y_i^2 - \Sigma_{i=1}^{n} e_i^2}{\Sigma_{i=1}^{n} y_i^2} = 1 - \frac{\Sigma_{i=1}^{n} e_i^2}{\Sigma_{i=1}^{n} y_i^2} = 1 - \frac{RSS}{TSS} = 1 - \frac{e'e}{\tilde{y}'\tilde{y}}$$

b)

Formal proof of $R^2 = corr^2(\mathbf{y}, \hat{y}) = \rho_{y,\hat{y}}^2 = \frac{ESS}{TSS}$

➔ First take the square root of $\rho_{y,\hat{y}}^2$

Tarik Benli 15-719-818, Ramon Gmür 16-705-220

$$\rho_{y,\hat{y}} = \frac{\sigma^2_{y,\hat{y}}}{\sqrt{\sigma^2_y * \sigma^2_{\hat{y}}}} = \frac{\Sigma^n_{i=1}(y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\Sigma^n_{i=1}(y_i - \bar{y})^2 * \Sigma^n_{i=1}(\hat{y}_i - \bar{y})^2}} = \frac{\Sigma^n_{i=1}(y_i + \hat{y}_i - \hat{y}_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\Sigma^n_{i=1}(y_i - \bar{y})^2 * \Sigma^n_{i=1}(\hat{y}_i - \bar{y})^2}}$$

$$= \frac{\Sigma^n_{i=1}\left(y_i\hat{y}_i + \hat{y}_i^2 - \hat{y}_i^2 - \bar{y}\hat{y}_i - \bar{y}y_i + \hat{y}_i\bar{y} - \bar{y}\,\hat{y}_i + \bar{y}^2\right)}{\sqrt{\Sigma^n_{i=1}(y_i - \bar{y})^2 * \Sigma^n_{i=1}(\hat{y}_i - \bar{y})^2}}$$

$$= \frac{\Sigma^n_{i=1}((y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + (\hat{y}_i - \bar{y})^2)}{\sqrt{\Sigma^n_{i=1}(y_i - \bar{y})^2 * \Sigma^n_{i=1}(\hat{y}_i - \bar{y})^2}}$$

$$= \frac{\Sigma^n_{i=1}(\hat{y}_i - \bar{y})^2}{\sqrt{\Sigma^n_{i=1}(y_i - \bar{y})^2 * \Sigma^n_{i=1}(\hat{y}_i - \bar{y})^2}} = \sqrt{\frac{\Sigma^n_{i=1}(\hat{y}_i - \bar{y})^2}{\Sigma^n_{i=1}(y_i - \bar{y})^2}} = \sqrt{\frac{ESS}{TSS}}$$

➔ And now square it: $\frac{ESS}{TSS} = \rho^2_{y,\hat{y}} = corr^2(\mathbf{y}, \hat{y}) = R^2$

One can interpret R^2 as the squared correlation coefficient between the true value y_i and the estimated value $\hat{y}_i$. In a regression model R^2 measures how good the estimated value explains the true value. In other words, it explains the variation in the estimated $\hat{y}_i$ and its true value y_i. Thus, this can be translated to the correlation between these two variables as shown above. One can generally say: The higher R^2 or $\rho^2_{y,\hat{y}}$ the better the model can predict the true values where the R^2 is in a range between 0 and 1 and the correlation between -1 and 1.

c) By transforming our X variables linearly, one cannot gain more information out of our variables than before. For each datapoint the true value of y_i and the estimated value of $\hat{y}_i$ do not change at all by linear transformation. It neither change the relation nor the underlying data. Since we have showed above that $R^2 = \rho^2_{y,\hat{y}}$, one can see that the correlation between y_i and $\hat{y}_i$ is not affected either and the calculated R^2 remains the same.

d) Intuitively the residual sum squared always decreases if one adds another regressor into the model (see formal proof 1.e) and so R^2 will increase. That is because with a new regressor one might gather more information from the data, since the estimated values may be predicted more precisely. Only in a special case, if the added regressor is perfectly correlated with an already existing one in the model, the RSS would stay the same, since one cannot gather more information by adding this new but perfectly correlated variable.

Tarik Benli 15-719-818, Ramon Gmür 16-705-220

e)

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\Sigma_{i=1}^{n} e_i^2}{\Sigma_{i=1}^{n} y_i^2} = 1 - \frac{\Sigma_{i=1}^{n} \left(y_i - x_i\hat{\beta} - \bar{e}\right)^2}{\Sigma_{i=1}^{n} y_i^2}$$

If one adds an additional regressor x_i, the $\hat{\beta}$ will increase and therefore the RSS is getting smaller as a direct consequence of this. Furthermore, R^2 will increase. In other words, the precision of the prediction is increasing until the error term is hypothetically equal to zero.

f) Working with R^2 could lead to multiple types of misinterpretation:

- The **type** of data implies different values of R^2. While time-series data often have a high R^2, the exact opposite is often true for cross-section data.
- Different functional forms for y_i can **change** R^2 (for example: using $\log(y_i)$ often increases R^2)
- Adding explanatory variables to a model always **increase** R^2
    - This may lead one towards overfitting the model with too many variables. So, one can increase the R^2 even by just adding a large set of totally random predictors.

Furthermore, it is only an in-sample measurement: That means R^2 only measures the precision within the sample. A better way to measure the precision between samples is for example the measurement techniques of Cross-Validation.

Tarik Benli 15-719-818, Ramon Gmür 16-705-220

2. Empirical Question

a) Big school dummy

i) The coefficient on *classize* is 0.134 and statistically significant. In this case, this means that the average marks in a grammar tests rises by 0.134 per additional student in class.

ii) The new coefficient on *classize* shrinks to 0.102, the one of the new dummy *big school* is 1.246 (both are statistically significant). This means that, while an additional student only brings a rise in the average grammar test marks of 0.102 by controlling for school size, the bigger schools have higher averages.

b) Natural log

i) The coefficient is -0.0603 on *classsize* (-0.0007 in the log-model) and -0.335 on *pct_dis* (-0.0048 in the log model). The effect of *classize* is now even smaller than before and now slightly negative. The effect of the percentage of disadvantaged kids is also negative. With a rise of the amount of disadvantaged kids of 1%, the average mark decreases by -0.335 points. The model with the logged grammar scores shows the coefficients as an (approximated) percentage change with respect to the constant. The coefficients in the log model show percentage changes, meaning with one unit change in *classize* or *pct_dis*, the grammar marks change by 0.07% and 0.5%.

ii) In the regression of grammar scores on *classize* and *pct_dis*, the coefficient of the latter variable can be interpreted as follows: If pct_dis rises by one unit, the average grammar mark decreases by 0.335 points controlling for class size.

c) Small size dummy

i) The coefficient of small size tells us that small classes score 2.56 points higher on grammar tests than big classes, controlling for the percentage of disadvantaged kids. In terms of economical significance, this coefficient seems rather large, comparing it with previous coefficients and with respect to the constant. In Stata, the two-sided test $\beta_2 = 0$ yields a p-value of 0.2016. Assuming that a small sizes class has a positive effect on grammar scores, we also use a one-sided test $\beta_2 \leq 0$, yielding a p-value of 0.1008 (which in this scenario is exactly 50% of the two-sided p-value). This means we cannot reject our hypotheses.

We recommend using the two-sided hypothesis test, because with *small_size* we are looking at an extreme attribute of class sizes and should not assume that the effect of *small_size* is positive (only 8 classes fall into the category small sized). This lack of

observations for small class sizes explains why the results are far from significant. Calculation of the hypothesis test that $\beta_2 = 0$ by hand:

1: $H_0: \beta_2 = 0, H_A: \beta_2 \neq 0$

2: Degrees of freedom: $N - K = 1967 - 2 = 1965$

$s^2 = \frac{1}{N-K}\sum_{i=1}^{n} e_i^2$

t-value: $\frac{2.560}{2.004} = 1.277$

3: At the 5% level: $P(t_{(1965)} \leq \overline{t_{0.975}}) = 0.975 \rightarrow \overline{t_{0.975}} = 1.96$

$|1.277| > 1.96 \rightarrow H_0$ cannot be rejected

ii) First, we need to regress the grammar scores only on *pct_dis* and then take the residuals from this model (see table below: 1). Then we regress *small_size* on *pct_dis* too and also take the residuals (see table below: 2). Now we can regress the residuals from the first model one the residuals on the second model. The resulting coefficient is the coefficient of grammar score on small sized classes (and is exactly the same as in i).

Y-VARIABLE:	(mrkgrm)	(small_size)	(residuals1)
VARIABLES	1	2	3
pct_dis	-0.327***	9.57e-06	
	(0.00977)	(0.000110)	
residuals2			2.560
			(2.003)
Constant	77.11***	0.00394*	-6.64e-09
	(0.183)	(0.00206)	(0.128)
Observations	1,967	1,967	1,967
R-squared	0.363	0.000	0.001

iii) To show that $\hat{\beta}_1 = \bar{y} - \bar{X}_{-1}\hat{\beta}_{-1}$, we need the means of *mrkgrm*, *small_size* and *pct_dis*. Now by deducting the means of *small_size* and *pct_dis* multiplied with their respective coefficient from the mean of *mrkgrm*, we get $\hat{\beta}_1$:

$$\overline{mrkgrm} - \overline{small_size} \times 2.5597 - \overline{pct_dis} \times (-0.32677) = 77.099$$

iv) The correct interpretation is that small classes have an average grammar score that is 3.65% higher than in bigger classes.

d) Many disadvantaged dummy

i) The joint hypothesis that $\beta_3 = 0$ and $\beta_4 = 0$ has an F-value of 401.85 (p-value = 0). Calculating this by hand shows the same results (small difference due to rounding):

$$\frac{(R_U^2 - R_R^2)/q}{(1 - R_U^2)/(N - K)} = \frac{(0.3005 - 0.0142)/2}{(1 - 0.3005)/(1966\text{-}3)} = 401.72$$

In conclusion, we reject the joint hypothesis that $\beta_3 = 0$ and $\beta_4 = 0$. Many disadvantaged kids (alone and combined with class size) affects grammar scores.

ii) The effect of having 10 additional students in a class with less than 10% disadvantaged kids is –1.1 (since *many_dis* is a dummy with value 0, both β_3 and β_4 are not relevant in this specific case here)

e) Separated regressions

The table below shows the results separating classes with high and low percentages of disadvantaged kids as well as the results from d). The coefficient on *classize* in (2) is exactly the same as in (3), which makes sense because both only consider class size for classes with less than 10% disadvantaged kids. For classes with more than 10% disadvantaged kids, one can calculate that the separate regressions (1) and (2) and the combined regression (3) also yield the same results:

$$63.81 + classize \times 0.159$$

$$= 79.52 + classize \times (-0.110) + (-15.71) + (classize \times many\_dis) \times 0.269$$

In conclusion, model (3) is a combination of models (1) and (2)

VARIABLES	(1) high dis	(2) low dis	(3) d)
classize	0.159***	-0.110***	-0.110***
	(0.0373)	(0.0255)	(0.0291)
many_dis			-15.71***
			(1.346)
classize × many_dis			0.269***
			(0.0438)
Constant	63.81***	79.52***	79.52***
	(1.103)	(0.821)	(0.937)
Observations	858	1,109	1,967
R-squared	0.021	0.017	0.301

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

f) Region dummies

The region dummies cannot all be included in the same model because of multicollinearity (dummy variable trap). If one wants to include region dummies, one needs to be omitted.

g) The table below shows the results for regression separated by region. While the coefficients of *pct_dis* are all negative and in the same range, the coefficients on *classize* are not.

VARIABLES	(1) Reg1	(2) Reg2	(3) Reg3	(4) Reg4	(5) Reg5	(6) Reg6
classize	-0.0901**	-0.0550	0.168**	0.00741	0.0212	-0.0758
	(0.0451)	(0.0823)	(0.0669)	(0.0459)	(0.0429)	(0.0541)
pct_dis	-0.249***	-0.252***	-0.213***	-0.490***	-0.319***	-0.404***
	(0.0218)	(0.0309)	(0.0275)	(0.0389)	(0.0195)	(0.0232)
Constant	81.01***	77.17***	69.65***	80.18***	75.24***	79.62***
	(1.311)	(2.511)	(2.226)	(1.479)	(1.557)	(1.883)
Observations	255	195	267	276	574	400
R-squared	0.344	0.266	0.257	0.382	0.373	0.460

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

We propose a model that includes all the regions. To do so, we use the region dummies and omit region 5 (the region with most observations):

VARIABLES	mrkgrm
classize	0.00654
	(0.0218)
pct_dis	-0.309***
	(0.0103)
Reg1	3.501***
	(0.431)
Reg2	0.969**
	(0.467)
Reg3	0.399
	(0.410)
Reg4	3.258***
	(0.419)
Reg6	0.249
	(0.360)
Constant	75.55***
	(0.801)
Observations	1,967
R-squared	0.400

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

This model shows positive coefficients for all regions, suggesting a positive effect of class size on grammar scores. The effect for region 5 is shown in the coefficient of *classize*, for the other regions one needs to add the effect of the respective region dummy.

h) Subsample with only one class

i) The coefficient of *sc_boys* is -0.302 and statistically significant, meaning that with one additional boy in the school, the average grammar scores decrease by 0.302 points. The coefficient of *classize* is 0.0961, suggesting that the average grammar scores increase by this much with one additional student in class. However, this effect is not statistically significant anymore.

ii) The coefficient of *sc_boys* is -0.206. This means, controlling for the number of girls per school, an additional boy in the school decreases the average grammar scores by 0.206 points, which is slightly less than before.

iii) From the estimation in h-ii) one cannot say anything about the exact effect of one pupil in general, because the effects differ with respect to gender. However, we can expect the variance in the number of girls and boys to be roughly the same:

$$Cov(sc\_boys, sc\_girls) \approx Var(sc\_boys) = Var(sc\_girls)$$

The standard deviations in model h-ii) suggest this expectation to be true. Also, the correlation between *sc_boys* and *sc_girls* is with 0.7918 close to 1.

Now we can compare the two models:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \times sc\_girls + \hat{\beta}_2 \times sc\_boys + \epsilon$$

$$\hat{y} = \hat{\gamma}_0 + \hat{\gamma}_1 \times (sc\_girls + sc\_boys) + \mu$$

Solving the second model for $\hat{\gamma}_1$, we get:

$$\hat{\gamma}_1 = \frac{\mathrm{Cov}(\hat{y},\ \mathrm{sc\_girls})}{\mathrm{Var}(\mathrm{sc\_girls})} \frac{\mathrm{Var}(\mathrm{sc\_girls})}{\mathrm{Var}(\mathrm{sc\_girls} + \mathrm{sc\_boys})} + \frac{\mathrm{Cov}(\hat{y},\ \mathrm{sc\_boys})}{\mathrm{Var}(\mathrm{sc\_boys})} \frac{\mathrm{Var}(\mathrm{sc\_boys})}{\mathrm{Var}(\mathrm{sc\_girls} + \mathrm{sc\_boys})}$$

Combining this equation with auxiliary models only including *sc_boys* or *sc_girls* solved for their respective coefficients, we get the following formula:

$$\hat{\gamma}_1 = \frac{\mathrm{Var}(\mathrm{sc\_girls}) + \mathrm{Cov}(\mathrm{sc\_grils},\ \mathrm{sc\_boys})}{\mathrm{Var}(\mathrm{sc\_girls}\ + \mathrm{sc\_boys})} \hat{\beta}_1 + \frac{\mathrm{Var}(\mathrm{sc\_boys}) + \mathrm{Cov}(\mathrm{sc\_grils},\ \mathrm{sc\_boys})}{\mathrm{Var}(\mathrm{sc\_girls}\ + \mathrm{sc\_boys})} \hat{\beta}_2$$

Using the expectation with the similar variances and the correlations stated at the beginning of h-iii), we get the following approximation:

$$\hat{\gamma}_1 \approx \frac{1}{2}\left(\hat{\beta}_1 + \hat{\beta}_2\right) = \frac{1}{2}(0.096 - 0.206) = -0.055$$

Compared to the coefficient-value of the regression of *mrkgrm* on *classize* (-.0475), it seems that one can say something about the effect of increasing the class size by one pupil, even though the effect is different for boys and girls.

i) It is very unlikely that assumption 2 holds. There are definitely omitted variables that influence the grammar scores. Possible examples are cultural background, family income, school types (public or private) or the degree of preparation the schools provide for this standardized grammar tests.

3. Log-file

See attachment