Problem Set 2

MOEC0021 Empirical Methods

Fenqi Guo Wenjie Tu

Fall Semester 2020

R-Squared and Hypothesis Testing

1. Theory - Playing around with R-Squared

$$y = \beta X + \epsilon, \quad \epsilon \sim i.i.d(0, \sigma^2 I)$$

where X includes a constant term.

1(a)

$$\begin{aligned}
TSS &= \sum \tilde{y}_i^2 \\
&= \sum (y_i - \bar{y})^2 \\
&= \sum (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\
&= \sum \left((y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \right) \\
&= \underbrace{\sum (y_i - \hat{y}_i)^2}_{RSS} + \underbrace{\sum (\hat{y}_i - \bar{y})^2}_{ESS} + 2 \sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\
&= RSS + ESS + 2 \sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})
\end{aligned}$$

Recall that $\hat{\beta}_{OLS}$ is the estimator that minimizes the RSS (assuming K regressors),

$$\hat{\beta}_{OLS} \in \arg\min \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_K x_{iK})^2$$

FOC w.r.t. β_0,

$$2 \sum (y_i - \hat{y}_i)(-1) = 0 \iff \sum (y_i - \hat{y}_i) = 0$$

FOC w.r.t. β_j, $j \in [1, K]$,

$$2 \sum (y_i - \hat{y}_i)(-x_{ij}) = 0 \iff \sum (y_i - \hat{y}_i) x_{ij} \quad \forall j \in [1, K]$$

$$\begin{aligned}\sum(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}) &= \sum(y_i-\hat{y}_i)\hat{y}_i - \bar{y}\underbrace{\sum(y_i-\hat{y}_i)}_{0}\\ &= \sum(y_i-\hat{y}_i)(\beta_0+\beta_1 x_{i1}+\cdots+\beta_K x_{iK})\\ &= \beta_0\underbrace{\sum(y_i-\hat{y}_i)}_{0}+\beta_1\underbrace{\sum(y_i-\hat{y}_i)x_{i1}}_{0}+\cdots+\beta_K\underbrace{\sum(y_i-\hat{y}_i)x_{iK}}_{0}\\ &= 0\end{aligned}$$

Notice that it is crucial to have a constant term in order to obtain this result.

$$TSS = ESS + RSS \implies ESS = TSS - RSS$$

$$R^2 = \frac{ESS}{TSS} = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

1(b)

$$\begin{aligned}\mathrm{Cov}(y,\hat{y}) &= \mathrm{Cov}(\hat{y}+e,\hat{y})\\ &= \mathrm{Cov}(\hat{y},\hat{y}) + \underbrace{\mathrm{Cov}(e,\hat{y})}_{0}\\ &= \mathrm{Var}(\hat{y})\end{aligned}$$

$$\begin{aligned}\mathrm{Corr}^2(y,\hat{y}) &= \frac{\mathrm{Cov}^2(y,\hat{y})}{\mathrm{Var}(y)\mathrm{Var}(\hat{y})}\\ &= \frac{\mathrm{Var}^2(\hat{y})}{\mathrm{Var}(y)\mathrm{Var}(\hat{y})}\\ &= \frac{\mathrm{Var}(\hat{y})}{\mathrm{Var}(y)}\\ &= \frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2}\\ &= \frac{ESS}{TSS}\\ &= R^2\end{aligned}$$

Intuition: you can see the R^2 as a measure of how well your estimates correlate with real data.

1(c)

Regressors are doubled while coefficients are halved, implying that the residuals from two models are the same, and so are the RSS and consequently the R^2.

Denote the residuals from (M1) $y = X\beta + \epsilon$ by e and the residuals from (M2) $y = \tilde{X}\beta + \nu$.

$$e = y - X\hat{\beta}$$

$$
\begin{aligned}
u &= y - \tilde{X}\tilde{\beta} \\
&= y - 2X \cdot \frac{1}{2}\beta \\
&= y - X\beta
\end{aligned}
$$

$$
R^2_{M1} = 1 - \frac{e'e}{\tilde{y}'\tilde{y}} = 1 - \frac{u'u}{\tilde{y}'\tilde{y}} = R^2_{M2}
$$

1(d)

The OLS minimizes the sum of squared residuals. When you add an extra explanatory variable, the minimum cannot be higher than before. When you have already minimized a multi-dimentional function, e.g., with respect to two dimension, minimizing it also with respect to a third one can only deliver a lower minimum. Therefore, the RSS will always be lower with an extra regressor, and as result, the R^2 will be higher.

1(e)

Consider the regression $y = X\beta + \varepsilon$ and add an additional explantory variable, so that we get $y = X\beta + \gamma z + \nu$. The estimated equations are $y = X\hat{\beta} + e$ and $y = X\tilde{\beta} + z\tilde{\gamma} + u$ respectively, so that:

$$
\begin{aligned}
e &= y - X\beta \\
&= y - X(X'X)^{-1}X'y \\
&= (I - X(X'X)^{-1}X')y \\
&= M_X y
\end{aligned}
$$

$$
\begin{aligned}
u &= y - X\tilde{\beta} - z\tilde{\gamma} \\
&= y - X(X'X)^{-1}X'(y - z\tilde{\gamma}) - z\tilde{\gamma} \\
&= (I - X(X'X)^{-1}X')y - (I - X(X'X)^{-1}X')z\tilde{\gamma} \\
&= M_X y - M_X z\tilde{\gamma} \\
&= e - M_X z\tilde{\gamma}
\end{aligned}
$$

This implies that:

$$
\begin{aligned}
u'u &= (e - M_X z\tilde{\gamma})'(e - M_X z\tilde{\gamma}) \\
&= e'e - \underbrace{e'M_X z}_{\text{scalar}}\tilde{\gamma} - \tilde{\gamma}\underbrace{z'M'_X e}_{\text{scalar}} + \tilde{\gamma}^2 z'M'_X M_X z \\
&= e'e - 2\tilde{\gamma}z'M'_X e + \tilde{\gamma}^2 z'M_X z \\
&= e'e - 2\tilde{\gamma}z'M'_X y + \tilde{\gamma}^2 z'M_X z
\end{aligned}
$$

As we have seen in class $\tilde{\gamma} = (z'M_X z)^{-1}z'M_X y$. Plugging this into the previous expression,

$$
\begin{aligned}
u'u &= e'e - 2\tilde{\gamma}z'M'_X y + \tilde{\gamma}^2 z'M_X z \\
&= e'e - 2\frac{(z'M'_X y)^2}{z'M'_X z} + \frac{(z'M'_X y)^2}{z'M'_X z} \\
&= e'e - \frac{(z'M'_X y)^2}{z'M'_X z}
\end{aligned}
$$

$$
u'u < e'e \implies R^2_{ii} > R^2_i
$$

1(f)

The coefficient of determination, R^2, says that $R^2\%$ of the variation in the dependent variable can be "predicted" by variations in the independent variable. "Predicted" is not the same as "explained". Models can have excellent predictive power without necessarily having explanatory power (and sometimes one is all we need). The R^2 measure is not a measure of how "good" a model is because it cannot tell you whether:

- the included variables are statistically significant;
- the regressors are the true cause of movements in the dependent variable;
- there are omitted variables.

2. Empirical Question - Hypothesis Testing

```
library(stargazer) # print output
library(car) # hypothesis
```

```
d.class <- read.csv("class_size_pset.csv")
```

2(a)

```
# generate big_school dummy
d.class$big_school <- ifelse(d.class$n_classes > 2, 1, 0)
```

```
model1.1 <- lm(mrkgrm ~ classize, data = d.class)
model1.2 <- lm(mrkgrm ~ classize + factor(big_school), data = d.class)

stargazer(model1.1, model1.2, header = F,
          title = "Regression results in 2(a)",
          keep.stat = c("n", "rsq"), omit = "Constant",
          dep.var.caption = "grammar scores", single.row = T,
          covariate.labels = c("class size", "big school"))
```

Table 1: Regression results in 2(a)

	grammar scores	
	mrkgrm	
	(1)	(2)
class size	0.134*** (0.025)	0.102*** (0.027)
big school		1.246*** (0.337)
Observations	1,967	1,967
R^2	0.014	0.021
Note:		*p<0.1; **p<0.05; ***p<0.01

(i) An additional class member is associated with an increase in grammar test scores by 0.134 points.

(ii) An increase in class size by one member is associated with an increase in grammar test scores by 0.102 points, controlling for whether a school has more than 2 classes.

2(b)

```
# drop big_school dummy
d.class$big_school <- NULL

# generate natural log of grammar scores
d.class$ln_mrkgrm <- log(d.class$mrkgrm)

model2.1 <- lm(mrkgrm ~ classize + pct_dis, data = d.class)
model2.2 <- lm(ln_mrkgrm ~ classize + pct_dis, data = d.class)

stargazer(model2.1, model2.2, header = F, keep.stat = c("n", "rsq"), no.space = F,
          title = "Regression results in 2(b)", single.row = T, digits = 4)
```

Table 2: Regression results in 2(b)

	Dependent variable:	
	mrkgrm	ln_mrkgrm
	(1)	(2)
classize	-0.0603^{***} (0.0211)	-0.0007^{**} (0.0003)
pct_dis	-0.3349^{***} (0.0102)	-0.0048^{***} (0.0001)
Constant	79.0520^{***} (0.7043)	4.3677^{***} (0.0101)
Observations	1,967	1,967
R^2	0.3652	0.3686
Note:		*p<0.1; **p<0.05; ***p<0.01

(i) In the level-level specification (column 1), the coefficient on class size means that an additional member would have a negative impact on grammar test scores by 0.06 points on average. In the log-level specification (column 2), the coefficient on class size implies that an additional member in class is associated with a decrease in grammar test score by 0.1%.

To compare the two, the easiest way is to look at the effect around the mean value of grammar test score in log-level specification.

```
effect1 <- coef(model2.1)["classize"]
effect2 <- mean(d.class$mrkgrm) * coef(model2.2)["classize"]

cat(sprintf("Column 1: %.3f
            \nColumn 2: %.3f", effect1, effect2))
```

```
## Column 1: -0.060
##
## Column 2: -0.051
```

(ii) Since *pct_dis* is in percentage points, we are in a level-log-specification setting. The coefficient reads: 1 percentage point increase in the share of disadvantaged kids is associated with a drop in average test scores by 0.34 points.

2(c)

$$mrkgrm_i = \beta_1 + \beta_2 smallsize_i + \beta_3 pctdis_i + \varepsilon_i$$

```
# generate small_size dummy
d.class$small_size <- ifelse(d.class$classize <= 10, 1, 0)

model3.1 <- lm(mrkgrm ~ small_size + pct_dis, data = d.class)

stargazer(model3.1, header = F, keep.stat = c("n", "rsq"),
          title = "Regression results in 2(c) i", single.row = T)
```

Table 3: Regression results in 2(c) i

	Dependent variable:
	mrkgrm
small_size	2.560 (2.004)
pct_dis	-0.327^{***} (0.010)
Constant	77.099^{***} (0.183)
Observations	1,967
R^2	0.363
Note:	*p<0.1; **p<0.05; ***p<0.01

```
mean_score <- mean(d.class$mrkgrm, na.rm = T)
effect <- coef(model3.1)["small_size"]
relative_effect <- (effect / mean_score) * 100

cat(sprintf("Effect relative to the mean is %.3f percent", relative_effect))
```

(i)

```
## Effect relative to the mean is 3.521 percent
```

Holding the percentage of disadvantaged kids constant, test scores in classes with fewer than 10 students are on average 2.56 points higher than in classes with more than 10 students. This difference corresponds to 3.52% of the average score, which seems quite economically insignificant. In terms of statistical significance, we cannot reject the null hypothesis that $\beta_2 = 0$ against any alternative.

```
# for one-sided test with alternative beta2>0,
# we can compute p-value using function pt
p_value <- pt(coef(summary(model3.1))[2, 3], df.residual(model3.1), lower.tail = F)
cat(sprintf("p-value for one-sided test is %.3f", p_value))
```

```
## p-value for one-sided test is 0.101
```

```
# calculate t statistic by hand
t_stat <- coef(summary(model3.1))[2, 1] / coef(summary(model3.1))[2, 2]
cat(sprintf("t-stat for coefficeint small_class is %.3f", t_stat))
```

```
## t-stat for coefficeint small_class is 1.277
```

t-statistic is smaller than critical values, we cannot reject the null hypothesis.

(ii) $\hat{\beta}_2$ can be acquired through the following steps:

1. Regress *mrkgrm* on *pct_dis*, obtain residual $\hat{e}_1$
2. Regress *small_size* on *pct_dis*, obtain residual $\hat{e}_2$
3. Regress $\hat{e}_1$ on $\hat{e}_2$, obtain $\hat{\beta}_2$

```
## partitioned regression
# regress mrkgrm on pct_dis
reg1 <- lm(mrkgrm ~ pct_dis, data = d.class)

# regress small_size on pct_dis
reg2 <- lm(small_size ~ pct_dis, data = d.class)

# regress residuals from reg1 on residuals from reg2
reg3 <- lm(reg1$residuals ~ reg2$residuals)

# Comparing beta2 calculated by partitioned regression with beta2 calculated by model5
print(paste("$\beta_2$ is", round(reg3$coefficients['reg2$residuals'], digits = 4)))
```

[1] "β_2 is 2.5597"

```
stargazer(
  model3.1, reg3, header = F, keep.stat = c("n", "rsq"), omit = c("Constant", "pct_dis"),
  dep.var.labels.include = F, dep.var.caption = "grammar scores", single.row = T,
  column.labels = c("standard", "partitioned"), digits = 4,
  title = "Standard and partitioned regressions",
  covariate.labels = c("small class", "residuals")
)
```

```
# Calculate the mean value for each variable
mrkgrm.bar <- mean(d.class$mrkgrm)
small_size.bar <- mean(d.class$small_size)
pct_dis.bar <- mean(d.class$pct_dis)

# Get the coefficients from model5
beta2 <- coef(model3.1)['small_size']
beta3 <- coef(model3.1)['pct_dis']
```

Table 4: Standard and partitioned regressions

	grammar scores	
	standard	partitioned
	(1)	(2)
small class	2.5597 (2.0039)	
residuals		2.5597 (2.0034)
Observations	1,967	1,967
R^2	0.3631	0.0008
Note:		*p<0.1; **p<0.05; ***p<0.01

```
# intercept computed "by hand"
beta1 <- mrkgrm.bar - (beta2 * small_size.bar + beta3 * pct_dis.bar)

# intercept estimated by model3.1
intercept <- coef(model3.1)['(Intercept)']

cat(sprintf("$\\hat{\\beta}_1$ is %.4f
            \nIntercept is %.4f", beta1, intercept))
```

(iii) $\hat{\beta}_1$ is 77.0992

Intercept is 77.0992

```
model3.2 <- lm(ln_mrkgrm ~ small_size + pct_dis, data = d.class)

stargazer(model3.2, header = F, keep.stat = c("n", "rsq"), single.row = T,
          title = "Regression result from model 3.2")
```

Table 5: Regression result from model 3.2

	Dependent variable:
	ln_mrkgrm
small_size	0.037 (0.029)
pct_dis	-0.005^{***} (0.0001)
Constant	4.345^{***} (0.003)
Observations	1,967
R^2	0.367
Note:	*p<0.1; **p<0.05; ***p<0.01

```
exp(coef(model3.2)["small_size"])-1
```

iv

```
## small_size
##  0.0372102
```

On average, classes with less than 10 students have 3.72% higher scores than classes with more than 10 students.

2(d)

```
# generate many_dis dummy
d.class$many_dis <- ifelse(d.class$pct_dis > 10, 1, 0)

model4.1 <- lm(mrkgrm ~ classize*many_dis, data = d.class)

stargazer(model4.1, header = F, single.row = T,
          keep.stat = c("n", "rsq"), omit = "Constant",
          title = "Regression result in 2(d)")
```

Table 6: Regression result in 2(d)

	Dependent variable:
	mrkgrm
classize	-0.110^{***} (0.029)
many_dis	-15.713^{***} (1.346)
classize:many_dis	0.269^{***} (0.044)
Observations	1,967
R^2	0.301
Note:	*p<0.1; **p<0.05; ***p<0.01

```
# test joint hypotheses on beta3 = 0 and beta4 = 0
linearHypothesis(model4.1, c("many_dis = 0", "classize:many_dis = 0"), test = "F")
```

i

```
## Linear hypothesis test
##
## Hypothesis:
## many_dis = 0
## classize:many_dis = 0
##
## Model 1: restricted model
## Model 2: mrkgrm ~ classize * many_dis
##
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1   1965 97259
## 2   1963 69006  2     28253 401.85 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$F = \frac{\frac{R_U^2 - R_R^2}{q}}{\frac{1-R_U^2}{N-K}}$$

- q is the number of restricted coefficients.
- N is the number of observations.
- K is the number of unrestricted coefficients.
- $N - K$ is the degrees of freedom.

```
# compute F-statistic manually
restricted <- lm(mrkgrm ~ classize, data = d.class)
unrestricted <- lm(mrkgrm ~ classize*many_dis, data = d.class)
r2_restricted <- summary(restricted)$r.squared
r2_unrestricted <- summary(unrestricted)$r.squared
df_unrestricted <- summary(unrestricted)$df[2]
n_restrictions <- 2
f_stat <- ((r2_unrestricted-r2_restricted)/n_restrictions) /
  ((1-r2_unrestricted)/df_unrestricted)
print(paste("The F-stat is ", round(f_stat, digits = 4)))
```

```
## [1] "The F-stat is  401.8476"
```

Since *F-statistics* is way larger than 1, we can safely reject the null hypothesis (H_0: $\beta_3 = 0$ **AND** $\beta_4 = 0$). We therefore conclude that there exists a linear relationship between *mrkgrm* and *many_dis* **OR** there exists a linear relationship between *mrkgrm* and *many_dis*$\times$*classize*.

ii 10 additional students in a class with less than 10% disadvantaged pupils is associated with a decrease in $0.11 \times 10 = 1.1$ points in grammar score.

2(e)

```
model5.1 <- lm(mrkgrm ~ classize, data = subset(d.class, many_dis == 1))
model5.2 <- lm(mrkgrm ~ classize, data = subset(d.class, many_dis == 0))

stargazer(model4.1, model5.1, model5.2, header = F,
          keep.stat = c("n", "rsq"), omit = "Constant", single.row = T,
          title = "Comparison between 2(d) and 2(e)", no.space = F,
          column.labels = c("Interaction", "High", "Low"))
```

In the regression model from 2(d),

$$mrkgrm = \beta_0 + \beta_1 classize + \beta_2 manydis + \beta_3 classize \times manydis + \epsilon$$

- β_1 equals the coefficient on *classize* with subsample (*many_dis*=0).
- $\beta_1 + \beta_3$ equals the coefficient on *classize* with subsample (*many_dis*=1).

2(f)

Table 7: Comparison between 2(d) and 2(e)

	Dependent variable:		
		mrkgrm	
	Interaction	High	Low
	(1)	(2)	(3)
classize	−0.110*** (0.029)	0.159*** (0.037)	−0.110*** (0.026)
many_dis	−15.713*** (1.346)		
classize:many_dis	0.269*** (0.044)		
Observations	1,967	858	1,109
R^2	0.301	0.021	0.017
Note:			*p<0.1; **p<0.05; ***p<0.01

```
# Generate a dummy for each region
dummy <- as.factor(d.class$regioncode)
dummies <- model.matrix(~dummy)
d.class <- cbind(d.class, dummies)
```

We cannot include all of them. Since there are six categories, we can include only five of them to avoid multicolinearity.

2(g)

```
# regress grammar scores on classize and pct_dis for region1
model7.1 <- lm(mrkgrm ~ classize + pct_dis,
data = subset(d.class, regioncode == 'Reg1'))

# regress grammar scores on classize and pct_dis for region2
model7.2 <- lm(mrkgrm ~ classize + pct_dis,
data = subset(d.class, regioncode == 'Reg2'))

# regress grammar scores on classize and pct_dis for region3
model7.3 <- lm(mrkgrm ~ classize + pct_dis,
data = subset(d.class, regioncode == 'Reg3'))

# regress grammar scores on classize and pct_dis for region4
model7.4 <- lm(mrkgrm ~ classize + pct_dis,
data = subset(d.class, regioncode == 'Reg4'))

# regress grammar scores on classize and pct_dis for region5
model7.5 <- lm(mrkgrm ~ classize + pct_dis,
data = subset(d.class, regioncode == 'Reg5'))

# regress grammar scores on classize and pct_dis for region6
model7.6 <- lm(mrkgrm ~ classize + pct_dis,
```

```
data = subset(d.class, regioncode == 'Reg6'))

stargazer(
  model7.1, model7.2, model7.3, model7.4, model7.5, model7.6, header = F,
  column.labels = c('Region1', 'Region2', 'Region3', 'Region4', 'Region5', 'Region6'),
  font.size = 'small', keep.stat = c('n', 'rsq'), no.space = T, omit = "Constant"
)
```

Table 8:

	Dependent variable:					
	mrkgrm					
	Region1	Region2	Region3	Region4	Region5	Region6
	(1)	(2)	(3)	(4)	(5)	(6)
classize	-0.090^{**}	-0.055	0.168^{**}	0.007	0.021	-0.076
	(0.045)	(0.082)	(0.067)	(0.046)	(0.043)	(0.054)
pct_dis	-0.249^{***}	-0.252^{***}	-0.213^{***}	-0.490^{***}	-0.319^{***}	-0.404^{***}
	(0.022)	(0.031)	(0.027)	(0.039)	(0.020)	(0.023)
Observations	255	195	267	276	574	400
R^2	0.344	0.266	0.257	0.382	0.373	0.460
Note:						*p<0.1; **p<0.05; ***p<0.01

While in region 1, it seems that adding a student to the class may have a negative and statistically significant effect on scores, in region 3, the opposite appears to be true. In other regions, coefficients are not statistically different from zero. In any case, since these are separate regressions, we cannot conclude anything about whether the effects of class size differ across regions. To do so, we should include *classize*×*region* interaction term to the model

```
model7.7 <- lm(mrkgrm ~ pct_dis + classize*factor(regioncode), data = d.class)

stargazer(model7.7, header = F, keep.stat = c("n", "rsq"),
          title = "Regression result in 2(f)", single.row = T)
```

```
linearHypothesis(
  model7.7, hypothesis.matrix =
    c("classize:factor(regioncode)Reg2 = 0", "classize:factor(regioncode)Reg3 = 0",
      "classize:factor(regioncode)Reg4 = 0", "classize:factor(regioncode)Reg5 = 0",
      "classize:factor(regioncode)Reg6 = 0"), test = "F"
)
```

```
## Linear hypothesis test
##
## Hypothesis:
## classize:factor(regioncode)Reg2 = 0
## classize:factor(regioncode)Reg3 = 0
## classize:factor(regioncode)Reg4 = 0
## classize:factor(regioncode)Reg5 = 0
## classize:factor(regioncode)Reg6 = 0
##
## Model 1: restricted model
```

Table 9: Regression result in 2(f)

	Dependent variable:
	mrkgrm
pct_dis	-0.307^{***} (0.010)
classize	-0.097^{*} (0.053)
factor(regioncode)Reg2	-2.569 (2.378)
factor(regioncode)Reg3	-8.387^{***} (2.403)
factor(regioncode)Reg4	-4.548^{**} (2.062)
factor(regioncode)Reg5	-7.224^{***} (1.990)
factor(regioncode)Reg6	-6.283^{***} (2.306)
classize:factor(regioncode)Reg2	0.0001 (0.085)
classize:factor(regioncode)Reg3	0.184^{**} (0.081)
classize:factor(regioncode)Reg4	0.154^{**} (0.071)
classize:factor(regioncode)Reg5	0.131^{**} (0.066)
classize:factor(regioncode)Reg6	0.109 (0.076)
Constant	81.885^{***} (1.504)
Observations	1,967
R^2	0.403
Note:	*p<0.1; **p<0.05; ***p<0.01

```
## Model 2: mrkgrm ~ pct_dis + classize * factor(regioncode)
##
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1   1959 59177
## 2   1954 58885  5    291.22 1.9327 0.08588 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-statistic is 1.932. We cannot reject the null hypothesis with much confidence.

2(h)

```
model8.1 <- lm(mrkgrm ~ classize + sc_boys, data = subset(d.class, n_classes == 1))
model8.2 <- lm(mrkgrm ~ classize, data = subset(d.class, n_classes == 1))
model8.3 <- lm(mrkgrm ~ sc_girls + sc_boys, data = subset(d.class, n_classes == 1))

stargazer(model8.1, model8.2, model8.3, header = F, keep.stat = c("n", "rsq"),
          omit = "Constant", single.row = T)
```

i Here we only consider the subsample of *n_classes = 1*, we thus have this equation *classize = sc_boys + sc_girls*. *classize* is dependent on *sc_boys*. Class size will increase simultaneously as the number of boys increases. One more boy in school will decrease the average mark in grammar test by 206 point. However, an marginal increase in class size might result from an marginal increase in the number of boys or girls. Here we assume the marginal effect results from an additional girl. One more girl in the class will decrease the average mark in grammar test by 0.096 points.

Table 10:

	Dependent variable:		
	mrkgrm		
	(1)	(2)	(3)
classize	0.096 (0.110)	−0.048 (0.083)	
sc_girls			0.096 (0.110)
sc_boys	−0.302** (0.153)		−0.206* (0.115)
Observations	240	240	240
R^2	0.018	0.001	0.018
Note:			*p<0.1; **p<0.05; ***p<0.01

ii Since *sc_girls* and *sc_boys* are independent, we can say that one more boy in school will decrease the average mark in grammar test by 0.206 point.

```
lh <- linearHypothesis(model8.3, c("sc_boys=sc_girls"))
t.stat <- round(sqrt(lh$F[2]), digits = 4)
print(paste("The t-stat is ", t.stat))
```

iii

```
## [1] "The t-stat is  1.9771"
```

Since the coefficient on *sc_boys* in model8.3 can be read as the effect of having an additional boy while holding the number of girls constant, this is equivalent to the effect of increasing the overall class size by one person, or by one boy, to be more precise. The coefficient on *sc_girl* instead would tell us the effect of increasing the overall class size by one girl. To test whether one effect is larger than the other, we can perform a simple t-test under the null hypothesis $\beta_{scboys} - \beta_{scgirls} = 0$. The *t-statistic* would be:

$$t = \frac{\hat{\beta}_{scboys} - \hat{\beta}_{scgirls}}{\sqrt{Var(\hat{\beta}_{scboys} - \hat{\beta}_{scgirls})}}$$

Since *t-statistic* is 1.977, we can reject the null hypothesis at 95% confidence level.

2(i)

It is easy to think about that class size may be endogenous in this context. One can think of omitted variable problems and assume that class sizes vary depending on whether schools are located in cities or rural areas, which in turn may affect directly school performance for several reasons.