Final Assignment

Elisa Quadrini

1. Introduction

This project is based on a dataset, named *"Women In STEM Education Fields"*, which I created in order to perform the analysis of the growing trend of women enrolling in STEM education programs.

The dataset has been created by selecting variables from the Eurostat website, each of which has been customized using a specific function in the Eurostat website in order to visualize data regarding female individuals, referred to the years 2019, 2020, 2021. The decision to include three different years was driven by the limited sample size; otherwise, the dataset would have consisted of only 38 observations.

The provenance of this data is explained by the UNESCO OECD Eurostat (UOE) joint data collection methodology: Eurostat collects and disseminates data from the EU Member States, candidate countries and EFTA countries.

The data have been collecting with a time frequency of one year.

1.1 Description of the dataset

The dataset consists of 114 observations, with 38 for each year (2019, 2020, and 2021). Each observation is labeled according to the country where the data was collected.

The response variable that will be used through all the analysis is **tertEdustem**, that stands for "Tertiary Education in STEM fields". It is a quantitative variable that represents the number of female graduates in one of the STEM discipline as a share of all STEM graduates in the tertiary education level (ISCED 5-8). STEM fields are classified as ISCED-F 05 (natural sciences, mathematics and statistics), 06 (information and communication technologies) or 07 (engineering, manufacturing and construction).

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	661	1855	4966	3428	65726

This is the summary of my response variable; as we can see, it is a numerical variable that takes values from 0 to 65726 (number of women).
To address the potential problem of negative fitted values, which are not meaningful in this context since the response variable cannot take negative values, I decided to apply a logarithmic transformation. Since the minimum value of the variable is 0, I added 1 before applying the transformation to avoid undefined values.

Now I proceed to describe the predictors included in my analysis:

- **Area**: categorical; represents the European area in which is situated the country where the observation i has been collected and consists of three levels (est is used as the baseline):

est	nord	ovest	sud
39	24	30	21

- **Early Childhood Education**: the indicator measures the percentage of the female children aged three to the starting age of compulsory primary education who participated in early childhood education and care (level 0). Note that the starting age of compulsory primary education varies across European countries (ranging from 5 to 7 years according to Eurostat).
- **At Most Lower Sec Education (25/34)**: the indicator is defined as the percentage of female people of age 25 to 34 who has successfully completed less than primary, primary and lower secondary education (levels 0-2).
- **At Most Lower Sec Education (35/44)**: the indicator is define as the percentage of female people of age 35 to 44 who has successfully completed less than primary, primary and lower secondary education (levels 0-2).
- **At Least Upper Sec Education (20/24)**: the indicator is defined as the percentage of female people aged 20-24 who have successfully completed at least upper secondary education (upper secondary, post-secondary non-tertiary and tertiary education (levels 3-8)).
- **At Least Upper Sec Education (25/64)**: the indicator is defined as the percentage of female people aged 25-64 who have successfully completed at least upper secondary education (upper secondary, post-secondary non-tertiary and tertiary education (levels 3-8)).
- **Tertiary Educational Attainment**: the indicator measures the percentage of the female population aged 25-34 who have successfully completed tertiary studies (e.g. university, higher technical institution, etc.).
- **Employment Rates of Recent Grad**: the indicator presents in percentage the employment rates of female people aged 20 to 34 fulfilling the following conditions: currently employed, having attained at least upper secondary education, not having received any education or training in the four weeks preceding the survey and having successfully completed their highest educational attainment 1, 2 or 3 years before the survey.
- **Students From Abroad Enrolled (STEM)**: this variable represents the number of mobile female students from abroad enrolled in the tertiary education level. Its range goes from 299 to 265354 people for country.
- **Female Teachers in Tertiary Edu**: this indicator represents the percentage of female teachers out of the total number of teachers in tertiary education.
- **Public Educational Expenditure (Mln)**: this is the annual amount of public educational expenditure expressed in million of euros, from the pre-primary to the tertiary education. Its range goes from 645.6 to 181083.5 mln € for country.
- **Inability To Face Unexpected Financial Exp**: this is the percentage of the totality of households that report the inability to face unexpected financial expenditures expressed in percentage.
- **Average Weekly Hours of Work**: this is the average number of usual weekly hours of work of female employees aged 15 to 64 years old. Its range goes from 25.70 to 43.10 weekly hours of work for country.

All the predictors expressed in percentage have a value range that goes from 0 to 100.

1.2 Missing values

In my original dataset there were several missing values, which I chose to handle by imputing the mean of the respective variable. This decision was supported by several factors.

First, the structure of my dataset, which consists of three observations for each European country, makes every row essential to maintaining the balance between countries: removing rows with missing values could disrupt this balance.

Second, by imputing the mean, the distribution of each variable remained unchanged when comparing the dataset with missing values removed to the dataset where missing values were replaced with the mean.

1.3 References

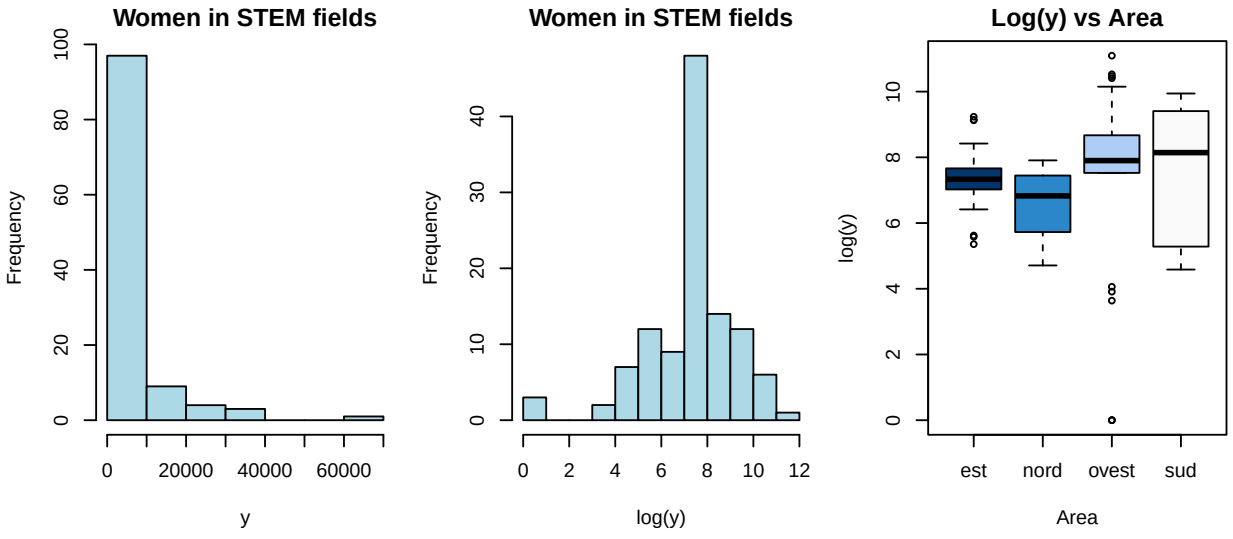
Creation of the dataset:
https://ec.europa.eu/eurostat/web/education-and-training/database.

2. Applied goals

The goal of the project is to evaluate the efficacy of a linear regression model, based on the available variable, in predicting the number of women graduated in STEM education fields in a European country. Although I expect to have some issues due to the potential data correlation (collected by countries on three different years), I think it's crucial to understand the link between the dependent variable and the independent ones in order to develop a useful tool such as a linear regression model to predict which countries can drive the increasing phenomenon described in chapter 1.

3. Graphical representation of the variables

The following plots regard the representation of my response variable, before and after the logarithmic transformation, in order to visualize how it improves in terms of distribution:

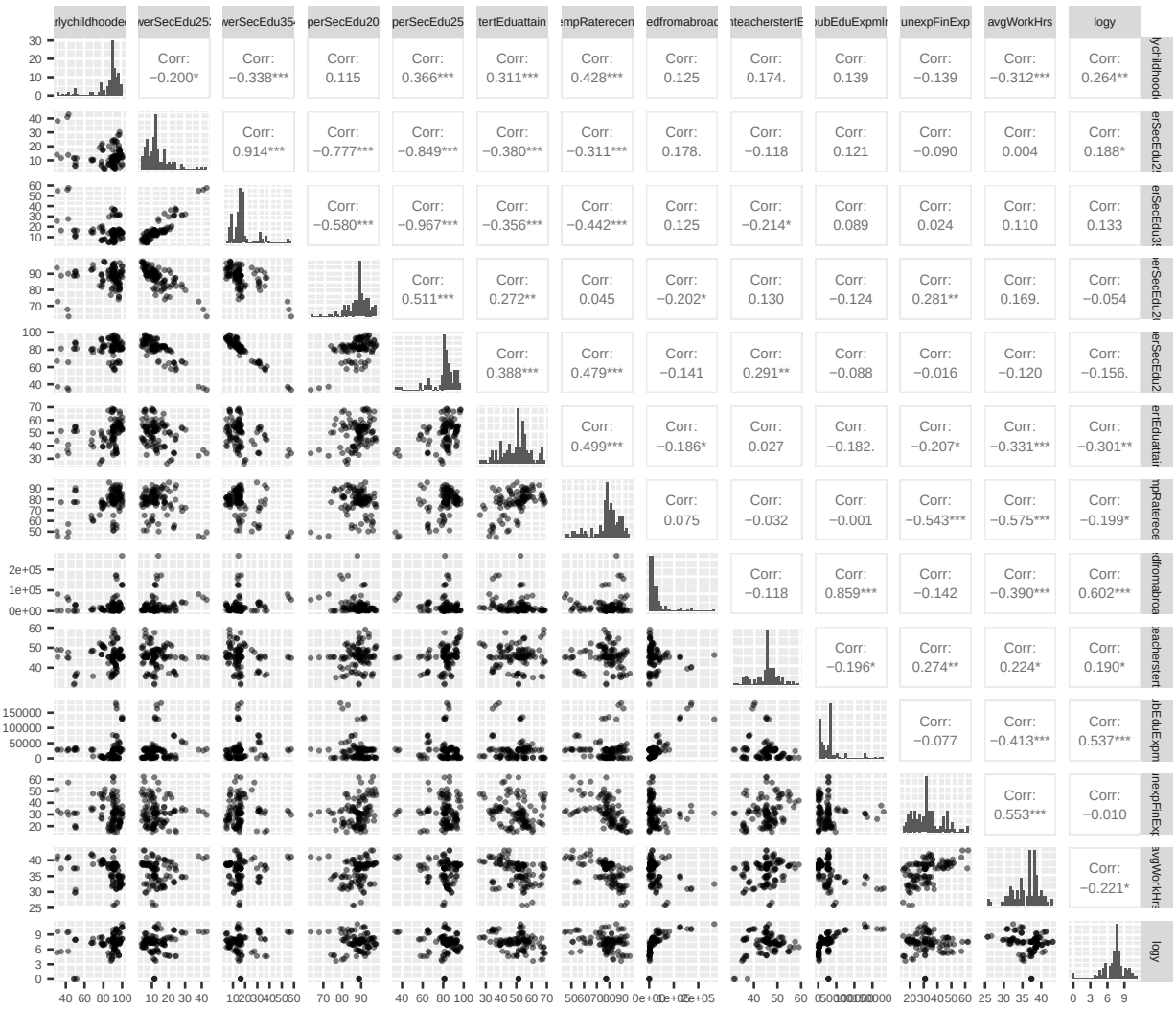


On the left side there's an histogram which represents the frequency of my continue response: as we can see, the great majority of the observed y_i are between the values 0 and 10000 in a left-skewed distribution, that doesn't seem to follow a gaussian pattern. Then there's the plot of the logarithm of the response variable. Observing the plot, it becomes evident that the distribution is right-skewed, with most values concentrated around log(y) = 7 and log(y) = 8 but it seems more similar to a gaussian distribution.

On the right side, the boxplot examines the relationship between log(y) and the categorical variable area. The visualization highlights regional differences, with higher variability in the "sud" region and several extreme values, particularly in the "est" and "ovest" areas.

Overall, the analysis suggests that geographical location may influence the response variable, indicating the presence of spatial patterns in the distribution of women in STEM fields. However, further statistical analysis is required to confirm this relationship.

Now I will show the so called *Correlation scatterplot matrix*, which shows in the upper panels the correlation coefficients for each pair of variables (with the associated significance levels), on the diagonal the histogram of the distribution of each variable and in the lower panels the scatterplots showing relationships between variable pairs:



From the scatterplots, and as confirmed by the correlation coefficients, strong positive correlations are

observed, particularly between the variables *upperSecEdu2024* and *upperSecEdu2564*, as well as *lowerSecEdu2534* and *lowerSecEdu3544*. This was expected, as these variables essentially contain the same information but for different age groups.

Similarly, the strong negative correlations between *upperSecEdu2024* or *upperSecEdu2564* and *lowerSecEdu2534* or *lowerSecEdu3544* are also not surprising, as they represent opposing levels of educational attainment within the population.

Additionally, there is a high positive correlation between *pubEduExpmln* and *enrolledfromabroadstem*, suggesting a relationship between public educational expenditure and the number of mobile female students from abroad enrolled in STEM's tertiary education level.

Furthermore, the scatterplots indicate that some variables appear to have a relationship with the response variable *logy*. Specifically, *enrolledfromabroadstem* and *pubEduExpmln* exhibit a logarithmic pattern, which may require further investigation during the diagnostic analysis of which will be the selected model.

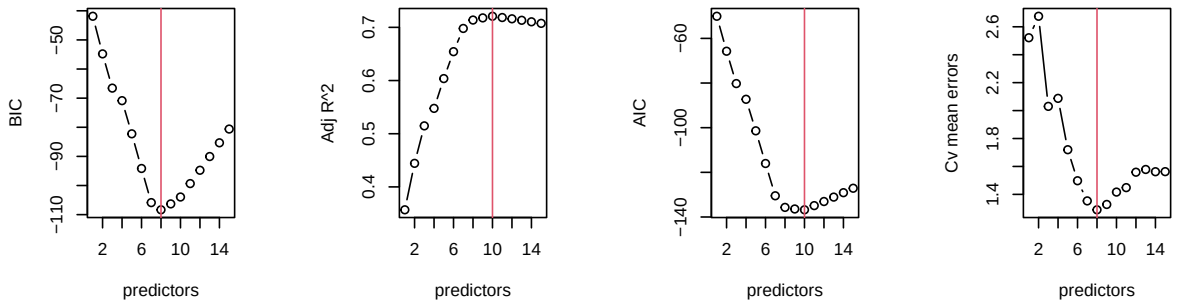
Lastly, the presence of highly correlated variables suggests potential collinearity issues, which will be analyzed further in the following sections.

4. Variable selection

First of all we proceed by computing a variable selection based on the Best Subset Selection method, which enables us to assess which variables should be included in the final model without incoming in the issue of not considering one variable due to its higher t-test p-value.
In my analysis, the parameters have been estimated by the OLS method and I decided to center all the predictors of the model by subtracting the mean from the variable values for a more precise interpretation of the coefficients.

```
women[,2:13] = scale(women[,2:13], center = TRUE, scale = FALSE) #centering the variables

ols = regsubsets(logy ~. , nvmax = 15, data = women) #best subset selection
sum = summary(ols)

aic_values = numeric(nrow(sum$which)) #AIC
for (i in 1:nrow(sum$which)) {
  aic_values[i] = sum$bic[i] - (i+2)*log(n) + 2*(i+2)
}

p = 15  #cv errors
k = nrow(women)
set.seed(1)
folds = sample(1:k, nrow(women), replace = FALSE)
cv.errors = matrix(NA, k, p, dimnames = list(NULL, paste(1:p)))
for(j in 1:k) {
  best.fit = regsubsets(logy ~. , nvmax = 15, data = women[folds != j,])
  max_models = nrow(summary(best.fit)$which)

  for (i in 1:15) {
    mat = model.matrix(as.formula(best.fit$call[[2]]), women[folds==j,])
    coefi = coef(best.fit, id = i)
    xvars = names(coefi)
    pred = mat[, xvars]%*%coefi
    cv.errors[j,i] = mean((women$logy[folds == j] - pred)^2)
  }
```

```
}
cv.mean = numeric(max_models)
for (i in 1:max_models){
  cv.mean[i] = mean(cv.errors[,i])
}
```



It is crucial to keep in mind the principle that a less complex model is generally preferable when the difference in the values of the criteria used for model comparison is very small. In my case, R^2 adjusted and AIC select the model with 10 predictors, while the BIC and the CV error select the one with 8 predictors.
However, even for the R^2 adjusted and AIC criteria, the improvement of the 10-predictor model compared to the 8-predictor model is marginal, therefore, I prefer to consider the model with only 8 predictors. Furthermore, my goal is to compute a prediction based on my fitted model, so it is more reasonable to select the best model according to the CV error.

It seems that we have selected the best model, but we have to take into account a very potential misleading issue: the collinearity issue. So, before fitting the model and printing the variable selected by the procedure, let's further investigate this issue.

4.1 Collinearity issue

From the exploratory analysis, we observed that some variables are highly correlated, both positively and negatively. This could lead to a problem in the variable selection procedure: even though it doesn't consider the single t-test as a selection criteria, it could fail to not consider a significant variable due to its high correlation with others.

Such collinearity may interfere because it increases the variance of the estimated parameters, which in turn can lead to wider confidence intervals. To assess this issue, let's compute the Variance Inflation Factor (VIF).

- lowerSecEdu2534 : 20.09
- lowerSecEdu3544 : 56.57
- upperSecEdu2564 : 35.88

As we can see, the variables *upperSecEdu2564*, *lowerSecEdu2534* and *lowerSecEdu3544* that were strongly correlated have a collinearity issue. So let's combine them into a new variable, named *bestEdu*, that represents the educational gap between women with higher education (upper secondary or above) and those with lower education (only lower secondary).

```
women_new = subset(women, select = -c(lowerSecEdu2534, lowerSecEdu3544, upperSecEdu2024,
                                       upperSecEdu2564))
```

```
women_new$uppSecEdu = rowMeans(cbind(women$upperSecEdu2024, women$upperSecEdu2564))
women_new$lwrSecEdu = rowMeans(cbind(women$lowerSecEdu2534, women$lowerSecEdu3544))

women_new$bestEdu = women_new$uppSecEdu - women_new$lwrSecEdu
```

	GVIF	Df	GVIF^(1/(2*Df))
area	13.01	3	1.534
earlychildhoodedu	1.659	1	1.288
tertEduattain	2.348	1	1.532
empRaterecent	2.934	1	1.713
enrolledfromabroadstem	4.876	1	2.208
femteacherstertEdu	2.063	1	1.436
pubEduExpmln	5.066	1	2.251
unexpFinExp	1.974	1	1.405
avgWorkHrs	3.15	1	1.775
bestEdu	2.957	1	1.72

There's no more collinearity issue now, as for the categorical variable we consider the vale of the GVIf. As we changed some variables of the model, we have to implement a new variable selection: it selects again the model with 8 predictors (the same ones as before), which shows that only two levels of the categorical variable *area* are significant: "nord" and "ovest". So I proceed by create 2 dummy variables:

```
women_new$area_ovest = ifelse(women_new$area == "ovest", 1, 0)
women_new$area_ovest = as.factor(women_new$area_ovest)

women_new$area_nord = ifelse(women_new$area == "nord", 1, 0)
women_new$area_nord = as.factor(women_new$area_nord)
```

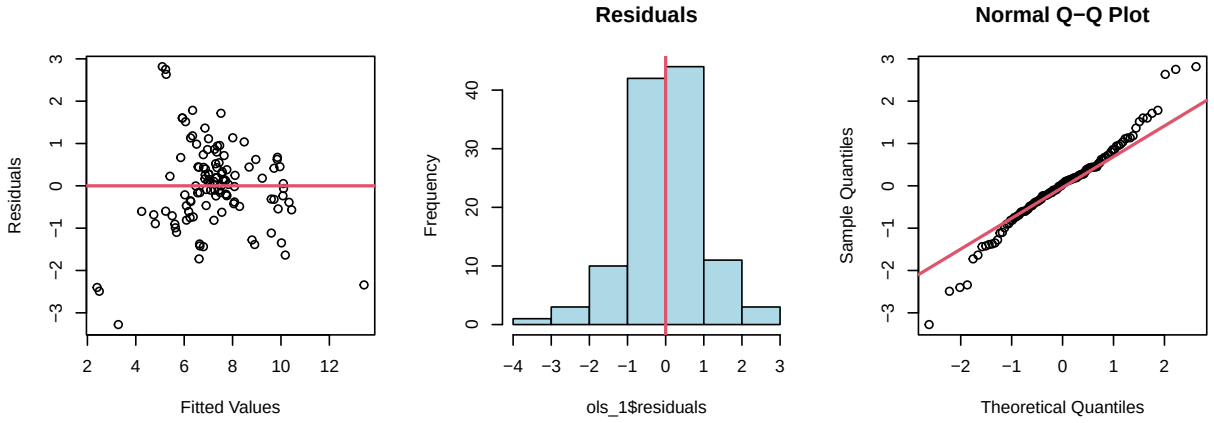
These two new variables will be included in the model, defined as:

$$\begin{cases} y_i = \beta_0 + \sum_{j=1}^{6} \beta_j x_{i,j} & x_7 = 0, x_8 = 0 \\ y_i = \beta_0 + \beta_7 + \sum_{j=1}^{6} \beta_j x_{i,j} & x_7 = 1, x_8 = 0 \\ y_i = \beta_0 + \beta_8 + \sum_{j=1}^{6} \beta_j x_{i,j} & x_7 = 0, x_8 = 1 \end{cases}$$

where $x_7 = 1$ if area $=$ ovest, $x_8 = 1$ if area $=$ nord. We now finally can fit the model selected by the best subset selection procedure, using the CV errors criteria, and we proceed by testing the model hypothesis through the diagnostics analysis:

```
ols_1 = lm(logy ~ earlychildhoodedu + empRaterecent + enrolledfromabroadstem +
           femteacherstertEdu + unexpFinExp + avgWorkHrs + area_ovest + area_nord, data = women_new)
```

5. Diagnostics



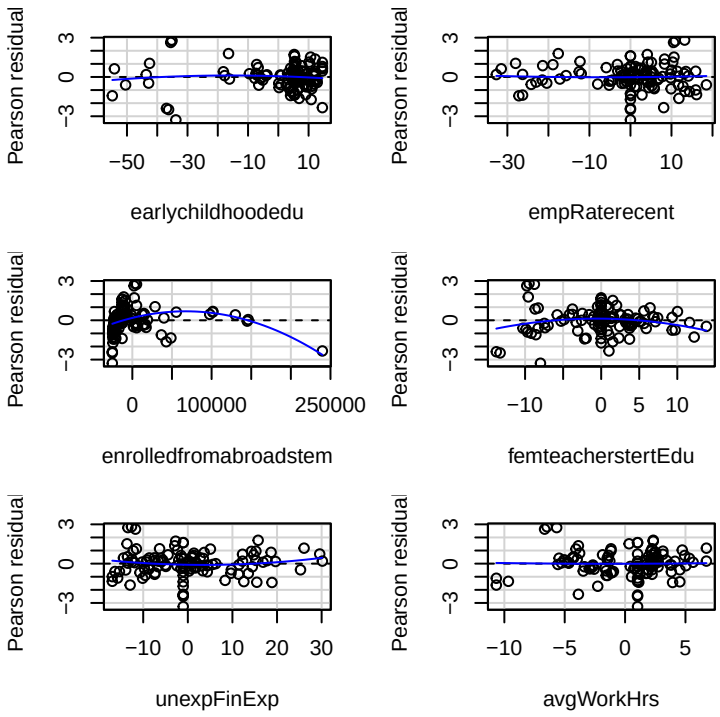

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(ols_1)
## W = 0.97395, p-value = 0.0252
```

From these plots, we check:

- **normality assumption**: the histogram shows that the residuals are approximately symmetric, with a peak around zero, suggesting a roughly normal distribution. However, there are some values in the tails that might lead to a deviation from the normality distribution.
  This is enhanced by the normal Q-Q plot: most points lie close to the reference line, supporting the normality assumption. However, some deviations in the tails suggest the presence of heavy tails, which indicates mild violations of normality. The Shapiro-Wilk test confirms this considerations: the p-value is lower than 0.05, leading to the rejection of the null hypothesis of normal distribution of the residuals.
- **homoschedasticity** of the errors: we see from the graphical representation of the residuals with respect to the fitted values that the residuals don't seem to have a constant variance: the plot has the shape similar to a double outward bow. This is not surprising, as the dataset used to fit the model consists of **panel data**: this type of data that depends on both spatial location and temporal period in which the observations were collected. This structure may lead to a potential violation of the assumption of error independence, resulting in heteroscedasticity and possibly even autocorrelation of the errors.

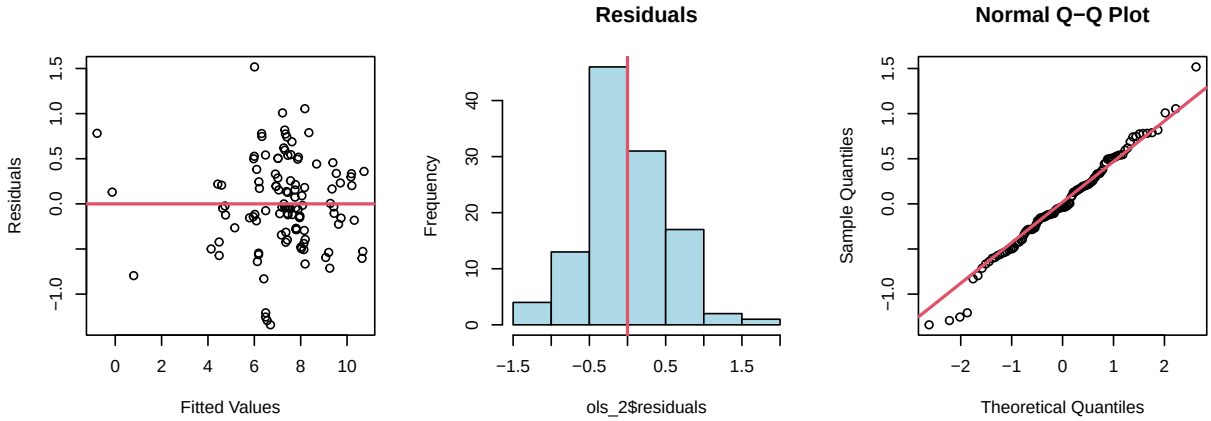
From the residuals plots in the following page we check:

- **linearity assumption**: if the residuals exhibit a non-random pattern, it suggests a potential violation of the linearity assumption. The response variable has been transformed by applying a logarithmic transformation in the exploratory analysis, but from the residuals plots we see that even the variables *enrolledfromabroadstem* and *femteacherstertEdu* have a non linear pattern, so we apply the logarithmic transformation of these variables and then we compute again the variable selection to detect if this model is again the best model. Note that in order to compute the log transformation, I added to these variables their minimum value + 1, otherwise they would assume values <=0 :



Now the best model with respect to the CV errors is the following, with 7 predictors:

```
ols_2 = lm(logy ~  empRaterecent + enrolledfromabroadstem_l + femteacherstertEdu_l
           + pubEduExpmln + avgWorkHrs + area_ovest + area_nord, data = women_new)
```




```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(ols_2)
## W = 0.98763, p-value = 0.3848
```

We can see that the model has improved as there is no longer a violation of the normality assumption, although the residuals' distribution seems to be left-skewed. However, heteroscedasticity remains in the

residuals and it cannot be resolved through any transformation of either the covariates or the response variable. I have tested multiple transformations, but none were effective as the residuals plots remained the same.

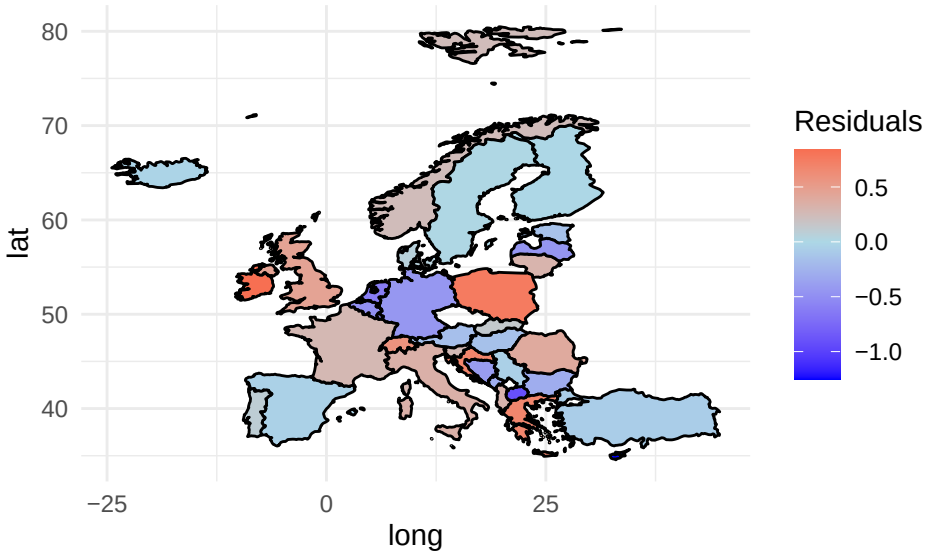
This fact is probably due to the structure of my dataset: my observations are likely not independent not only because they were collected at the country level, but also because for each country there are three different observations. As a result, I expected some correlation between the residuals, which could explain the presence of heteroscedasticity.

To confirm this hypothesis, let's check whether we can detect any spatial pattern in the residuals.

5.1 Spatial pattern

The following map visualizes the spatial distribution of residuals across different European countries. The color scale represents the magnitude and direction of the residuals:

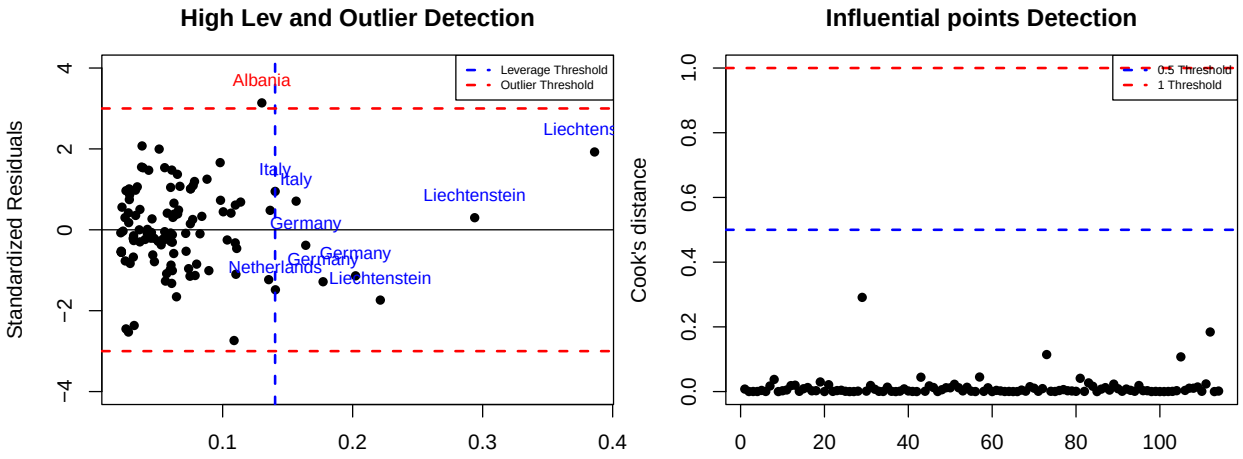
- red shades indicate positive residuals, meaning that the model underestimated the response variable in these countries.
- blue shades indicate negative residuals, meaning that the model overestimated the response variable.
- intermediate shades suggest that the model's fitted values were closer to the observed values in those regions.



From this visualization, we can observe that some regional trends in the residuals indicate spatial correlation. In particular, the fitted values for some countries that fall under the two dummy variables in my model, area_nord and area_ovest, appear to be close to the observed values of the response variable. However other countries, such as the UK and Ireland, deviate more from the observed values, despite also being located in the northwest region.

Additionally, Germany and the nearby countries display a recognizable pattern, and also do several eastern European countries. All these considerations suggest that spatial correlation exists in the residuals, and it has not been fully accounted for in the current model.

5.2 Outliers detection



The left plot illustrates the detection of leverage and outlier for each observation. The blue dashed line denotes the threshold beyond which points are considered high leverage points: $2(p+1)/n$ where "n" represents the number of observations and "p" denotes the number of predictors in my model.

The plot indicates the presence of several observations with high leverage. However, this is not a concern as these are not influential points because there are no points above the 0.5 threshold in the influential points detection plot (right side). It is not surprising that the observations with the highest leverage values are the Liechtenstein ones, because this country has some values for the predictors and for the response variable that are really far away from the center of the data.

For what concerns the outliers, the plot reveals only one observation (from the country Albania) that is above the red threshold $|r_i| > 3$ but even in this case it is not a problem because it is not an influential point. For this reason, since the presence of high leverage points and outliers does not significantly affect the model fit, we can continue to use all observations in the dataset for predicting the response without worrying about misleading results. This also can be a good news for our fitting, as the model seems to well cover widely dispersed data.

6. Intepretation of the coefficients and their uncertainties

The best overall model is:

$Y = \beta_0 + \beta_1 X_{empRaterecent} + \beta_2 X_{enrolledfromabroadstem_l} + \beta_3 X_{femteacherstertEdu_l} + \beta_4 X_{pubEduExpmln} + \beta_5 X_{avgWorkHrs} + \beta_6 X_{areaovest} + \beta_7 X_{areanord}$

```
##                              Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)             -1.232417e+00 0.3626350836 -3.398504 9.555797e-04
## empRaterecent           -2.580729e-02 0.0059725633 -4.320974 3.513947e-05
## enrolledfromabroadstem_l  7.712237e-01 0.0395298995 19.509884 7.242761e-37
## femteacherstertEdu_l      6.920333e-01 0.1232649210  5.614195 1.597692e-07
## pubEduExpmln              1.096572e-05 0.0000018551  5.911123 4.186267e-08
## avgWorkHrs               -8.467306e-02 0.0242491534 -3.491794 7.008060e-04
## area_ovest1              -7.883697e-01 0.1766223169 -4.463591 2.017817e-05
## area_nord1               -5.363747e-01 0.1666644226 -3.218291 1.711327e-03
```

The coefficients' interpretation depend on which transformation I have computed on my covariates and my response variable to improve the linearity of the model, in particular all my independent variables have been centred in their mean, the response is logarithmic and so are other two variables, enrolledfromabroadstem_l and femteacherstertEdu_l. Note that I centered all my independent variables before applying to some of them the logarithmic transformation. As a result, both the transformed enrolledfromabroadstem and femteacherstertEdu could take negative values. To address this issue, I added the absolute value of their minimum value plus 1, since the logarithm is not defined at 0: this adjustment does not affect the model's validity, as linear models are invariant under scale transformations.

Let's now analyse the effect of each beta on the response:

- β_0: the intercept is estimated as -1.2324. Since the response variable is in the logarithm scale, when all the predictors are at their mean value (except the ones that have been transformed with the log, that assume the value of $\bar{x_i}$ - min(x_i)) and the country's area is not the west or the north one (baseline of the dummy variables), the estimated number of women graduated in one of the STEM fields is $e^{-1.2324}$ - 1 = 0.292 - 1 = -0.708, since the transformation I computed was log(y+1). This value does not seem to make sense in relation to the unit of measure of my response, suggesting that these predictor values may not correspond to a realistic scenario.
- β_1: a one unit increase in the employment rate of recent graduates, that is a 1% increase in the rate, is associated with a decrease of 0.0258 in logy, holding all other variables constant.
- β_2: a one unit increase in the logarithm of the number of female students from abroad enrolled in STEM tertiary education fields ($log(x_2) + 1$), is associated with a 0.7712 increase in logy, holding all other variables constant.
- β_3: a one unit increase in the log-transformed percentage of female teachers in tertiary education, that is $log(x_3) + 1$, corresponds to a 0.692 increase in logy, holding all other variables constant.
- β_4: a one unit increase in public educational expenditure, so one million euros increment, leads to a very small but positive increase of 1.096572×10^{-05} in logy, holding all other variables constant.
- β_5: a one hour increase in the average weekly working hours of female employees leads to a 0.0847 decrease in logy, holding all other variables constant.
- β_6: being in a Western European region is associated with a decrease of 7.88 in logy compared to the reference category, that could be the East or the South one since the baseline value for x_6 and x_7 is 0, holding all other variables constant.
- β_7: similarly to β_6, being in a Northern European region leads to a 5.36 decrease in logy compared to the reference category, holding all other variables constant.

For what concern the coefficients' uncertainties, the Standard Error of each coefficient measures the variability in the estimated effect of the variable on logy. A smaller standard error implies greater precision in estimating the predictor's effect.

In my model, all variables exhibit relatively small standard errors, with the most precise predictor being the pubEduExpmln with $SE(\hat{\beta_4}) = 0.0000018551$.

6.1 Confidence intervals and single t-test

Through the confint function we compute the $(1 - \alpha)\% = 95\%$ confidence interval for the betas:

```
##                                2.5 %        97.5 %
## (Intercept)             -1.951376e+00 -5.134573e-01
## empRaterecent           -3.764848e-02 -1.396611e-02
```

```
## enrolledfromabroadstem_l  6.928519e-01  8.495956e-01
## femteacherstertEdu_l      4.476486e-01  9.364180e-01
## pubEduExpmln              7.287808e-06  1.464364e-05
## avgWorkHrs               -1.327494e-01 -3.659675e-02
## area_ovest1              -1.138541e+00 -4.381988e-01
## area_nord1               -8.668031e-01 -2.059462e-01
```

From the output, we observed that all the variables are statistically significant because the respective confidence interval does not include zero. This observation is further supported by the results of the single t-test:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

From the table in the previous page, there are the values of the t statistic that determine if a parameter is statistically significant, indicating whether the predictor has an effect on the response variable. We can see that all the variables included in the model are significant at a $\alpha = 5\%$ level, since the greater pvalue available in the summary is the one for area_nord and is equal to $0.001711 < 0.05$.

6.2 Testing a group of predictors

Since we want to determine whether area_nord, which has the highest p-value among all the variables in the model, is relevant to the fit, we will use the ANOVA procedure to test whether a model with all 7 predictors is equivalent to a model without area_nord. If the two models are found to be equivalent, we can justify dropping the non-significant variable in favor of a less complex model:

$$\begin{cases} H_0 : Y = \beta_0 + \beta_1 X_{empRaterecent} + \beta_2 X_{enrolledfromabroadstem_l} + \beta_3 X_{femteacherstertEdu_l} + \beta_4 X_{pubEduExpmln} \\ \quad +\beta_5 X_{avgWorkHrs} + \beta_6 X_{areaovest} \\ H_1 : Y = \beta_0 + \beta_1 X_{empRaterecent} + \beta_2 X_{enrolledfromabroadstem_l} + \beta_3 X_{femteacherstertEdu_l} + \beta_4 X_{pubEduExpmln} \\ \quad +\beta_5 X_{avgWorkHrs} + \beta_6 X_{areaovest} + \beta_7 X_{areanord} \end{cases}$$

```
## Analysis of Variance Table
##
## Model 1: logy ~ empRaterecent + enrolledfromabroadstem_l + femteacherstertEdu_l +
##     pubEduExpmln + avgWorkHrs + area_ovest
## Model 2: logy ~ empRaterecent + enrolledfromabroadstem_l + femteacherstertEdu_l +
##     pubEduExpmln + avgWorkHrs + area_ovest + area_nord
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    107 31.289
## 2    106 28.504  1    2.7851 10.357 0.001711 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There's evidence to reject the null hypothesis and so to keep the full model with area_nord.

6.3 Goodness of fit

```
##          R2 Adjusted_R2    Cv_error   train_MSE
##   0.9326211   0.9281716   1.3533152   0.2500337
```

As measures of goodness of fit, I chose to consider the R^2 and R^2 adjusted as measures of internal validation, and the Cross Validation error, defined as $Cv_n = \frac{1}{114}\sum_{i=1}^{114}(\frac{y_i-\hat{y}_i}{1-h_{ii}})^2$, which is a measure of external validation compared with the training MSE $= \frac{1}{n}\sum_{i=1}^{114} e_i^2$. This choice is due to the fact that the goal of my analysis is to find out if my model is good for computing prediction based on new observation, and the cross validation error provides an average measure over the n observations of how well the model predicts fitted values based on the test set, which is a subset of my dataset composed by 1 observation that was not included in the model fitting.

The difference between the CV error and the training MSE is 1.10, meaning that the model performs much better in fitting values on the training data than predict response's values based on new data. It is enough to tell if my model is not good in giving prediction? We will find out in chapter 7, using a new observation available in the Eurostat website.

In my study, 93.2% of the variability in log(y) is explained by my model without considering the number of predictors, and 92.8% when accounting for them. The two values are very similar, indicating that my model is highly effective in explaining the variability of the response, regardless of the number of predictors. So it's not surprising that the model is less accurate in predicting new values because it's R^2 adjusted is very high.

7 Prediction

In order to test if my model is able to compute a realistic prediction of a new observation, I selected from the Eurostat website a data observation for the country Sweden and the year 2022:

```
data_new = data.frame(empRaterecent = 86.3 - mean(df$empRaterecent),
                      enrolledfromabroadstem_l = log(16700 - mean(df$enrolledfromabroadstem)
                                                     - min(women_new$enrolledfromabroadstem) + 1),
                      femteacherstertEdu_l = log(96.4 - mean(df$femteacherstertEdu)
                                                 - min(women_new$femteacherstertEdu) + 1),
                      pubEduExpmln = 40456.7 - mean(df$pubEduExpmln),
                      avgWorkHrs = 36.9 - mean(df$avgWorkHrs),
                      area_ovest = as.factor(0),
                      area_nord = as.factor(1))

predict(ols_2, newdata = data_new, interval = "prediction")
```

```
##        fit     lwr      upr
## 1 8.472877 7.38396 9.561794
```

```
real_value = log(2016 + 1)
real_value
```

```
## [1] 7.609367
```

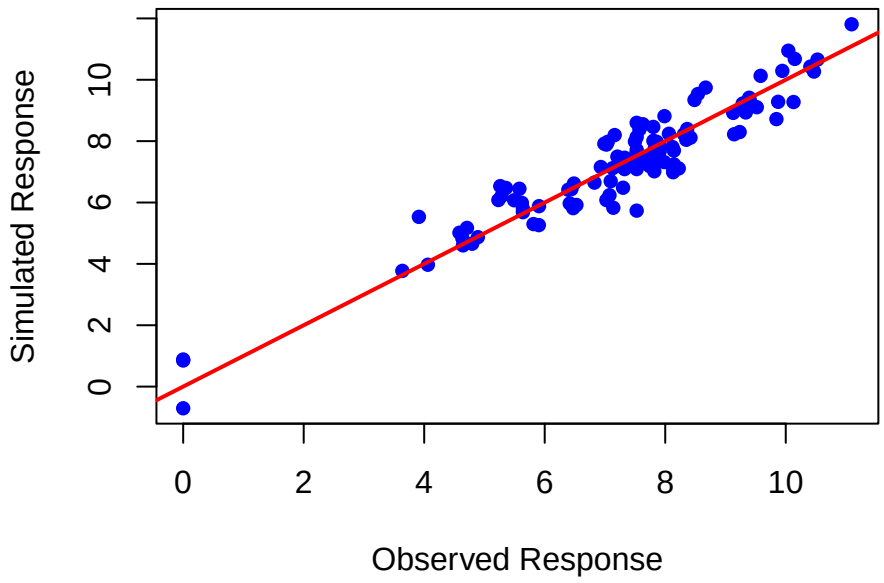
Despite the cross-validation error being larger than the training error, the model still provides a reliable prediction as the real value of the response is included in the 95% prediction interval. The $MSE = (7.609 - 8.473)^2$ is 0.746 for this new observation; this value is smaller than the CV error, but it's less stable because depends on the value of the new y_i.

Is this a reasonable value for Sweden? The predicted value is higher than the response mean (7.274), as well as the actual observed value. This can be attributed to the characteristics of Northern European countries, particularly their substantial investment in the educational system, so the answer is yes.

8. Simulation

Now let's simulate new responses $\hat{y}$ using the estimated coefficients beta as the true parameters, adding random noise based on the model's error distribution.

```
set.seed(12)
beta = coefficients(ols_2)
X = model.matrix(ols_2)
y_hat = X %*% beta + rnorm(n, 0, sigma(ols_2))
y_obs = ols_2$model[, 1]
plot(y_obs, y_hat, xlab="Observed Response", ylab="Simulated Response", pch=16, col="blue")
abline(a=0, b=1, col="red", lwd=2)
```



The red line represents the ideal y = x relationship and as we can see from the plot, my model's estimates are accurate because the points lie on the bisector. This is surprising because during the diagnostic phase I was unable to fully resolve the violation of the assumption of independent residuals, meaning that some heteroscedasticity in the errors remains.

9. Conclusion

What I found out in my analysis is that it's reasonable to use my model to give some prediction, despite it could be improved for example by considering other variables representing socio-political factors, or some geographic pattern that I haven't taken into account such as the wideness of each country or the total population of the country.

What I did not expect was how well my model fit the observed data, leading to the conclusion that a linear model is still well performing in fitting some data, in my case panel data, that violate some crucial assumption of the model, such as the uncorrelation of the errors, and it still can be used for making inference analysis.