ASSIGNMENT 1

Elisa Quadrini

Study on women attaining STEM education fields

The dataset used in the analysis is the following one:

```
## # A tibble: 6 x 16
##   GEO    Time Area  Tertiary_Edu_STEM Early_Childhood_Edu At_Most_Lower_Sec_Ed~1
##   <chr> <dbl> <chr>             <dbl>               <dbl>                  <dbl>
## 1 Belg~  2019 ovest              1869                98.2                   14.8
## 2 Bulg~  2019 est                1084                79.7                   17.5
## 3 Czec~  2019 est                2521                86.4                    7
## 4 Denm~  2019 nord               2515                97.7                   17.6
## 5 Germ~  2019 ovest             25060                94.2                   13.2
## 6 Esto~  2019 nord                366                91.5                   13.1
## # i abbreviated name: 1: ‘At_Most_Lower_Sec_Edu(25/34)‘
## # i 10 more variables: ‘At_Most_Lower_Sec_Edu(35/44)‘ <dbl>,
## #   ‘At_Least_Upper_Sec_Edu(20/24)‘ <dbl>,
## #   ‘At_Least_Upper_Sec_Edu(25/64)‘ <dbl>, Tertiary_Edu_Attain <dbl>,
## #   Employment_Rates_Recent <dbl>, ‘Enrolled_From_Abroad(STEM)‘ <dbl>,
## #   Fem_Teachers_Tertiary_Edu <dbl>, ‘Public_Edu_Exp(Mln)‘ <dbl>,
## #   Unexpected_Financial_Exp <dbl>, Average_Weakly_Hrs_Work <dbl>
```

The response variable is a continous one, and is named as “Tertiary_Edu_STEM”; then there’s a set of variables which are 2 character and 13 numerical variables, which are going to be the predictors of the model. We can denote that in the data collection, there’re panel data, which are difficult to manage in a multiple linear regression analysis because of the risk of lack of independence of the residuals, deriving from the autocorrelation of the variables, but we’re going to analyze this issue after the inspection of the missing values.

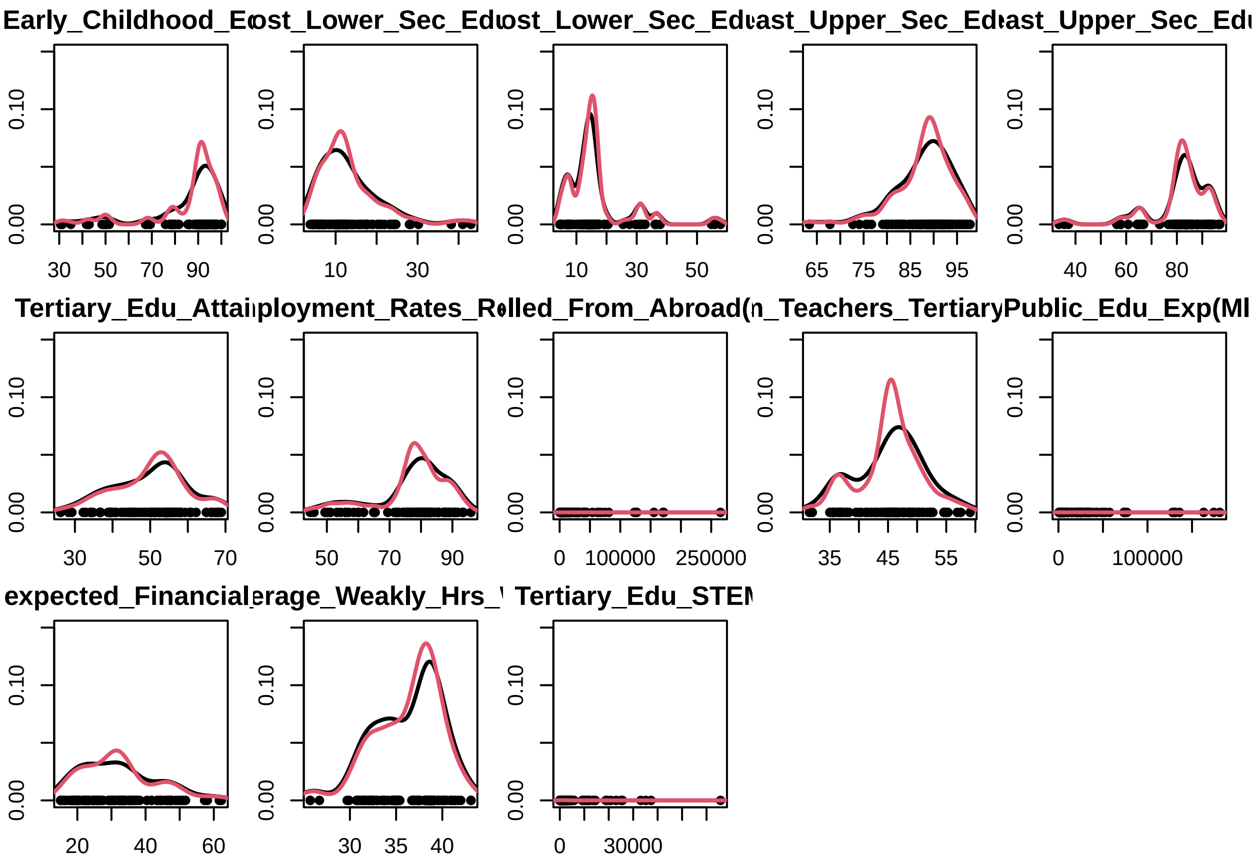
Missing values

First of all, we need to investigate if in the dataset are there some missing values, and manage them in the most appropriate way in order to create a valuable dataset to compute our analysis:

```r
library(mice)
md.pattern(women, plot = FALSE)
```

Using this comand we can easily visualize if are there any missing values and where are they collocated. Each column represents a variable, and the numbers at the bottom represent the number of missing values in the respective column, and at the bottom-right point there’s the total number of missing values.

After considering all the possible options, the best approach is to replace each missing value with a single fake data, which in this case is going to be the average or the median of the complete data, depending on how well the imputed dataset would fit the incomplete one:



Autocorrelation issue

We now have to check another important assumption of the multiple linear regression model, which is the assumption of indipendence of the residuals. It could be violated by the panel data, so first of all we need to check with the autocorrelation function, the autocorrelation of each variable:

```r
library(stats)

#autocorrelation for each numerical variable
acf_results <- list()

for (col in colnames(women_imp_mean)) {
  if (is.numeric(women_imp_mean[[col]])) {
    acf_values <- acf(women_imp_mean[[col]], plot = FALSE)
    acf_results[[col]] <- acf_values$acf
  }
}

#ACF at lag 1
for (col in names(acf_results)) {
  if(acf_results[[col]][2] > 0.25 || acf_results[[col]][2] < -0.25) {
    cat("ACF(1) for", col, ":", round(acf_results[[col]][2], 4), "\n")
  }
}
```

```
## ACF(1) for Time : 0.9737
## ACF(1) for Employment_Rates_Recent : 0.3354
## ACF(1) for Unexpected_Financial_Exp : 0.3394
```

```
#ACF at lag 2
for (col in names(acf_results)) {
  if(acf_results[[col]][3] > 0.25 || acf_results[[col]][3] < -0.25) {
    cat("ACF(2) for", col, ":", round(acf_results[[col]][3], 4), "\n")
  }
}
```

```
## ACF(2) for Time : 0.9474
```

The ACF is the value of the autocorrelation corresponding to a certain lag from the current observation: this means that ACF(1) represent the correlation between the current observation and the previous one and so on.

These are the value of the ACF that underline a moderate autocorrelation of the variables Employment_Rates_Recent and Unexpected_Financial_Exp. The variable Time is not going to be part of the model so its coefficient is not relevant.

So this analysis underlines that we could have some issues if we use the data corresponding to the years 2019, 2020, 2021 in our model, but in order to valuate the relevance of this issue or the possibility or not to continue to use the assumptions of the model, we can compute the Durbin-Watson test:

```
library(dplyr)
library(lmtest)

#model with the continous variables
women_filtered <- women_imp_mean %>% select(-GEO, -Time, -Area)

model <- lm(Tertiary_Edu_STEM ~ ., data = women_filtered)

#Test Durbin-Watson
dwtest(model)
```

```
##
##  Durbin-Watson test
##
## data:  model
## DW = 1.7408, p-value = 0.07538
## alternative hypothesis: true autocorrelation is greater than 0
```

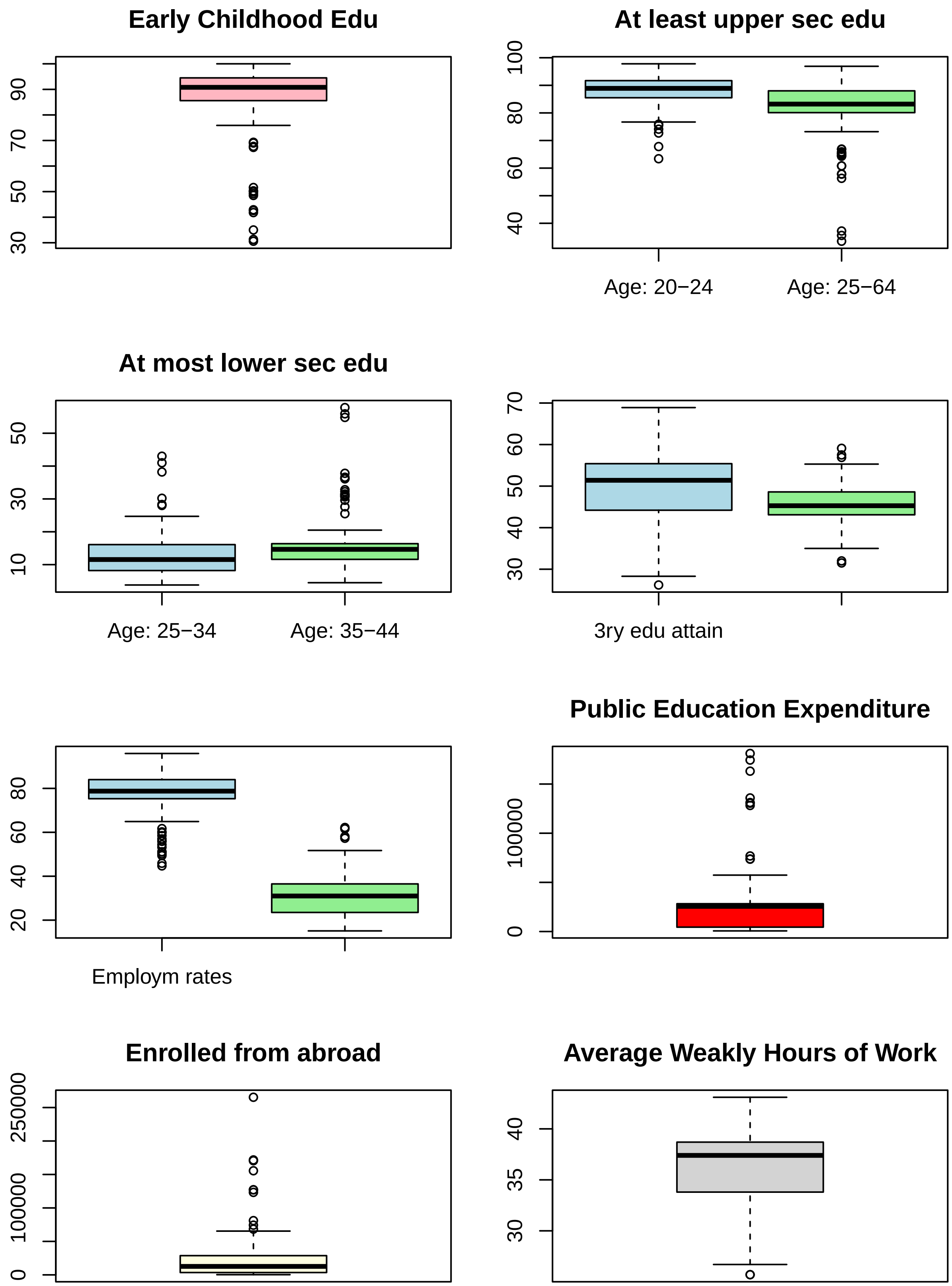
The Durbin-Watson (DW) statistic indicates the level of autocorrelation in the residuals.

General rule: 2.0 → No autocorrelation (ideal), < 2.0 → Positive autocorrelation (potential issue), > 2.0 → Negative autocorrelation A DW value that is slightly below 2, is suggesting moderate positive autocorrelation in the residuals.

The p-value indicates whether we can reject the null hypothesis of "no autocorrelation" in the residuals; if $p < 0.05$, there is statistical evidence of autocorrelation, but in our case it is 0.07538 so we can assume that there's no autocorrelation in the data.

Exploratory analysis

Univariate plots

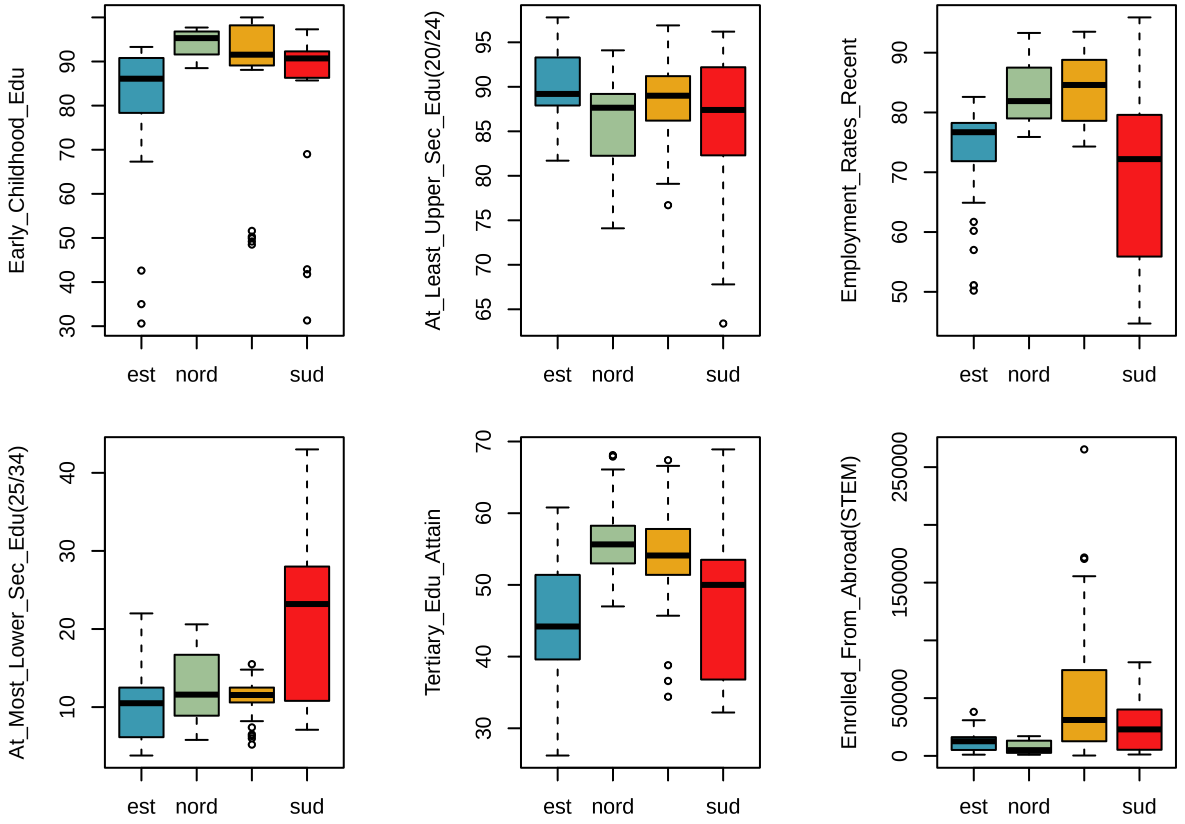


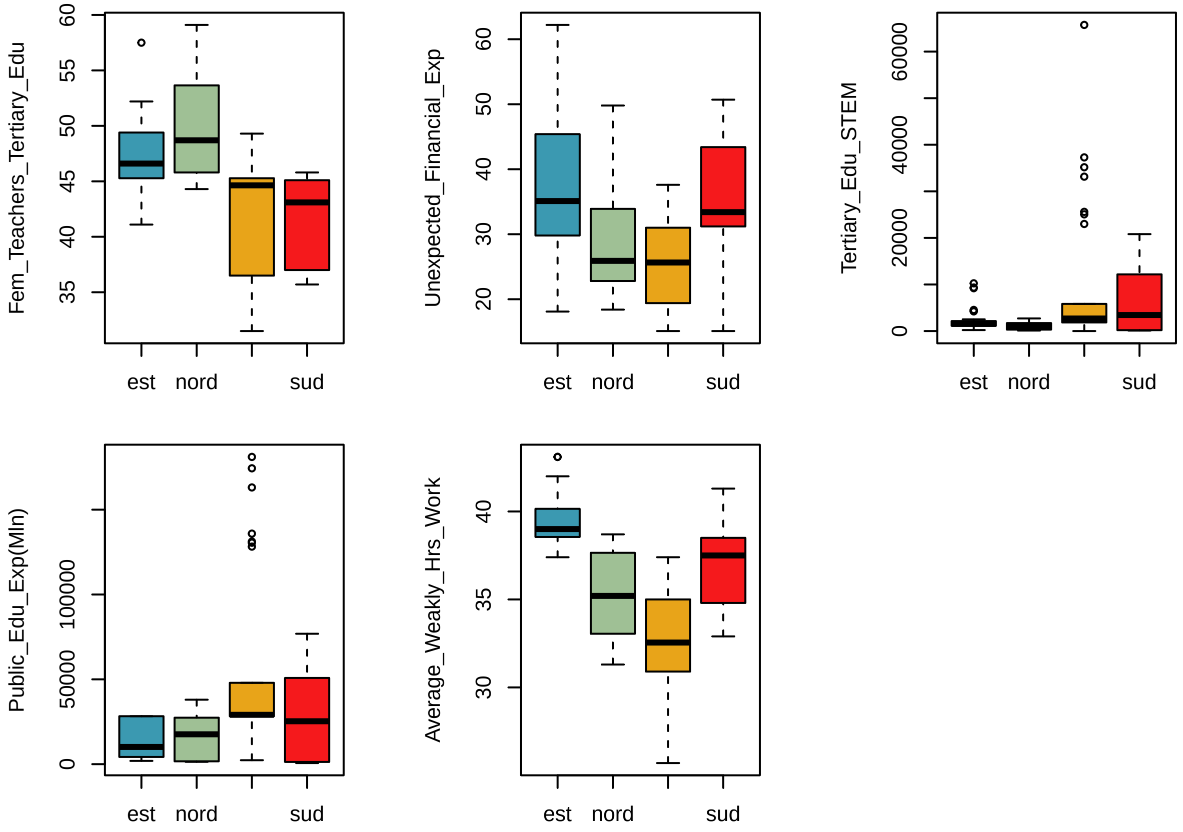
We can see above the univariate plots of the continuos covariates of the model, and what we can notice is that are there several outliers points that later on will need to be managed. From a roughly view of these distribution, we can presume that these outliers come from a specific trend depending on where these data have been collected. For example, is not unusual to think that in the north countries there are more students enrolled from abroad than in the other European countries.

Let's investigate these trends by considering the boxplots with respect to the three levels of the category "Area":

```
women_imp_mean$Area = as.factor(women_imp_mean$Area)
table(women_imp_mean$Area)
```

```
##
##   est  nord ovest   sud
##    39    24    30    21
```

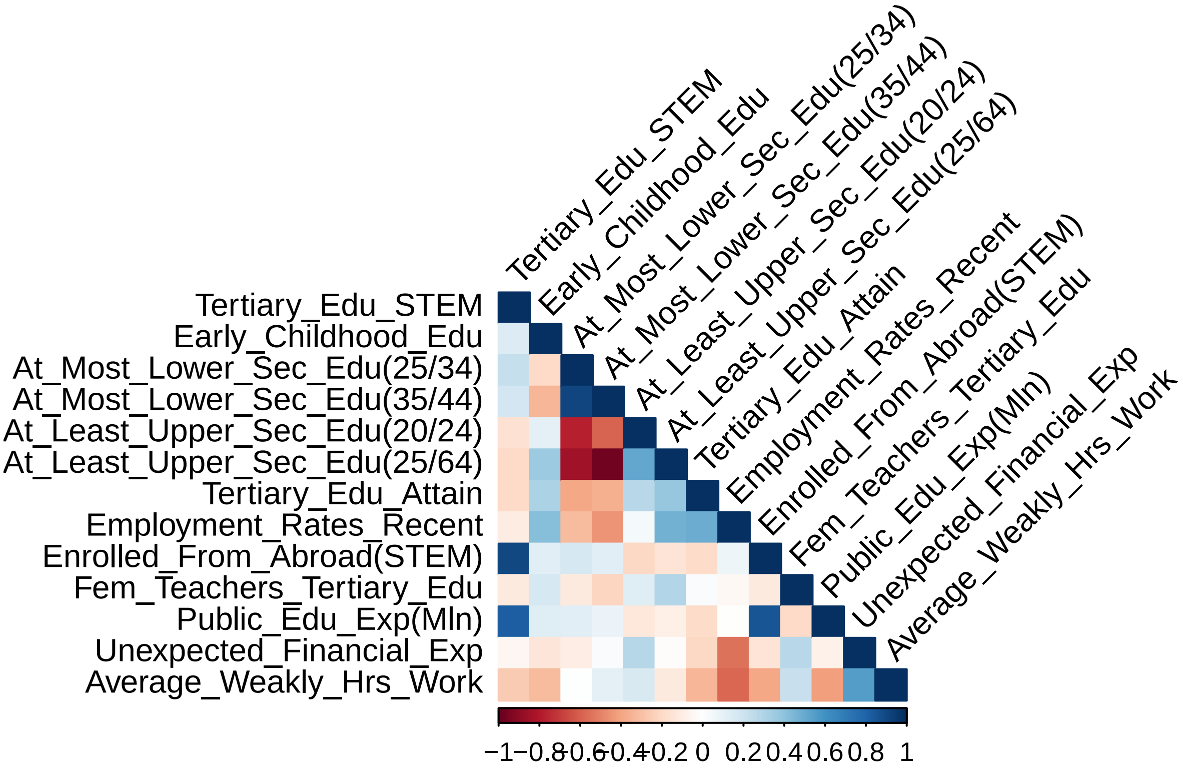




Here we can see the marginal distribution of the single covariates and the response with respect to the dummy variable "Area"; we notice that some distribution are slightly different, but we need to take in count that the population of these areas is not equal.

Now we can inspect the correlation between the variables and the response through the matrix of sample correlation:

```r
library(corrplot)
cor_matrix <- cor(women_filtered[, 1:13])
print(cor_matrix)
corrplot(cor_matrix, method = "color", type = "lower", tl.col = "black", tl.srt = 45)
```



From this rappresentation, we can see that there are some variable that are highly correlated; this might be a problem in the construction of our model, so later on we need to manage them.

Multiple Linear Regression

Let's fit a multiple linear regression using all the variables of our model, with an interaction term:

```
women_all = women_imp_mean[, 3:16]
colnames(women_all) <- make.names(colnames(women_all))
ols = lm(Tertiary_Edu_STEM ~. - Public_Edu_Exp.Mln. + Public_Edu_Exp.Mln.*Area, data= women_all)
summary(ols)
```

We denote that these estimates of the coefficients doesn't make sense with respect to the unit of measure of our response variable, which is a positive number.

So we decide to center all the continous variables in their means in order to obtain more significative values of the estimates:

```
vars_to_center <- c("Public_Edu_Exp.Mln.", "Early_Childhood_Edu",
                    "At_Most_Lower_Sec_Edu.25.34.", "At_Most_Lower_Sec_Edu.35.44.",
                    "At_Least_Upper_Sec_Edu.20.24.", "At_Least_Upper_Sec_Edu.25.64.",
                    "Tertiary_Edu_Attain", "Employment_Rates_Recent",
                    "Enrolled_From_Abroad.STEM.", "Fem_Teachers_Tertiary_Edu",
                    "Unexpected_Financial_Exp", "Average_Weakly_Hrs_Work")
```

```r
women_all[vars_to_center] <- scale(women_all[vars_to_center], center = TRUE, scale = FALSE)

ols_centered=lm(Tertiary_Edu_STEM ~. -Public_Edu_Exp.Mln. +Public_Edu_Exp.Mln.*Area, data = women_all)

summary(ols_centered)
```

```
##
## Call:
## lm(formula = Tertiary_Edu_STEM ~ . - Public_Edu_Exp.Mln. + Public_Edu_Exp.Mln. *
##     Area, data = women_all)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8601.7 -1628.5  -206.8  1576.2  7760.9
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    4.740e+03  9.134e+02   5.190 1.19e-06 ***
## Areanord                       1.913e+03  1.563e+03   1.224 0.224127
## Areaovest                     -9.509e+02  1.354e+03  -0.702 0.484213
## Areasud                       -3.855e+02  1.888e+03  -0.204 0.838620
## Early_Childhood_Edu            1.850e+01  2.408e+01   0.768 0.444262
## At_Most_Lower_Sec_Edu.25.34.   6.183e+02  1.833e+02   3.374 0.001075 **
## At_Most_Lower_Sec_Edu.35.44.  -5.022e+02  2.179e+02  -2.305 0.023342 *
## At_Least_Upper_Sec_Edu.20.24.  4.520e+02  1.147e+02   3.941 0.000155 ***
## At_Least_Upper_Sec_Edu.25.64. -2.754e+02  1.454e+02  -1.894 0.061248 .
## Tertiary_Edu_Attain            1.408e+02  4.803e+01   2.931 0.004231 **
## Employment_Rates_Recent       -2.988e+01  6.310e+01  -0.473 0.636947
## Enrolled_From_Abroad.STEM.     2.232e-01  1.736e-02  12.858  < 2e-16 ***
## Fem_Teachers_Tertiary_Edu     -2.135e+02  8.776e+01  -2.433 0.016867 *
## Unexpected_Financial_Exp      -1.598e+01  4.005e+01  -0.399 0.690814
## Average_Weakly_Hrs_Work        5.684e+02  1.704e+02   3.336 0.001213 **
## Public_Edu_Exp.Mln.            1.221e-02  4.941e-02   0.247 0.805432
## Areanord:Public_Edu_Exp.Mln.  -6.960e-02  7.061e-02  -0.986 0.326819
## Areaovest:Public_Edu_Exp.Mln.  1.118e-03  5.403e-02   0.021 0.983537
## Areasud:Public_Edu_Exp.Mln.    1.716e-01  5.875e-02   2.921 0.004354 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2886 on 95 degrees of freedom
## Multiple R-squared:  0.9197, Adjusted R-squared:  0.9045
## F-statistic: 60.44 on 18 and 95 DF,  p-value: < 2.2e-16
```

When we fit a model with an interaction term and a dummy variable, like in this case Area has 4 levels and there’s the interaction between Area and Public_Edu_Exp(Mln), we expect that our response variable is estimated by 4 different values of the intercept depending on the Area’s levels, and the slope of the continue variable Public_Edu_Exp(Mln) is not constant, but it assumes 4 different values for the same reason as the intercept, with the adding of the other p-1 betas parameters.

Interpreting the parameters

The intercept:

```
## (Intercept)
##    4740.446
```

The intercept represents the expected number of women graduating in STEM fields in the reference category ("est") when all other predictor variables are at their mean values (since we centered the continuous predictors). It provides a baseline estimate of the number of women in STEM in the eastern region, assuming the average conditions for all other variables.

Areanord = 1,913: in the North, the number of women in STEM is expected to be 1,913 higher than in the East. However, the p-value (0.224) suggests this is not statistically significant.

Areaovest = -950.9: in the West, the number of women in STEM is 950 fewer than in the East, but this effect is also not significant (p = 0.484).

Areasud = -385.5: in the South, there are 385 fewer women than in the East, and this effect is also not significant (p = 0.839).

The regional differences in STEM graduates are not statistically significant, except for Areanord, which shows a moderate increase.

Values of the Public Education Expenditure:

```
## Public_Edu_Exp.Mln.
##          0.01220521
```

The estimate suggests that a 1 million increase in public education spending in the east is associated with an increase of only 0.012 women in STEM, which is negligible. The p-value (0.805) confirms that this effect is not statistically significant.

Public education spending, in general, does not appear to have a strong direct impact on STEM graduation rates.

Interaction Between Region and Public Education Expenditure:

Areanord:Public_Edu_Exp.Mln. (-0.0696): in the North, the effect of public education spending is slightly negative, but not significant (p = 0.327).

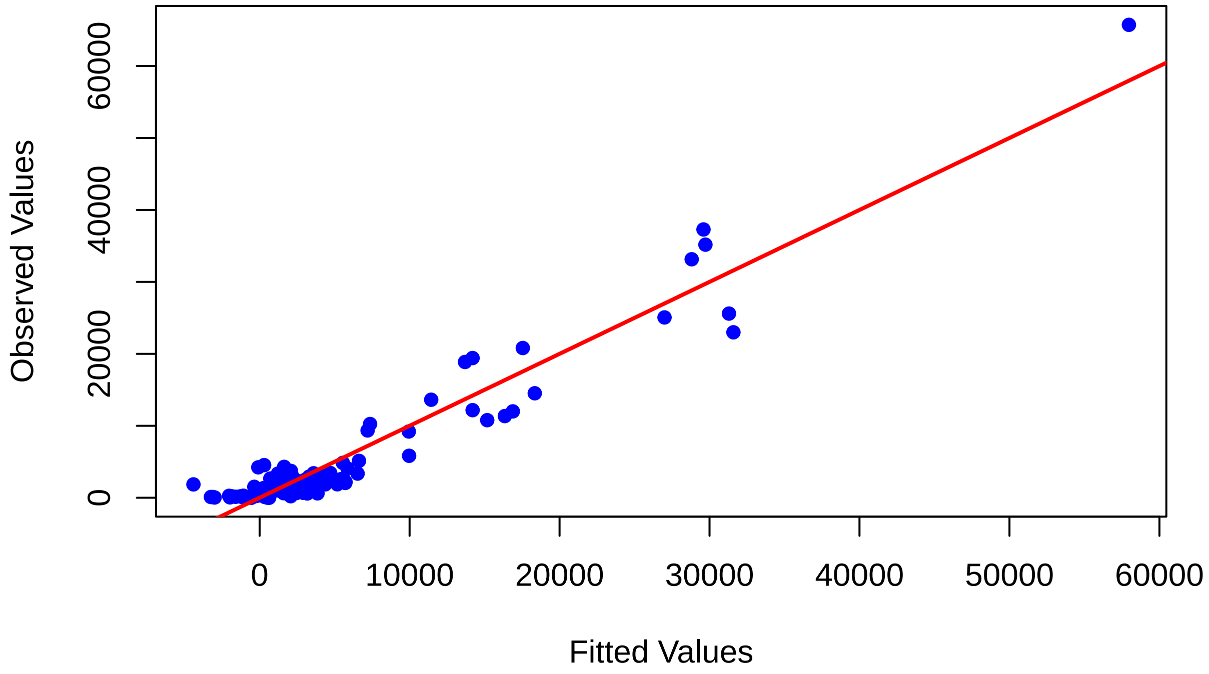
Areaovest:Public_Edu_Exp.Mln. (0.001118): in the West, the interaction effect is very close to zero, and not significant (p = 0.984).

Areasud:Public_Edu_Exp.Mln. (0.1716): in the South, the effect of public education spending is positive and significant (p = 0.004).

Public education spending has a significant effect only in the South, where higher investment correlates with more women in STEM.

Graphical presentation of the model:

We can notice in the graphic below that the points are well explained by the red line, which is the regression line created by our model. The observed values are the real values of the response variable in our dataset.



Sigma and R-Squared

```
print(summary(ols_centered)$sigma)
```

```
## [1] 2886.365
```

```
r_squared <- summary(ols_centered)$r.squared
adj_r_squared <- summary(ols_centered)$adj.r.squared
print(c(R2 = r_squared, Adjusted_R2 = adj_r_squared))
```

```
##          R2 Adjusted_R2
##   0.9196865   0.9044692
```

Sigma (Residual Standard Error): this is the average prediction error of the model.

Multiple R-squared = 0.9197 this indicates that 92% of the variability in STEM graduates is explained by the model.

Adjusted R-squared = 0.9045 (90.4%) It accounts for the number of predictors. Since it is very close to R-squared, it confirms that the model is not overfitting and remains reliable.

The model explains most of the data's variability with acceptable prediction errors. The small difference between R-squared and Adjusted R-squared indicates a well-balanced model.

Testing the betas

```
##
## Coefficient: (Intercept)
##   Estimate: 4740.446
##   t-value: 5.189894
##   p-value (two-sided): 1.191322e-06
##   p-value (one-sided, H1: beta > 0 ): 5.956608e-07
##
## Coefficient: Areanord
##   Estimate: 1912.933
##   t-value: 1.223605
##   p-value (two-sided): 0.2241274
##   p-value (one-sided, H1: beta > 0 ): 0.1120637
##
## Coefficient: Areaovest
##   Estimate: -950.853
##   t-value: -0.7022939
##   p-value (two-sided): 0.4842131
##   p-value (one-sided, H1: beta < 0 ): 0.7578934
##
## Coefficient: Areasud
##   Estimate: -385.4903
##   t-value: -0.2042183
##   p-value (two-sided): 0.8386198
##   p-value (one-sided, H1: beta < 0 ): 0.5806901
```

This can be iterated for all the estimates of the beta_j.

We performed one-sided hypothesis tests based on the sign of each estimated coefficient (beta_j). If beta_j > 0, we tested H1: beta_j > 0; if beta_j < 0, we tested H1: beta_j < 0, ensuring that the alternative hypothesis aligns with the estimated effect's direction.

This is the single t-test, an hypothesis test with the null hypothesis equal to beta_j=0. It evaluates the significance of each variable in the construction of the model; if the p-value is high, it means that the variable is not so significant in the estimate of the response, but it could be significant with respect to interactions with the other variables.

Test of a group of predictors

We now test the full model with all the predictors against a reduced model without the interaction term and all that predictors whose p-value of the single t-test indicated a not so relevant influence in the model:

```
## Analysis of Variance Table
##
## Model 1: Tertiary_Edu_STEM ~ (Area + Early_Childhood_Edu + At_Most_Lower_Sec_Edu.25.34. +
##     At_Most_Lower_Sec_Edu.35.44. + At_Least_Upper_Sec_Edu.20.24. +
##     At_Least_Upper_Sec_Edu.25.64. + Tertiary_Edu_Attain + Employment_Rates_Recent +
##     Enrolled_From_Abroad.STEM. + Fem_Teachers_Tertiary_Edu +
##     Public_Edu_Exp.Mln. + Unexpected_Financial_Exp + Average_Weakly_Hrs_Work) -
##     Public_Edu_Exp.Mln. - Early_Childhood_Edu - Employment_Rates_Recent -
##     Unexpected_Financial_Exp
## Model 2: Tertiary_Edu_STEM ~ (Area + Early_Childhood_Edu + At_Most_Lower_Sec_Edu.25.34. +
##     At_Most_Lower_Sec_Edu.35.44. + At_Least_Upper_Sec_Edu.20.24. +
##     At_Least_Upper_Sec_Edu.25.64. + Tertiary_Edu_Attain + Employment_Rates_Recent +
##     Enrolled_From_Abroad.STEM. + Fem_Teachers_Tertiary_Edu +
##     Public_Edu_Exp.Mln. + Unexpected_Financial_Exp + Average_Weakly_Hrs_Work) -
##     Public_Edu_Exp.Mln. + Public_Edu_Exp.Mln. * Area
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1    102 1242043009
## 2     95  791454858  7 450588151 7.7264 2.248e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If the p-value < 0.05, we reject the null hypothesis, meaning that the eliminated variables jointly have a significant effect.

Prediction of the response

```r
# New dataset with hypothetical new values of the predictors
newdata <- data.frame(
  Area = "nord",
  Public_Edu_Exp.Mln. = 900,
  Early_Childhood_Edu = 80,
  At_Most_Lower_Sec_Edu.25.34. = 30,
  At_Most_Lower_Sec_Edu.35.44. = 25,
  At_Least_Upper_Sec_Edu.20.24. = 40,
  At_Least_Upper_Sec_Edu.25.64. = 35,
  Tertiary_Edu_Attain = 50,
  Employment_Rates_Recent = 70,
  Enrolled_From_Abroad.STEM. = 15,
  Fem_Teachers_Tertiary_Edu = 45,
  Unexpected_Financial_Exp = 20,
  Average_Weakly_Hrs_Work = 38
)

# Interval of prediction at level 95%
predict(ols_centered, newdata = newdata, interval = "prediction", level = 0.95)
```

```
##        fit      lwr      upr
## 1 39134.53 15888.88 62380.18
```

The range of prediction is the value upr-lwr; it is big but it aligns with the unit of measures of the response, so it can be a good prediction. The value "fit" represents the new prediction of the response.