Investigating the relationship between metal exposure experienced by the child and the mother and its effect on the child’s IQ between the ages 6-11.

STAT 331 Final Project

Date: 2022-12-06

Group #: 11

Names:
Pival Dhingra
Sagun Malwatkar
Prachi Patel
Reva Puri

Summary:

The objective of the report was to investigate the relationship between metal exposure experienced by the child and the mother and its effect on the child's IQ between the ages 6-11. In order to carry out this investigation, we fit several models using methods such as LASSO, stepwise, forward, and backwards. In order to select the best model to summarize the findings, we selected the full model as it had the highest R-squared and value and lowest BIC value of 120.0868. This model was the best fit for our objective because compared to the other models, this model fit the data and helped explain our findings relative to the objective. From our findings we concluded that there is not a significant relationship between metal exposure experienced by the child and the mother and its effect on the child's IQ.

Objective:

Our goal for the analysis is to investigate the relationship between metal exposure experienced by the child and the mother and its effect on the child's IQ between the ages 6-11.

Exploratory Data Analysis:

The data set that was explored covered 3 data frames called phenotypeNA with 1301 rows and 7 columns, exposomeNA with 1301 rows and 223 columns and covariatesNA with 1301 rows and14 columns. The focus was on the variable IQ (belonged to phenotypeNA) and all the metal exposures (in exposomeNA). Therefore, those were the features of the study. Looking at the data, we found out that every variable was continuous except the variable that measured the exposure of Thallium in mother and child. That variable was categorical with values: "Detected" for when the element is found in the mother or child and "Undetected" for when it is not found in either of them. Looking at the summary of the data set with the necessary variables present, it was found that the mean and median of all the variables were fairly close to each other, which might be because of absence of outliers. Also, for IQ there were 10 missing values, marked as "NA". Hence, 10 rows were removed while plotting graphs. Further investigation of the distributions of each of the variables led to the conclusion that the variables were not symmetric about their mean and median. For mothers' metal exposures, we can clearly see that lots of variables are clustered around certain values. For IQ, we can say that around 7% of children have an IQ of 32 and clearly is not normal either. However, for the metal exposure in children, most of the distributions seem normal but are slightly skewed (left or right). That indicates the direction of the outliers present, contradicting our previous analysis of outliers. Additionally, upon analyzing the multi-collinearity between the exposure or metals, IQ and age of the child through scatterplots and the scanning procedure of multicollinearity elimination (*Appendix 36*), we established that there was no multicollinearity present. Figures 1.0-1.9 located at the bottom of the document shows all the graphs with scatter plots between IQ and the metals, along with the shape of their distributions and collinearities which are below 50% except the “Age” variable. It was also found that some of the correlations had negative values.

Methods:

The aim of this study was to explore the relationship between the metal exposure experienced by the child and the mother and the child’s IQ, targeting ages 6-11. We fit multiple models, namely, the full complete model, forward selection, stepwise selection, VIF and backward elimination. In order to select an appropriate model for our data, we looked at the BIC values and adjusted R^2 values for each of the models. The R^2 values were relatively similar for all the models. Since unexplained variation in the dependent variable and the number of independent variables (complexity) increase with BIC, we chose the full model because it has the least BIC value of 120.0868. Figures 2.0 and 2.1 show a summary of the AIC, BIC, R^2, Adjusted R^2 and p-value for all the models. As you can see in figure 2.0, the full model and the reduced model have the same values for all the parameters as the reduced model and the full model turned out to be the same. This is because none of the covariates in the had a VIF that was greater than 5 and hence, all the covariates were retained. The statistical methods that were used are appropriate as AIC, BIC, and adjusted R^2 values were taken into account before selecting an appropriate model. The full model has the least AIC and BIC values compared to the rest of the models. The AIC is primarily used to estimate future predictions, and hence was not of interest while selecting the model as it did not align with our goal of the investigation. Rather than understanding the future predictions, we wanted to understand the relationship between IQ and metal exposure, and hence to select the best model, we took into consideration the R squared and BIC values. The AIC for the full model is also the lowest amongst all the other models, which further makes it a best fit to select as it can also help with future predictions of the relationship between metal exposure and child IQ. While creating the different models, we used extensions of the basic multiple linear regression model where we chose reduced/minimal model. The reduced model was to find a relationship between the child’s IQ and the child’s age. Although the model that we selected doesn’t fit the data well, it is the best model that could be used to interpret the data as there is not a significant relationship observed between metal exposure and child IQ. The necessary assumptions required to fit the model are met. The model assumes linearity, equal variance, independence and normality, and hence all the assumptions of a multiple linear regression model are met.

The comparison between full and reduced models in terms of R square, adjusted R square, AIC, BIC and p-value of the F-test.

Category	Full_mod	Reduced_mod
R2	0.5207	0.5207
adj. R2	0.5124	0.5124
AIC	7570.7880	7570.7880
BIC	7694.7041	7694.7041
F pval	0.0000	0.0000

Figure 2.0: A comparison between the full and reduced models

The comparison between differnet methods to analyse the data

	Full_mod	FWD	STEP	BACK
R2	0.52061	0.52002	0.52002	0.52002
adj_R2	0.51268	0.51437	0.51437	0.51437
AIC	68.24673	7626.89000	7558.64000	7558.64000
BIC	120.08680	7766.50000	7646.41000	7646.41000
MSPE for LASSO method	41.57002	NA	NA	NA
MSPE for RIDGE method	41.75840	NA	NA	NA
F pval	0.00000	0.00000	0.00000	0.00000

Figure 2.1: comparison between different models used to select the best model

Results:

After thorough analysis, we concluded that there is no significant relationship between the IQ of a child and the amount of metal the child and their mother were exposed to, based on the fit of the model.

As mentioned previously, we fit multiple models to our dataset, namely, the full complete model, forward selection, stepwise selection, VIF and backward elimination. From the calculation results according to Figure 2.1, we see that the p-values for all of the models is less than 0.0001. That indicates that there is enough evidence to reject the null hypothesis, namely, that no linear relationship exists between the child's IQ (response variable) and the metal exposure for any of the metals for the mother and child. For additional analysis and scope of improvement, the full model was chosen based on the BIC value and the adjusted R^2 and AIC.
Furthermore, from the residual plot for the full model in Figure 2.3, we can see that the independent variables (i.e., the metal exposures) record the deterministic element well for our required goal, which is to investigate the relationship between child's IQ and metal exposures. The pattern in the residuals does not seem doubtful or problematic.

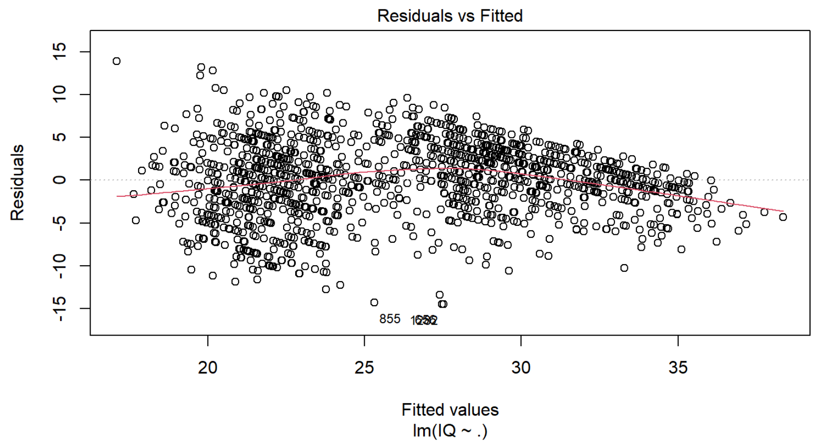

Figure 2.2: Residuals Vs Fitted values for the full model.

The results of further analysis and attempts to improve the full model can be seen in Figure 2.0. We found the adjusted R^2 value of 51.24% for the full model which suggests that the variation in the response variable around its mean is not explained well by the model. In fact only around 51.24% of the variation in the IQ of a child around the mean IQ is explained by the model. The spread of the data around the fitted regression line is more than we would like for our fitted model. Furthermore, the p-value of the fitted model is found to be less than 0.0001. Thus, there exists enough evidence to reject the null hypothesis surrounding the model, namely, that no linear relationship exists between the child's IQ (response variable) and the metal exposure for any of the metals for the mother and child. The BIC of 7894.70 also implies the presence of variation in the child's IQ unexplained by the metal exposure factors that were considered and a complex model with more explanatory variables. We ran a VIF score screening procedure to decrease the number of explanatory variables because theoretically the amount of metal in a mother's body could be collinear with the amount of metal in the child's body. This could be because of their living conditions or it could even depend on when the data was observed with respect to the child's age. The VIF score screening did not result in a reduced model. We found the result to be the same full model since there were no independent variables with a higher VIF score. Therefore, we have the same adjusted R^2 value, BIC and p-value as shown in Figure 2.0.

In conclusion, after fitting multiple models (full model, forward selection, stepwise selection, backward elimination), the model with least BIC value (full model) showed no significant relationship between the IQ of a child and the amount of metal the child and their mother were exposed to.

Discussion:

One of the main limitations of this study is that it is not possible to segregate the metals individually and test for their effects. There will always be one or another metal present at the time of the exposure. Additionally, since the data was taken from different regions of Europe, there will be a difference in atmosphere for all the candidates, and hence not all the units of study will be homogenous in terms of exposure. Another limitation for the study is that there could be other factors influencing a child's IQ other than the effects of metals such as socioeconomic status, and educational background, which can not be eliminated during the testing of the child's IQ. Hence, not all the data presented for the child's IQ is solely attributed to the effects of metals. Another limitation to the study could be the lack of follow-up to test the child's IQ. Furthermore, another limitation in the study is the fact that we don't know the level of metal exposure experienced by the child and the mothers. Some could have experienced more/less exposure but since there is no control over this, it is hard to identify the level of exposure. Hence the level of metal exposure is not homogenous amongst all the participants, both children and mothers of the HELIX study. This could greatly affect the data and skew the data in many directions. Lastly, a limitation

The objective of the study was to investigate the relationship between metal exposure experienced by the child and the mother and its effect on the child's IQ between the ages 6-11. After our investigation, we found that there is no significant relationship between the effect of metals exposure on the child's IQ between the ages 6-11. This conclusion seems reasonable given the results from the best model selected as well as the methods used to perform explanatory data analysis. In the explanatory data analysis it was revealed that there was no significant correlation between the different metals exposures and child IQ, which is also in line with the results from the selected model. These findings make sense and seem reliable since the metals exposures that were used in the experiment were ordinary metals that we encounter in our day-to-day lives. None of the metals that we worked with in the data set were harmful metals that had a direct impact on the human brain or metals that could affect pregnant women. This means that if the metals did not impact the brain in any manner, then it would not have affected the IQ of children either, which is exactly what we found through our analysis. There was no significant correlation between metal exposure and child IQ, and the very little correlation that was observed could be contributed to the limitations in the study. For instance, one of the limitations in the study was that other factors such as socio-economic status and educational background that could affect child IQ can not be eliminated when testing a child for their IQ. This could explain the minimal relationship observed between metal exposure and child IQ.

Additional Figures:

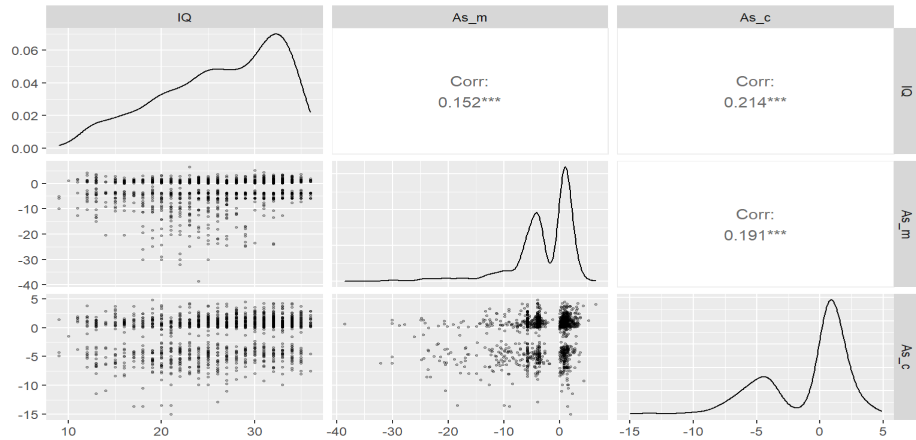

Figure 1.0: The distributions of IQ and Arsenic's exposure to mother (As_m) and child (As_c) along with correlations.

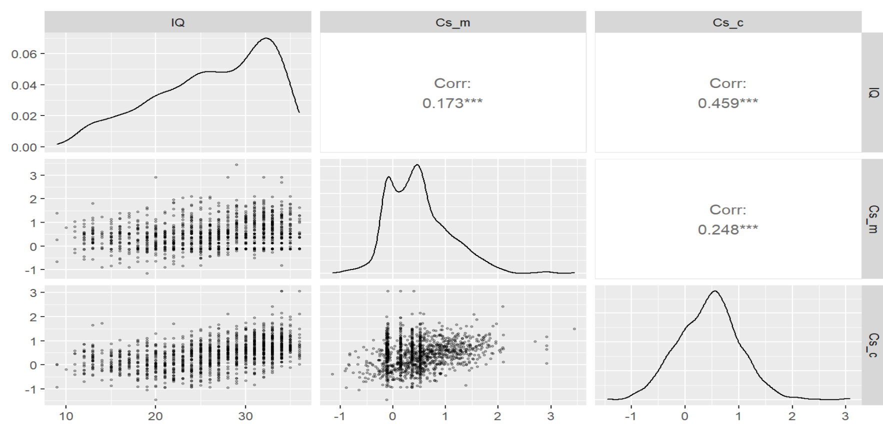

Figure 1.2: The distributions of IQ and Caesium's exposure to mother (Cs_m) and child (Cs_c) along with correlations.

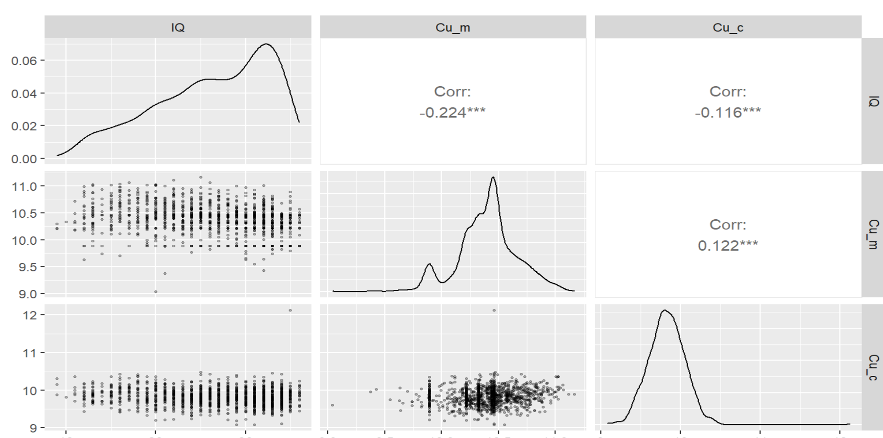

Figure 1.3: The distributions of IQ and Copper's exposure to mother (Cu_m) and child (Cu_c) along with correlations.

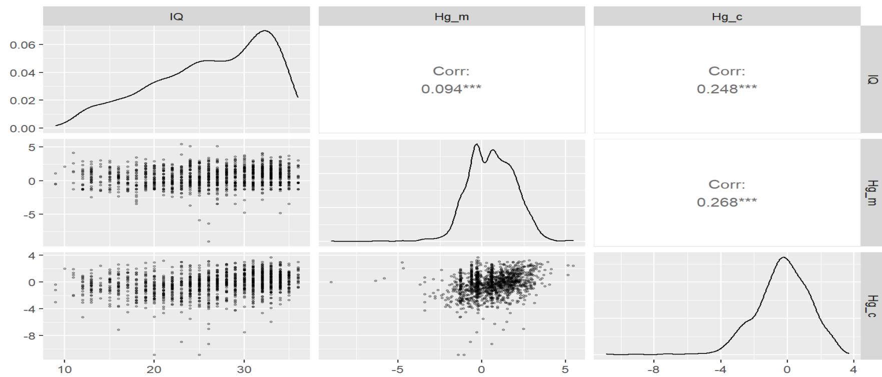

Figure 1.4: The distributions of IQ and Mercury's exposure to mother (Hg_m) and child (Hg_c) along with correlations.

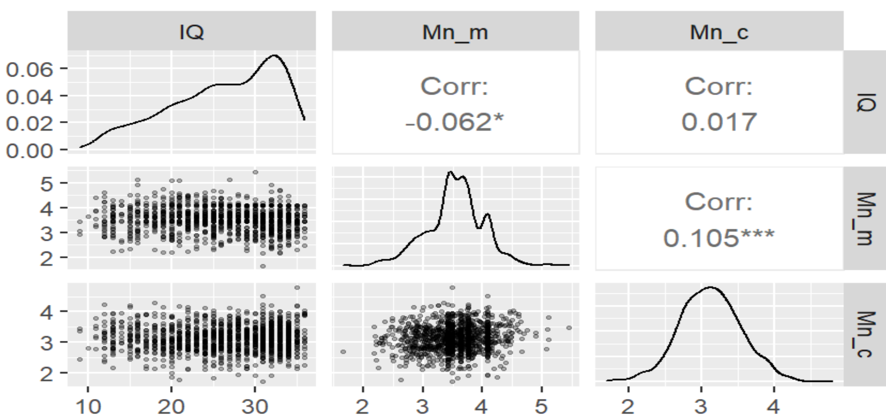

Figure 1.5: The distributions of IQ and Manganese's exposure to mother (Mn_m) and child (Mn_c) along with correlations.

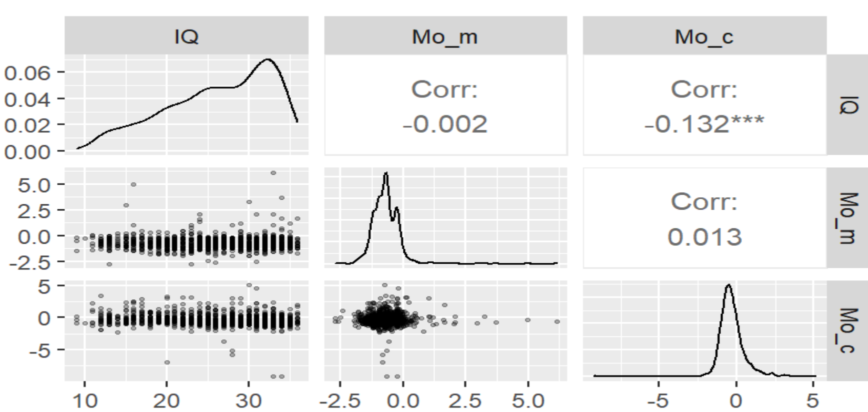

Figure 1.6: The distributions of IQ and Molybdenum exposure to mother (Mo_m) and child (Mo_c) along with correlations.

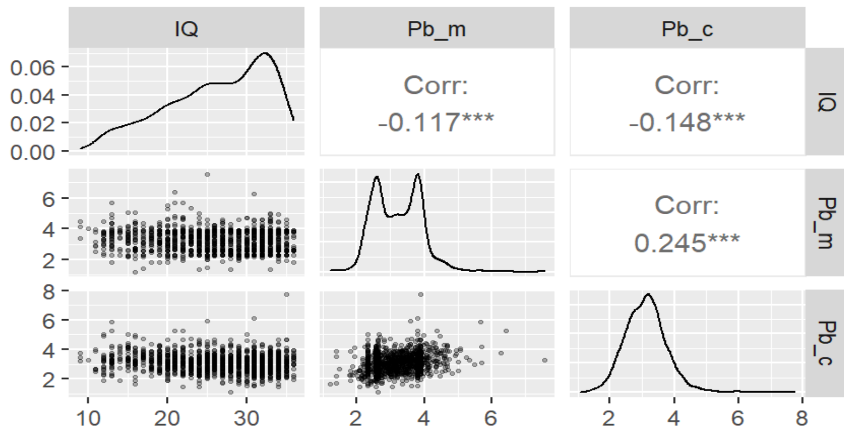

Figure 1.7: The distributions of IQ and Lead exposure to mother (Pb_m) and child (Pb_c) along with correlations.

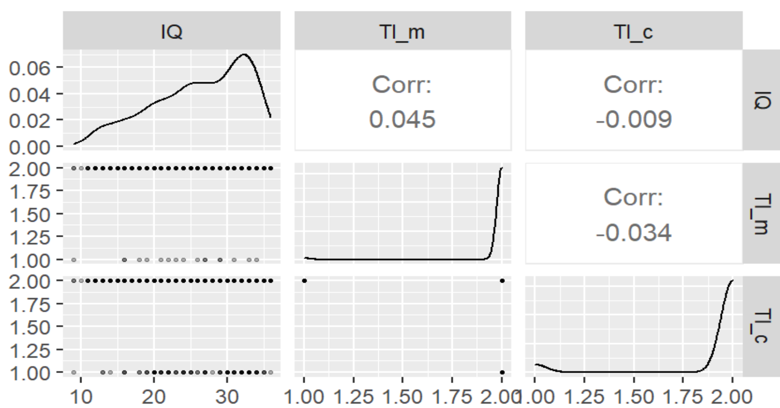

Figure 1.8: The distributions of IQ and Thallium exposure to mother (Tl_m) and child (Tl_c) along with correlations.

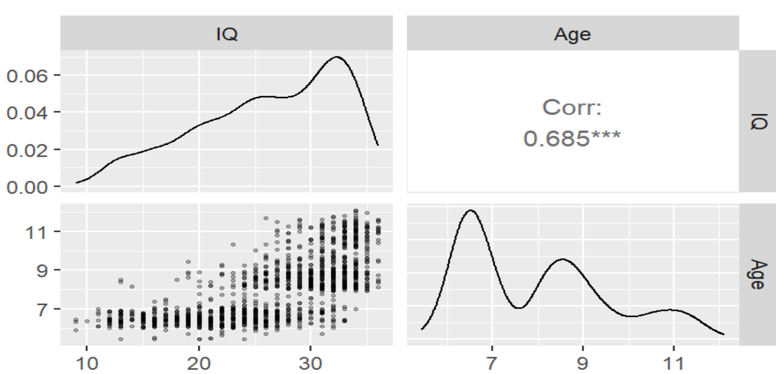

Figure 1.9: The distributions of IQ and Age of the child along with correlation to IQ.