STAT 331 Final Project

"How well does a mother's lifestyle behaviour affect her child's IQ and what is the relative impact that postnatal and prenatal behaviour have on this?"

Contributors:
Aania Raheem, Muhammad Bilal Khan, Kazi Rahman, Srijan Chaudhuri

Summary:

This report explores the question of how well data on a mother's lifestyle can predict the intelligence quotient of her child at 6-11 years of age. Analysis was further split into prenatal and postnatal lifestyle of the mother and prediction accuracy was observed.

Available data from the HELIX study was used to build multiple models, all of which was regressed on the raven score of the child. A total of 5 models were constructed and prediction errors from each model were taken into account to conclude on the most optimal model and the relative strengths of each model. Model selection was done based on metrics describing model goodness of fit.

At the end of our analysis, our optimal model has an MSPE value of X and whatever metric there is. Our optimal model only considered features under the Lifestyle family of the dataset. When compared with a model fitted with all the possible features in the dataset, our optimal model still fell short but only within hopefully a range of +-1 or whatever.

Objective:

The main goal of our research was to reach a conclusive decision on if a mother's lifestyle predicts her child's IQ well and which period of lifestyle predicts the child's IQ most accurately; postnatal or prenatal?

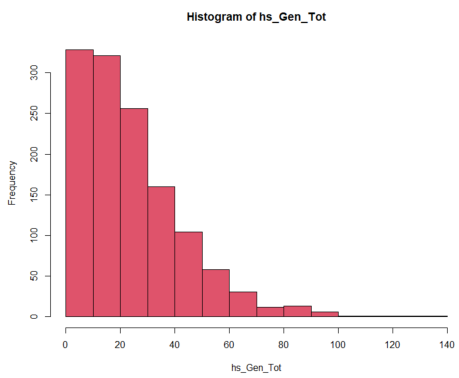
Our analysis was aiming to uncover the effect of a mother's external and non-genetic attributes and habits on a child's IQ by measuring such an effect on the sample of women provided from the HELIX study. The final conclusion we aimed to produce was a holistic comparison of the linear models regressing the raven score on lifestyle covariates, prenatal lifestyle covariates and postnatal lifestyle covariates, as well as reference models created for relative comparison. This allowed us to quantify the relative strength of our models of interest in prediction.

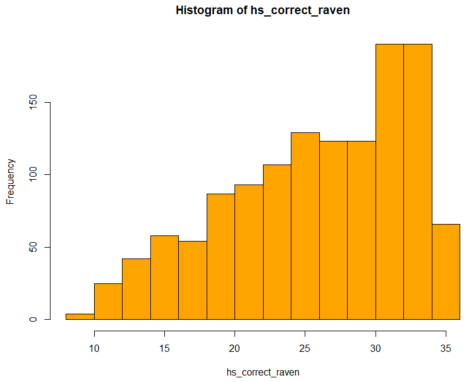
Exploratory Data Analysis:

Data Insights:

An initial run-through of the provided data helped us identify two possible variables that could be used as measures of IQ and act as our response variable:

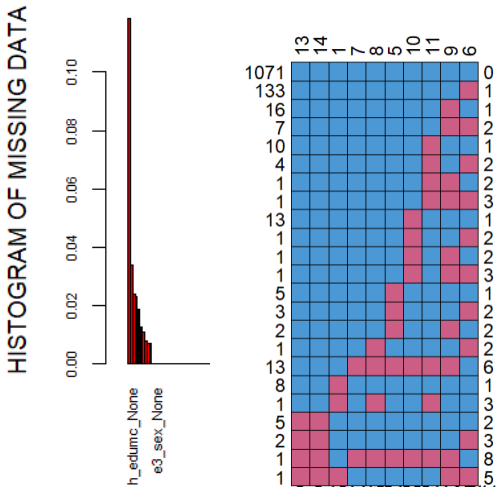
- hs_correct_raven : the number of correct answers on the raven test, taken by the child at age 6-11
- hs_Gen_tot : the neural behaviour of a child at age 6-11





A brief examination of plots, distribution and definition of the two variables ultimately led us to select hs_correct_raven as our response variable of interest. This was due both to its definition as a measure of IQ and also because of the variability of the raven scores as opposed to the neural behaviour measures, which seemed quite skewed and tending to the left. As researchers, we were interested in if our lifestyle covariates could explain the variability behind the raven scores.

Data Cleaning Summaries:



On an initial run-through of data, it was found that many rows contained missing values for a variety of covariates.

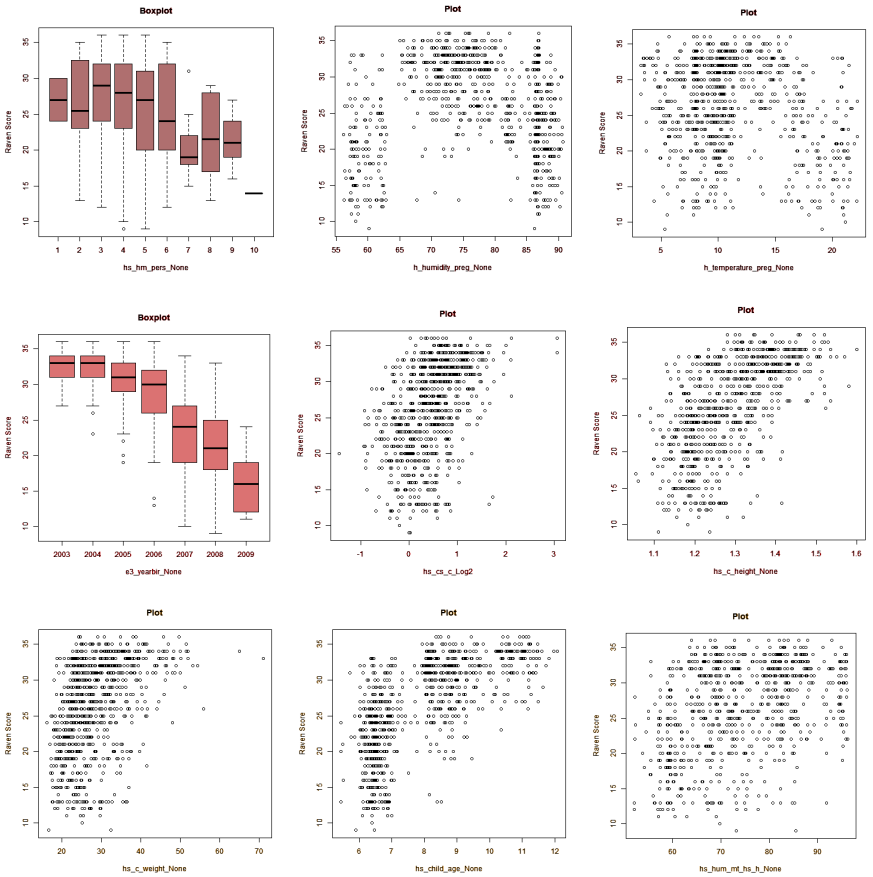
Here is the distribution of NA values. According to the histogram, most NA's are concentrated at a specific covariate. This is actually hs_wgtgain_None (maternal weight gain during pregnancy) which is NA for 11.8% of all rows. This is a significant portion of the 17.67% of rows with at least one NA. The rest affect < 5% of the rows.

The second plot shows the subset of covariates that allow for any NA's in at least one row, excluding any covariate columns which do not produce NA's. The top labels are the indices of covariates, indexed according to their arrangement in our created dataset, which contains all available covariates excluding any phenotype variables that are not hs_correct_raven. The left labels are the number of rows that have NA's in that specific arrangement, and the right labels are the number of covariates that have an NA value for those specific amounts of rows. Column 6 is hs_wgtgain_None; we can see that it contributes to most combinations of NA covariates.

Measuring Covariate Transformations:

While investigating the distributions of different covariates, we analysed the type of relationship between hs_correct_raven and variables of interest. Since a full model was being considered, all possible covariates excluding phenotype data were examined.

This process was such that we plotted/boxplotted each relevant continuous/categorical covariate against hs correct raven, and visually identified interesting patterns.



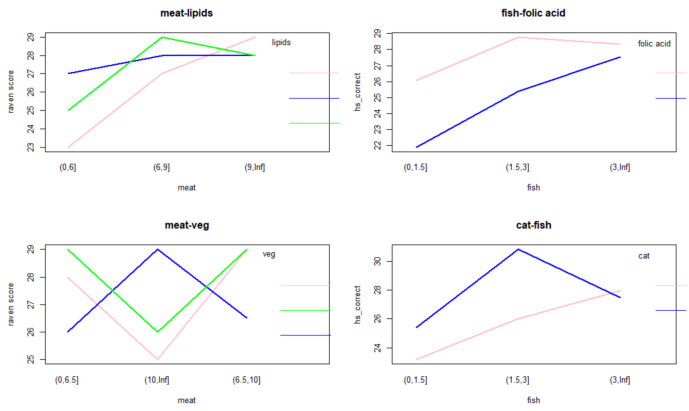
Out of the 196 non-lifestyle covariates graphed, we have 20 that displayed a notable pattern. Note that the lifestyle covariates were considered in the next section.

The first four are examples of covariates possibly having a quadratic relationship with raven score, what with the data roughly resembling a parabola.

The last five are examples of covariates that seem to have a possible logarithmic relation to the raven score.

Examining Possible Interactions and Covariate Relations:

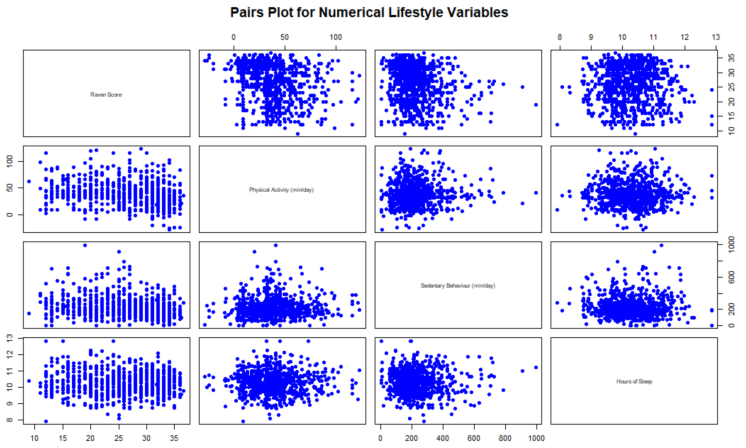
Plots for comparison of covariates, as well as interaction plots were examined to investigate relationships between covariates.



We find that, for example, in the fish-folic acid interaction plot, that the slopes between amount of fish consumed and raven score seem relatively constant for each category of folic acid, thus we have grounds to assume an interaction is not likely.

On the other hand, our meat-veg plot shows different parities of slopes between amount of mean consumed and raven score for each category of vegetables consumed, providing

evidence towards a possible interaction. This type of analysis was done with many more covariate pairs and two-way anova tables were also computed to further investigate the interactions.



To consider the relation between the numerical covariates under lifestyle, as well as their relation to raven score, pairs plots were examined.

We see here that all the plots show some level of randomness, especially the ones between the covariates, implying little to no correlation and no transformative relation that would imply an interaction.

Methods

General Strategy:

We chose to fit 5 different models to answer our research question. The models were separated as follows:

1. Model with all possible covariates under the lifestyle family (Lifestyle Model)
2. Model with all possible covariates under lifestyle and postnatal (Postnatal Model)
3. Model with all possible covariates under lifestyle and prenatal (Prenatal Model)
4. Model with all possible covariates in the dataset (Full Model)
5. Model with a single constant (Degenerate Model)

To choose the best model amongst the 5 different models described we observed metrics such as the mean square prediction error, R^2, adjusted R^2, AIC and BIC for a holistic comparison. Before we could run predictions on our models, we needed to perform feature selection and look for any possible interaction associated with our covariates of interest. To perform feature selection and avoid post selection bias we split our data into three parts.

Our data was split into 3 datasets: Train (40%), Validation (40%), and Test (20%). Feature selection was performed using the train dataset, whereas the validation dataset was used to fit models. We further implemented cross validation into the validation dataset to avoid any overfitting and to allow comparisons to be made between potential intermediary models. Prediction accuracy was tested only on the test data set to produce a final account of strength of model.

Data Cleaning:

A necessary step before the performance of feature selection was to address the missing data. We started off by assuming that there was no measurement error in the data collection, since we are not in contact with the team that performed data collection. We also assumed that data was formatted correctly. Hence, the only data points that require cleaning are ones with NA's. We decided upon three standard strategies dealing with NA's; list-wise deletion, mean/mode imputation and multiple imputation.

According to literature, MICE (Multiple Imputation with Chained Equations) requires the assumption of MAR (Missing at Random), that the NA's in a certain covariate may only be dependent on observed data, not unobserved data. We analysed the data to check if this assumption can be made, and then proceeded with multiple imputation as our primary cleaning method. Our algorithm for checking the MAR assumption is for each covariate with some NA's, convert it into a binary (1:NA, 0: o/w) and regressing this on the rest of the data and extract p-values of no association with the model. If these p-values are below our threshold (0.15), we can conclude MAR. More is explained in the data cleaning chapter of the appendix.

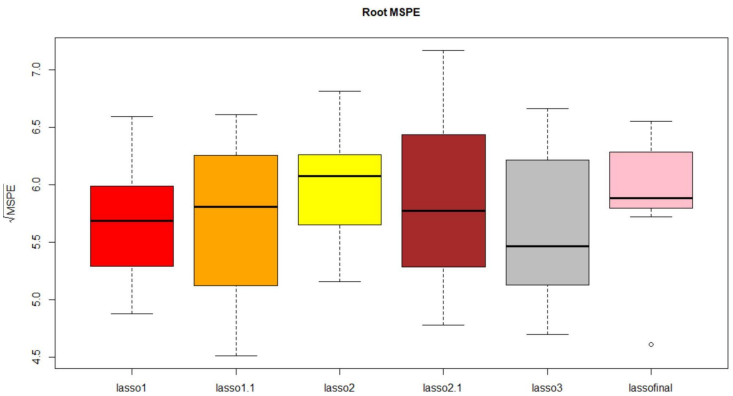
From our EDA, we encountered hs_wgtgain_None; it has a very large and divergent fraction of its data as NA (11.8%). Removing this covariate would introduce bias in our sample, since while it contributes most to the NA values, the covariate still has 88% of its data intact. Since so much data is present, it is highly likely that there exist dependencies on the probability of an NA in hs_wgtgain_None and the values of some other covariates. Thus, we will not remove the column. Since all other covariates have much lower ratios of NA values and have their NA's follow an insignificant distribution, it is safe to continue with an imputation method.

Feature Selection:

In addition to covariate-relation plots as discussed in EDA, we took advantage of stepwise, backward and forward selection, as well as LASSO to narrow down our covariates for our different models. By eliminating features, we hoped to improve the performance of our model as well as reduce potential multicollinearity between covariates.

We found that between multiple different selection methods, models containing variables screened by LASSO performed the best in terms of low prediction error. This was especially true for our lifestyle, postnatal and prenatal models.

Under the consideration of interaction terms for particular models, the volume of data played an important role in informing decisions. Since the available amount of data couldn't facilitate the consideration of all possible interactions. Thus, we took an intuitive approach, performing much of the interaction examination we did in EDA on covariate pairs that could be plausible considering previous knowledge and literature. The general strategy involved considering the p values produced by two-way anova and interpreting the graphs accordingly.



Models were then fit with varying interaction terms and compared to those with just main effects, as shown in the following plot.

It describes the distribution of root mean square prediction errors for each fold during k-fold cross validation for potential lifestyle models.

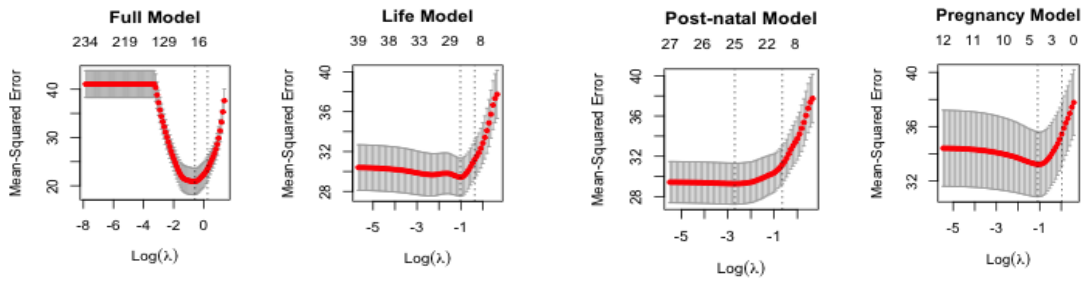
Note that these models included interactions either during variable selection or after selection was performed on main effects. Models 2, 2.1, 3 and final pertain to models including interactions.

We see that generally the models including interactions perform just as well or even worse than our main effects model in terms of prediction errors, which gave us grounds to believe interactions weren't wise to include.

Furthermore, upon inspection of the models where variable selection was performed on main effects and interactions, it seemed that most of the time interactions were being removed by LASSO or they remained and showed very high p-values of association under an OLS fit. We also encountered the issue of data volume and the data being unable to represent most of the categorical covariates' interaction terms.

Thus interactions were dropped in our final models.

We do this for each subset of reduced covariates we have and get four models for full, life, postnatal and pregnancy. Once the 5 different models were analysed with the various different feature selection algorithms, we chose the best model from Model-1,2,3 to compare against the best possible model that could be made, which is Model 4. To choose the best model, we observed the adjusted R^2,AIC,BIC and MSPE.



Out of all the different feature selection algorithms, LASSO has done the best for all 5 of our models. The above graph is a demonstration of our lambda values generated. These reduced covariates are used as the new set of covariates for training.

Selection for the Full Model:

Since the Full Model required the assessment of 237 covariates, it would be unwise to perform automatic selection on a model with all possible covariates and interactions. This is especially true since our EDA highlighted many quadratic and logarithmic relationships in between covariates and the outcome. We performed two types of transformations, pertaining to Models 2, 3 and 4; we used scatterplots from our EDA to subjectively choose quadratic/logarithmic transformations for covariates and employed them on Model 3.

For Model 2, instead of only applying specific transformations to some covariates, we applied a quadratic to all covariates and added them to main effects. The rationale behind this

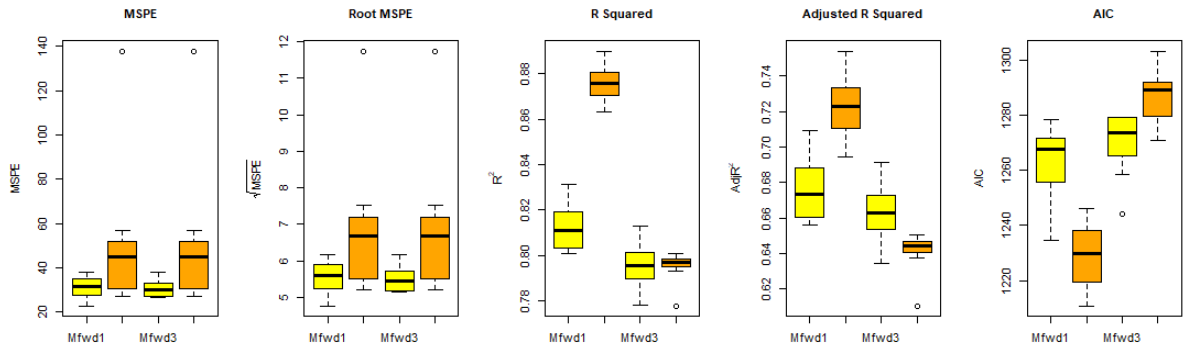
was the remark that all non-linear relationships can be approximated by a quadratic relationship. This implied the inclusion of a quadratic term for all covariates, hence removing the chances that we miss a nonlinear relationship in our model.

For Model 4, we split all covariates into 11 buckets s.t covariates in each bucket are subjectively related (descriptions have implied interaction, ie. 'Calcium concentration in mother's blood' and 'Calcium concentration in child's blood). Then we regress all buckets individually (main effects and all interactions) on the outcome and extract all interaction terms with a p-value < 0.1, implying that this specific interaction and the outcome have some statistically significant relationship. These terms act as the interaction terms in Model 4, along with main effects. This model is significant since the buckets rule out 1000's of interactions that have been quickly, subjectively been calculated as unrelated due to their holistic nature. It is the only efficient method of adding interaction terms to a model that can have ~ 56'000 terms.

Modelling:

After reducing our covariate set during feature selection, we proceeded to fit our models under OLS. We took this approach to ensure our estimates were unbiased and that we could produce interpretable prediction metrics. Our breadth of knowledge in OLS estimators as compared to LASSO estimates also contributed to this decision to guarantee conclusions and investigations of models were being done in an informed manner.

For the full model, four forward-select models were generated alongside our LASSO model, to judge its behaviour alongside LASSO. Model 1 regressed the outcome on all 237 main effects. Model 2 regressed on all main effects and on the quadratic transformations of all main effects. Model 3 regressed on the main effects, but with the appropriate transformations identified in the EDA. Model 4 regressed on main effects plus identified full-model interactions identified in the variable selection section. The models were trained on the training data subset, and their summaries were compared on the validation data subset. These boxplots come from the MSPE, R2, Adj R2 and AIC values from each fold of our k-fold cross validation, k=7:



All models have a similar rmspe, the range of median(rmspe) being [5.596, 6.686]. Model 2 succeed in every other metric by a significant margin; it captures a median R2 of 0.876, a median adjusted R2 of 0.723, and a median AIC of 1229.894. Since our R2 and adjusted R2 both imply a large capture of variability, this implies that the penalty from more covariates in model 2 is weaker than the variability captured by those covariates. Hence, since model 2

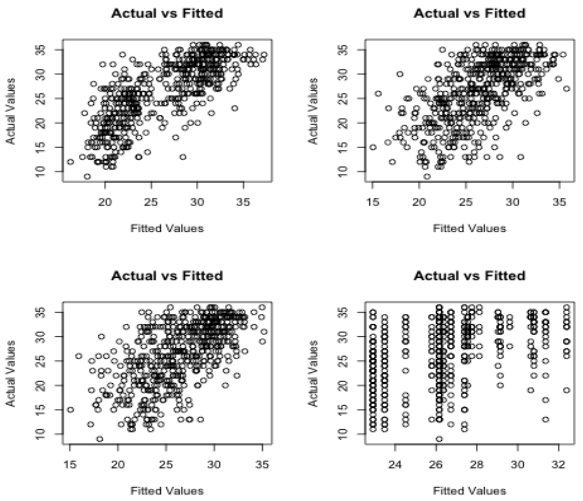
contains the quadratic main effects, we can conclude that most of the covariates hold a quadratic component in their relationship with Raven score.

After choosing Model 2, we tested it on our testing data by predicting outcomes and calculating an RMSPE value of 4.914. This implies that any predicted Raven Score coming from Model 2 may be off by a score of 4.914. The LASSO/OLS model on main effects has a test RMSPE of 4.2862. The primary goal of the full model is to predict outcome values as well as possible. Hence, we proceed with the LASSO/OLS model as our final full model.

Diagnosing Assumptions:

Validating assumptions are an important part of modelling. The major assumptions we diagnose are Linearity, Independence, Normality, and Equal Variance.

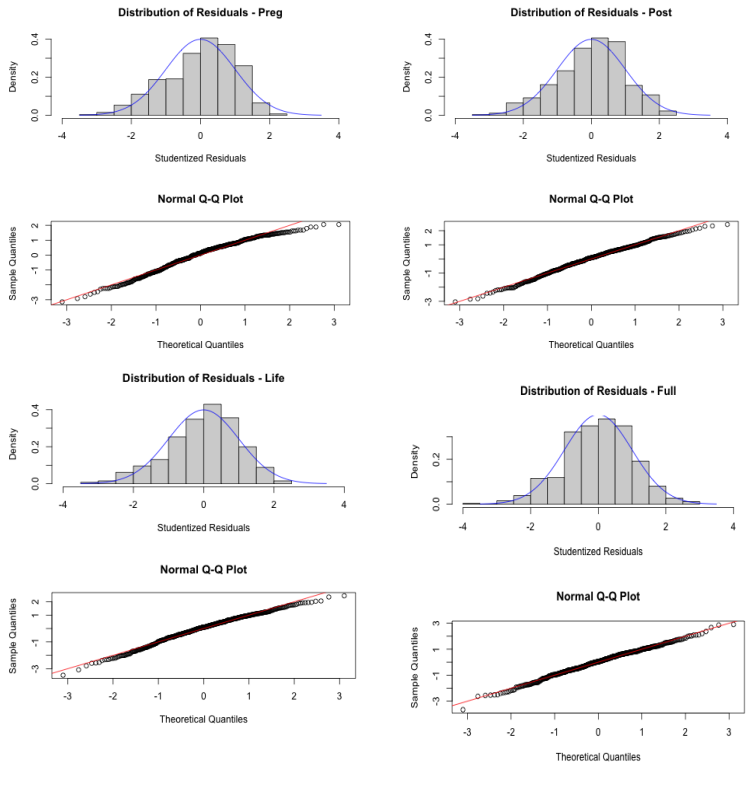
Linearity:



The plots are for Full (top-right), Lifestyle (top-left), Post-natal (bottom-right) and Pregnancy (bottom left).

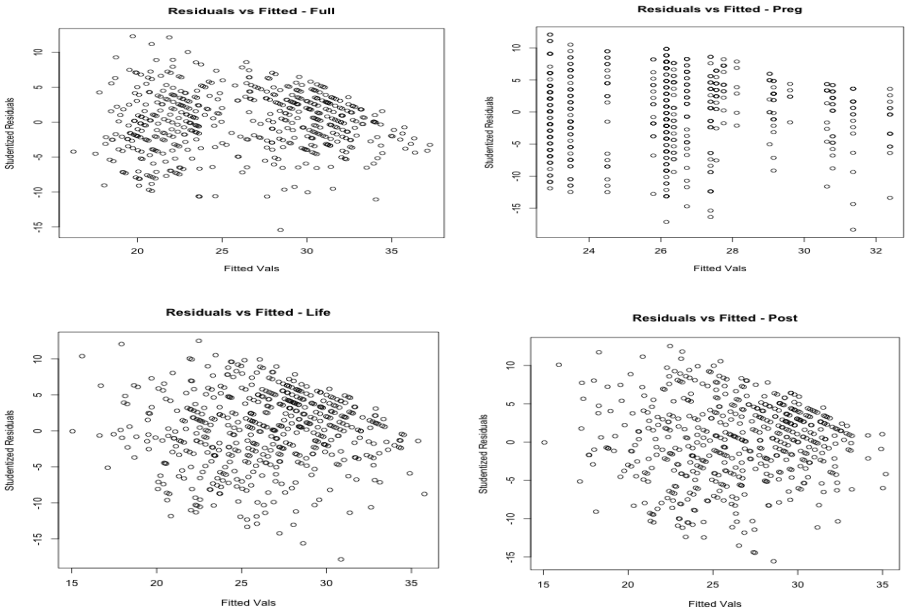
We can see the points are distributed reasonably symmetrically across the diagonals for all our four final models (excluding the degenerate model). Hence our assumption of linearity in the models is satisfied.

Normality:



We can see the distribution of residuals approximately outlining the normal distribution curve and falling on the straight line in the qq-plot apart from at the ends. This verifies our assumption on the normality of all our four models.

Homoscedasticity:



In all four plots of the studentized residuals vs fitted values, we do not detect any conspicuous patterns. Hence as we can see it looks like the residual terms are distributed with equal variance, and our assumption of homoscedasticity is justified.

Independence:

The data was obtained from a study conducted in Europe. Thus we had no control over the collection or the methods used when collecting the given data. As the study was published on a well reputed research site, it is safe to assume that data was collected using good practice, therefore we can assume independence.

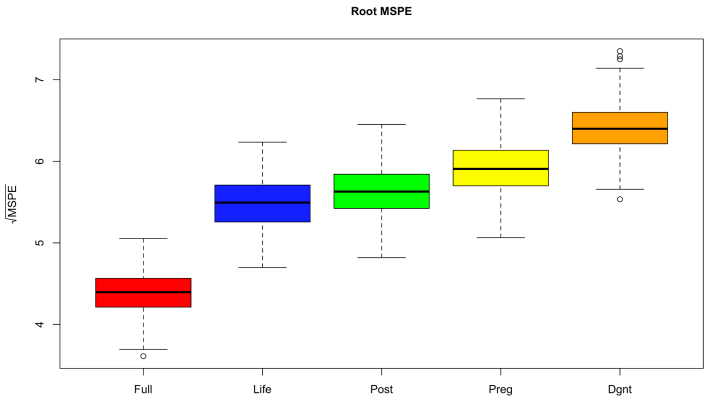
Thus we have all our assumptions satisfied and our use of linear models is justified.

Results

Table of Model Metrics:

	AIC	BIC	R^2	Adj R^2	MSPE	Root MSPE	No. of covariates
Full	3023.990	3117.659	0.571	0.554	18.372	4.286	93 covariates
Lifestyle	3253.822	3402.841	0.366	0.323	25.257	5.026	35 covariates
Post	3272.154	3459.491	0.366	0.310	26.056	5.105	27 covariates
Preg	3341.672	3371.476	0.165	0.157	32.454	5.697	10 covariates
Worst	3425.72	3434.235	0	0	37.595	6.131	0 covariates

Boxplots of root mean square prediction errors for each fold of cross validation:



Discussion:

Implications of Results:

Our final results show that the best possible lifestyle model we could fit can predict the raven score for a child to within 5 points, which is only slightly higher than that of our "best" model, which can predict it to within 4.28 points. However, as the boxplot displays, the general distribution of root mean square prediction errors for our life model shows that it tends towards higher errors of prediction than the full model and seems equally as far from our worst model as our best model. The adjusted R^2 and AIC values corroborate the mediocrity of the lifestyle model as it has a higher AIC value than the full model despite it having much fewer covariates.

This implies that lifestyle covariates don't contribute very much to the predictive attribute of the full model and ultimately a mother's non-genetic habits and lifestyle behaviour isn't a very good predictor of her child's IQ. On the other hand, we see that whatever source of prediction lifestyle covariates do contribute towards raven scores, it seems that postnatal behaviour has a greater effect than prenatal behaviour, since it has generally a lower distribution of root MSPE values and its adjusted R^2 and AIC is very close to that of lifestyle.

Limitations:

The above results could have been affected by limitations such as the fact that the postnatal period of the lifestyle family contains more covariates than the prenatal period and that prenatal only contains categorical covariates. Additionally, we fail to account for the genetic effect on a child's intelligence from the mother, which definitely will contribute towards the variability of the raven score. Another limitation towards our result was that we couldn't fully rely on the raven score to accurately depict a child's IQ. We also have 17% of rows containing some NA value, 12% of which comes from a single covariate. Since we did not learn how to handle NA's in our class we had to resort to online resources and dealt with them as best we could. Lack of data volume was also an issue, as it affected model fitting and validation, as well as difficulty in accounting for many covariates or interactions. Since we are not in contact with the data collection team, we are limited to assuming no measurement/formatting errors.