**Part A. Q2**

**Methods**

A multiple linear regression model is used to investigate the estimated relationship between MSEL early learning composite score and folic acid intake during the first trimester of pregnancy. A base univariate regression model of MSEL score on folic acid intake is subsequently fitted with each one of the potential confounders, namely, those that are statistically significantly associated with both MSEL score and acid intake, according to the p-values in Table 1. A 10% change in the beta estimate for folic acid is set a priori to determine confounders. After adjusting for confounders, any potential effect modifier that interact with the association between MSEL and folic acid intake are added to the updated model to test for statistical significance. Stratified analysis by stratum of the significant effect modifier(s) is then conducted.

Before performing further analyses, variable year of birth is standardized by rescaling it with a mean of 0 and standard deviation of 1 to account for the small range of the original variable relative to its mean and any resulting potential loss of accuracy.

To assess the assumptions of linear regression and complete model diagnostics, advanced residual analysis is used to generate three scatterplots of estimated residuals versus predicted values, jackknife residuals versus predicted values, and studentized residuals versus predicted values. Tolerance and variance inflation are checked to account for any potential issues with collinearity. To identify outliers and influential points, jackknife residual is tested for its statistical significance using an adjusted α = 0.00005767012for multiple comparisons with n = 867, k = 3, and df = 862 and statistics such as leverage, cook’s D, DFBETAS, and BFFIT are also assessed. Model is then readjusted as needed.

**Results**

With a base model of MSEL score on folic acid intake, two potential confounders, child’s year of birth and maximum education at home, are added separately to create two additional models to determine their effect on the beta estimate for folic acid in the base model. Maximum education at home is kept in the model as it changes the beta estimate for folic acid in the base model by more than 10%. Then, with a new base model of MSEL score on folic acid and maximum education at home, child’s year of birth is added to the model to test its confounding effects. Child’s year of birth is subsequently kept in the model since it changes the beta estimate by more than 10%. Thus, a regression model of MSEL score on folic acid intake, adjusting for maximum education at home and child’s year of birth, is used for further analysis.

This model is then tested for effect modification by testing the statistical significance of interaction terms between folic acid intake and binary variables in the study, namely, maximum education at home, child’s sex, and child’s genotype. No effect modification by maximum education at home is found since the interaction term is not statistically significant (p=0.2699). Child’s sex is also not a statistically significant effect modifier (p=0.8971). However, child’s genotype is a significant effect modifier (p=0.0454). As such, two separate models for genotypes CC/CT and TT are developed for stratified analysis.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Obs.Statistics** | **Jackknife Residual** | **Leverage** | **Cook’s D** | **DFBETA** | **BFFIT** |
| **367** | 4.7588(p<.00001)\* | 0.0033 | 0.019 | 0.0814\* | -0.2758\* |
| **725** | 2.0805(p=.037729) | 0.0168\* | 0.018 | -0.2510\* | -0.2724\* |

Table 2: Residual Analysis[[1]](#footnote-1) n=867, k=3, df=862, adjusted α=0.00005767012

Further model diagnostics shows no issues regarding collinearity since tolerance is > 0.1 and variance inflation <10. Residual analysis results are presented in Table 2 above for potentially problematic observations 367 and 725 as identified from Cook’s D arrow plots. Jackknife residual significance, leverage, cook’s D, DFBETAS, and BFFIT are tested for their effects. Problematic entries include a statistically significant jackknife residual (p<.00001) for observation 367, indicating that it’s a significant outlier; the leverage value for observation 725 is greater than 2(k+1)/n, indicating large leverage and potentially influential; the absolute values of DFBETA values for both observations 367 and 725 are greater than , indicating large influence on slope estimate; and lastly, BFFIT values for both observations 367 and 725 are greater than , indicating influence on the fit of the regression line. As a result, the model is then refit after removing these two problematic observations. The new model has a beta estimate for folic acid intake has a 7% difference from that of the original model.

The assumptions of linear regression are assessed before and after refitting the stratified models by removing the influential points. The refitted models satisfy the assumptions better than the original model. Specifically, linearity is satisfied because the average residual is approximately 0 across the range of predicted values and evenly spread around the reference line. Independence is satisfied by assumption of data structure and the description of data that each subject is independent from one another with regards to their MSEL score are. Normality is approximately satisfied. Even though the Shapiro-Wilk null hypothesis of normality is rejected (p<0.0001), as were the other three tests of normality, the histogram shows a bell-shaped curve, and the normal probability plot follows an approximately straight line. Furthermore, since mean and median of the residuals differ by less than 20% of 1SD, skewness and kurtosis are both <|1|, normality can be assumed overall. And lastly, homoscedasticity is approximately satisfied because the variance of the residuals is approximately constant across the range of predicted values. Overall, the assumptions of linear regression are satisfied, and no additional variable transformations are performed.

For the final model refitted by removing influential points from the original model, stratified by child’s genotype, there is a slight improvement on model r-squared from 0.0504 to 0.1196. This r-square means that folic acid intake, maximum education at home, and child’s birthyear taken together explain 11.96% of the total variation in MSEL score. The global F-test for each model is statistically significant (p<0.001), as well as each individual parameter estimate.

The following Table 3 summarizes these results, truncated to three decimal places. Adjusting for maximum education at home and child’s birthyear, for a child with CC/CT genotype, 1𝜇g of folic acid intake is associated with 0.005 increase in MSEL score (p=0.215), while for a child with TT genotype, 1𝜇g of folic acid intake is significantly associated with 0.017 increase in MSEL score (p=0.0114).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ParametersStatistics** | **Estimate** | **Standard Err.** | **Pr > |t|** | **Pr > F** | **Model R2** |
| **Intercept** | 65.376 | 2.065 | <.0001 | <.0001 | .0557 |
| **Folic acid intake** | 0.005 | 0.002 | 0.215 |  |  |
| **Maximum Education** | 9.597 | 1.883 | 0.0002 |  |  |
| **Child’s birthyear** | 3.405 | 0.921 | 0.0006 |  |  |

Table 3.1: Summary Results from Final Regression Model, Genotype = CC/CT

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ParametersStatistics** | **Estimate** | **Standard Err.** | **Pr > |t|** | **Pr > F** | **Model R2** |
| **Intercept** | 58.180 | 5.75 | <.0001 | 0.0045 | 0.1196 |
| **Folic acid intake** | 0.017 | 0.006 | 0.0114 |  |  |
| **Maximum Education** | 12.357 | 5.853 | 0.0372 |  |  |
| **Child’s birthyear** | 2.049 | 3.032 | 0.5006 |  |  |

Table 3.2: Summary Results from Final Regression Model, Genotype = TT

**Code**

%***reg***(folicacid\_t1);

%***reg***(folicacid\_t1 yob);

%***reg***(folicacid\_t1 maxedu\_bs);

%***reg***(folicacid\_t1 maxedu\_bs yob);

%***reg***(folicacid\_t1 maxedu\_bs prepregbmi\_cat);

%***reg***(folicacid\_t1 maxedu\_bs yob prepregbmi\_cat);

**data** charge;

set charge;

zyob = yob;

**run**;

**proc** **standard** data = charge mean = **0** std = **1** out=charge;

var zyob;

**run**;

**proc** **means** data = charge;

var zyob;

**run**;

**proc** **glm** data = charge;

class maxedu\_bs (ref = "Some college or less");

model msel\_elcs = folicacid\_t1 maxedu\_bs zyob folicacid\_t1\*maxedu\_bs / solution clparm;

title "Check if educ is effect modifier";

**run**;

**proc** **glm** data = charge2;

class maxedu\_bs (ref = "Some college or less");

class mthfr677 (ref = "CC or CT genotype");

model msel\_elcs = folicacid\_t1 maxedu\_bs zyob mthfr677 folicacid\_t1\*mthfr677 / solution clparm;

title "Check if genotype is effect modifier";

**run**;

**proc** **glm** data = charge2;

class maxedu\_bs (ref = "Some college or less");

class child\_male (ref = "Female");

model msel\_elcs = folicacid\_t1 maxedu\_bs zyob child\_male folicacid\_t1\*child\_male / solution clparm;

title "Check if sex is effect modifier";

**run**;

**proc** **sort** data = charge2;

by mthfr677;

**run**;

**proc** **glm** data = charge2;

model msel\_elcs = folicacid\_t1 maxedu\_bs zyob / solution clparm;

title "stratified regression analysis by genotype";

by mthfr677;

**run**;

**proc** **reg** data = charge;

model msel\_elcs = folicacid\_t1 maxedu\_bs zyob / r influence tol vif;

output out = out r = resid p = pred student = student rstudent = jackknife;

title "check collinearity and model assumption";

**run**;

**proc** **sgplot** data=out;

scatter y=resid x=pred ;

refline **0** / axis = y;

title "resd vs pred";

**run**;

**proc** **sgplot** data=out;

scatter y=student x=pred ;

refline **0** / axis = y;

title "student vs pred";

**run**;

**proc** **sgplot** data=out;

scatter y=jackknife x=pred ;

refline **0** / axis = y;

title "jackknife vs pred";

**run**;

**proc** **univariate** data = out normal plot;

var resid;

**run**;

**data** charge2;

set charge;

if subject\_id = **4179** then delete;

if subject\_id = **9010** then delete;

**run**;

**proc** **reg** data = charge2;

model msel\_elcs = folicacid\_t1 maxedu\_bs zyob / r influence tol vif;

output out = out r = resid p = pred student = student rstudent = jackknife;

title "model";

**run**;

**proc** **sgplot** data=out;

scatter y=resid x=pred ;

refline **0** / axis = y;

title "resd vs pred";

**run**;

**proc** **sgplot** data=out;

scatter y=student x=pred ;

refline **0** / axis = y;

title "student vs pred";

**run**;

**proc** **sgplot** data=out;

scatter y=jackknife x=pred ;

refline **0** / axis = y;

title "jackknife vs pred";

**run**;

**proc** **univariate** data = out normal plot;

var resid;

**run**;

**Output**

**Table

Description automatically generatedGraphical user interface, text, application

Description automatically generated**

**Graphical user interface, application

Description automatically generated**

**Table

Description automatically generated**

**Graphical user interface, text, application

Description automatically generated**

**Graphical user interface, table

Description automatically generated with medium confidence**

**Text

Description automatically generated with medium confidence**

**Graphical user interface, table

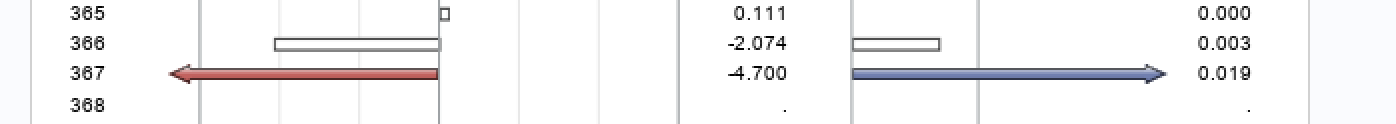
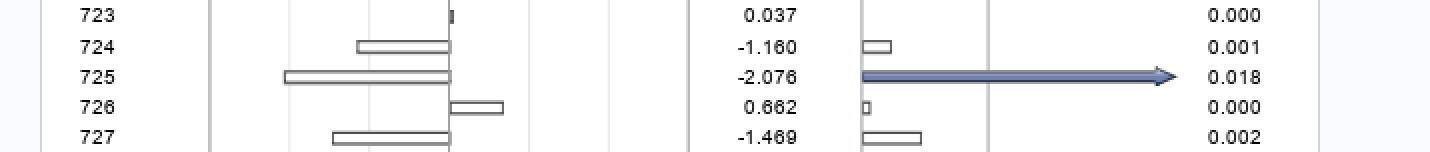
Description automatically generated**

**Graphical user interface, application, table

Description automatically generated**

**Graphical user interface, application, table

Description automatically generated**

****

**Diagram

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Chart, line chart

Description automatically generated**

**Graphical user interface, application, table

Description automatically generated**

**Diagram

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Table

Description automatically generated**

**Table

Description automatically generated**

**Chart

Description automatically generated**

1. Problematic entries are marked with an asterisk. Results obtained from SAS output, see appendix for details. [↑](#footnote-ref-1)