**Part B. Q1**

**Methods**

To develop a prediction model for birthweight in the subset of non-smoking mothers from the National Maternal and Infant Health Survey dataset, stepwise selection method with an entry level selection value and a stay selection value of 0.15 is used, with additional inclusion of higher order terms such as quadratic terms and interaction terms in the initial maximum model. After reaching a satisfactory model, model diagnostics are subsequently performed to assess for model fit and outliers. Including correlation analysis, model collinearity and model regression assumption analysis.

**Results**

After deleting missing values and centering the numeric predictor variables, collinearity is checked using correlation analysis. There is high correlation between mother’s pre-pregnancy weight and mother’s pre-pregnancy BMI category (r=0.81). No other issues of collinearity are found. Given that weight and BMI essentially provide the same information, weight is dropped from further analysis because BMI tends to have fewer missing data and because BMI is an important effect modifier of interest in the primary analysis. The squared term for weight is subsequently dropped as a result.

Stepwise selection is used to produce a prediction model. Ten variables included in the resulting model (r-squared=13.31%, p<.0001) are mother’s age at delivery (p=0.7428), weight gain during pregnancy (p=0.0366), duration of pregnancy (<.0001), mother’s age at delivery squared (p=0.0092), duration of pregnancy squared (p=0.0050), ever taken a prenatal class (p=0.5610), mother’s pre-pregnancy BMI category (p=0.0693), infant’s sex (p=0.0002), along with two interaction terms, namely weight gain during pregnancy\*infant’s sex (p=0.0820), and mother’s age at delivery\*mother’s pre-pregnancy BMI category (p=0.0452). No additional adjustment is needed as all the lower order terms contained in the higher order terms are all present. Interaction terms are kept in the model for further analyses. Collinearity analysis shows no apparent issues since all values of tolerance are greater than 0.1 and variance inflation values are less than 10.

Even though there exist various potentially influential points indicated by the jackknife residual arrow plots, the model mostly satisfies the assumptions of linear regression, so no further variable adjustment is processed. In particular, 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 birth weight measurement. 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 an approximately 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 reasonably small, 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 and data manipulation are performed.

**Code**

libname library "X:\";

**data** b;

set library.nmihs88\_train\_278;

if smoking = **0**;

junkid= **1**;

**run**;

**proc** **means** data = b NMISS N;

var momage educ prenatalClass smoking parity wtprepreg BMIprepregCat wtGain gestWeeks male bwt;

**run**;

**data** b;

set b;

if parity ne **.** AND bwt ne **.**;

**run**;

**proc** **means** data = b NMISS N;

var momage educ prenatalClass smoking parity wtprepreg BMIprepregCat wtGain gestWeeks male bwt;

**run**;

**proc** **means** data = b;

by junkid;

var momage wtprepreg wtgain gestweeks;

output out=sumstats mean=meanmomage meanwtprepreg meanwtgain meangestweeks;

**run**;

**data** b;

merge sumstats b;

by junkid;

momage = momage-meanmomage;

wtprepreg = wtprepreg - meanwtprepreg;

wtgain = wtgain - meanwtgain;

gestweeks = gestweeks - meangestweeks;

momage2 = momage\*\***2**;

wtprepreg2 = wtprepreg\*\***2**;

wtgain2 = wtgain\*\***2**;

gestweeks2 = gestweeks\*\***2**;

**run**;

**proc** **corr** data=b best=**5**;

var momage wtprepreg wtgain gestweeks momage2 wtprepreg2 wtgain2 gestweeks2 educ prenatalclass smoking parity bmiprepregcat male;

title "Pairwise correlations among X's: checking collinearity";

**run**;

**proc** **glmselect** data=b;

class educ;

class prenatalclass;

class parity;

class bmiprepregcat;

class male;

model bwt = momage wtgain gestweeks

momage2 wtgain2 gestweeks2

educ prenatalclass parity bmiprepregcat male

male\*momage male\*wtgain male\*gestweeks

bmiprepregcat\*momage bmiprepregcat\*wtgain bmiprepregcat

/ selection=stepwise details=all hierarchy=single showpvalues sle=**0.15** sls=**0.15** select=sl stop=sl;

title "Stepwise regression with interactions";

**run**;

**data** b;

set b;

wtgain\_male = wtgain\*male;

momage\_bmiprepregcat = momage\*bmiprepregcat;

**run**;

**proc** **reg** data = b;

model bwt = momAge wtGain gestWeeks momage2 gestweeks2 prenatalClass BMIprepregCat male wtgain\_male momage\_bmiprepregcat / tol vif;

title "check collinearity issues";

**run**;

**proc** **glm** data = b;

class prenatalclass (ref="0");

class bmiprepregcat (ref="0");

class male (ref="0");

model bwt = momAge wtGain gestWeeks momage2 gestweeks2 prenatalClass BMIprepregCat male wtgain\_male momage\_bmiprepregcat

/ solution clparm;

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

title "check model assumptions";

**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**

Graphical user interface, application, table

Description automatically generated with medium confidenceTable

Description automatically generated with medium confidenceTable

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generated

Table

Description automatically generatedTable

Description automatically generated

Table

Description automatically generated

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, line chart

Description automatically generated

Chart, diagram

Description automatically generatedGraphical user interface, chart, scatter chart

Description automatically generated

Table

Description automatically generatedChart, line chart

Description automatically generatedTable

Description automatically generated