/\*Applied Regression Group Project - Away-From-Home Meals and Body Mass Index in the NHANES Dataset\*/

/\* We are assessing the association between number of away-from-home meals and BMI while controlling for age, gender, race, education, income, physical activity, alcohol consumption, and history of blood pressure as potential confounders. Specifically, we hypothesize that an increase in the number of away-from-home meals is associated with an increase in BMI. We are using the 2015-2016 NHANES data as our data source for this research question.

/\*Data Import and Cleaning\*/

/\*While looking through the 2015-2016 NHANES data, we identified six data files that contained the appropriate variables that we needed for our hypothesis: BPQ\_I, DEMO\_I, DBQ\_I, BMX\_I, PAQ\_I, and ALQ\_I. We imported the XPT data files into SAS and created temporary files in the work directory using the LIBNAME statement.\*/

libname final xport "/home/jmc23920/sasuser.v94/DEMO\_I.XPT";

data DEMO\_I; set final.DEMO\_I;

run;

libname final xport "/home/jmc23920/sasuser.v94/BPQ\_I.XPT";

data BPQ\_I; set final.BPQ\_I;

run;

libname final xport "/home/jmc23920/sasuser.v94/DBQ\_I.XPT";

data DBQ\_I; set final.DBQ\_I;

run;

libname final xport "/home/jmc23920/sasuser.v94/BMX\_I.XPT";

data BMX\_I; set final.BMX\_I;

run;

libname final xport "/home/jmc23920/sasuser.v94/ALQ\_I.XPT";

data ALQ\_I; set final.ALQ\_I;

run;

libname final xport "/home/jmc23920/sasuser.v94/PAQ\_I.XPT";

data PAQ\_I; set final.PAQ\_I;

run;

/\*We then merged and sorted each of the six data files by a variable, SEQN, which is common in all six data files. SEQN is the respondent sequence number - an identifier for each respondent.\*/

proc sort data = work.DEMO\_I;

by SEQN;

run;

proc sort data = work.BPQ\_I;

by SEQN;

run;

proc sort data = work.DBQ\_I;

by SEQN;

run;

proc sort data = work.BMX\_I;

by SEQN;

run;

proc sort data = work.ALQ\_I;

by SEQN;

run;

proc sort data = work.PAQ\_I;

by SEQN;

run;

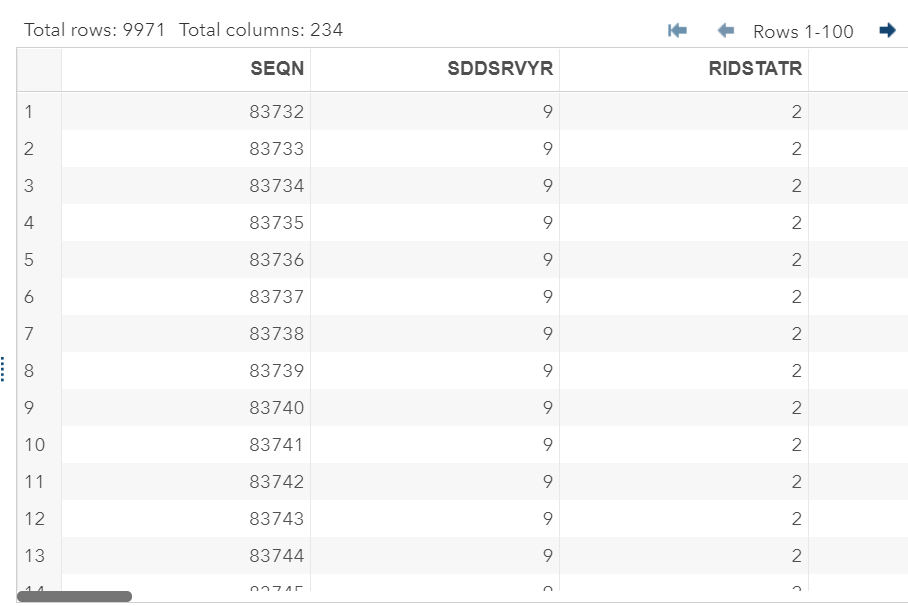
/\*Next, we merged the six data files using SEQN, and labeled our merged data file "complete."\*/

data complete;

merge work.DEMO\_I work.BPQ\_I work.DBQ\_I work.BMX\_I work.ALQ\_I work.PAQ\_I;

by SEQN;

run;



/\*Using the PROC CONTENTS statement, we checked the contents of our merged data file to familiarize ourselves with the data. Our merged data file contained 9,971 observations and 234 variables. Next, we created a new dataset containing only our needed variables:

SEQN (unique identifier for each respondent);

RIAGENDR (gender);

RIDAGEYR (age in years at screening);

RIDRETH1 (race/ethnicity);

INDFMPIR (ratio of family income to poverty);

WTINT2YR (full sample 2-year interview weight);

WTMEC2YR (full sample 2-year MEC exam weight);

SDMVPSU (pseudo-cluster/PSU, masked for confidentiality);

SDMVSTRA (pseudo-strata, masked for confidentiality);

BPQ020 (ever told that he/she had high blood pressure);

DBD895 (number of meals not prepared at home in the last week);

BMXBMI (BMI);

DMDEDUC2 (education level, adults 20+ only);

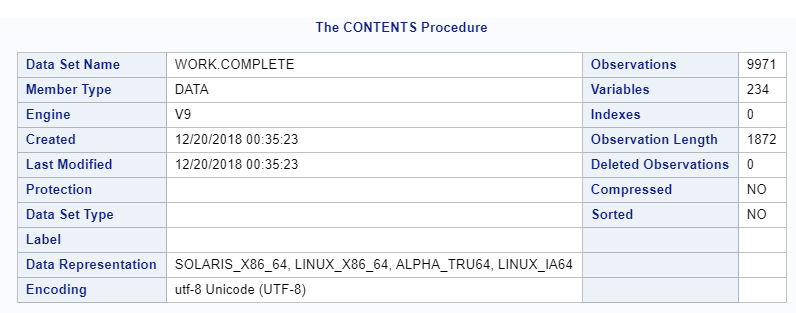
PAQ665 (moderate-intensity recreational activities in a given week); and

ALQ120Q (frequency of drinking alcohol in the past 12 months)

Keeping only these variables gave us 9,971 observations and 15 variables.\*/

proc contents data = complete;

run;



data final; set complete;

keep SEQN RIAGENDR RIDAGEYR RIDRETH1 INDFMPIR WTINT2YR WTMEC2YR SDMVPSU SDMVSTRA

DBD895 DMDEDUC2 BPQ020 BMXBMI ALQ120Q PAQ665;

run;



/\*We then ran PROC UNIVARIATE on our exposure outcome, DMD895, to examine the distribution of the individuals who provided the number of their meals not prepared at home in a given week. We only ran the functions on respondents aged between 18 and 65, since our research question aims to focus only on adults and since we plan on analyzing only individuals meeting this age criteria. We excluded individuals who answered "Don't Know" (coded as 9999). We also excluded individuals who specified a number more than 21 meals (coded as 5555) because the dataset did not provide a specific number for them. PROC UNIVARIATE showed us that the distribution of DMD895 was right skewed.\*/

proc univariate data = final;

var DBD895;

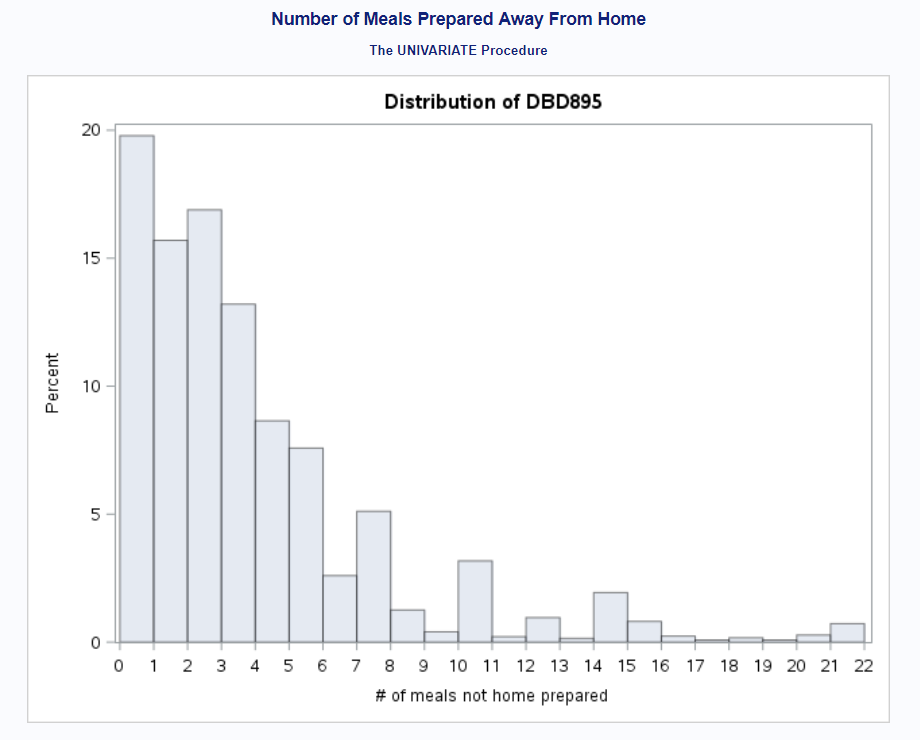
where (DBD895 < 22) and

(18 <= RIDAGEYR <= 65);

histogram / endpoints = 0 to 22 by 1;

title 'Number of Meals Prepared Away From Home';

run;



/\*We then ran PROC FREQ and PROC UNIVARIATE on our outcome variable, BMXBMI, to examine its distribution as well. Again, we only ran PROC UNIVARIATE on respondents aged between 18 and 65. PROC FREQ showed us that most individuals fell under the category of "Normal Weight." PROC UNIVARIATE showed us that the distribution of the responses was right skewed.\*/

proc format;

value bmicatf 1 = "underweight"

2 = "normal weight"

3 = "overweight"

4 = "obese";

run;

data final; set final;

if 0 < BMXBMI < 18.5 then BMICAT = 1;

else if 18.5 <= BMXBMI < 25 then BMICAT = 2;

else if 25 <= BMXBMI < 30 then BMICAT = 3;

else if BMXBMI >= 30 then BMICAT = 4;

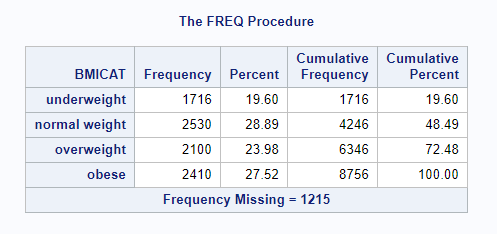
format BMICAT bmicatf.;

run;

proc freq data = final;

table bmicat;

run;



proc univariate data = final;

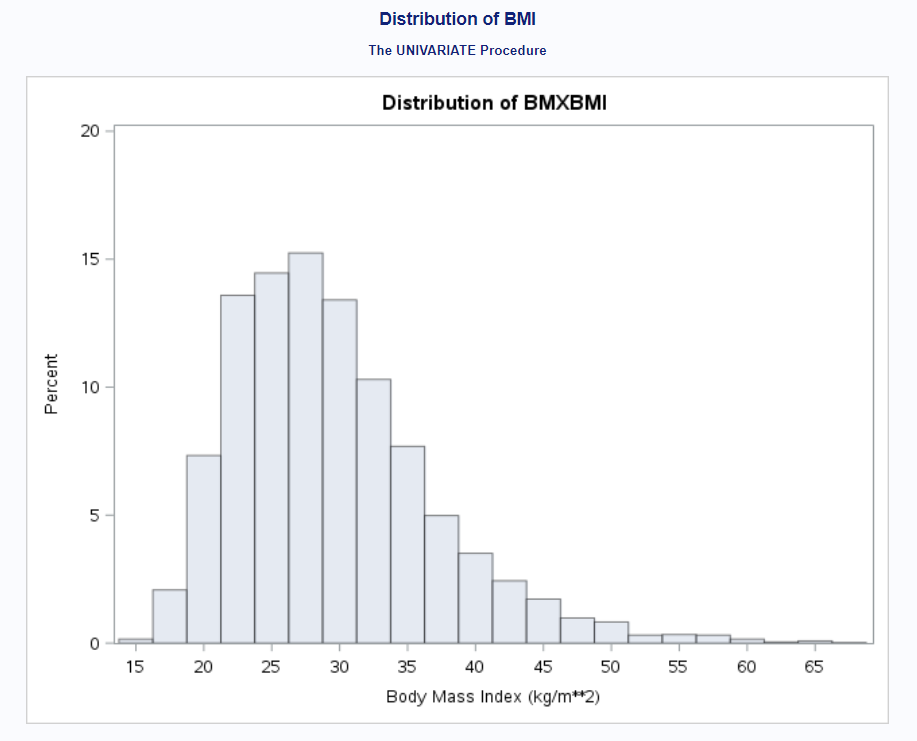
var BMXBMI;

where 18 <= RIDAGEYR <= 65;

histogram;

title 'Distribution of BMI';

run;



/\*Since the outcome variable was slightly right skewed, we chose to run a log transformation and called the newly created variable log\_bmi. We then ran another PROC UNIVARIATE function to check that our outcome variable was now normally distributed; based on the output, it seemed that the transformation worked.\*/

data final; set final;

log\_bmi = log(BMXBMI);

run;

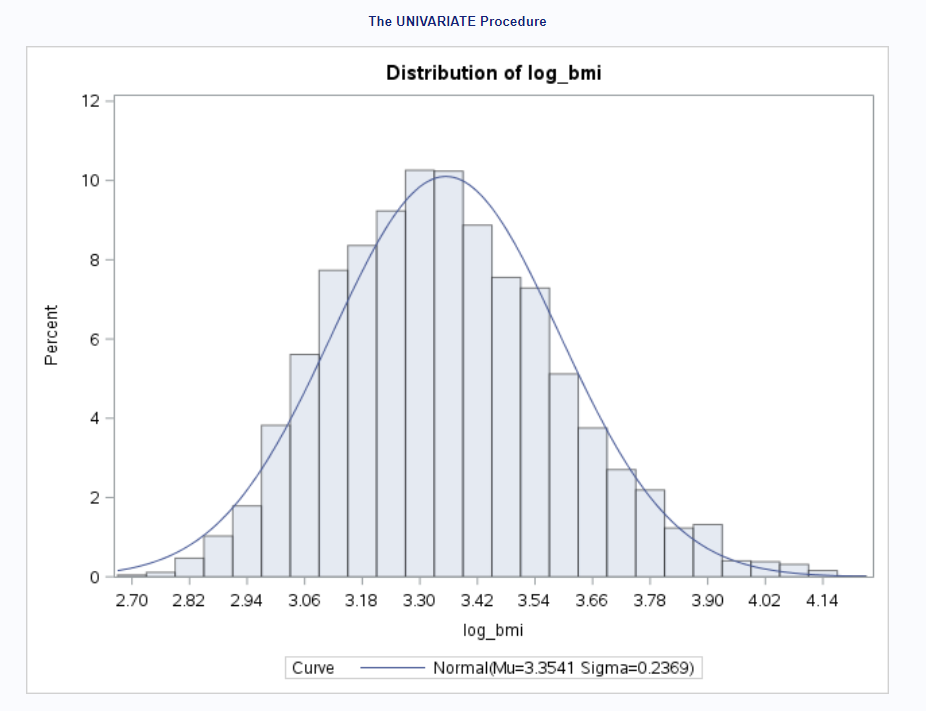
proc univariate data = final;

var log\_bmi;

where 18 <= RIDAGEYR <= 65;

histogram/normal;

run;



/\*We were then ready to visualize our data. We first created a scatterplot between our exposure variable (DBD895) and our outcome variable (log\_bmi). Our scatterplot did not tell us much, but from visual inspection, there appeared to be no association between our exposure and outcome variables.\*/

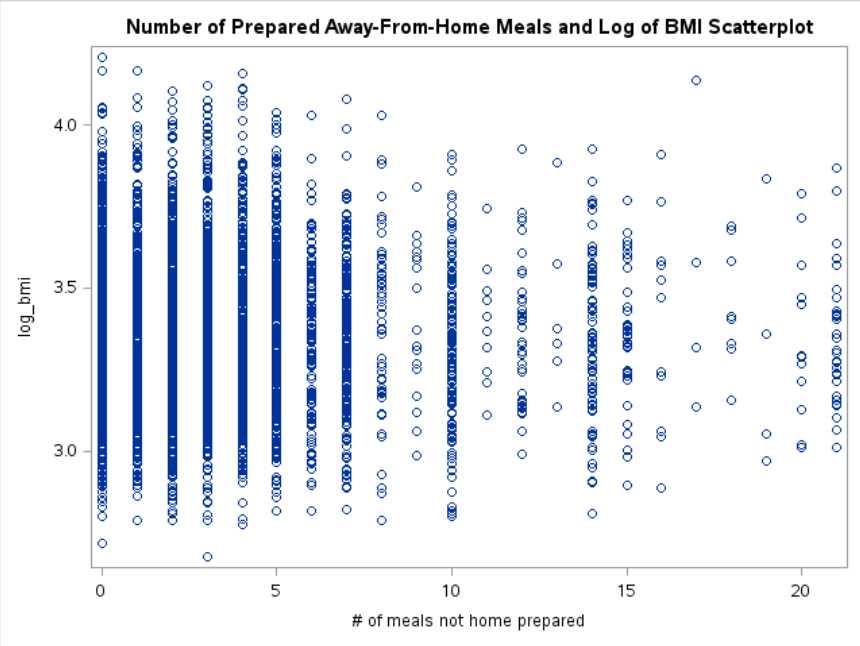
proc sgplot data = final;

scatter x = DBD895 y = log\_bmi;

where (DBD895 < 22) and

(18 <= RIDAGEYR <= 65);

run;



/\*Model Selection\*/

/\*We then ran a backward model selection process to determine what our best model would be, using an alpha level of 0.10. In running this model selection process, we made sure to account for the sampling weights, the data clustering, and the complex stratified study design. We also included individuals that met the age criteria; we excluded those that refused or answered "Don't Know" for the covariates. Our final model left us with 7 predictors.\*/

/\*Manual Backward Selection with SLS = 0.10\*/

/\*Step 1 -- All Covariates\*/

proc surveyreg data = final;

strata SDMVSTRA;

cluster SDMVPSU;

class RIAGENDR RIDRETH1 (ref = '3') BPQ020 DMDEDUC2 PAQ665;

model log\_bmi = DBD895 RIAGENDR RIDAGEYR RIDRETH1 INDFMPIR BPQ020

DMDEDUC2 ALQ120Q PAQ665 DBD895\*RIAGENDR / solution;

weight WTMEC2YR;

where 18 < RIDAGEYR <= 65 and

BPQ020 NOT in (7, 9) and

DBD895 < 22 and

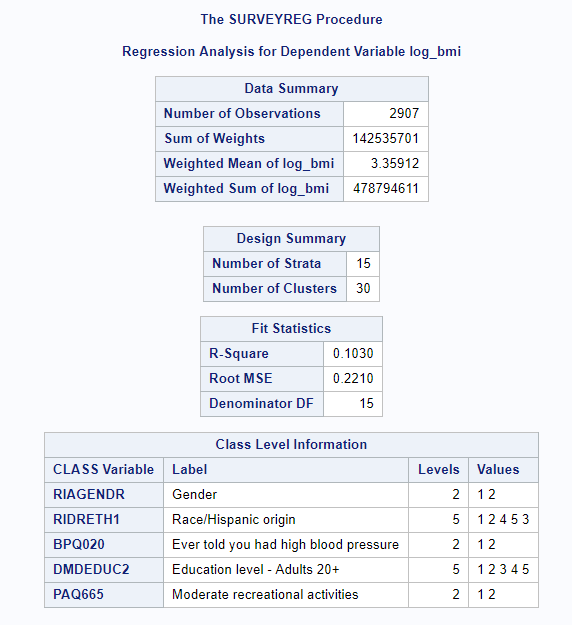
DMDEDUC2 NOT in (7, 9) and

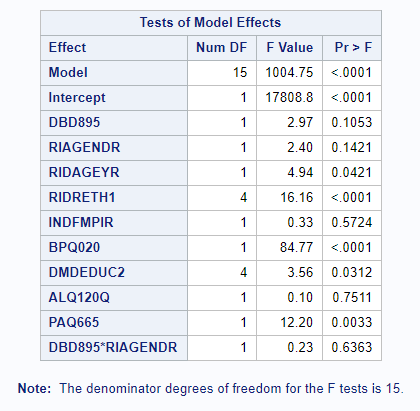
ALQ120Q NOT in (777, 999) and

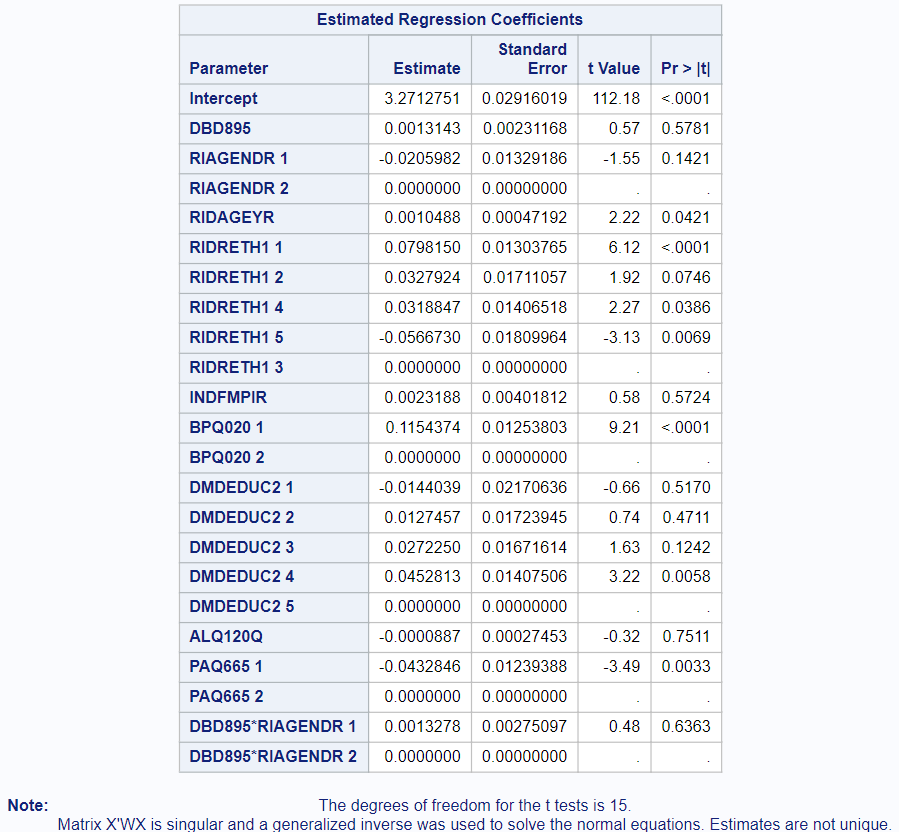
PAQ665 NOT in (7, 9) and

INDFMPIR NOT in (77, 99);

run;







/\*Step 2 -- Dropped ALQ120Q (p = 0.7511)\*/

proc surveyreg data = final;

strata SDMVSTRA;

cluster SDMVPSU;

class RIAGENDR RIDRETH1 (ref = '3') BPQ020 DMDEDUC2 PAQ665;

model log\_bmi = DBD895 RIAGENDR RIDAGEYR RIDRETH1 INDFMPIR BPQ020

DMDEDUC2 PAQ665 DBD895\*RIAGENDR / solution;

weight WTMEC2YR;

where 18 < RIDAGEYR <= 65 and

BPQ020 NOT in (7, 9) and

DBD895 < 22 and

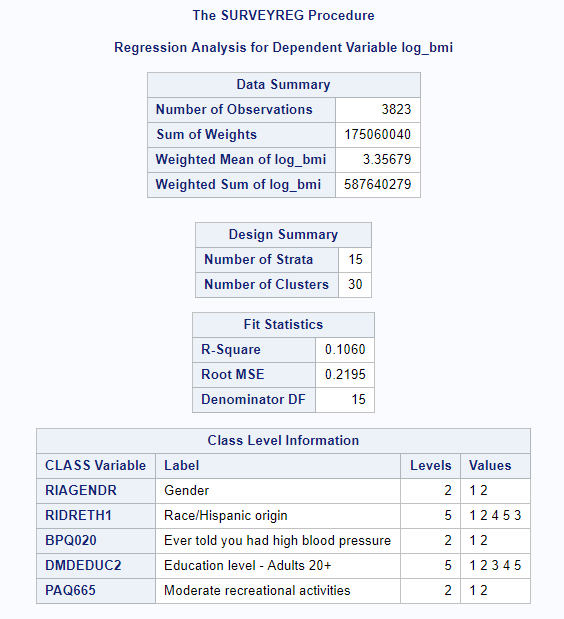
DMDEDUC2 NOT in (7, 9) and

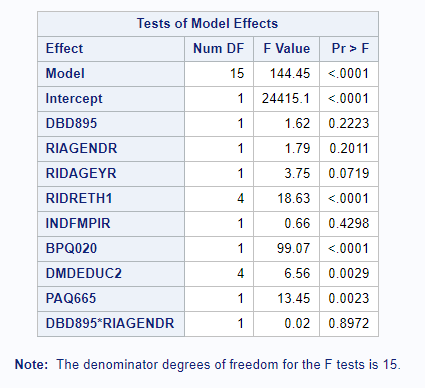
ALQ120Q NOT in (777, 999) and

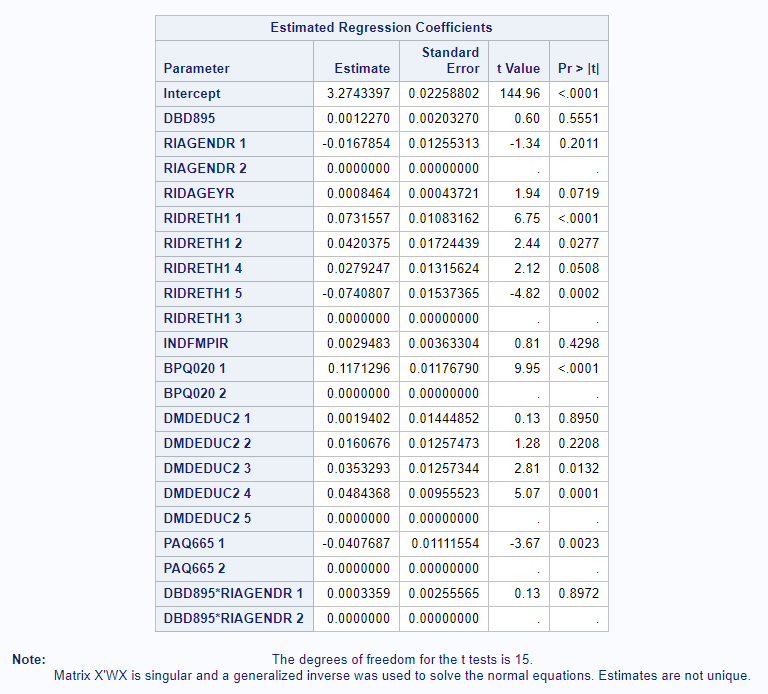
PAQ665 NOT in (7, 9) and

INDFMPIR NOT in (77, 99);

run;







/\*Step 3 -- Dropped ALQ120Q and DBD895\*RIAGENDR (p = 0.8972)\*/

proc surveyreg data = final;

strata SDMVSTRA;

cluster SDMVPSU;

class RIAGENDR RIDRETH1 (ref = '3') BPQ020 DMDEDUC2 PAQ665;

model log\_bmi = DBD895 RIAGENDR RIDAGEYR RIDRETH1 INDFMPIR BPQ020

DMDEDUC2 PAQ665 / solution;

weight WTMEC2YR;

where 18 <= RIDAGEYR <= 65 and

BPQ020 NOT in (7, 9) and

DBD895 < 22 and

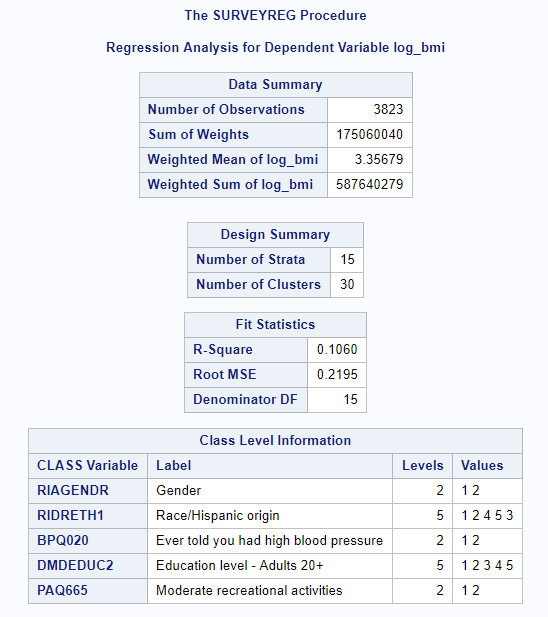
DMDEDUC2 NOT in (7, 9) and

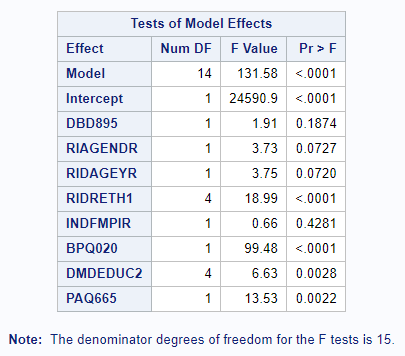
ALQ120Q NOT in (777, 999) and

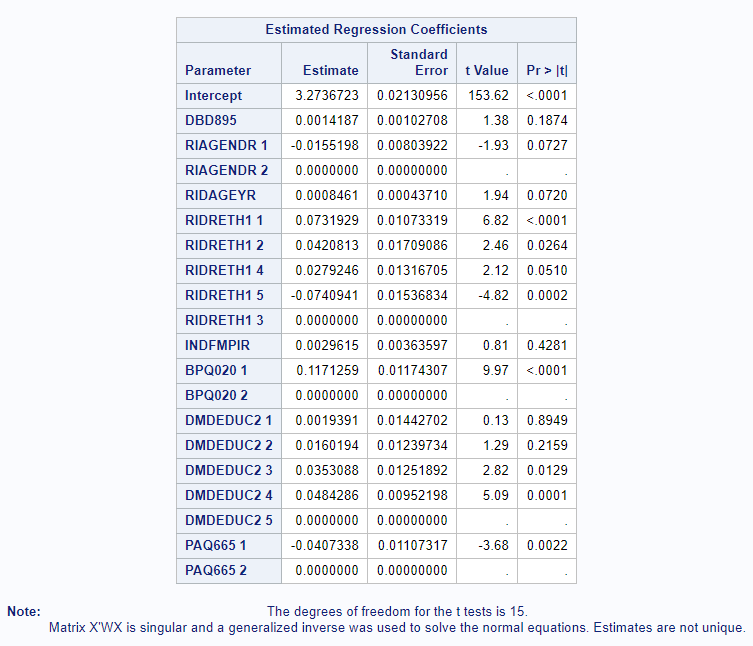
PAQ665 NOT in (7, 9) and

INDFMPIR NOT in (77, 99);

run;







/\*Step 4 -- Dropped ALQ120Q, DBD895\*RIAGENDR, and INDFMPIR (p = 0.4281) -- Final Model\*/

/\*Our final model included 7 predictors, all of which met the alpha level of 0.10. The r-squared value was 0.1047. 4,204 observations were used in the final model.\*/

proc surveyreg data = final;

strata SDMVSTRA;

cluster SDMVPSU;

class RIAGENDR RIDRETH1 (ref = '3') BPQ020 DMDEDUC2 PAQ665;

model log\_bmi = DBD895 RIAGENDR RIDAGEYR RIDRETH1 BPQ020 DMDEDUC2 PAQ665 /

solution;

weight WTMEC2YR;

where 18 <= RIDAGEYR <= 65 and

BPQ020 NOT in (7, 9) and

DBD895 < 22 and

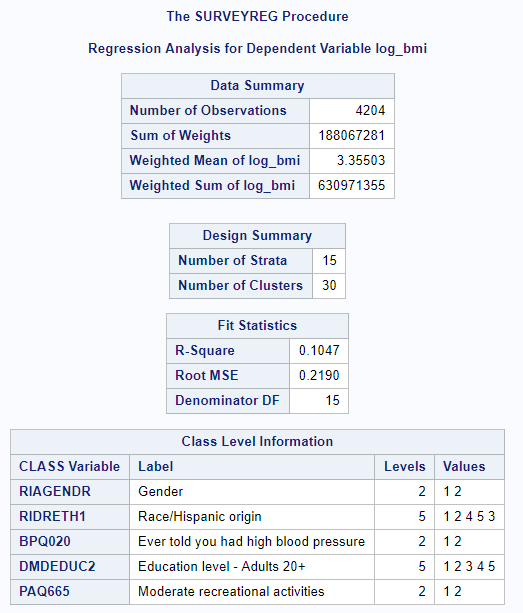
DMDEDUC2 NOT in (7, 9) and

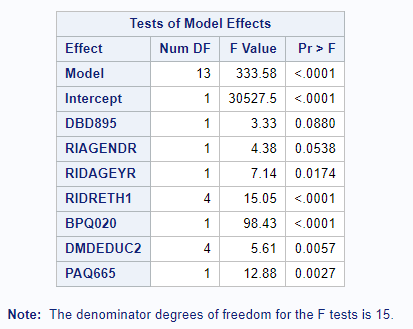
ALQ120Q NOT in (777, 999) and

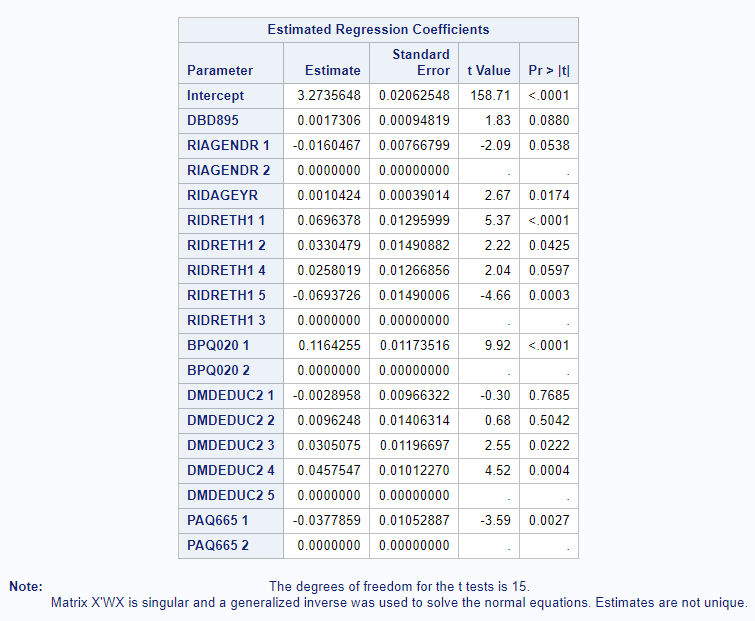
PAQ665 NOT in (7, 9) and

INDFMPIR NOT in (77, 99);

run;







/\*Regression Diagnostics\*/

/\*We ran a series of regression diagnostics to test whether our model met the assumptions for linear regression models.\*/

/\*Normality of Residuals - We tested for normality of the residuals. Because the p-value for any of the tests (e.g., Kolmogorov-Smirnov, Anderson-Darling, etc.) was less than 0.05, we rejected the normal assumption of the residuals.\*/

proc glm data = final;

class RIAGENDR RIDRETH1 BPQ020 DMDEDUC2 PAQ665;

model log\_bmi = DBD895 RIAGENDR RIDAGEYR RIDRETH1 BPQ020 DMDEDUC2 PAQ665;

where 18 <= RIDAGEYR <= 65 and

BPQ020 NOT in (7, 9) and

DBD895 < 22 and

DMDEDUC2 NOT in (7, 9) and

ALQ120Q NOT in (777, 999) and

PAQ665 NOT in (7, 9) and

INDFMPIR NOT in (77, 99);

output out=out p=pred r=r;

run; quit;

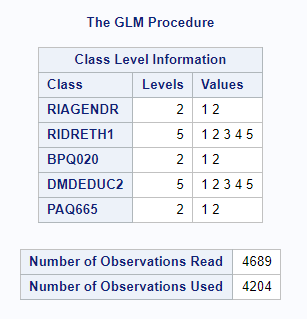
proc univariate normal;

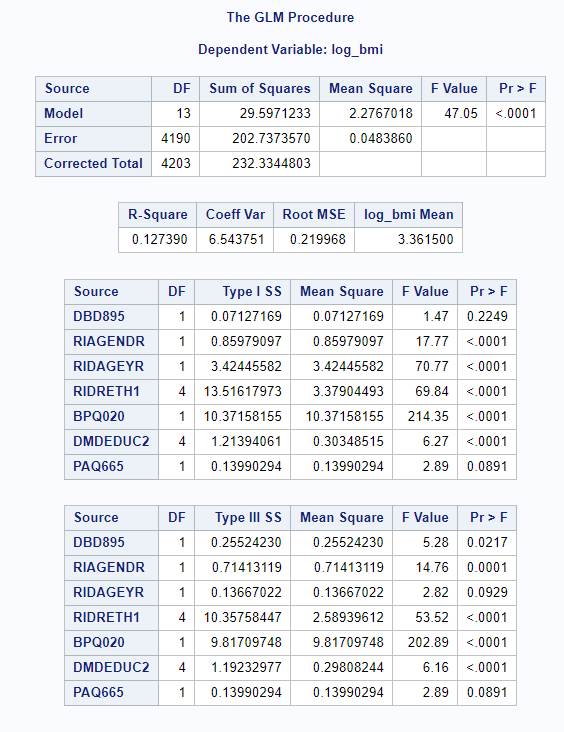
var r;

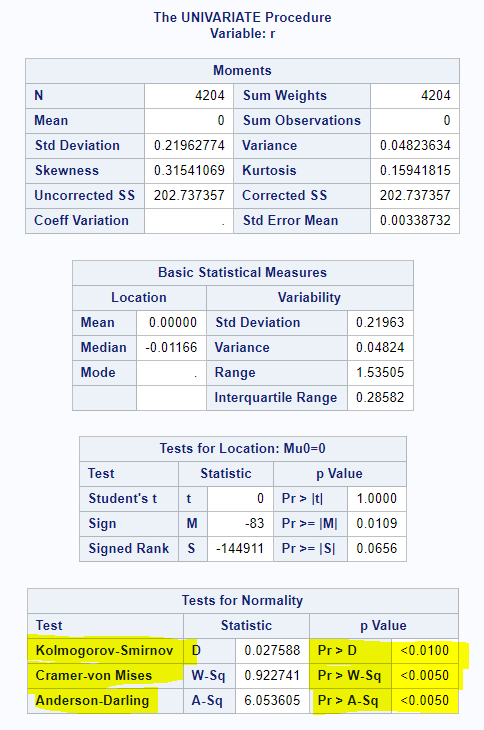
qqplot r;

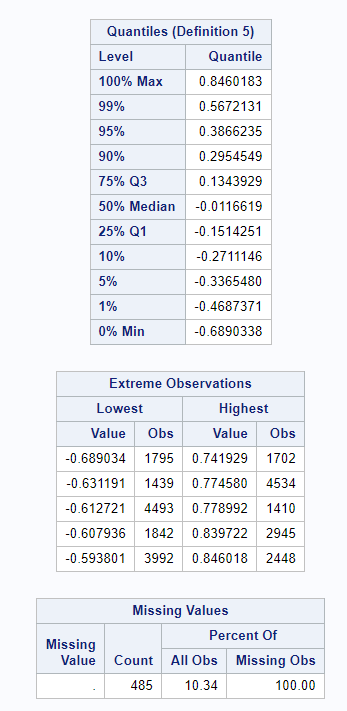
histogram r/normal kernel;

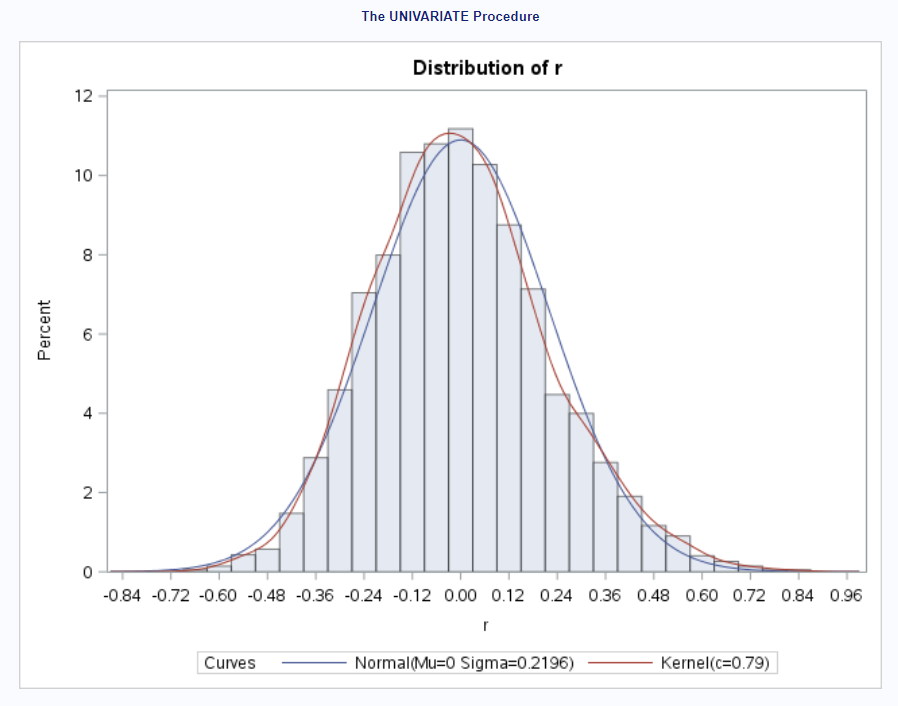
run;

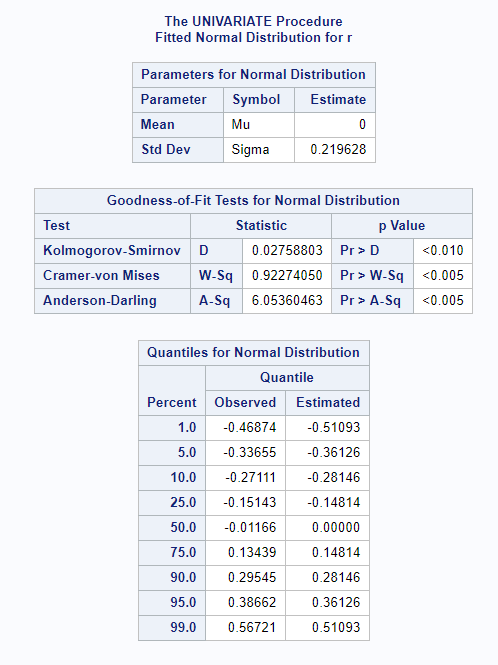


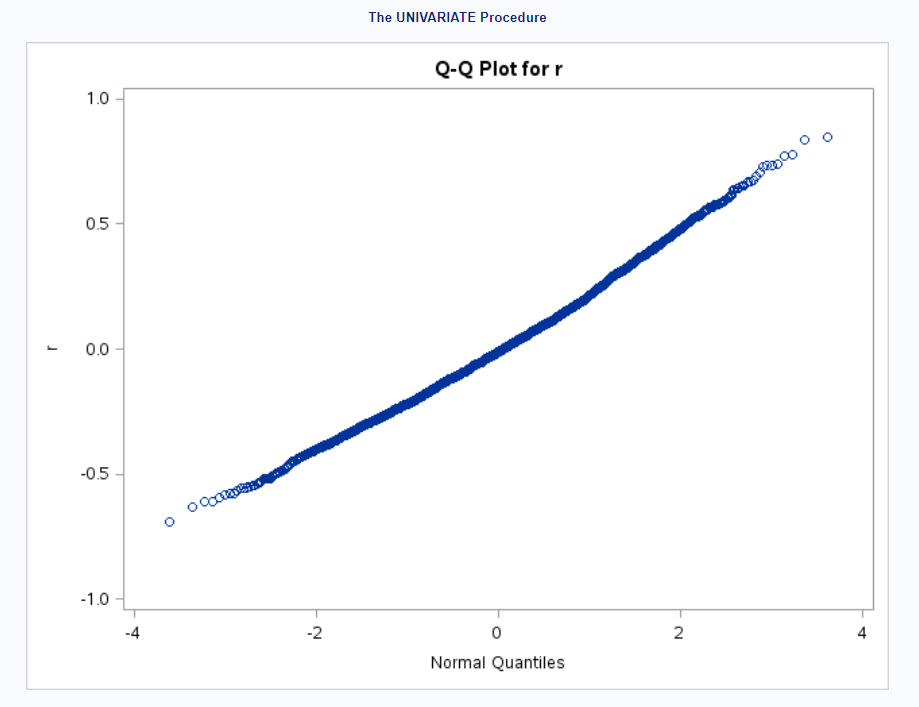












/\*Homogeneity of Residual Variance - We checked for homoscedasticity. Because the p-value for the test was less than 0.05, we rejected the homoscedasticity assumption.\*/

proc reg data = final;

model log\_bmi = DBD895 RIAGENDR RIDAGEYR RIDRETH1 BPQ020 DMDEDUC2 PAQ665/spec;

where 18 <= RIDAGEYR <= 65 and

BPQ020 NOT in (7, 9) and

DBD895 < 22 and

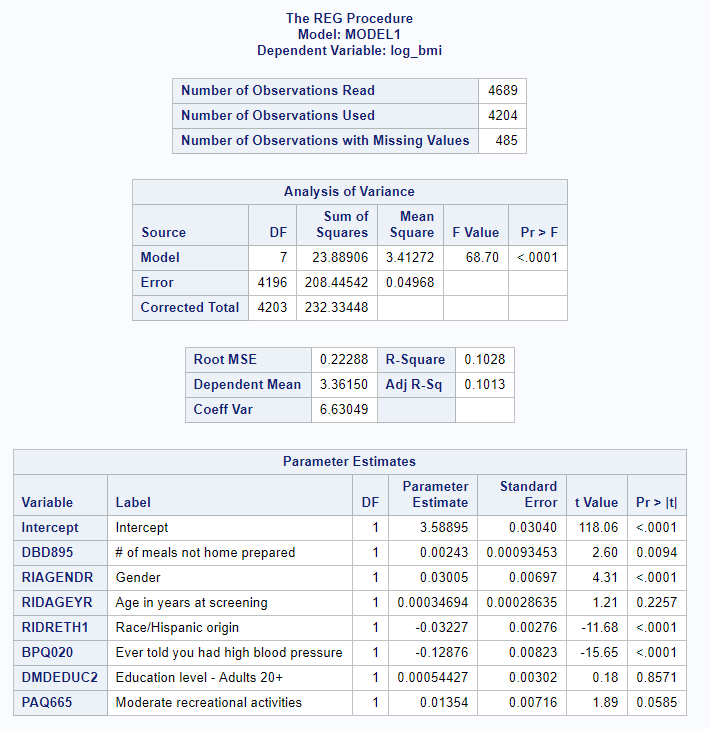
DMDEDUC2 NOT in (7, 9) and

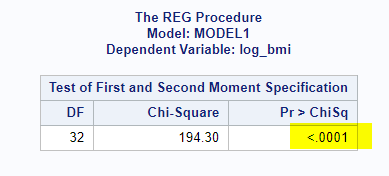
ALQ120Q NOT in (777, 999) and

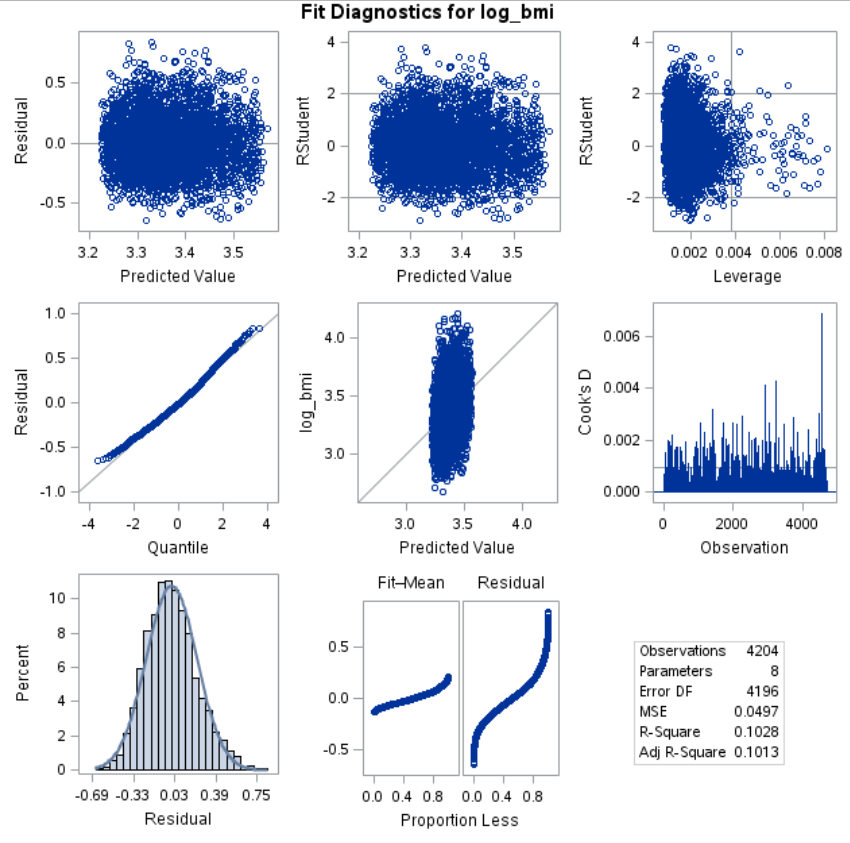
PAQ665 NOT in (7, 9) and

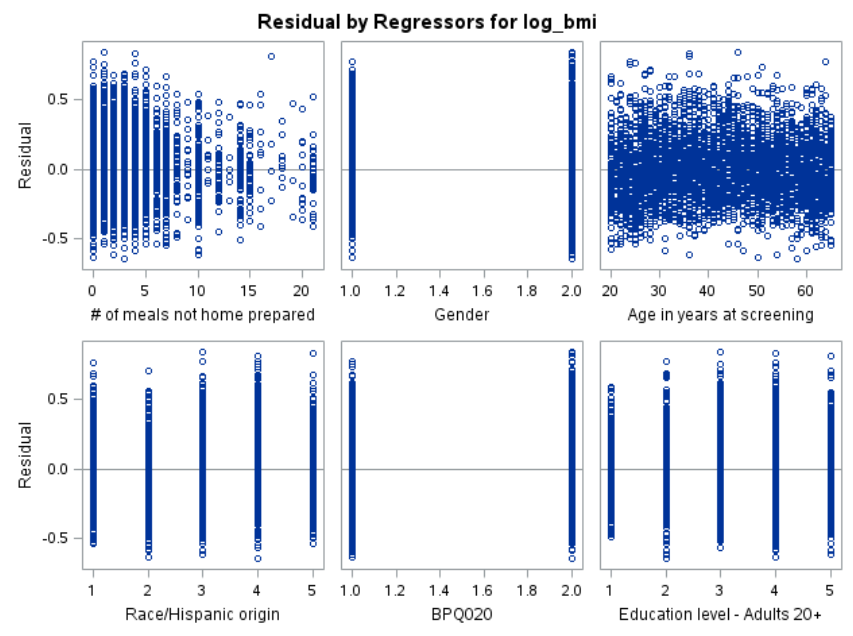
INDFMPIR NOT in (77, 99);

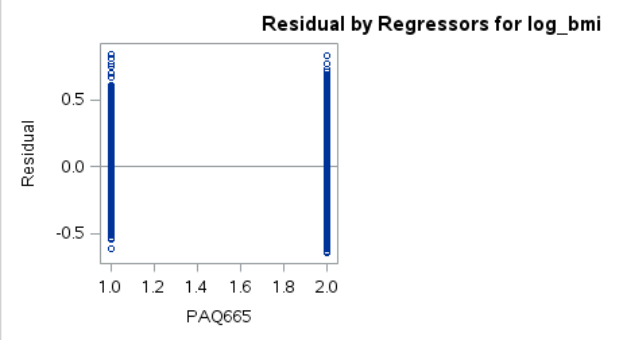
run;











/\*Multicollinearity - We tested for multicollinearity among the predictors. Upon examination of the

Type II tolerance values, none of them were below 0.1, suggesting that we may not need to be as concerned with multicollinearity.\*/

proc glm data = final;

class RIAGENDR RIDRETH1 BPQ020 DMDEDUC2 PAQ665;

model log\_bmi = DBD895 RIAGENDR RIDAGEYR RIDRETH1 BPQ020 DMDEDUC2 PAQ665 /

tolerance;

where 18 <= RIDAGEYR <= 65 and

BPQ020 NOT in (7, 9) and

DBD895 < 22 and

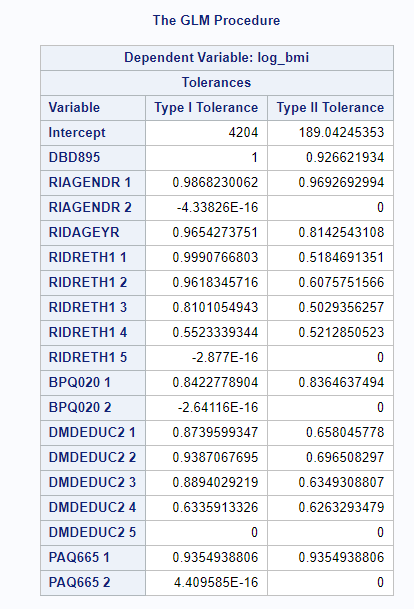
DMDEDUC2 NOT in (7, 9) and

ALQ120Q NOT in (777, 999) and

PAQ665 NOT in (7, 9) and

INDFMPIR NOT in (77, 99);

run;



/\*Outliers/Influential Points - We tested for outliers and influential observations. PROC GLM used 4,204 observations.\*/

proc glm data = final;

class RIAGENDR RIDRETH1 BPQ020 DMDEDUC2 PAQ665;

model log\_bmi = DBD895 RIAGENDR RIDAGEYR RIDRETH1 BPQ020 DMDEDUC2 PAQ665;

where 18 <= RIDAGEYR <= 65 and

BPQ020 NOT in (7, 9) and

DBD895 < 22 and

DMDEDUC2 NOT in (7, 9) and

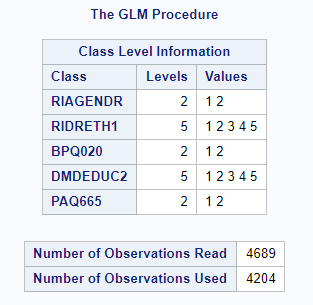
ALQ120Q NOT in (777, 999) and

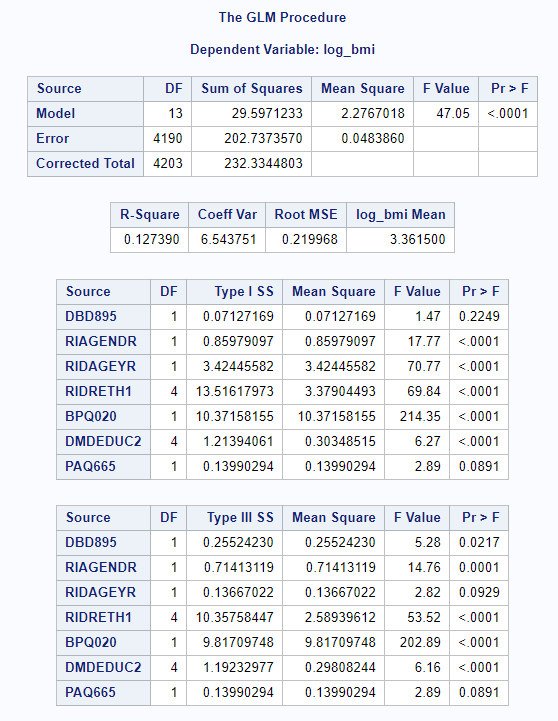
PAQ665 NOT in (7, 9) and

INDFMPIR NOT in (77, 99);

output out=out r=r rstudent=rs h=leverage cookd=cook covratio=covratio dffits=dffit;

run;





/\*Outliers - In examining outliers specifically, we looked for observations that had an RStudent value that was greater than or equal to the absolute value of 2.5. 64 observations met this criterion.\*/

proc print data = out n;

where (rs >= 2.5 OR rs <= -2.5) AND rs ne .;

run;

/\*Influential Points - In examining influential points specifically, we looked for observations that

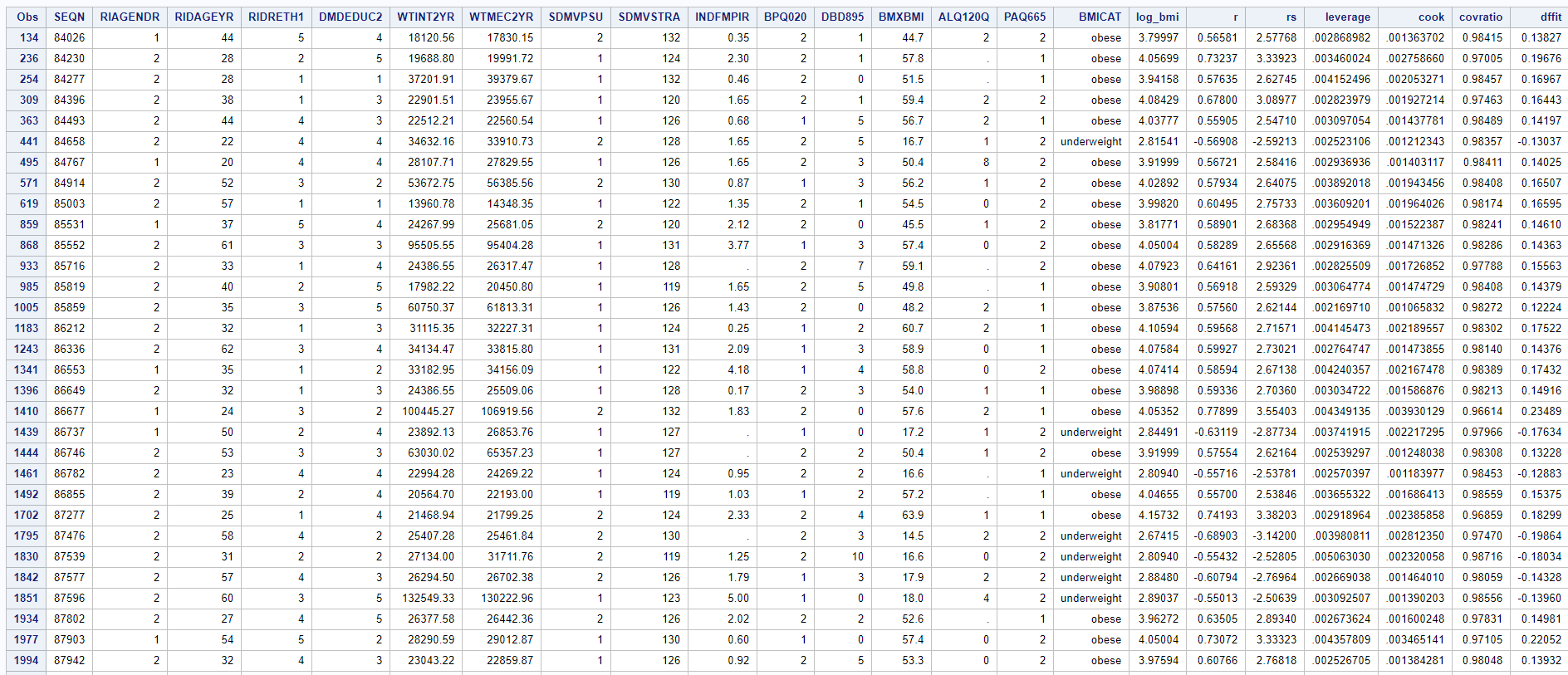
had a Cook's D value greater than (4/4,206). 200 observations met this criterion.\*/

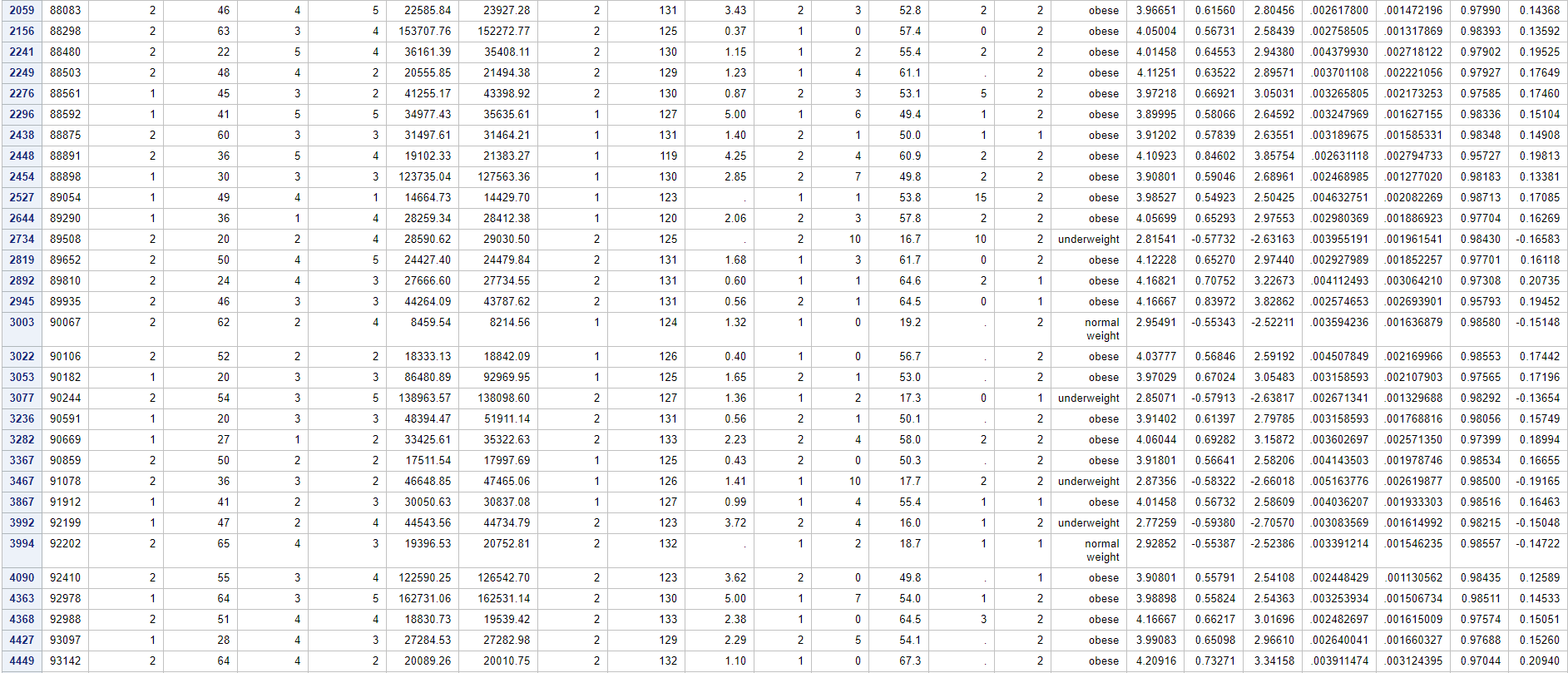
proc print data = out n;

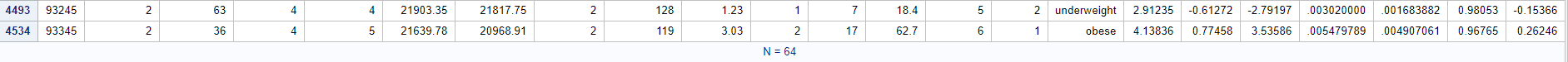
where cook > (4/4206);

run;

**OUTPUT FOR OUTLIERS:**







**OUTPUT FOR INFLUENCE POINTS:**

