Baruch COLLEGE
The City University of New York

Impact of Education Level and Smoking Status of Pregnancy Women:

Baby Weight Dataset Exploration

Xiaoying Zhao

Special Topics in Statistical Analysis
Course Final Project
Professor Gene Fisch
December 20, 2019

Table of Contents

Introduction **3**

Research Question 1:

Classification of Smoking Status Using Logistic Regression **4**

Rationale *4*

Statistical Model *4*

SAS Program Output *5*

Conclusion *6*

Research Question 2:

Effects of Mother's Education Level on Pregnancy Weight Gain Using Permutation Test **7**

Rationale *7*

Statistical Model *7*

SAS Program Output *8*

Conclusion *9*

Research Question 3:

Smoking and Education effects on Baby Weight Using ANOVA with Cross-Validation **10**

Rationale *10*

Statistical Model *10*

SAS Program Output *11*

Conclusion *16*

Appendix: SAS Program **17**

Introduction

This project established three research questions based on the 1997 birth weight data from National Center for Health Statistics to study the mother's effects on baby weight. The project focused on exploring the impacts of education level and smoking status of expectant mother on infant's weight and mother's weight gain.

Firstly, project defined possible factors to classify whether mother was a smoker by using logistic regression. Then it studied the effect of mother's education level on mother's weight gain via application of permutation test (Wilcoxon Rank Sum test). Finally, project conducted analysis on effect of mother's education level and smoking status on infant weight by applying ANOVA and cross-validation analysis. Each question was discussed by its rationale, statistical model, SAS output, and conclusion.

Project found that infant weight and gender, and mother's race, marriage status and education level had contributions to classify mother's smoking status. Average weight gains during pregnancy for mothers with different education levels were statistically significantly different. Both mother's education level and smoking status had major influence on baby weight.

Research Question 1:

Classification of Smoking Status Using Logistic Regression

Rationale

Everyone knows that smoking can cause fertility problems. Also, expectant mother smoked before or during pregnancy can contribute to premature birth, infant organ damage, and birth defects. The first question focused to find out significant factors which have contribution to identify pregnancy smoking status. The first research question was interested in exploring possible factors to identify a smoking expectant mother by using data available in Baby Birth dataset. The outcomes of this question were either this mother was a smoker or not. Therefore, logistic regression, a statistical model used to model binary dependent variable, was adopted here. As for logistic regression, it did not require a linear relationship between dependent and independent variables, and it also did not require residuals needed to follow a normal distribution. Assumptions of logistic regression were presented below:

1. Dependent variable should only have two outcomes.
2. Individual observation should be independent.
3. Little or no multicollinearity among the independent variables.
4. Independent variables should be linearly related to the log odds.
5. Large sample size was required.

Statistical Model

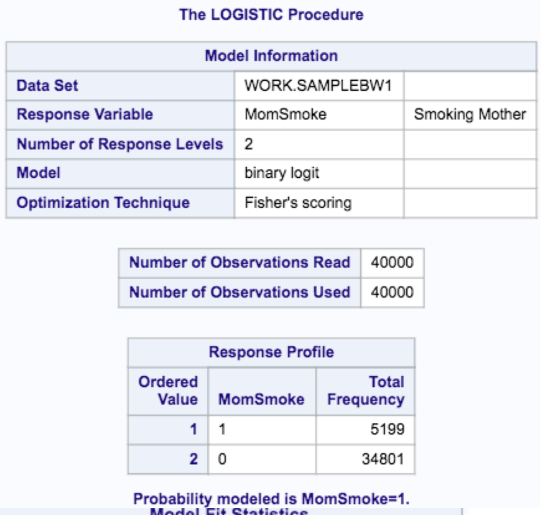
The whole dataset had 8 variables, including variable ‘Smoke’, which was indicator of smoking mother (1 = yes, 0 = no). It had two possible outcomes, meaning it was a dichotomous outcome variable. Logistic regression was used to model dichotomous outcome variables. Since the project wanted to define the pregnancy smoking status, variable ‘Smoke’ was the dependent variable. The candidate independent variables to build this model included the followings:

Independent variable	Definition
Weight	Infant’s birth weight (gm)
Black	Indicator of black mother (1=yes)
Married	Indicator of married mother (1=yes)
Boy	Indicator of boy
MomEdLevel	Indicator of Mother’s education level (0 = high school, 1 = some college, 2 = college, 3 = less than high school)

Only variable ‘Weight’ was continuous variable, the rest 4 variables were categorical variables. The logistic regression model for the first research questions was temporarily presented as:

$$Smoke = \alpha + \beta_1 Weight + \beta_2 Black + \beta_3 Married + \beta_4 Boy + \beta_5 MomEdLevel$$

SAS Program Output



The LOGISTIC Procedure

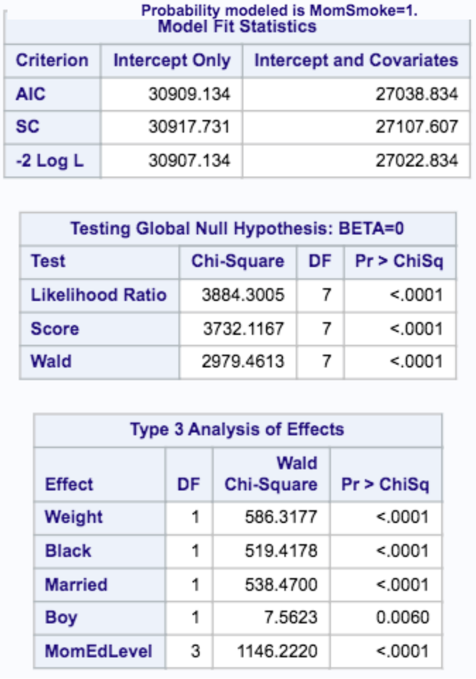
Model Information		
Data Set	WORK.SAMPLEBW1	
Response Variable	MomSmoke	Smoking Mother
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	40000
Number of Observations Used	40000

Response Profile		
Ordered Value	MomSmoke	Total Frequency
1	1	5199
2	0	34801

Probability modeled is MomSmoke=1.

To generate the model, 40000 was selected out of original dataset using simple random sampling. 5199 mothers were smoker in the sample, and the rest of sample (n = 34801) was non-smoker. SAS is modeling variable MomSmoke using a binary logit model and that the probability of mother being smoker is being modeled.

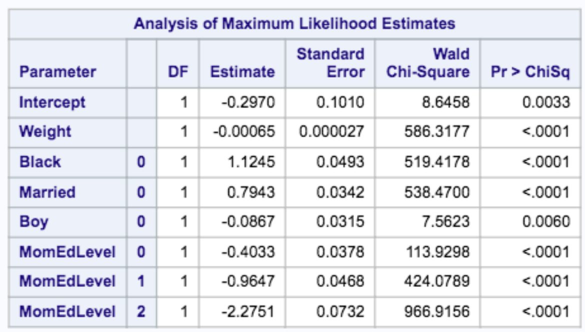


Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	30909.134	27038.834
SC	30917.731	27107.607
-2 Log L	30907.134	27022.834

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	3884.3005	7	<.0001
Score	3732.1167	7	<.0001
Wald	2979.4613	7	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Weight	1	586.3177	<.0001
Black	1	519.4178	<.0001
Married	1	538.4700	<.0001
Boy	1	7.5623	0.0060
MomEdLevel	3	1146.2220	<.0001

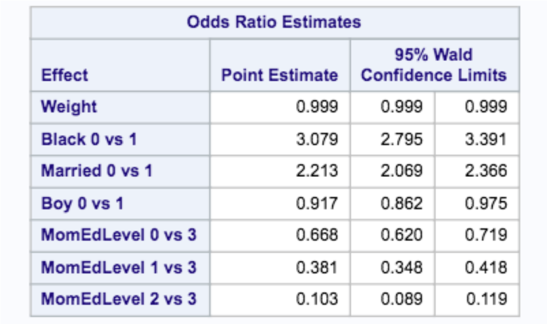
Three tables on the left were used to evaluate the model fit. The table of model fit statistics described the overall fit of the model. Akaike’s Information Criterion (AIC) was common statistics used to compares a set of models. Here, only one model was generated, so there was no comparison. The global null hypothesis test was used to compare current model with the empty model(β_i= 0, i =1,2,3,4). The global null hypothesis was β_i= 0, meaning the model is an empty model. Probabilities that β_i= 0 under three different statistics were close to 0, meaning that the null was rejected and that the model should not be empty. The third table was analysis of effects, which indicated the hypothesis tests for each variable in the model individually. All p-values presented that all variables were statistically significant to improve the model fit and should be included in the model.



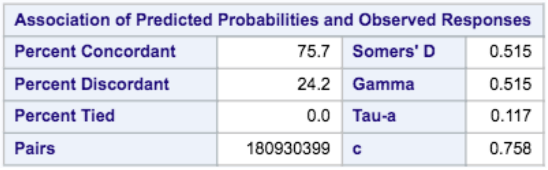
Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.2970	0.1010	8.6458	0.0033
Weight		1	-0.00065	0.000027	586.3177	<.0001
Black	0	1	1.1245	0.0493	519.4178	<.0001
Married	0	1	0.7943	0.0342	538.4700	<.0001
Boy	0	1	-0.0867	0.0315	7.5623	0.0060
MomEdLevel	0	1	-0.4033	0.0378	113.9298	<.0001
MomEdLevel	1	1	-0.9647	0.0468	424.0789	<.0001
MomEdLevel	2	1	-2.2751	0.0732	966.9156	<.0001

The table above presented coefficients of each parameters. All coefficients were statistically significant in the model. Take variable ‘Black’ and ‘Married’ as examples. For every mother was black, the log odds of being smoker increased by 1.1245. For every mother was married, the log odds of being smoker increased by 0.7934.

Interpretation of coefficients for levels of MomEdLevel was slightly different from previous variables. Recall that variable ‘MomEdLevel’, education level of mother, had 4 levels (0 = high school, 1 = some college, 2 = college, 3 = less than high school). If a mother had attended high school with MomEdLevel=0, versus less than high school with MomEdLevel=3, decreased by 0.4033.



Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Weight	0.999	0.999	0.999
Black 0 vs 1	3.079	2.795	3.391
Married 0 vs 1	2.213	2.069	2.366
Boy 0 vs 1	0.917	0.862	0.975
MomEdLevel 0 vs 3	0.668	0.620	0.719
MomEdLevel 1 vs 3	0.381	0.348	0.418
MomEdLevel 2 vs 3	0.103	0.089	0.119



Association of Predicted Probabilities and Observed Responses			
Percent Concordant	75.7	Somers' D	0.515
Percent Discordant	24.2	Gamma	0.515
Percent Tied	0.0	Tau-a	0.117
Pairs	180930399	c	0.758

The table of odds ratio estimates transformed coefficients into odds ratios, which were exponentiated coefficients. For a baby being a boy, the odds of being a smoking mother increased by a factor of 0.917.

The final table presented the test for the overall effect of mother’s education level. It displayed measures of association between predicted probability and the observed responses. There were over 180 million pairs compared.

Conclusion

We can conclude that infant weight and gender, and mother’s race, marriage status and education level had contributions to classify mother’s smoking status. But contributors were not limited to those variables in Baby Weight Dataset.

Research Question 2:

Effects of Mother's Education Level on Pregnancy Weight Gain Using Wilcoxon Rank Sum Test

Rationale

Managing mother's weight during pregnancy became popular healthcare topic. Gaining too little or too much weight during pregnancy could be harmful to both infant and expectant mother. Mother's weight gain during pregnancy was an important health index. Education created better health. People who were well-educated tended to have better income and more resources to achieve and maintain the healthy life. Did women with different education background gain different weight during pregnancy? This question was the second research question of this project. Non-parametric permutation test, Wilconxon Rank Sum Test was adopted for this question. This test was preferred because:

1. Mother's education levels were independent of each other.
2. Mother's weight gain was continuous and might not be normally distributed.

Statistical Model

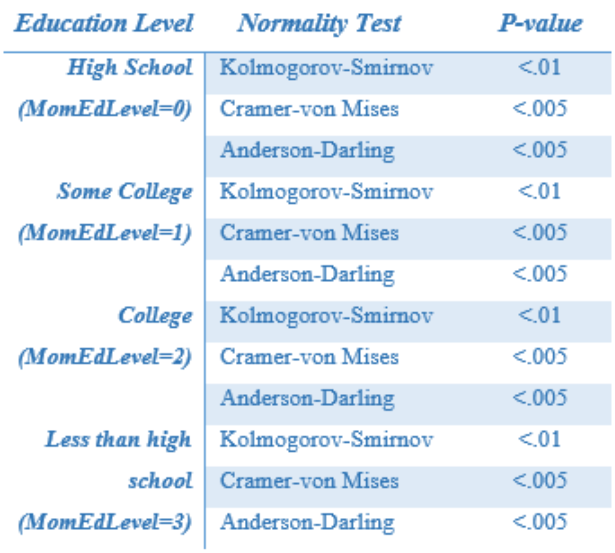
The Birth Weight dataset contained the variables required for permutation test.

Variables	Definition
MomEdLevel	Indicator of Mother's education level (0 = high school, 1 = some college, 2 = college, 3 = less than high school)
MomWtGain	Mother's weight gain during pregnancy (this number must be added to 30 to obtain actual weight gain)

The dataset for this research question contained 10000 observations, which were randomly selected from the original baby weight dataset.

The model will be presented as MomWtGain ~ MomEdLevel. In this case, dependent variable was MomWtGain, and independent variable was MomEdLevel.

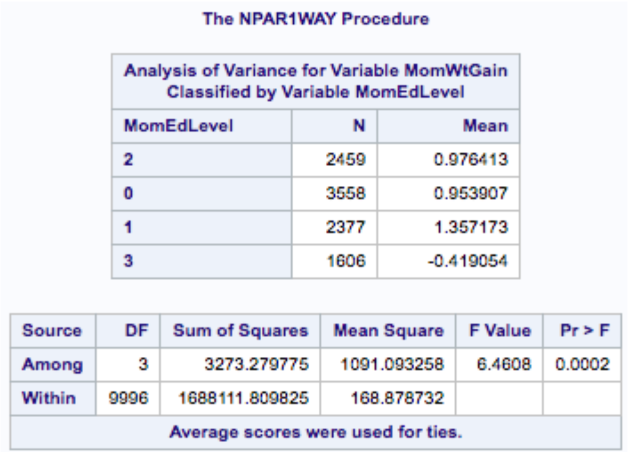
SAS Program Output



Education Level	*Normality Test*	*P-value*
High School (MomEdLevel=0)	Kolmogorov-Smirnov	<.01
	Cramer-von Mises	<.005
	Anderson-Darling	<.005
Some College (MomEdLevel=1)	Kolmogorov-Smirnov	<.01
	Cramer-von Mises	<.005
	Anderson-Darling	<.005
College (MomEdLevel=2)	Kolmogorov-Smirnov	<.01
	Cramer-von Mises	<.005
	Anderson-Darling	<.005
Less than high school (MomEdLevel=3)	Kolmogorov-Smirnov	<.01
	Cramer-von Mises	<.005
	Anderson-Darling	<.005

Normality should be the first test to be performed on the dataset. The summary of normality test for each education level group was presented on the left.

Three different normality tests were performed on individual education level group. The null hypothesis was that the data follows normal distribution. Probability that null hypothesis was true for individual education level was close to zero, indicating none of these subgroups follow normal distribution. Hence, a Wilcoxon Rank Sum test should be performed here.



The NPAR1WAY Procedure

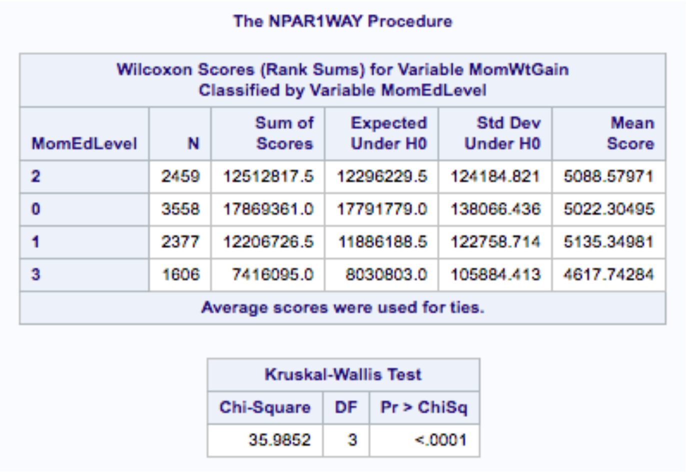
Analysis of Variance for Variable MomWtGain Classified by Variable MomEdLevel

MomEdLevel	N	Mean
2	2459	0.976413
0	3558	0.953907
1	2377	1.357173
3	1606	-0.419054

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Among	3	3273.279775	1091.093258	6.4608	0.0002
Within	9996	1688111.809825	168.878732		

Average scores were used for ties.

SAS procedure NPARIWAY performed the non-parametric tests. Tables on left were produced with ANOVA option. For each level of education, it displayed the number of observations and the mean of variable MomWtGain. The average weight gain for mothers who had attended less than high school was (30-0.41=29.59lb), which was the lowest average weight gain among 4 different education levels. The average weight gain for mothers who had attended high school was (30+1.36=31.36lb), which was the highest. Although difference in means among 4 levels was not huge, probability of F-test was close to zero, indicating that education levels accounted for a significant portion of the variability in the dependent variable mother’s weight gain during pregnancy.



The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable MomWtGain
Classified by Variable MomEdLevel

MomEdLevel	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
2	2459	12512817.5	12296229.5	124184.821	5088.57971
0	3558	17869361.0	17791779.0	138066.436	5022.30495
1	2377	12206726.5	11886188.5	122758.714	5135.34981
3	1606	7416095.0	8030803.0	105884.413	4617.74284

Average scores were used for ties.

Kruskal-Wallis Test

Chi-Square	DF	Pr > ChiSq
35.9852	3	<.0001

Tables above were produced with Wilconxon option. The first table was a summary of the Wilcoxon scores for variable MomWtGain by Education levels. Ktuskal-Wallis test was the most important component in the test. The null hypothesis for this test was mothers' weight gains during pregnancy were same among different education level. Since the probability that the null hypothesis was true was close to zero, we reject the null hypothesis, meaning mothers' weight gains during pregnancy were statistically significantly different among different education level.

Conclusion

Consequently, we can conclude that mothers' weight gains during pregnancy were statistically significantly different among different education level. In this research question, only categorical variable education level was involved, which does not apply that mother's education level was the only factor to influence the weight gain during pregnancy. There were more factors involved than variables presented in the original babe dataset.

Research Question 3:

Smoking and Education effects on Baby Weight Using ANOVA with Cross-Validation

Rationale

The third research question emphasized on the effect of mother's smoking status and education level on newborn weights. Since there were four different levels of education levels and 2 levels of smoking status, Analysis of Variance (ANOVA) test were adopted to compare average infant weight among different groups. After completing the ANOVA test, cross-validation was adopted to assess the performance of statistical analysis and to predict infant weight based on mother's education level and smoking status.

ANOVA was a statistical model used to compare two or more means. In this research question, four infant weight means were being examined. The null hypothesis in this case was all four-population means were exactly equal, meaning there were no difference among four education levels and two smoking status. The following assumption should be met:

1. Observation should be independent.
2. Normality: outcomes must follow a normal distribution.
3. Homogeneity: variances within subgroup population should be equal.

Cross-Validation was a model validation method to evaluate the performance of the ANOVA and prevent analysis from limitation problems, such as overfitting and underfitting. Cross-Validation required one validation dataset and one training dataset. These two datasets should be randomly selected from the same distribution.

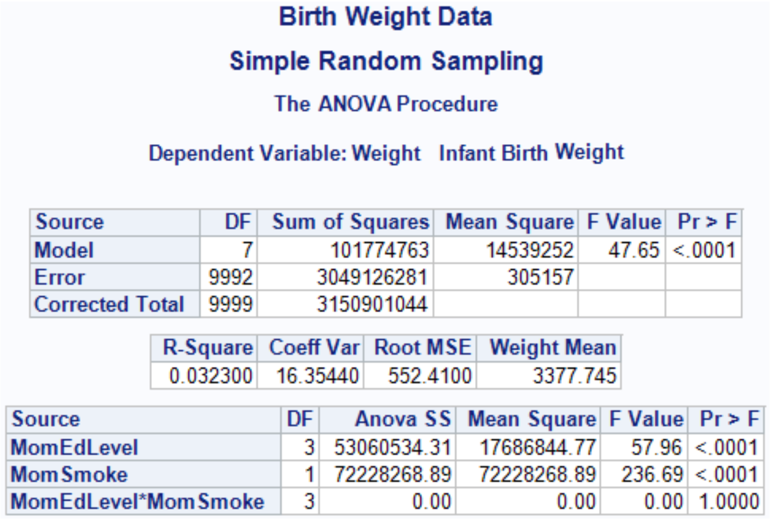
Statistical Model

For this research question, ANOVA was adopted to explore the difference on average infant weight with mothers who had different education level. Then cross validation was used to confirm the result from the ANOVA test. Sample for the ANOVA test was randomly selected from original dataset, totally 10000 analytical observations. The model for ANOVA was expressed as

$$Weight = MomEdLevel\ \ MomSmoke\ \ MomEdLevel \times MomSmoke$$

The outcomes of initial cross-validation presented left-skewed distribution. After several exploration of data transformation, outcomes finally followed normal distribution.

SAS Program Output



Birth Weight Data

Simple Random Sampling

The ANOVA Procedure

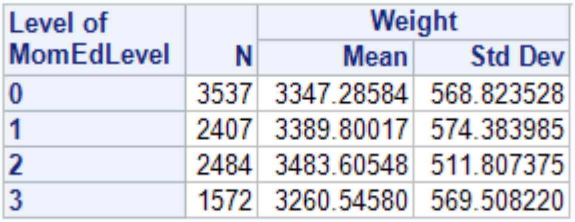
Dependent Variable: Weight Infant Birth Weight

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	101774763	14539252	47.65	<.0001
Error	9992	3049126281	305157		
Corrected Total	9999	3150901044			

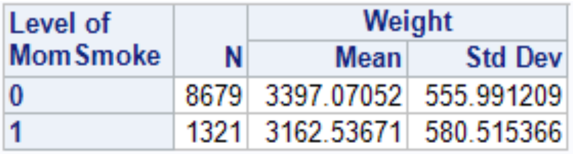
R-Square	Coeff Var	Root MSE	Weight Mean
0.032300	16.35440	552.4100	3377.745

Source	DF	Anova SS	Mean Square	F Value	Pr > F
MomEdLevel	3	53060534.31	17686844.77	57.96	<.0001
MomSmoke	1	72228268.89	72228268.89	236.69	<.0001
MomEdLevel*MomSmoke	3	0.00	0.00	0.00	1.0000

ANOVA results were shown above. The first table presented the test of overall all model. Probability that the null hypothesis was true was close to zero, meaning that null hypothesis was rejected and that average infant weights for groups with different education level and smoking status should not be same. R-square represented the model fit. Although the model was statistically significant, only 3% of the variance for infant weights that were explained by education level and smoking status of mother. Individual variable in the model were statistically significant, but the interaction of these two independent variables was not significant.



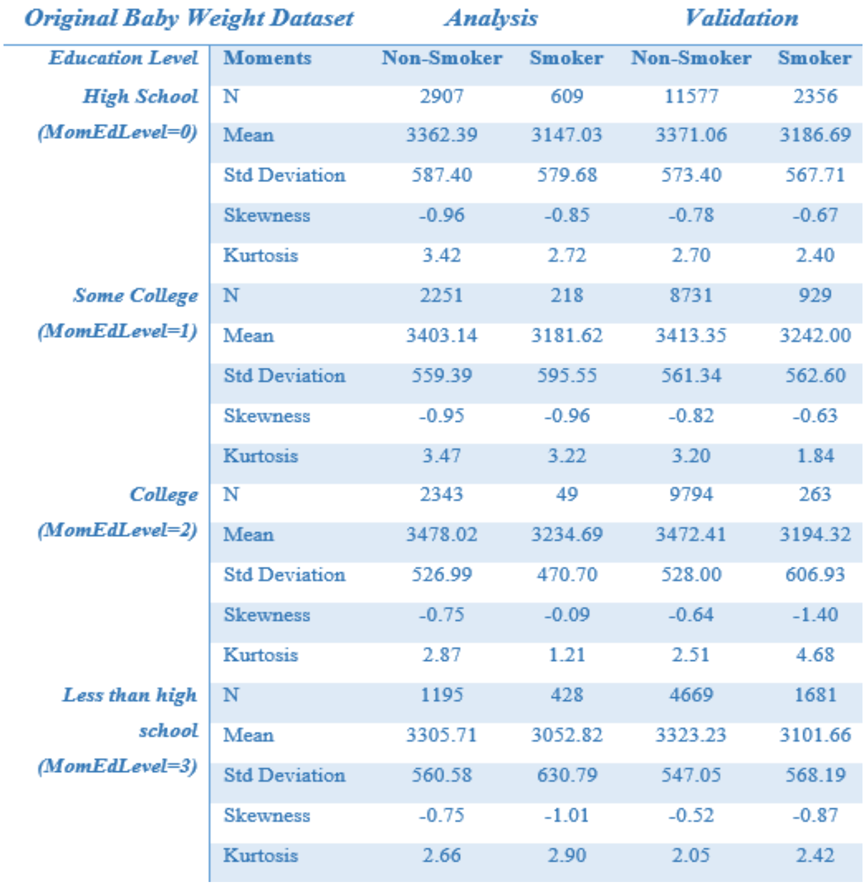
Level of MomEdLevel	N	Weight	
		Mean	Std Dev
0	3537	3347.28584	568.823528
1	2407	3389.80017	574.383985
2	2484	3483.60548	511.807375
3	1572	3260.54580	569.508220



Level of MomSmoke	N	Weight	
		Mean	Std Dev
0	8679	3397.07052	555.991209
1	1321	3162.53671	580.515366

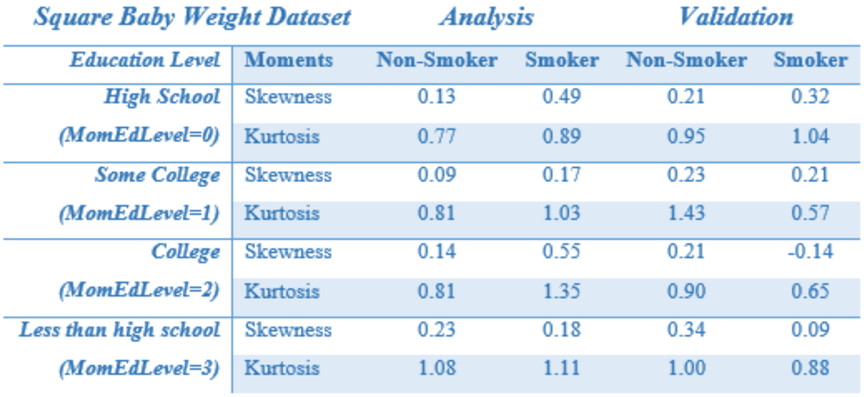
Average infant weights with mothers had different education levels and average infant weights with mother's different smoking status were presented on the left. Firstly, average infant weights for different groups were different, corresponding to the ANOVA result. Secondly, sample sizes of each subgroups were not close. Especially, the sample size for non-smoking mother was 8679, but the sample size of smoking mother was 1321, indicating a huge gap of over 7000 observations, which had impact on the analysis.

Since there were two categorical variables involved, one with 4 levels, another with 2 levels, there were total 8 subgroups. 20000 observations were randomly selected from original Baby Weight dataset and equally divided into two datasets to conduct cross-validation test: analysis and validation. cross-validation analysis would be presented in table format based on SAS program output from the perspectives of basic moments and statistical tests.



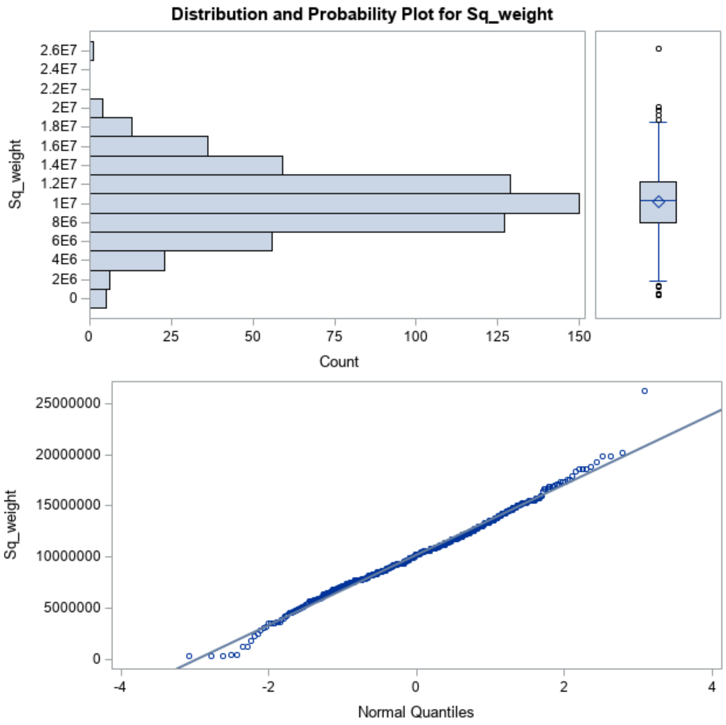
Original Baby Weight Dataset		*Analysis*		*Validation*	
Education Level	**Moments**	**Non-Smoker**	**Smoker**	**Non-Smoker**	**Smoker**
High School	N	2907	609	11577	2356
(MomEdLevel=0)	Mean	3362.39	3147.03	3371.06	3186.69
	Std Deviation	587.40	579.68	573.40	567.71
	Skewness	-0.96	-0.85	-0.78	-0.67
	Kurtosis	3.42	2.72	2.70	2.40
Some College	N	2251	218	8731	929
(MomEdLevel=1)	Mean	3403.14	3181.62	3413.35	3242.00
	Std Deviation	559.39	595.55	561.34	562.60
	Skewness	-0.95	-0.96	-0.82	-0.63
	Kurtosis	3.47	3.22	3.20	1.84
College	N	2343	49	9794	263
(MomEdLevel=2)	Mean	3478.02	3234.69	3472.41	3194.32
	Std Deviation	526.99	470.70	528.00	606.93
	Skewness	-0.75	-0.09	-0.64	-1.40
	Kurtosis	2.87	1.21	2.51	4.68
Less than high	N	1195	428	4669	1681
school	Mean	3305.71	3052.82	3323.23	3101.66
(MomEdLevel=3)	Std Deviation	560.58	630.79	547.05	568.19
	Skewness	-0.75	-1.01	-0.52	-0.87
	Kurtosis	2.66	2.90	2.05	2.42

The table above presented the moments of all 8 subgroups, including sample size, mean, standard deviation, skewness, and kurtosis, for both analysis and validation sub datasets. We can see from the table that average infant weights within education level for analysis and validation were close to each other. Average infant weights within same sub dataset (analysis or validation) for different education levels were different. Skewness of each subgroups was from -0.52 to -0.95, indicating data followed a left-skewed distribution. Also, kurtosis for each group was around 3. These two feature moments of each subgroup showed that the data required transformation. To transform left-skewed dataset, the most common techniques were square and cube. New variable squared infant weight (Sq_Weight) was generated from the original data set.

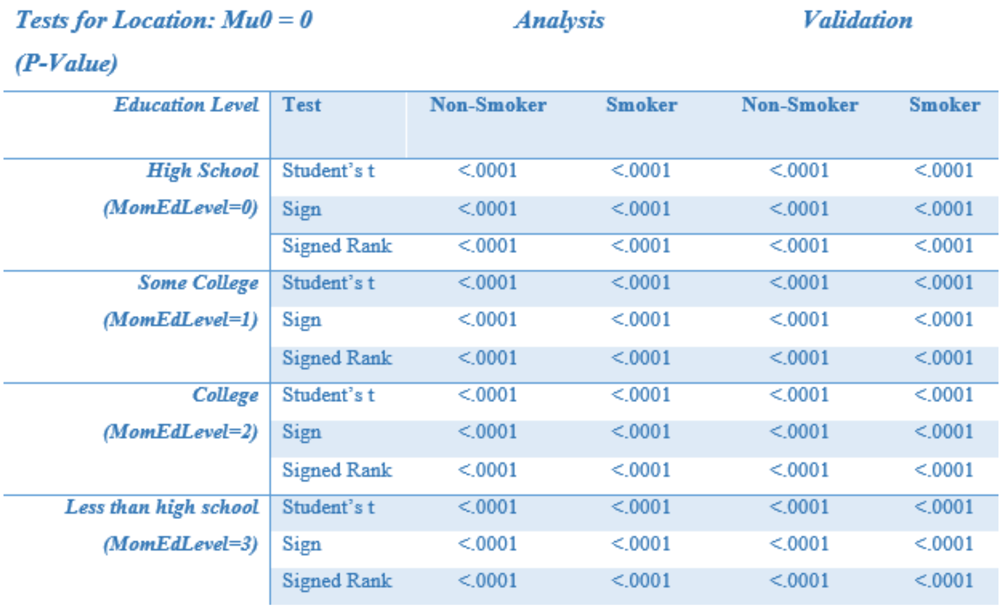


Square Baby Weight Dataset		*Analysis*		*Validation*	
Education Level	**Moments**	**Non-Smoker**	**Smoker**	**Non-Smoker**	**Smoker**
High School	Skewness	0.13	0.49	0.21	0.32
(MomEdLevel=0)	Kurtosis	0.77	0.89	0.95	1.04
Some College	Skewness	0.09	0.17	0.23	0.21
(MomEdLevel=1)	Kurtosis	0.81	1.03	1.43	0.57
College	Skewness	0.14	0.55	0.21	-0.14
(MomEdLevel=2)	Kurtosis	0.81	1.35	0.90	0.65
Less than high school	Skewness	0.23	0.18	0.34	0.09
(MomEdLevel=3)	Kurtosis	1.08	1.11	1.00	0.88

Cross-validation was reapplied to square transformed baby weight dataset. Feature moments of skewness and kurtosis for different subgroups were presented at the table on the left.

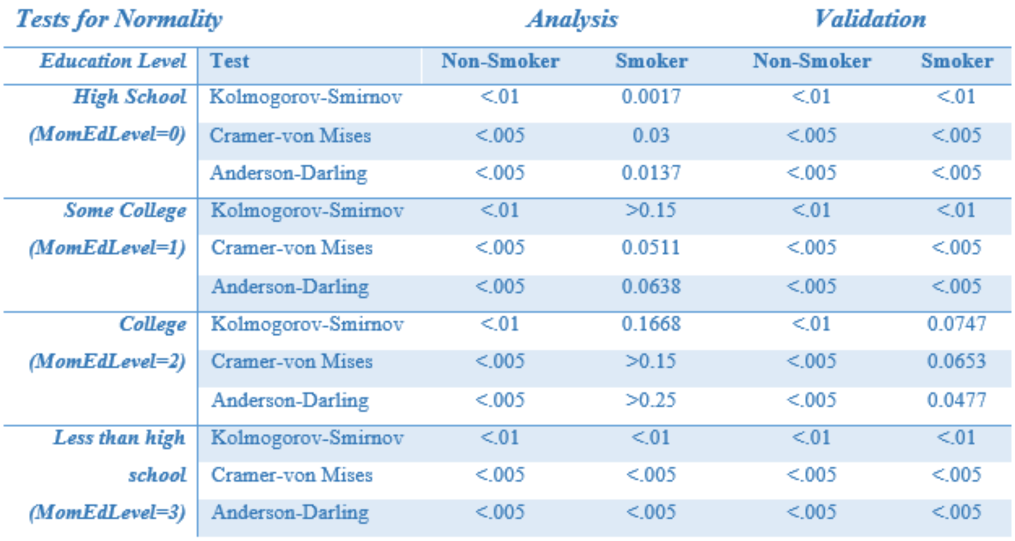


All skewness indexes were close to 0, and all kurtosis indexes were close to 1, meaning shape of every subgroup used to study the third question was symmetric, which also corresponded to the distribution plot and normal q-q plot of every subgroups. Take subgroups of square baby weight with mothers who attended high school and did not smoke as an example. The distribution plot and normality plot were on the left. Both indicated this subgroup followed normal distribution, so did others.



Tests for Location: Mu0 = 0 (P-Value)		Analysis		Validation	
Education Level	Test	Non-Smoker	Smoker	Non-Smoker	Smoker
High School (MomEdLevel=0)	Student's t	<.0001	<.0001	<.0001	<.0001
	Sign	<.0001	<.0001	<.0001	<.0001
	Signed Rank	<.0001	<.0001	<.0001	<.0001
Some College (MomEdLevel=1)	Student's t	<.0001	<.0001	<.0001	<.0001
	Sign	<.0001	<.0001	<.0001	<.0001
	Signed Rank	<.0001	<.0001	<.0001	<.0001
College (MomEdLevel=2)	Student's t	<.0001	<.0001	<.0001	<.0001
	Sign	<.0001	<.0001	<.0001	<.0001
	Signed Rank	<.0001	<.0001	<.0001	<.0001
Less than high school (MomEdLevel=3)	Student's t	<.0001	<.0001	<.0001	<.0001
	Sign	<.0001	<.0001	<.0001	<.0001
	Signed Rank	<.0001	<.0001	<.0001	<.0001

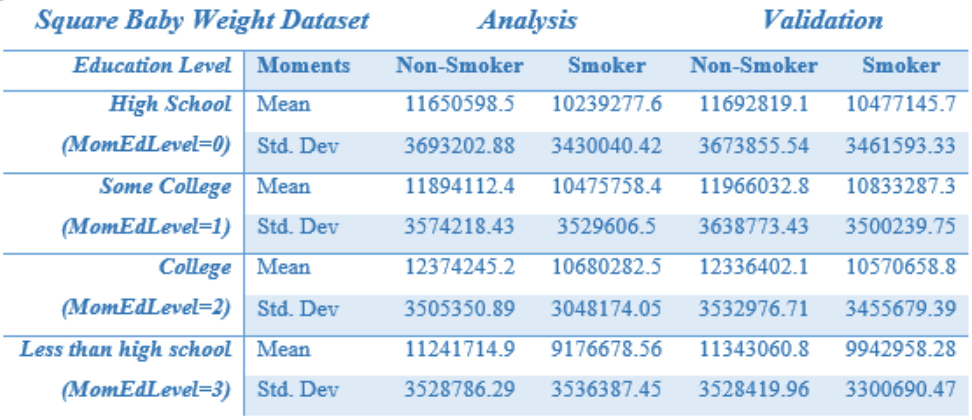
The table above was the summary of tests for location for every individual subgroup. Probability that the null hypothesis was true, which was mu = 0, was close to zero for every individual groups in both analysis dataset and validation dataset, meaning that the null hypothesis was rejected, and that mu was not equal to zero. All average infant weights for individual sub population were not equal to zero, which corresponded to the true that no baby weighted zero gram.



Tests for Normality		Analysis		Validation	
Education Level	Test	Non-Smoker	Smoker	Non-Smoker	Smoker
High School (MomEdLevel=0)	Kolmogorov-Smirnov	<.01	0.0017	<.01	<.01
	Cramer-von Mises	<.005	0.03	<.005	<.005
	Anderson-Darling	<.005	0.0137	<.005	<.005
Some College (MomEdLevel=1)	Kolmogorov-Smirnov	<.01	>0.15	<.01	<.01
	Cramer-von Mises	<.005	0.0511	<.005	<.005
	Anderson-Darling	<.005	0.0638	<.005	<.005
College (MomEdLevel=2)	Kolmogorov-Smirnov	<.01	0.1668	<.01	0.0747
	Cramer-von Mises	<.005	>0.15	<.005	0.0653
	Anderson-Darling	<.005	>0.25	<.005	0.0477
Less than high school (MomEdLevel=3)	Kolmogorov-Smirnov	<.01	<.01	<.01	<.01
	Cramer-von Mises	<.005	<.005	<.005	<.005
	Anderson-Darling	<.005	<.005	<.005	<.005

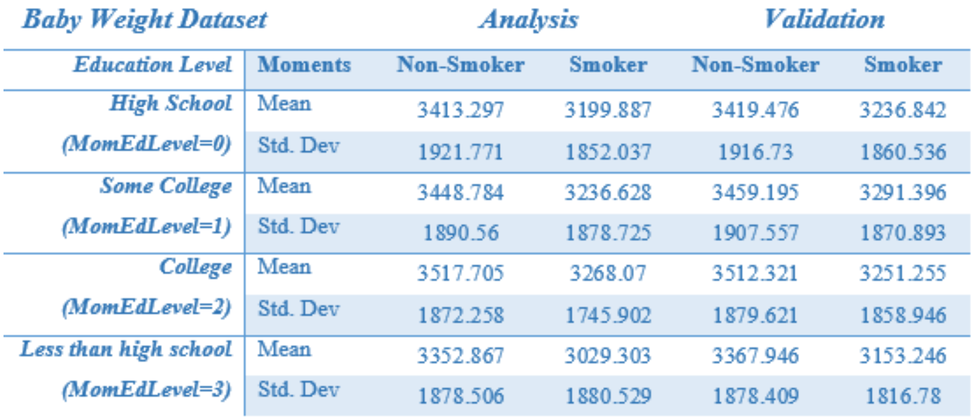
The table above was a summary of tests for normality for individual subgroups in both analysis and validation datasets. Majority subgroups have statistically significant results, meaning majority of subgroups did not follow normal distribution. Only groups of mothers had attended some college and smoked in analysis dataset and mothers who have attended college and smoked in both analysis and validation datasets presented different results.

Consequently, all the SAS output presented that mean and standard deviation of every subgroup became unbiased estimator of mean and standard deviation of associated subpopulation.



Square Baby Weight Dataset		*Analysis*		*Validation*	
Education Level	**Moments**	**Non-Smoker**	**Smoker**	**Non-Smoker**	**Smoker**
High School	Mean	11650598.5	10239277.6	11692819.1	10477145.7
(MomEdLevel=0)	Std. Dev	3693202.88	3430040.42	3673855.54	3461593.33
Some College	Mean	11894112.4	10475758.4	11966032.8	10833287.3
(MomEdLevel=1)	Std. Dev	3574218.43	3529606.5	3638773.43	3500239.75
College	Mean	12374245.2	10680282.5	12336402.1	10570658.8
(MomEdLevel=2)	Std. Dev	3505350.89	3048174.05	3532976.71	3455679.39
Less than high school	Mean	11241714.9	9176678.56	11343060.8	9942958.28
(MomEdLevel=3)	Std. Dev	3528786.29	3536387.45	3528419.96	3300690.47

The table above was the summary of mean and standard deviation for individual subgroup in both analysis and validation datasets. Within individual education level, means and standard deviations in analysis and validation datasets were close to each other separately. Indexes in the table were extremely large was they were square baby weight. After transformed back to the original weight, the summary was shown below.



Baby Weight Dataset		*Analysis*		*Validation*	
Education Level	**Moments**	**Non-Smoker**	**Smoker**	**Non-Smoker**	**Smoker**
High School	Mean	3413.297	3199.887	3419.476	3236.842
(MomEdLevel=0)	Std. Dev	1921.771	1852.037	1916.73	1860.536
Some College	Mean	3448.784	3236.628	3459.195	3291.396
(MomEdLevel=1)	Std. Dev	1890.56	1878.725	1907.557	1870.893
College	Mean	3517.705	3268.07	3512.321	3251.255
(MomEdLevel=2)	Std. Dev	1872.258	1745.902	1879.621	1858.946
Less than high school	Mean	3352.867	3029.303	3367.946	3153.246
(MomEdLevel=3)	Std. Dev	1878.506	1880.529	1878.409	1816.78

In the normal baby weight dataset, within individual education level, means and standard deviations in analysis and validation datasets were close to each other separately. It showed that the cross-validation model had a great performance and can be used to predict the baby weight in the subpopulation.

Conclusion

Base on the SAS output in the previous pages, we can conclude that both education level and smoking status of mother had huge impact to newborn weight. Average newborn weights of different subpopulation were different in perspective of mother’s education level and smoking status. Newborns whose mother had attended college and did not smoke had the heaviest average weight. Newborns whose mother smoked tended to have lighter average weight than newborns whose mother did not smoke.

Appendix: SAS Program

```
/*1. Logistic Regression*/

title1 'Birth Weight Data';
title2 'Simple Random Sampling';
proc surveyselect data=sashelp.BWeight
   method=srs n=40000 out=SampleBW1;
run;

Proc logistic data=SampleBW1 descending ;
     Class Black Married Boy MomEdLevel/param = ref;
     Model MomSmoke = Weight Black Married Boy MomEdLevel;
Run;


/* 2. Permutation */

proc surveyselect data=sashelp.BWeight
   method=srs n=10000 out=SampleBW2;
run;

proc npar1way data = SampleBW2 ANOVA Wilcoxon ;
     Class MomEdLevel ;
     Var MomWtGain;
     exact scores = data/mc n=9999 alpha=0.05;
Run;


/* 3. Resampling */

title1 'Birth Weight Data';
title2 'Simple Random Sampling';
proc surveyselect data=sashelp.BWeight
   method=srs n=10000 out=SampleBW1;
run;

Data ANOVAdataset;
Set SampleBW1;
Proc Sort data = SampleBW1;
     by MomEdLevel MomSmoke;
Run;

ods graphics on;
Proc ANOVA data = ANOVAdataset outstat=all;
Class MomEdLevel MomSmoke;
Model Weight = MomEdLevel MomSmoke MomEdLevel*MomSmoke ;
means MomEdLevel MomSmoke;
Run;
```

```
ods graphics off;


Data Analysis Validate;
Set sashelp.BWeight;
Retain k 10000 n 20000;
If RANUNI(99999) < k/n THEN DO;
k = k-1;
Output Analysis;
END;
Else Output Validate;
n = n-1;
DROP k n;
Run;

Proc Sort data = Analysis;
      by  MomEdLevel MomSmoke;
Run;

Ods GRAPHICS On;
Proc Univariate data = Analysis normal mode plot;
      title 'Cross-Validation Analysis Validation Sets: New RANUNI
number';
      title2 'Analysis Dataset by Education Level';
      by MomEdLevel MomSmoke ;
      var Weight;
      histogram / normal;
Run;
Ods GRAPHICS Off;

Proc Sort data = Validate;
      by MomEdLevel MomSmoke ;
Run;

Ods GRAPHICS On;
Proc Univariate data = Validate normal mode plot;
      title 'Cross-Validation Validate Validation Sets: New RANUNI
number';
      title2 'Validate Dataset by Education Level';
      by MomEdLevel MomSmoke;
      var Weight;
      histogram / normal;
Run;
Ods GRAPHICS Off;
/*Square transformation*/

Data BWeight;
set sashelp.BWeight;
Sq_weight = Weight**2;
RUN;

Proc Format;
```

```
      value Smoke 0 = 'Non-smoker' 1 = 'Smoker';
Run;

Data Analysis Validate;
Set BWeight;
Retain k 10000 n 20000;
If RANUNI(99999) < k/n THEN DO;
k = k-1;
Output Analysis;
END;
Else Output Validate;
n = n-1;
DROP k n;
Run;

Proc Sort data = Analysis;
      Format MomSmoke Smoke.;
      by MomEdLevel MomSmoke;
Run;

Ods GRAPHICS On;

Proc Univariate data = Analysis normal mode plot;
      title 'Cross-Validation Analysis Validation Sets: New RANUNI
number';
      title2 'Analysis Dataset by Smoke';
      by MomEdLevel MomSmoke;
      var Sq_weight;
      histogram / normal;
Run;
Ods GRAPHICS Off;

Proc Sort data = Validate;
      Format MomSmoke Smoke.;
      by MomEdLevel MomSmoke;
Run;

Ods GRAPHICS On;
Proc Univariate data = Validate normal mode plot;
      title 'Cross-Validation Validate Validation Sets: New RANUNI
number';
      title2 'Validate Dataset by Smoke';
      by MomEdLevel MomSmoke;
      var Sq_weight;
      histogram / normal;
Run;
Ods GRAPHICS Off;

/* cube transformation*/

Data BWeight3;
set sashelp.BWeight;
```

```
cr_wt=(weight)**(1/3);
RUN;

Proc Format;
      value Smoke 0 = 'Non-smoker' 1 = 'Smoker';
Run;

Data Analysis Validate;
Set BWeight3;
Retain k 16000 n 20000;
If RANUNI(99999) < k/n THEN DO;
k = k-1;
Output Analysis;
END;
Else Output Validate;
n = n-1;
DROP k n;
Run;

Proc Sort data = Analysis;
      Format MomSmoke Smoke.;
      by MomEdLevel MomSmoke;
Run;

Ods GRAPHICS On;

Proc Univariate data = Analysis normal mode plot;
      title 'Cross-Validation Analysis Validation Sets: New RANUNI
number';
      title2 'Analysis Dataset by Smoke';
      by MomEdLevel MomSmoke;
      var cr_wt;
      histogram / normal;
Run;
Ods GRAPHICS Off;

Proc Sort data = Validate;
      Format MomSmoke Smoke.;
      by MomEdLevel MomSmoke;
Run;

Ods GRAPHICS On;
Proc Univariate data = Validate normal mode plot;
      title 'Cross-Validation Validate Validation Sets: New RANUNI
number';
      title2 'Validate Dataset by Smoke';
      by MomEdLevel MomSmoke;
      var cr_wt;
      histogram / normal;
Run;
Ods GRAPHICS Off;
```

```
/* log transform*/

Data BWeight2;
set sashelp.BWeight;
Log_wt=log(weight);
RUN;

Proc Format;
      value Smoke 0 = 'Non-smoker' 1 = 'Smoker';
Run;

Data Analysis Validate;
Set BWeight2;
Retain k 10000 n 20000;
If RANUNI(99999) < k/n THEN DO;
k = k-1;
Output Analysis;
END;
Else Output Validate;
n = n-1;
DROP k n;
Run;

Proc Sort data = Analysis;
      Format MomSmoke Smoke.;
      by MomEdLevel MomSmoke;
Run;

Ods GRAPHICS On;

Proc Univariate data = Analysis normal mode plot;
      title 'Cross-Validation Analysis Validation Sets: New RANUNI
number';
      title2 'Analysis Dataset by Smoke';
      by MomEdLevel MomSmoke;
      var Log_wt;
      histogram / normal;
Run;
Ods GRAPHICS Off;

Proc Sort data = Validate;
      Format MomSmoke Smoke.;
      by MomEdLevel MomSmoke;
Run;

Ods GRAPHICS On;
Proc Univariate data = Validate normal mode plot;
      title 'Cross-Validation Validate Validation Sets: New RANUNI
number';
      title2 'Validate Dataset by Smoke';
      by MomEdLevel MomSmoke;
      var Log_wt;
      histogram / normal;
Run;
Ods GRAPHICS Off;
```