Predicting White Blood Cell Count

Isaac Mower

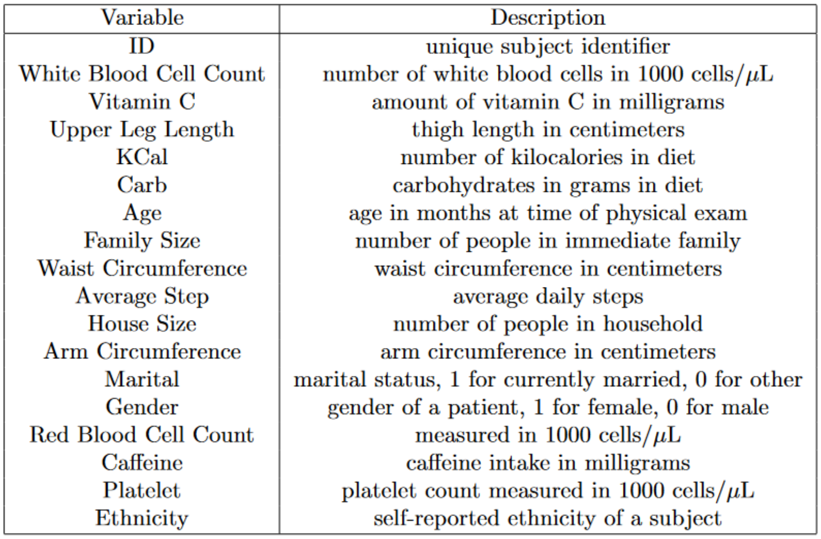
November 11th, 2023

Introduction

The National Health and Nutrition Examination Survey (NHANES) is an observational study run by the Centers for Disease Control and Prevention (CDC). The NHANES survey was designed to check up on the health and nutrition of children and adults in the United States of America (US). This survey is unique as it combines physical examinations and interviews (*Healthy People 2030*). According to the CDC, the NHANES survey is “used to determine the prevalence of major diseases and risk factors for diseases. Information will be used to assess nutritional status and its association with health promotion and disease prevention,” (*Centers for Disease Control and Prevention*). Simply put, this survey was used to help us understand our overall health as a nation as well as the health of smaller groups within the nation. Specifically, we are using the NHANES survey results to help us predict an individual’s white blood cell count. According to the University of Rochester, a high white blood cell count means that an individual may be currently fighting an illness while a low count could mean that the individual is more susceptible to disease (*University of Rochester Medical Center*). This paper aims to construct and validate a model that could be used to better predict White Blood Cell count from many variables. This will aid in helping doctors and researchers better understand how white blood cell populations are distributed, allowing them to better treat patients.

Data

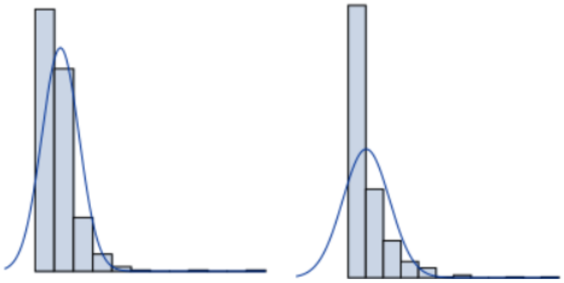
The NHANES survey observes and measures a total of 18 variables. These variables, as well as their descriptions and units of measure can be found in *Table 1*. Some variables of note are our predicted variable, white blood cell count, as well as red blood cell count and platelets.



Variable	Description
ID	unique subject identifier
White Blood Cell Count	number of white blood cells in 1000 cells/μL
Vitamin C	amount of vitamin C in milligrams
Upper Leg Length	thigh length in centimeters
KCal	number of kilocalories in diet
Carb	carbohydrates in grams in diet
Age	age in months at time of physical exam
Family Size	number of people in immediate family
Waist Circumference	waist circumference in centimeters
Average Step	average daily steps
House Size	number of people in household
Arm Circumference	arm circumference in centimeters
Marital	marital status, 1 for currently married, 0 for other
Gender	gender of a patient, 1 for female, 0 for male
Red Blood Cell Count	measured in 1000 cells/μL
Caffeine	caffeine intake in milligrams
Platelet	platelet count measured in 1000 cells/μL
Ethnicity	self-reported ethnicity of a subject

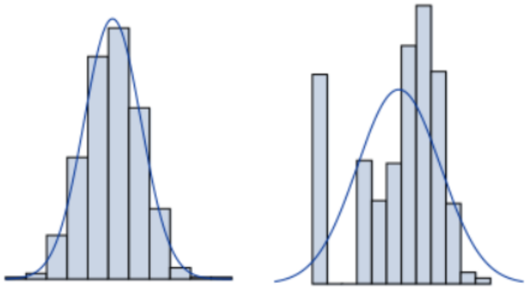
Table 1: a table displaying the variables in NHANES as well as their descriptions.

After looking at the histograms of each X and Y variable, I discovered that both Vitamin C and caffeine were right skewed, as evidenced by *Figures 1 and 2.* This right skewed data violates the assumptions of Ordinary Least Square (OLS) regression. To continue, this must be fixed.



Figures 1 and 2: Figure 1 (left) shows the distribution of Vitamin C. Figure 2 (right) shows the distribution of Caffeine.

I applied a $x^{\frac{1}{4}}$ transformation to Vitamin C. I chose the $x^{\frac{1}{4}}$ transformation because it was suggested by the Box Cox transformation. Below in *Figure 3* we can see that this transformation reshapes the data to look much more normal than before. Again, because of the Box Cox transformation, I also applied a $x^{\frac{1}{5}}$ transformation to caffeine. As shown below in *Figure 4*, aside from the spike on the left, the data is approximately normal. After trying other transformations, such as the tangent, natural log, and various exponentials, I discovered that the spike on the left of *Figure 4* never disappeared. After investigating the data set, I found that the spike is present because of the large amount of people that do not consume caffeine in a day. From this I concluded that the only way to remove the spike at zero was to remove more observations than a reasonable person would. So, I settled on the $x^{\frac{1}{5}}$ transformation as it made caffeine the most normal.



Figures 3 and 4: the histograms of the transformed variables Vitamin C (left) and Caffeine (right).

The collinearity diagnostics, namely the Variance Inflation Factor and the Condition Index, reported that we had multicollinearity issues with kilocalories and carbohydrates. There were also issues with family size and household size. Furthermore, there were multicollinearity issues with waist circumference and arm circumference.

I chose to drop carbohydrates, household size, and arm circumference. Carbohydrates was dropped from the model because it is a subset of kilocalories and kilocalories was more statistically significant with white blood cell count. Household size was dropped because it is measuring the same thing as family size and family size was more statistically significant. I chose to drop arm circumference because waist circumference is more significantly related to white blood cell count.

After transforming Vitamin C and caffeine as well as dropping carbohydrates, household size, and arm circumference, I found that all the assumptions were met. Those assumptions were the constant variance assumption and the assumption of the normality of the residuals. This means that we can move onto the next step, model selection.

Model and Validations

First, the data was randomly split into training and test data sets. The training data set was used to fit a model and I kept 20% of the data in the test data set for later validation. The split in the data allows us to determine how well the model can make predictions for new data observations and means.

Based on our knowledge of blood in the human body, I thought that red blood cell counts and platelets could have an interaction. Also, from our knowledge of human anatomy and physiology, we thought that there could be an interaction between upper leg length and waist circumference. I also believe there is an interaction between marital status and family size. After running OLS, I found that the interaction between red blood cell counts and platelets to be significant. The interaction between marital status and family size was also significant. Since the interaction between upper leg length and waist circumference was not significant, I ran a subset F-Test to see if it could be dropped from the model. The subset F-Tests yielded a p-value of 0.34, meaning that we could drop the interaction between upper leg length and waist circumference from our model.

After conducting backwards selection, stepwise selection, and all possible regression, I found that the best model included the intercept as well as these variables: Red Blood Cell Count; Platelet Count; the interaction between Red Blood Cell Count and Platelet Count; ethnicities Black and Latino; Marital Status; the interaction between Marital Status and Family Size; the Age of a subject. *Equations 1 and 2* show both the theoretical and predicted model.

These variables were selected for the model because they appeared in both the backwards and stepwise selection. This version of the model also had relatively low Akaike information criterion (AIC) as well as possessing a relatively high adjusted R-square when compared to other models.

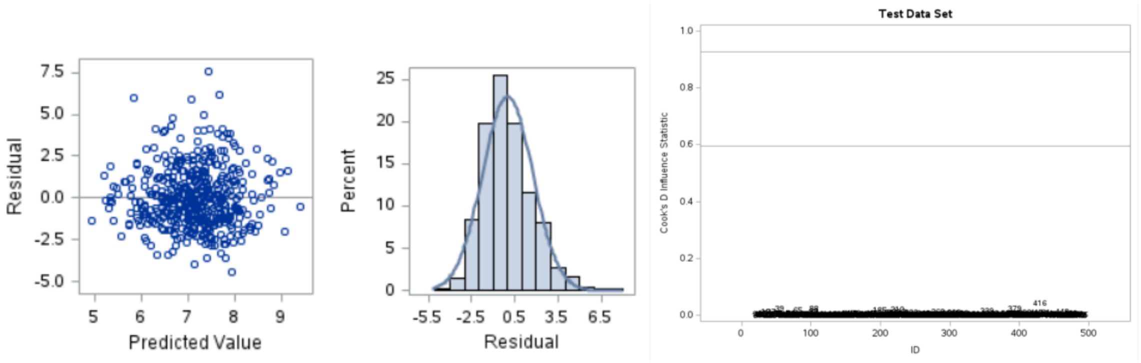
1) $$\text{WBC}=\beta_0+\beta_1(\text{RBC})+\beta_2(\text{Platelet})+\beta_3(\text{RBC})*(\text{Platelet})+\beta_4(\text{Latino})+\beta_5(\text{Black}) +\beta_6(\text{Marital})+\beta_7(\text{Marital})*(\text{Family Size})+\beta_8(\text{Age})+\epsilon$$

2) $$\widehat{WBC} = 12 - 1.2(RBC) - 0.01(Platelet) + 0.004(RBC) * (Platelet) -1.08(Black) - 0.22(Latino) - 0.14(Family\ Size) * (Marital) + 0.43(Marital) - 0.002(Age)$$

Equations 1 and 2: Equation 1 (top) is the theoretical model equation. Equation 2 (bottom) is the predicted model equation.

I combined all the data from the training and test data sets and reran OLS to get the best estimated regression function. That estimated regression function is shown in *Equation 2.* The statistically signifiacnt variables were Family Size, Age, and Ethnicity. The other variables were kept in the model because they provided more accurate results than models without those variables. *Equation 2* shows the predicted model equation as well as the predicted correlation coefficients for the variables. When we increase the Red Blood Cell count by one unit and hold all other variables constant, we predict that the White Blood Cell count will decrease 1.196 units. When we increase the Platelet count by one unit and hold all other variables constant, we predict that the White Blood Cell count will decrease 0.006 units. When we increase the age of a subject by one month and hold all variables constant, we predict that the White Blood Cell count will go down 0.002 units.

The final model confirms the constant variance assumption and confirms the assumption of the normality of residuals as evidenced by *Figures 5 and 6.* After checking the VIF and the Condition Index, I found that there were no issues with multicollinearity meaning that we can move on to checking for outliers. *Figure 7* shows the Cook's D Influence Statistic for the test data set. From this we can see that there are no outliers or influential points that we need to fix and there are no outliers or influential points that we could fix if we wanted to.



Figures 5,6, and 7:a residual vs. predicted value plot (top left) and a histogram showing the distribution of the residuals (top right) as well as the Cook's D Influence Statistic (bottom).

The model with only the intercept had a Mean Square Prediction Error (MSPR) of 10.57, while my model has an MSPR of 7.33. This validates that my model is better than the model with just an intercept. Furthermore my model has an Mean Square Error (MSE) of 3.06, since 7.33<10*3.06, my model is further validated as being better than a model with just an intercept. .

My model had an R-square of 0.151. Furthermore, the model's adjusted R-square had a value of 0.136. This means that my model can predict about 15.1% of the variation in the data. The other 84.9% of variation is due to confounding or chance. After adjusting for multiple variables, the model can only predict about 13.6% of variation in the data, leaving the other 86.4% due to random chance or confounding. This is very poor. This model is not worth the time to get predictions out of because of that low R-squared.

Conclusion

Predicting white blood cell counts is important because the count can tell if a patient is currently fighting an illness or is susceptible to an illness. Creating an accurate model to predict white blood cell counts from other variables can help doctors and researchers better identify what a patient is going through and can in turn better aid the patient in recovery. My model is not an accurate predictor of white blood cell counts. However, my model did find that white blood cell count has a significant relationship with platelet count, red blood cell count, and subject age. Because of this poor predictor, we need further research. Future research should be conducted specifically for white blood cell counts instead of including a variety of other factors that affect the nation's health, such as in the NHANES survey. Another future research direction should be towards how the platelets and red blood cells impact the white blood cell count. These future directions of research can help us better understand how white blood cell counts vary in different age groups and levels of sickness.

Citations:

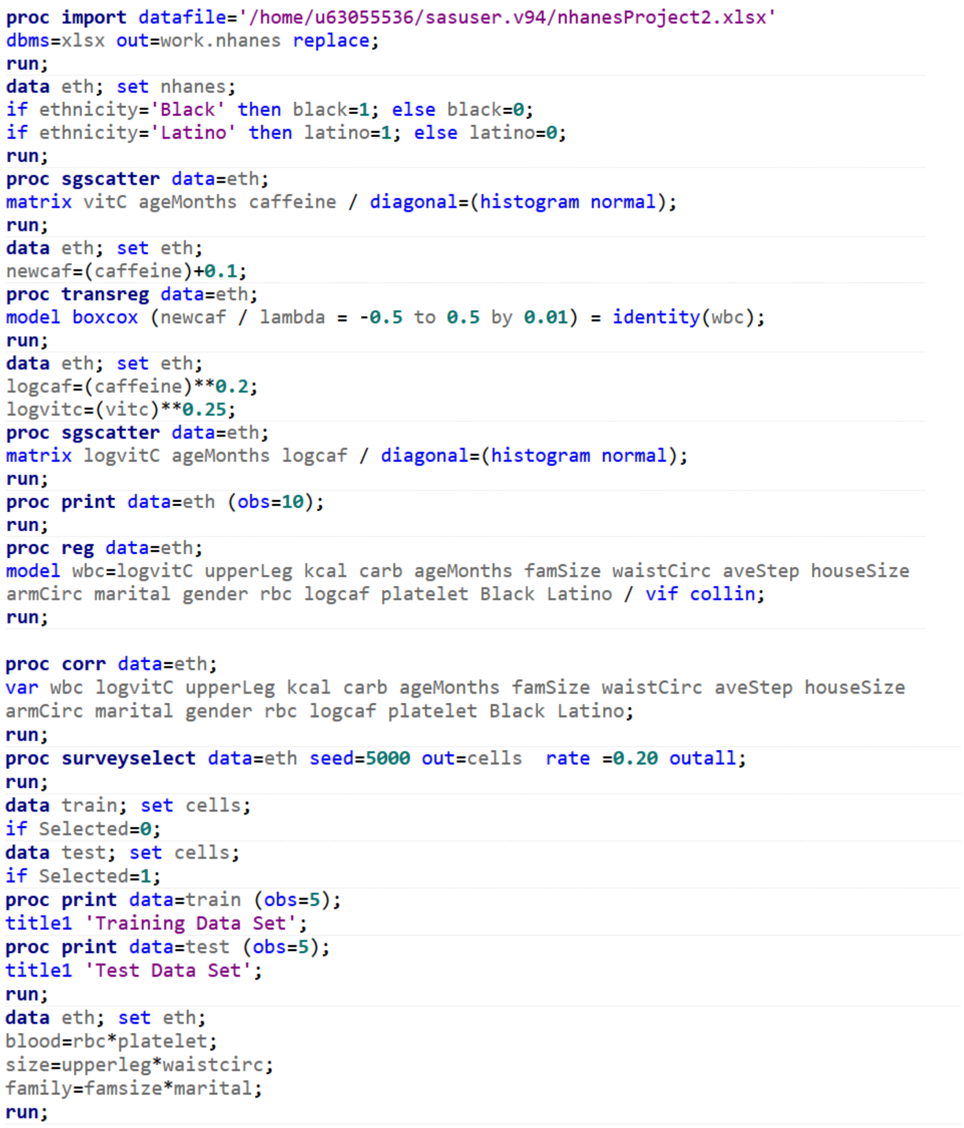
Centers for Disease Control and Prevention (CDC). (2023, May 31). *NHANES - about the National Health and Nutrition Examination Survey*. Centers for Disease Control and Prevention. https://www.cdc.gov/nchs/nhanes/about_nhanes.htm

Healthy People 2030 "National Health and Nutrition Examination Survey (NHANES)." *National Health and Nutrition Examination Survey (NHANES)*, https://health.gov/healthypeople/objectives-and-data/data-sources-and-methods/data-sources/national-health-and-nutrition-examination-survey-nhanes.

White Blood Cell count. White Blood Cell Count - Health Encyclopedia - University of Rochester Medical Center. (n.d.). https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=white_cellcount

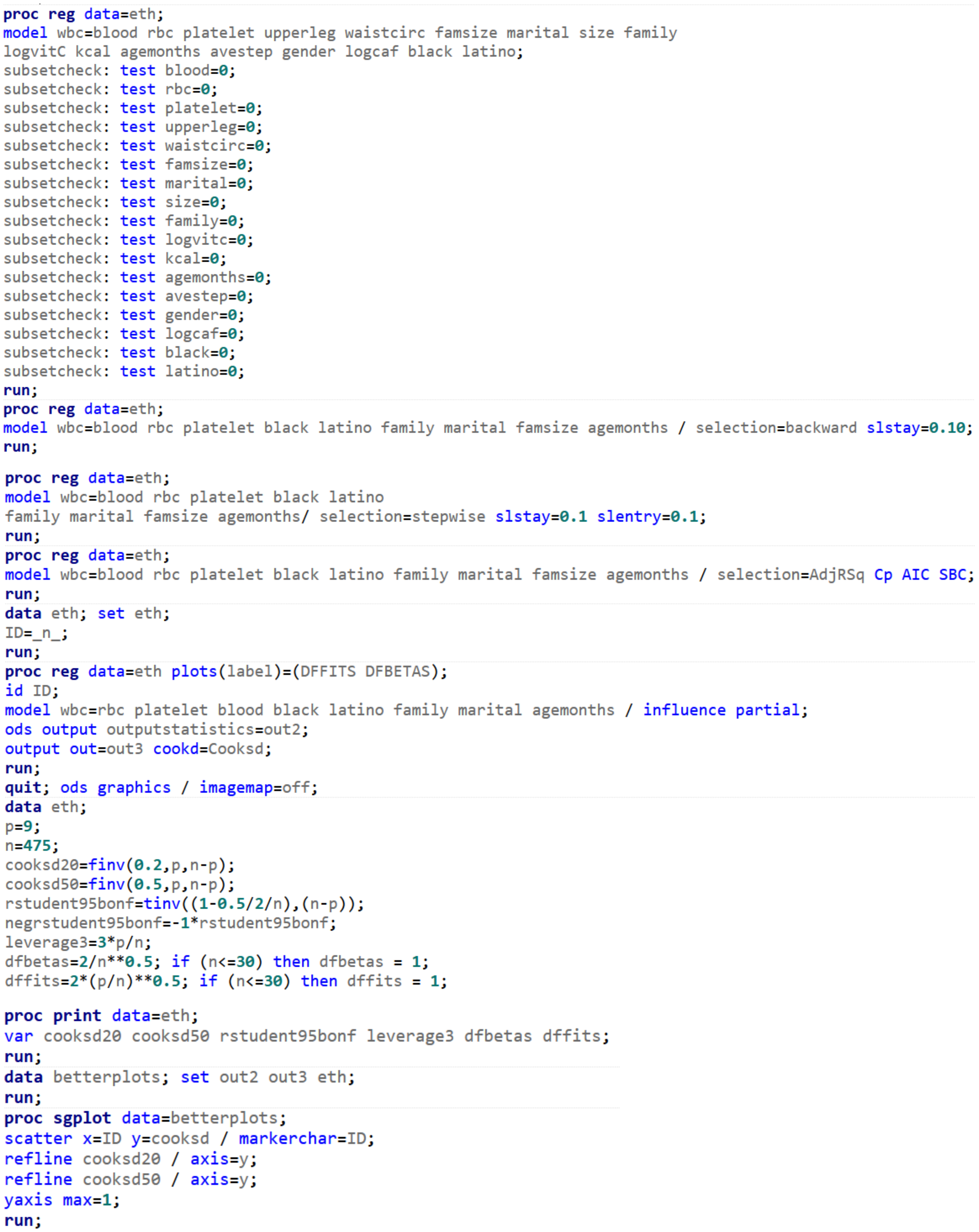
Appendix:

Below is the SAS code used to get tables, both numerical and graphical diagnostics, and p-values used in this essay.



```
proc import datafile='/home/u63055536/sasuser.v94/nhanesProject2.xlsx'
dbms=xlsx out=work.nhanes replace;
run;
data eth; set nhanes;
if ethnicity='Black' then black=1; else black=0;
if ethnicity='Latino' then latino=1; else latino=0;
run;
proc sgscatter data=eth;
matrix vitC ageMonths caffeine / diagonal=(histogram normal);
run;
data eth; set eth;
newcaf=(caffeine)+0.1;
proc transreg data=eth;
model boxcox (newcaf / lambda = -0.5 to 0.5 by 0.01) = identity(wbc);
run;
data eth; set eth;
logcaf=(caffeine)**0.2;
logvitc=(vitc)**0.25;
proc sgscatter data=eth;
matrix logvitC ageMonths logcaf / diagonal=(histogram normal);
run;
proc print data=eth (obs=10);
run;
proc reg data=eth;
model wbc=logvitC upperLeg kcal carb ageMonths famSize waistCirc aveStep houseSize
armCirc marital gender rbc logcaf platelet Black Latino / vif collin;
run;

proc corr data=eth;
var wbc logvitC upperLeg kcal carb ageMonths famSize waistCirc aveStep houseSize
armCirc marital gender rbc logcaf platelet Black Latino;
run;
proc surveyselect data=eth seed=5000 out=cells  rate =0.20 outall;
run;
data train; set cells;
if Selected=0;
data test; set cells;
if Selected=1;
proc print data=train (obs=5);
title1 'Training Data Set';
proc print data=test (obs=5);
title1 'Test Data Set';
run;
data eth; set eth;
blood=rbc*platelet;
size=upperleg*waistcirc;
family=famsize*marital;
run;
```



```
proc reg data=eth;
model wbc=blood rbc platelet upperleg waistcirc famsize marital size family
logvitC kcal agemonths avestep gender logcaf black latino;
subsetcheck: test blood=0;
subsetcheck: test rbc=0;
subsetcheck: test platelet=0;
subsetcheck: test upperleg=0;
subsetcheck: test waistcirc=0;
subsetcheck: test famsize=0;
subsetcheck: test marital=0;
subsetcheck: test size=0;
subsetcheck: test family=0;
subsetcheck: test logvitc=0;
subsetcheck: test kcal=0;
subsetcheck: test agemonths=0;
subsetcheck: test avestep=0;
subsetcheck: test gender=0;
subsetcheck: test logcaf=0;
subsetcheck: test black=0;
subsetcheck: test latino=0;
run;
proc reg data=eth;
model wbc=blood rbc platelet black latino family marital famsize agemonths / selection=backward slstay=0.10;
run;

proc reg data=eth;
model wbc=blood rbc platelet black latino
family marital famsize agemonths/ selection=stepwise slstay=0.1 slentry=0.1;
run;
proc reg data=eth;
model wbc=blood rbc platelet black latino family marital famsize agemonths / selection=AdjRSq Cp AIC SBC;
run;
data eth; set eth;
ID=_n_;
run;
proc reg data=eth plots(label)=(DFFITS DFBETAS);
id ID;
model wbc=rbc platelet blood black latino family marital agemonths / influence partial;
ods output outputstatistics=out2;
output out=out3 cookd=Cooksd;
run;
quit; ods graphics / imagemap=off;
data eth;
p=9;
n=475;
cooksd20=finv(0.2,p,n-p);
cooksd50=finv(0.5,p,n-p);
rstudent95bonf=tinv((1-0.5/2/n),(n-p));
negrstudent95bonf=-1*rstudent95bonf;
leverage3=3*p/n;
dfbetas=2/n**0.5; if (n<=30) then dfbetas = 1;
dffits=2*(p/n)**0.5; if (n<=30) then dffits = 1;

proc print data=eth;
var cooksd20 cooksd50 rstudent95bonf leverage3 dfbetas dffits;
run;
data betterplots; set out2 out3 eth;
run;
proc sgplot data=betterplots;
scatter x=ID y=cooksd / markerchar=ID;
refline cooksd20 / axis=y;
refline cooksd50 / axis=y;
yaxis max=1;
run;
```