Project 2 Paper - Stat 5100

Samuel Johnson

October 30, 2021

1 Introduction

White blood cells are an integral part of the human body. Although they make up just 1% of our total blood, they are largely responsible for fighting infectious diseases and keeping us safe (Gersten 2021).

In this paper we will examine a dataset containing various features in an effort to accurately predict the number of white blood cells in the human body.

Predicting the number of white blood cells in the body could be extremely useful. If a doctor could know beforehand that one of his patients was going to have a low white blood cell count, he might prescribe certain medications, suppliments, or diets that would prevent this low white blood cell count from ever occurring. This would strengthen the patient's immune system from infectious diseases.

1.1 Variables of Interest

We will be using a subset of the National Health and Nutrition Examination Survey (NHANES) dataset, which was collected by the National Center for Health Statistics (NCHS), to perform our analysis (CDC 2017). A description of the various variables is given below:

Variable	Description
ID	Unique subject identifier
wbc	Number of white blood cells (1000 cells/μL)
vitC	Vitamin C (mg) consumed on day one
upperLeg	Length of upper leg (cm)
kcal	Calories consumed on day one
carb	Carbohydrate (gm) consumed on day one
ageMonths	Age (months) on date of physical exam
famSize	Number of people in family
waistCirc	Waist circumference (cm)
aveStep	Average daily steps taken (averaged over 7 days, using a step counter)
houseSize	Number of people in household
armCirc	Arm circumference (cm)
married	Marital status (1 = currently married, 0 = other)
female	1 = female, 0 = male
rbc	Number of red blood cells (1000 cells/μL)
caffeine	Caffeine (mg) consumed on day one
platelet	Number of platelets SI (1000 cells/μL)
ethnicity	Self-reported ethnicity or race

Note that the $female$ and *married* attributes were originally named *gender* and *marital* respectively. They have been renamed to clarify their interpretation.

2 Linear Regression Assumptions

We begin our analysis by running a multiple linear regression model with all explanatory variables in order to check the assumptions of linear regression.

2.1 Constant Variance

The residual plot in figure 1 suggests that the residuals have constant variance but it is difficult to be certain. The Brown-Forsythe test of constant variance, on the other hand, returns a p-value of 0.051. This leads us to believe that the variance might not be constant and that we may need to perform one or more transformations.

2.2 Normality Distributed Residuals

The QQ plot in figure 1 shows that the residuals are definitely not normally distributed, with many dots lying above the line. Additionally, the correlation test of normality of residuals returned a value of 0.98649, which is less than the required 0.987 where $\alpha = 0.05$ and $n = 100$. This further demonstrates non-normality.

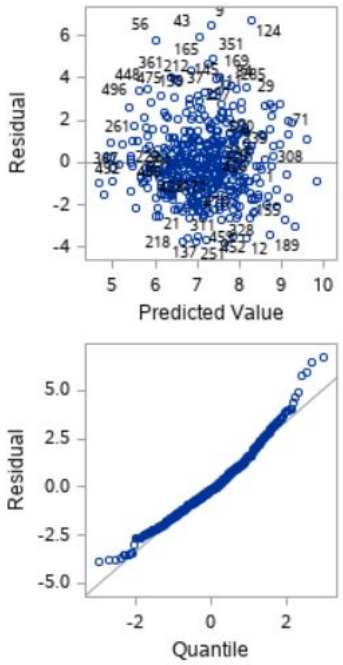

Figure 1: Residual Plot and QQ Plot

2.3 Residual Independance

We can see in figure 2 that the residuals move randomly up and down, without any recognizable pattern. This satisfies our assumption of independant residuals.

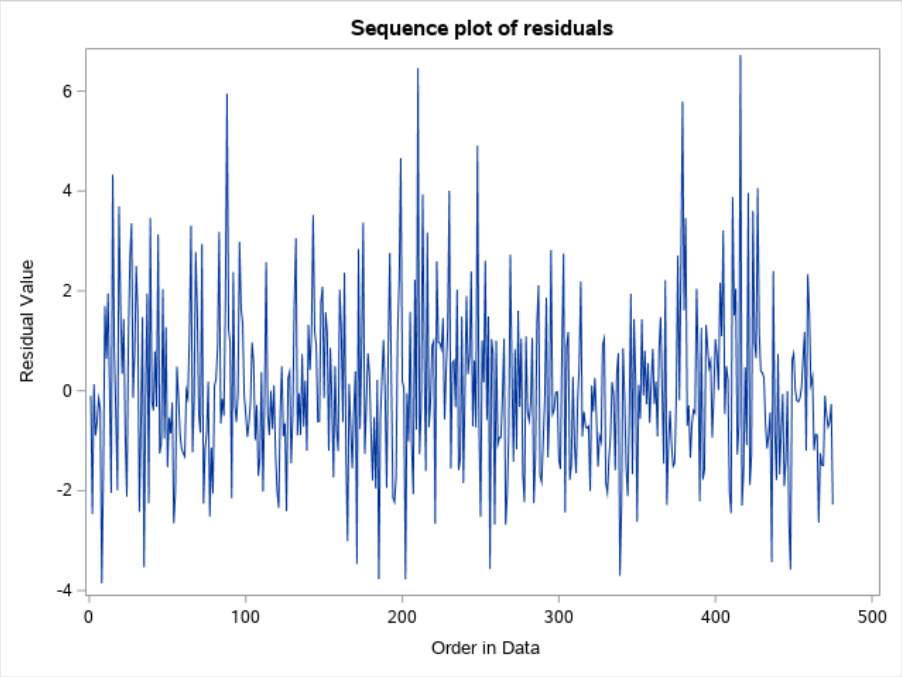

Figure 2: First Model Sequence Plot

2.4 Outliers & Influential Points

Figure 3 shows that there aren't any extremely influential points in our first model.

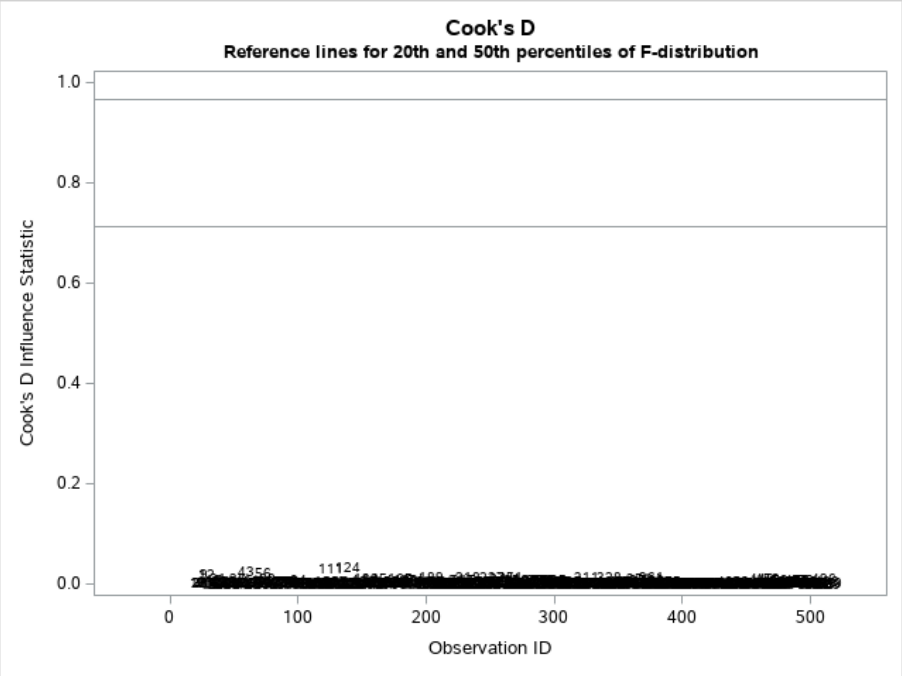

Figure 3: First Model Cook's D

Figure 4 shows that there are a few points with high leverage as well as a few outliers. We will perform some transformations later to try and resolve this issue.

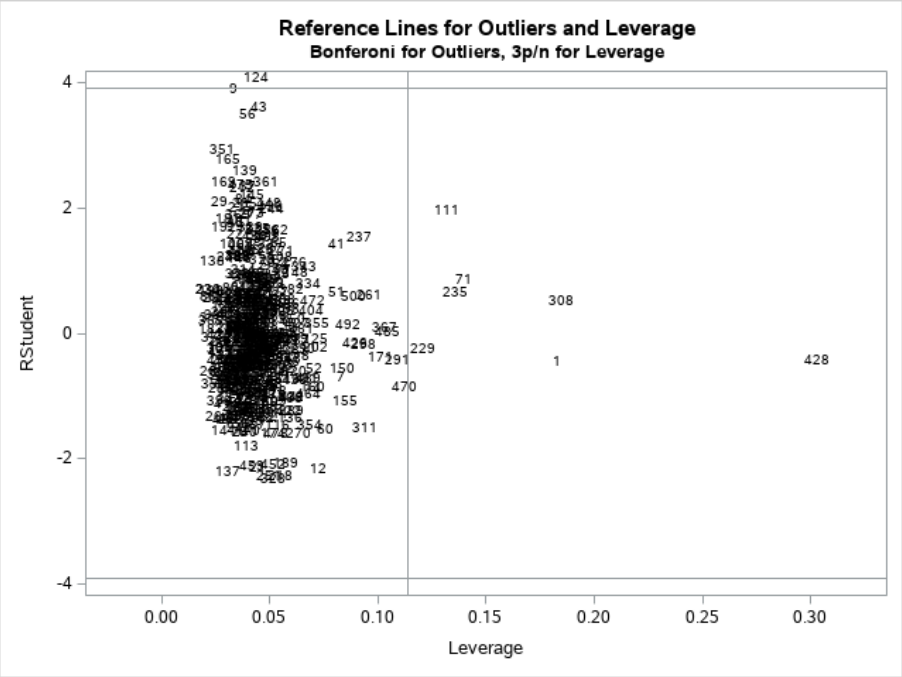

Figure 4: First Model Outliers & Leverage

2.5 Transforming Independant Variables

Upon examining histograms of each of the independant variables, we note that some of them are right skewed. Figures 5, 6, 7, 8, and 9 contain histograms of variables that exhibit right skewed behavior.

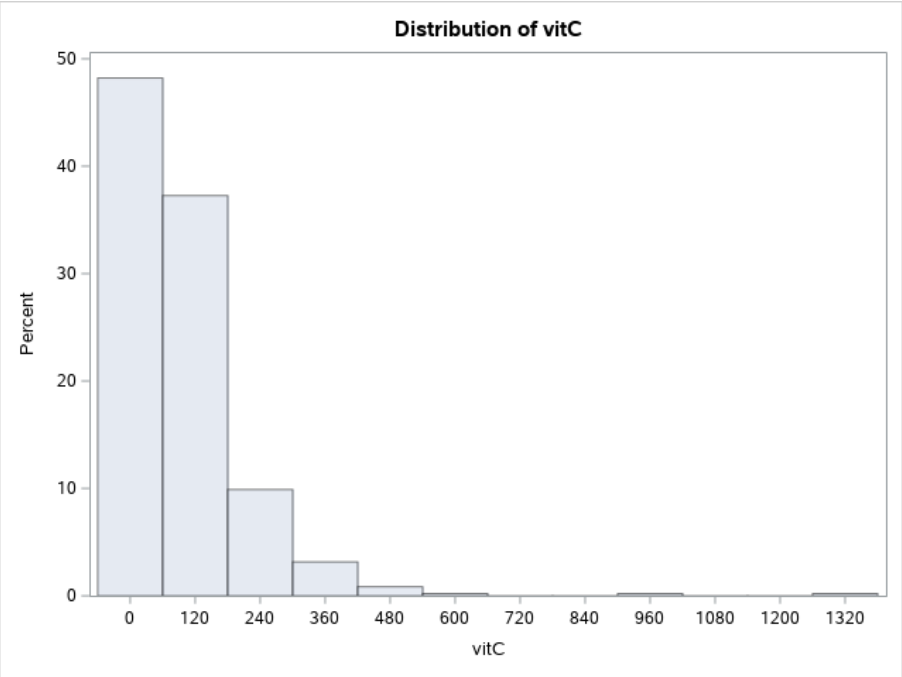

Figure 5: Vitamin C Histogram

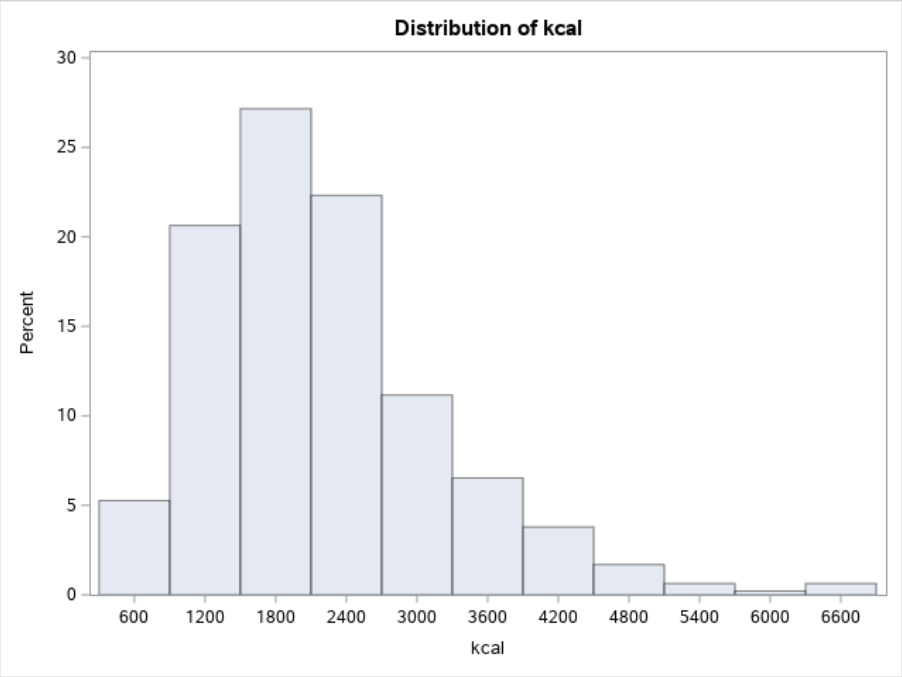

Figure 6: Calories Histogram

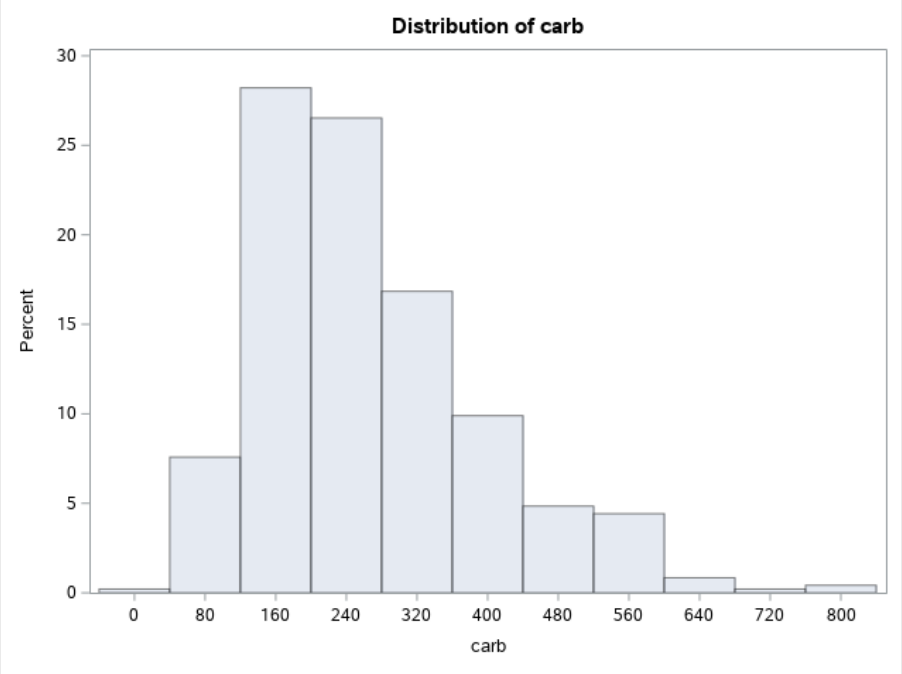

Figure 7: Carbohydrates Histogram

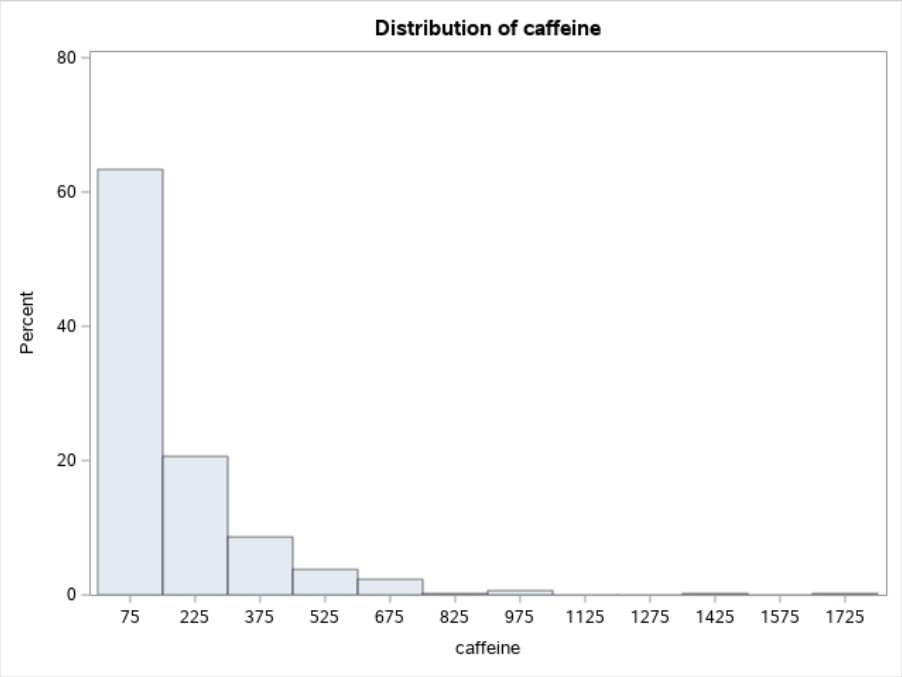

Figure 8: Caffeine Histogram

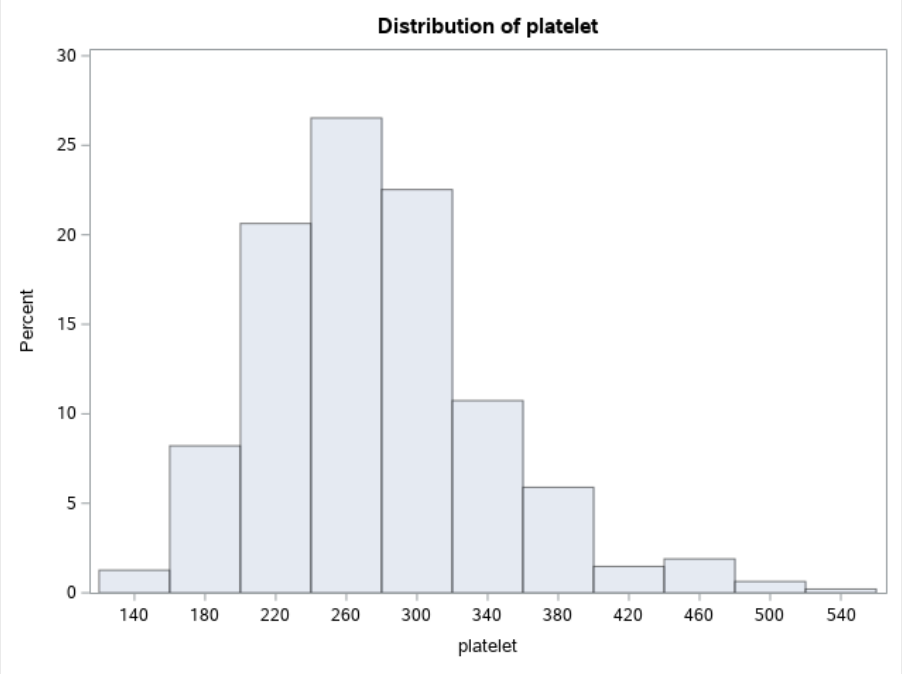

Figure 9: Platelet Histogram

To fix the right skewness problem, we took the square root of *vitC* and *caffeine*, as their minimum value was zero. We performed a log transformation on *kcal*, *carb*, and *platelet*.

2.6 Transforming the Dependant Variable

In an earlier section we noticed that the assumptions with constant variance and normality of residuals were not being met. As a result, we ran the box-cox method to determine a suitable transformation for *wbc* to see if that would fix our issue. Box cox returned a

recommended value of 0.2. Additionally, figure 10 shows that our dependant variable, *wbc*, is also right skewed.

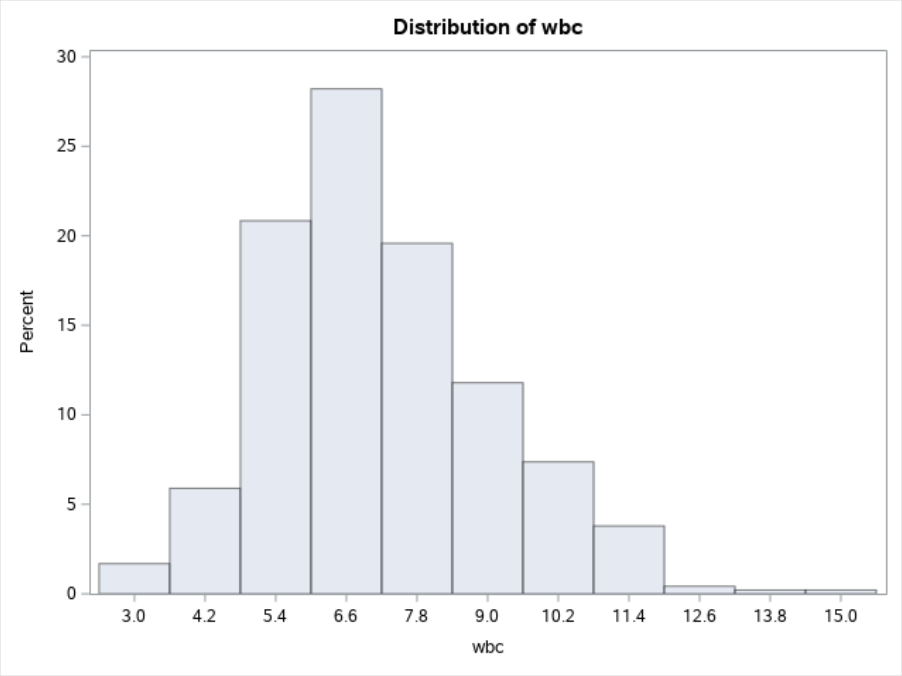

Figure 10: White Blood Cells Histogram

We performed the recommended transformation and created a new variable $wbcTrans = wbc^{0.2}$. After running another linear regression model, we check to see if our assumptions are now met.

The residual plot in figure 11 shows that the residuals now seem to have very constant variance. The Brown-Forsythe test of constant variance confirms this view with a p-value of 0.95531.

The QQ plot in figure 11 shows that the residuals now seem to be very normally distributed. The correlation test of normality confirms this with a value of 0.99771, which is higher than the minimum required value of 0.987 where $\alpha = 0.05$ and $n = 100$.

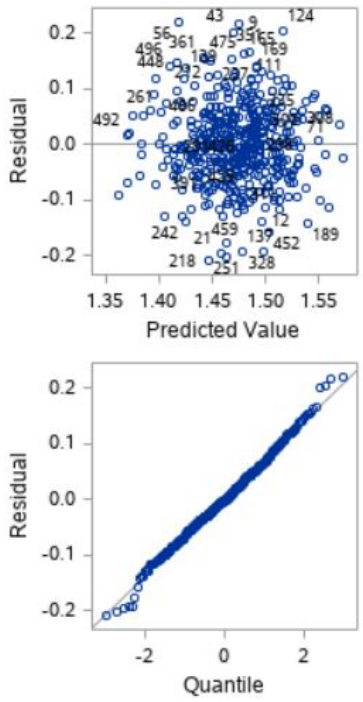

Figure 11: Transformed Model Residual Plot & QQ Plot

Transforming various variables in our model has satisfied our assumptions. We can now continue building our model in the next section.

3 Multicollinearity

After running VIF and collinearity diagnostics on our transformed model, we note that the pairs ($famSize$, $houseSize$) and ($waistCirc$, $armCirc$) exhibit multicollinearity. Several high VIF values can be seen in figure 12.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	1.11778	0.12812	8.72	<.0001	0
vitCSqrt		1	-0.00082364	0.00073890	-1.11	0.2656	1.21247
upperLeg	upperLeg	1	-0.00227	0.00107	-2.13	0.0334	1.66737
kcalLog		1	-0.01329	0.01478	-0.90	0.3689	4.46060
carbLog		1	0.02091	0.01337	1.56	0.1186	4.23144
ageMonths	ageMonths	1	-0.00009199	0.00001997	-4.61	<.0001	1.81561
famSize	famSize	1	-0.00232	0.00605	-0.38	0.7012	9.61421
waistCirc	waistCirc	1	0.00177	0.00042628	4.16	<.0001	2.64155
aveStep	aveStep	1	-0.00000152	8.648805E-7	-1.76	0.0794	1.09115
houseSize	houseSize	1	0.00140	0.00627	0.22	0.8232	9.83228
armCirc	armCirc	1	-0.00339	0.00134	-2.53	0.0116	2.68944
married	married	1	0.00079074	0.00698	0.11	0.9099	1.15693
female	female	1	0.00894	0.00965	0.93	0.3550	2.24013
rbc	rbc	1	0.00772	0.00810	0.95	0.3412	1.67925
caffeineSqrt		1	0.00007217	0.00047064	0.15	0.8782	1.24445
plateletLog		1	0.07442	0.01504	4.95	<.0001	1.14559
black		1	-0.04544	0.00989	-4.59	<.0001	1.32141
latino		1	-0.01995	0.00890	-2.24	0.0254	1.51031

Figure 12: Transformed Model VIF

Because of this collinearity, we will need to remove one variable from each of these pairs. To keep things simple, we will remove the two variables that are least significant in our transformed linear model.

$famSize$ has a p-value of 0.6651 and $houseSize$ has a p-value of 0.8552. Additionally, $waistCirc$ has a p-value of $< .0001$ while $armCirc$ has a p-value of 0.0090. We removed $houseSize$ and $armCirc$ from the model.

After removing these variables from the model, we run another regression model. We see in figure 13 that the high VIF values have dissappeared.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	1.11259	0.12846	8.66	<.0001	0
vitCSqrt		1	-0.00101	0.00073912	-1.36	0.1743	1.20114
upperLeg	upperLeg	1	-0.00242	0.00107	-2.26	0.0242	1.66160
kcalLog		1	-0.01235	0.01484	-0.83	0.4060	4.45769
carbLog		1	0.02102	0.01344	1.56	0.1185	4.22998
ageMonths	ageMonths	1	-0.00008267	0.00001918	-4.31	<.0001	1.65914
famSize	famSize	1	-0.00153	0.00228	-0.67	0.5024	1.35548
waistCirc	waistCirc	1	0.00098920	0.00029517	3.35	0.0009	1.25390
aveStep	aveStep	1	-0.00000163	8.674906E-7	-1.88	0.0611	1.08685
married	married	1	0.00080731	0.00699	0.12	0.9081	1.14686
female	female	1	0.01046	0.00963	1.09	0.2781	2.20828
rbc	rbc	1	0.00346	0.00797	0.43	0.6646	1.60834
caffeineSqrt		1	-0.00003001	0.00047121	-0.06	0.9492	1.23512
plateletLog		1	0.07255	0.01510	4.81	<.0001	1.14280
black		1	-0.04919	0.00982	-5.01	<.0001	1.28965
latino		1	-0.01902	0.00891	-2.14	0.0332	1.49814

Figure 13: VIF After Removal of houseSize and armCirc

4 Variable Interaction

Based on our knowledge of white blood cells, the following variables are likely to interact with each other and explain additional variation in the model. The columns *Variable 1* and *Variable 2* represent the two interacting variables. The column *New Variable* represents the name of the new interaction term.

Variable 1	Variable 2	New Variable
$log(carb)$	$log(kcal)$	$kcalLog\_carbLog$
$ageMonths$	$female$	$ageMonths\_female$
$married$	$waistCirc$	$married\_waistCirc$

We fit a model with the interaction variables that we created. The resulting p-values are given below:

$P(kcalLog\_carbLog) = 0.5407$
$P(ageMonths\_female) = 0.6114$
$P(married\_waistCirc) = 0.3873$

We can see that none of these variables are statistically significant, so we run a subset F-test to see if we can drop them all at once. Our hypotheses and results for

the test are given below. Note that we refer to the parameters of $kcalLog\_carbLog$, $ageMonths\_female$, and $married\_waistCirc$ as β_1, β_2, and β_3 respectively.

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$
H_a: $(\beta_1 \vee \beta_2 \vee \beta_3) \neq 0$
$P = 0.7142$

Where $\alpha = 0.05$, $P > \alpha$ and we fail to reject the null hypothesis that $\beta_1 = \beta_2 = \beta_3 = 0$. We conclude that the three interaction terms together aren't statistically significant and we drop each of them from the model.

5 Model Selection

We split the data and put 80% and 20% into the training and testing sets respectively. By withholding 20% of the data, we can evaluate how well our trained model can predict on unseen data.

We will now use the training set to perform backward selection, stepwise selection, and all possible regressions. We will analyze the results and determine which model is best.

5.1 Backward Selection

Backward selection at the 10% level generated a model with the following 7 explanatory variables:
$upperLeg$, $ageMonths$, $waistCirc$, $aveStep$, $plateletLog$, $black$, $latino$
$R^2_{adj} = 0.1666$

5.2 Stepwise Selection

Stepwise selection with an entry condition of 10% and an exit condition of 10% generated a model with the following 7 predictor variables:
$ageMonths$, $famSize$, $waistCirc$, $aveStep$, $female$, $plateletLog$, $black$
$R^2_{adj} = 0.1651$

5.3 All Possible Regressions

Next, we ran all possible regressions. We evaluated the results using R^2_{adj}, $C(p)$, AIC, and SBC.

The best model according to R^2_{adj} contained the following 11 variables:
$vitCSqrt$, $upperLeg$, $kcalLog$, $carbLog$, $ageMonths$, $famSize$, $waistCirc$, $aveStep$, $plateletLog$, $black$, $latino$
$R^2_{adj} = 0.1713$

The best model according to $C(p)$ contained the following 8 variables:
$upperLeg$, $ageMonths$, $famSize$, $waistCirc$, $aveStep$, $plateletLog$, $black$, $latino$
$R^2_{adj} = 0.1693$

The best model according to AIC contained the following 8 variables:
$upperLeg$, $ageMonths$, $famSize$, $waistCirc$, $aveStep$, $plateletLog$, $black$, $latino$

$R^2_{adj} = 0.1693$

The best model according to SBC contained the following 7 variables:
$upperLeg$, $ageMonths$, $waistCirc$, $aveStep$, $plateletLog$, $black$, $latino$
$R^2_{adj} = 0.1666$

5.4 The Best Model

We determine that the 7 variable model chosen by both backward selection and SBC is the best model for our purpose. It is the best for a few reasons. All of the variables are significant at the 5% level and it only has a slightly lower R^2_{adj} value than the top R^2_{adj} model, $(0.1713 > 0.1666)$ but more than makes up for it by having four fewer predictor variables. The theoretical equation of our model is given below:

$wbcTrans = \beta_0 + \beta_1 * upperLeg + \beta_2 * ageMonths + \beta_3 * waistCirc + \beta_4 * aveStep + \beta_5 * plateletLog + \beta_6 * black + \beta_7 * latino + \epsilon$

5.5 Best Model Assumptions

Now that we have chosen a *best* model, we need to recheck our assumptions one more time to ensure that we can make unbiased predictions and inference.

We can see in the upper half of figure 14 that the chosen model's residuals have constant variance. The Brown-Forsythe test of constant variance backs up this view with a p-value of 0.70812.

The bottom half of figure 14 shows that the residuals are normally distributed. Additionally, the correlation test of normality gives a value of 0.997, which is greater than the minimum required value of 0.987 where $\alpha = 0.05$ and $n = 100$.

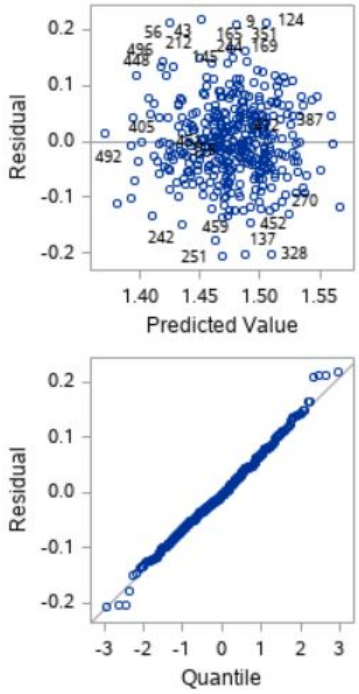

Figure 14: Best Model Residual Plots

Figure 15 shows us that not a single value reaches above the 20% threshold much less the 50% threshold of the f-distribution, showing us that the model has no highly influential points.

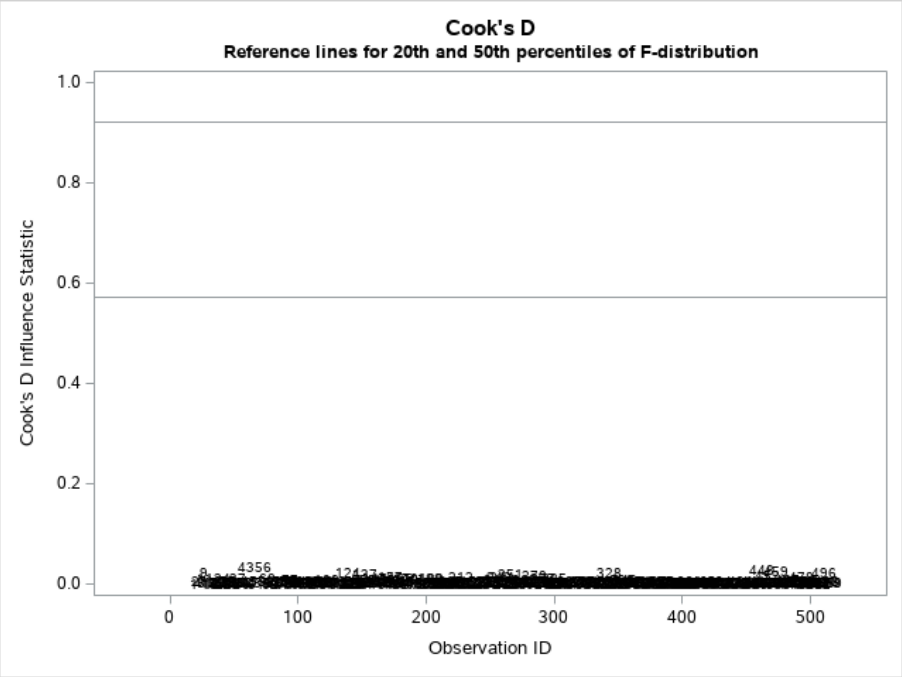

Figure 15: Best Model Cook's D

In figure 16 we can see that no points lie past either horizontal line or vertical line, demonstrating an absence of outliers and influential points.

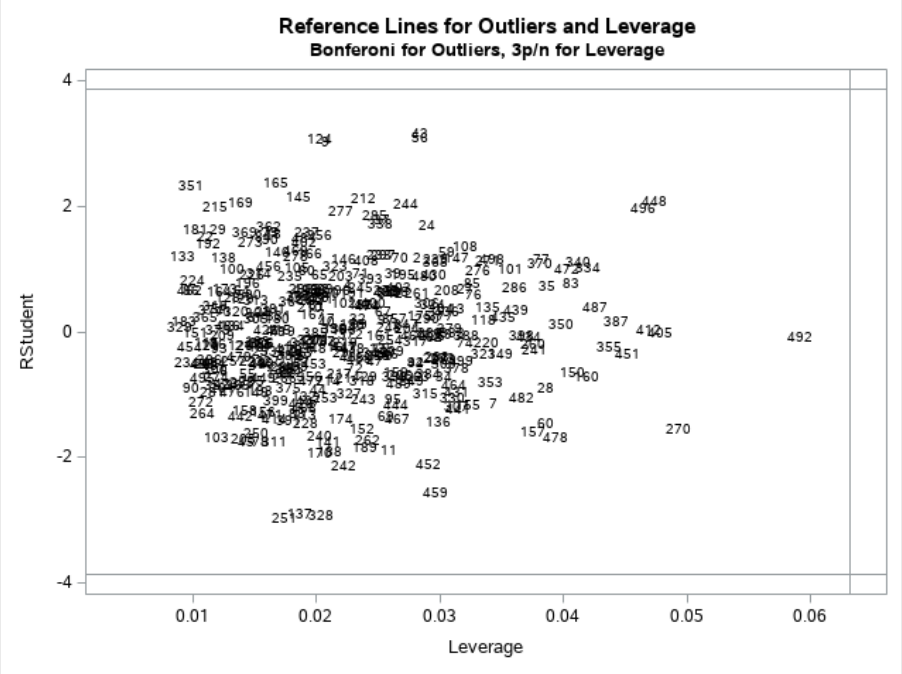

Figure 16: Best Model Outliers & Leverage

The sequence plot in figure 17 moves up and down randomly, suggesting that the observations are independant from one another.

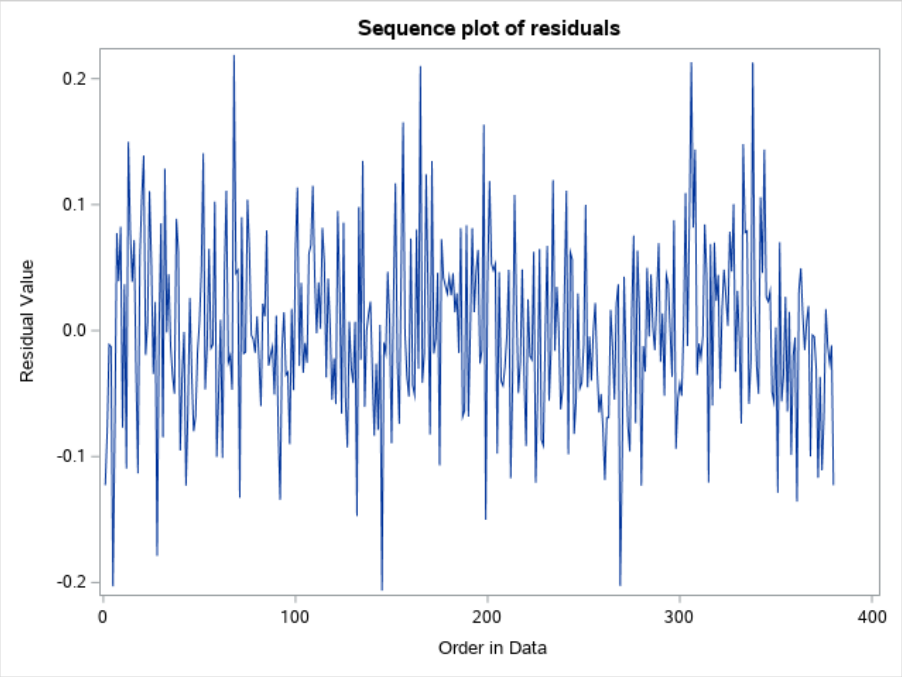

Figure 17: Best Model Sequence Plot

In conclusion, our final model satisfies the key assumptions of OLS, leaving us with unbiased coefficients and the ability to make inference.

6 Model Validation

We previously divided our dataset into training and testing portions. After fitting our model using the training data, we will use it to make predictions on our test data. Below we give the error rate (how far predictions are from actual values) for the training set (MSE) and the test set ($MSPR$).

$MSE = 0.00504$
$MSPR = 0.0047155$

The fact that the $MSPR$ is actually lower than the MSE shows that the model actually performed slightly better on the testing data than it did on the training data. This shows us that our model generalizes quite nicely on unseen data and has some degree of predictive ability.

We also created a model with just an intercept and no parameters. The intercept only model resulted in an $MSPR$ of 0.0060350, which is higher than the error rate of both our training set and our testing set. This shows that our chosen model beats the baseline of a simple intercept-only model.

7 Results

Now we combine our training and testing datasets together to fit our best model one more time and determine the optimal parameters. The resulting model where $wbcTrans = wbc^{0.2}$ is given below:

$\hat{wbcTrans} = 1.15371 - 0.00283 * upperLeg - 0.00008599 * ageMonths + 0.00092289 * waistCirc - 0.00000176 * aveStep + 0.07507 * plateletLog - 0.04926 * black - 0.01968 * latino$

We can now use our estimated equation to interpret the slope of several of our variables. For example:

If $black = 1$, then $wbsTrans$ is expected to be 0.04926 units lower than that of someone who is white, holding all other variables constant. For each unit increase in $upperLeg$, $wbcTrans$ is expected to decrease by 0.00283 units, holding all other variables constant. For each unit increase in $plateletLog$, $wbcTrans$ is expected to increase by 0.07507, holding all other variables constant.

Where $\alpha = 0.05$, all of the variables in our model are statistically significant on an *individual* basis because $P < \alpha$. Our final model generated an R^2 value of 0.1916 and an R^2_{adj} value of 0.1795. This means that our model explains 19.16% of the variation in our dependant variable, $wbcTrans$.

In figure 18, we can see a scatterplot of the actual wbc values on the x-axis plotted against the predicted wbc values on the y-axis. There is clearly a lot of variation between the actual value and the predicted value. This is what we would expect, given an R^2 value of just 0.1916.

Even though this model has low predictive ability, it was still worth our time. We did gain the ability of *inference* through this research. We can analyze the partial affects of any of our explanatory variables on $wbcTrans$.

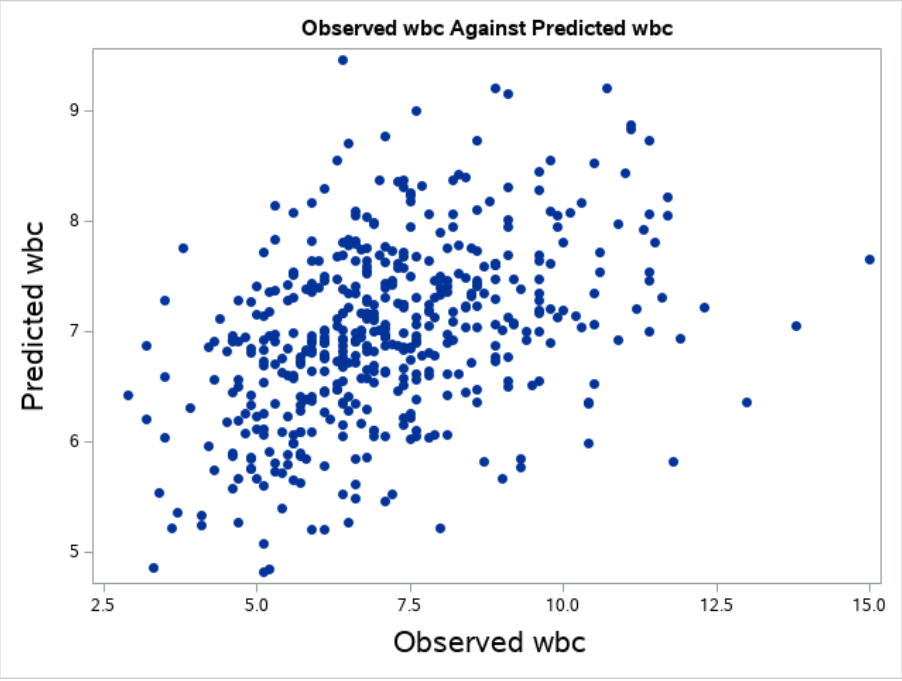

Figure 18: Best Model Actual vs Predicted Values

8 Conclusion

White blood cells are the body's protectors. If we can learn to predict how many there are, we might be able to anticipate when levels are about to decrease or locate individuals

with potentially low levels. With this information, doctors could prescribe medications, diets, or activities to boost the white blood cell count and fight off infections.

In our chosen model we located seven variables that have some predictive ability. They were $upperLeg, ageMonths, waistCirc, aveStep, plateletLog, black, latino$. Unfortunately, as of now our model isn't able to accurately predict the white blood cell count in the body, as it explains just 19.16% of its variation. More research is needed in the area if we are to accurately predict white blood cell levels.

More research is definitely needed in this area to improve predictions. One possible approach would be to put together a team of hematologists, doctors specializing in blood disorders, and several other types of doctors. The doctors would determine what variables might be useful. Then we could take their recommendations and collect relevant data in order to improve our model.

Another approach could be to analyze already existing publications or datasets on white blood cell prediction. This could shorten the research period and help us identify variables with high predictive value.

9 Bibliography

Gersten, Todd. "What Are White Blood Cells?" What Are White Blood Cells? - Health Encyclopedia - University of Rochester Medical Center, https://www.urmc.rochester.edu/encyclopedia/content.aspx?ContentID=35&ContentTypeID=160

Centers for Disease Control and Prevention. (2017, September 15). About the National Health and Nutrition Examination Survey. Centers for Disease Control and Prevention. Retrieved November 13, 2021, from https://www.cdc.gov/nchs/nhanes/about_nhanes.htm

10 Appendix: SAS Code

```
/* Project 2 */

/************************************ Introduction ********************************

/* Import the data */
proc import datafile='/home/u59308923/Assignments/Project 2 Data.xlsx'
dbms=xlsx
out=cells;
getnames=yes;
run;

/* We setup a dummy variable for ethnicity. */
data cells; set cells;
if ethnicity='Black' then black=1; else black=0;
if ethnicity='Latino' then latino=1; else latino=0;
run;

/* We take a look at the data. */
proc print data=cells(obs=5);
run;

/************************************ Data Exploration ****************************

/* Correlation Matrix */
proc corr data=cells;
run;

/* Make Scatterplot 1 */
proc sgscatter data=cells;
compare x=(vitC upperLeg kcal carb ageMonths) y=wbc;
run;

/* Make Scatterplot 2 */
proc sgscatter data=cells;
compare x=(famSize waistCirc aveStep houseSize armCirc married) y=wbc;
run;

/* Make Scatterplot 3 */
proc sgscatter data=cells;
compare x=(female rbc caffeine platelet black latino) y=wbc;
run;

/********************* First Model Outlier Examination ****************************

/* We run it to generate output. */
ods graphics on / imagemap=on;
proc reg data=cells plots(label) = (DFFITS DFBETAS);
id ID;
```

```
model wbc = vitC upperLeg kcal carb ageMonths famSize waistCirc aveStep houseSize arm
married female rbc caffeine platelet black latino / influence partial;
title1 'First Model Check for Outliers';
title2 'wbc = vitC upperLeg kcal carb ageMonths famSize waistCirc aveStep houseSize a
married female rbc caffeine platelet black latino';
ods output outputstatistics=out2;
output out=out3 p=pred r=resid cookd=CooksD;
run; quit;
ods graphics / imagemap=off;

/* Alternative thresholds for influential obs. and outlier diagnostics */
data temp;
p=18; /* p = # beta's (incl. intercept */
n = 475; /* n = sample size */
CooksD20 = finv(.20,p,n-p);
CooksD50 = finv(.50,p,n-p);
RStudent95Bonf = tinv((1-.05/2/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;

/* Printing our alternative thresholds. */
proc print data=temp;
var CooksD20 CooksD50 RStudent95Bonf  Leverage3 DFBETAS DFFITS;
title1 'Alternative thresholds';
run;

/* Combine several other datasets. */
data betterplots; set out2 out3 temp;
run;

/*Make Plot with Better Cook's D Reference Lines */
proc sgplot data=betterplots;
scatter x=ID y=cooksD / markerchar=ID;
xaxis label = 'Observation ID';
yaxis label = "Cook's D";
title1 "Cook's D";
title2 'Reference lines for 20th and 50th percentiles of F-distribution';
refline cooksD20 / axis=Y;  /*20th percentile*/
refline cooksD50 / axis=Y;  /*50th percentile*/
yaxis max=1;
run;

/*Make Plot with Better Studentized Deleted Residuals and Leverage Lines */
proc sgplot data=betterplots;
scatter x=HatDiagonal y=RStudent / markerchar=ID;
xaxis label = 'Leverage';
yaxis label = 'Studentized Deleted Residuals';
title1 'Reference Lines for Outliers and Leverage';
```

```
title2 'Bonferoni for Outliers, 3p/n for Leverage ';
refline RStudent95Bonf / axis=Y;  /*Upper limit outliers*/
refline NegRStudent95Bonf / axis=Y;  /*lower limit outliers*/
refline Leverage3 / axis=X;  /* limit leverage*/
yaxis max=4 min=-4;
run;

/********************************** Other Assumption Checking **********************

/* We create the macro for running diagnostics. */
%macro resid_num_diag(dataset,datavar,label='requested variable',predvar=' ',predlabe

/* We check the diagnostics of our first model. */
%resid_num_diag(dataset=out3, datavar=resid, label='Residual',
predvar=pred, predlabel='Predicted Value');
/* We note that the model suffers from non-constant variance and non-normal residuals

/* We check the lack of fit for our first model. */
proc rsreg data=cells;
model wbc = vitC upperLeg kcal carb ageMonths famSize waistCirc aveStep houseSize arm
married female rbc caffeine platelet black latino / lackfit covar=1 noopt;
title1 'First Model F-test for lack of fit';
run;

/* Look at sequence plot */
data temp; set out3;
order = _n_;
proc sgplot data=temp;
series x=order y=resid / lineattrs=(pattern=solid);
xaxis label='Order in Data';
yaxis label='Residual Value';
title1 'Sequence plot of residuals';
run;

/*
The above output demonstrates non-constant variance and non-normally distributed resi
As a result, we will run box cox to determine if a transformation is needed on wbc.
Note: The below code returned a recommended box-cox transformation of 0.2.
*/
proc transreg data=cells;
model boxcox(wbc / lambda=-2 to 2 by 0.1)
= identity(vitC upperLeg kcal carb ageMonths famSize waistCirc aveStep houseSize armC
married female rbc caffeine platelet black latino);
title1 'Box-Cox Transformation on Simple Model';
run;

/* Make Histograms for all variables see if they are skewed. */
proc univariate data=cells noprint;
hist wbc vitC upperLeg kcal carb ageMonths famSize waistCirc aveStep houseSize armCir
title1 'Predictor Variable Histograms';
```

```
run;

/*
We now know that skewness exists in many of our variables and that our assumptions ar
We attempt some transformations to fix these issues.
*/
data cells; set cells;

/* Square root on explanatory variables. */
vitCSqrt = sqrt(vitC);
caffeineSqrt = sqrt(caffeine);

/* Log on explanatory variables. */
kcalLog = log(kcal);
carbLog = log(carb);
plateletLog = log(platelet);

/* Transformation on predicted variable. */
wbcTrans = wbc**0.2;
run;

/*
We run another regression with our transformed variables to see if the issues have be
We also look for outliers again.
Note: We locate several multicollinear variables.
*/
proc reg data=cells
plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS);
id ID;
model wbcTrans = vitCSqrt upperLeg kcalLog carbLog ageMonths famSize waistCirc aveSte
armCirc married female rbc caffeineSqrt plateletLog black latino / partial vif collin
output out=out2 r=resid p=pred;
title1 'Transformed Linear Regression Model';
title2 'wbcTrans = vitCSqrt upperLeg kcalLog carbLog ageMonths famSize waistCirc aveS
armCirc married female rbc caffeineSqrt plateletLog black latino';
run;

/*
We check the diagnostics of our transformed model.
Note: The issues that previously existed have been resolved.
*/
%resid_num_diag(dataset=out2, datavar=resid, label='Residual',
predvar=pred, predlabel='Predicted Value');

/* We run the F-test for lack of fit on our transformed model. */
proc rsreg data=cells;
model wbcTrans = vitCSqrt upperLeg kcalLog carbLog ageMonths famSize waistCirc aveSte
armCirc married female rbc caffeineSqrt plateletLog black latino / lackfit covar=1 no
title1 'Transformed Model F-test for lack of fit';
run;
```

```
/****************************** Multicollinearity ******************************

/*
We previously removed several variables that had multicollinearity.
We run a regression without these removed variables.
We look for outliers again.
We note that multicollinearity has essentially dissappeared.
*/
proc reg data=cells
plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS);
id ID;
model wbcTrans = vitCSqrt upperLeg kcalLog carbLog ageMonths famSize waistCirc aveSte
married female rbc caffeineSqrt plateletLog black latino / partial vif collin;
output out=out3 r=resid p=pred;
title1 'Transformed Linear Model After MultiCollinearity Adjustment';
title2 'wbc = vitCSqrt upperLeg kcalLog carbLog ageMonths famSize waistCirc aveStep
married female rbc caffeineSqrt plateletLog black latino';
run;

/***************************** Interaction Terms ******************************

/* We split the data into training and testing sets. */
proc surveyselect data=cells seed=5000 out=cells2 rate=0.2 outall;
data train; set cells2;
if Selected=0;
data test; set cells2;
if Selected=1;
run;

/* Print training and testing data sets. */
proc print data=train (obs=5);
title1 'Training Data Set';
proc print data=test (obs=5);
title1 'Testing Data Set';
run;

/* Create a few interaction variables. */
data train; set train;
kcalLog_carbLog = kcalLog * carbLog;
ageMonths_female = ageMonths * female;
married_waistCirc = married * waistCirc;
run;

/*
Run a regression with our new interaction variables to check for significance.
Also run an subset f-test to determine whether to drop the interaction terms all at o
*/
proc reg data=train
```

```
plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS);
id ID;
model wbcTrans = vitCSqrt upperLeg kcalLog carbLog ageMonths famSize waistCirc aveSte
caffeineSqrt plateletLog black latino kcalLog_carbLog ageMonths_female married_waistC
subsetcheck: test kcalLog_carbLog = ageMonths_female = married_waistCirc = 0;
output out=out4 r=resid p=pred;
title1 'Regression Model with Interaction Terms';
title2 'wbc = vitCSqrt upperLeg kcalLog carbLog ageMonths famSize waistCirc aveStep m
caffeineSqrt plateletLog black latino kcalLog_carbLog ageMonths_female married_waistC
run;

/************************************ Model Selection ****************************

/* Remove the interaction terms from the model. Run backwards selection to decide on
/* We run another regression after removing multicollinear data. We look for outliers
proc reg data=train;
model wbcTrans = vitCSqrt upperLeg kcalLog carbLog ageMonths famSize waistCirc aveSte
caffeineSqrt plateletLog black latino / selection=backward slstay=0.10;
output out=out5 r=resid p=pred;
title1 'Backwards Selection';
run;

/* This is the model that was generated by backwards selection. We obtained r-squared
proc reg data=train;
model wbcTrans = upperLeg ageMonths waistCirc aveStep plateletLog black latino;
output out=out6 r=resid p=pred;
title1 'Model Generated by Backwards Selection';
run;

/* Run stepwise selection method. */
proc reg data=train;
model wbcTrans = vitCSqrt upperLeg kcalLog carbLog ageMonths famSize waistCirc aveSte
caffeineSqrt plateletLog black latino / selection=stepwise slentry=.10 slstay=.10;
output out=out7 r=resid p=pred;
title1 'Stepwise Selection';
run;

/* This is the model that was generated by stepwise selection. */
proc reg data=train;
model wbcTrans = ageMonths famSize waistCirc aveStep female plateletLog black;
output out=out8 r=resid p=pred;
title1 'Model Generated by Stepwise Selection';
run;

/* All possible regressions approach */
proc reg data=train;
model wbcTrans = vitCSqrt upperLeg kcalLog carbLog ageMonths famSize waistCirc aveSte
caffeineSqrt plateletLog black latino / selection=AdjRSq Cp AIC SBC;
output out=out9 r=resid p=pred;
title1 'All Possible Regressions';
```

```
run;

/******************** Best Model Outliers, Influential Points, & Diagnostics ********

/*
This is our chosen best model.
We run it to generate output.
*/
ods graphics on / imagemap=on;
proc reg data=train plots(label) = (DFFITS DFBETAS);
id ID;
model wbcTrans = upperLeg ageMonths waistCirc aveStep plateletLog black latino / infl
title1 'Best Model';
title2 'wbcTrans = upperLeg ageMonths waistCirc aveStep plateletLog black latino';
ods output outputstatistics=out2;
output out=out3 p=pred r=resid cookd=CooksD;
run; quit;
ods graphics / imagemap=off;

/* Alternative thresholds for influential obs. and outlier diagnostics */
data temp;
p=8; /* p = # beta's (incl. intercept */
n = 380; /* n = sample size */
CooksD20 = finv(.20,p,n-p);
CooksD50 = finv(.50,p,n-p);
RStudent95Bonf = tinv((1-.05/2/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;

/* Printing our alternative thresholds. */
proc print data=temp;
var CooksD20 CooksD50 RStudent95Bonf  Leverage3 DFBETAS DFFITS;
title1 'Alternative thresholds';
run;

/* Combine several other datasets. */
data betterplots; set out2 out3 temp;
run;

/*Make Plot with Better Cook's D Reference Lines */
proc sgplot data=betterplots;
scatter x=ID y=cooksD / markerchar=ID;
xaxis label = 'Observation ID';
yaxis label = "Cook's D";
title1 "Cook's D";
title2 'Reference lines for 20th and 50th percentiles of F-distribution';
refline cooksD20 / axis=Y;  /*20th percentile*/
refline cooksD50 / axis=Y;  /*50th percentile*/
```

```
yaxis max=1;
run;

/*Make Plot with Better Studentized Deleted Residuals and Leverage Lines */
proc sgplot data=betterplots;
scatter x=HatDiagonal y=RStudent / markerchar=ID;
xaxis label = 'Leverage';
yaxis label = 'Studentized Deleted Residuals';
title1 'Reference Lines for Outliers and Leverage';
title2 'Bonferoni for Outliers, 3p/n for Leverage ';
refline RStudent95Bonf / axis=Y;  /*Upper limit outliers*/
refline NegRStudent95Bonf / axis=Y;  /*lower limit outliers*/
refline Leverage3 / axis=X;  /* limit leverage*/
yaxis max=4 min=-4;
run;

/* We check the diagnostics of our best model. */
%resid_num_diag(dataset=out3, datavar=resid, label='Residual',
predvar=pred, predlabel='Predicted Value');

/* Look at the sequence plot of our best model. */
data temp; set out3;
order = _n_;
proc sgplot data=temp;
series x=order y=resid / lineattrs=(pattern=solid);
xaxis label='Order in Data';
yaxis label='Residual Value';
title1 'Sequence plot of residuals';
run;


/****************************** Model Validation ***********************************

/* We calculate the MSPR on the test set. */
data test; set test;
wbcTrans = wbc ** 0.2;
wbcTransHat = 1.20647 - 0.00244 * upperLeg - 0.00009358 * ageMonths + 0.00081307 * wa
0.00000239 * aveStep + 0.06685 * plateletLog - 0.04975 * black - 0.01913 * latino;
SqPredError = (wbcTrans - wbcTransHat)**2;
run;

/* Display the MSPR. */
proc means data=test mean;
var SqPredError;
title1 'MSPR for test set';
run;

/* Train an intercept only model. */
proc reg data=train;
model wbcTrans = ;
```

```
title1 'Intercept Only Model';
output out=out1 p=pred r=resid;
run;

/* Calculate MSPR on intercept only model. */
data interceptError; set test;
wbsTransHat = 1.47658;
sqPredError = (wbsTransHat - wbcTrans) ** 2;
run;

/* Display MSPR for intercept only model. */
proc means data=interceptError;
var sqPredError;
run;

/************************************ Final Model Results. ***********************

/* Run a regression with our best model and use all of the data. */
proc reg data=cells;
model wbcTrans = upperLeg ageMonths waistCirc aveStep plateletLog black latino;
title1 'Best Model Regression on All Data';
output out=out1 r=resid p=pred;
run;

/* We will now plot actual values against predicted values. First we put data on the
data out1; set out1;
wbcHat = pred ** 5;
run;

/* Plot actual values against predicted values. */
proc sgplot data=out1;
scatter x=wbc y=wbcHat / markerattrs=(symbol=CIRCLEFILLED);
xaxis label='Observed wbc' labelattrs=(size=15pt);
yaxis label='Predicted wbc' labelattrs=(size=15pt);
title1 'Observed wbc Against Predicted wbc';
run;
```