Paper 130-2018

Logistic and Linear Regression Assumptions: Violation Recognition and Control

Deanna Schreiber-Gregory, Henry M Jackson Foundation

ABSTRACT

Regression analyses are one of the first steps (aside from data cleaning, preparation, and descriptive analyses) in any analytic plan, regardless of plan complexity. Therefore, it is worth acknowledging that the choice and implementation of the wrong type of regression model, or the violation of its assumptions, can have detrimental effects to the results and future directions of any analysis. Considering this, it is important to understand the assumptions of these models and be aware of the processes that can be utilized to test whether these assumptions are being violated. Given that logistic and linear regression techniques are two of the most popular types of regression models utilized today, these are the are the ones that will be covered in this paper. Some Logistic regression assumptions that will reviewed include: dependent variable structure, observation independence, absence of multicollinearity, linearity of independent variables and log odds, and large sample size. For Linear regression, the assumptions that will be reviewed include: linearity, multivariate normality, absence of multicollinearity and auto-correlation, homoscedasticity, and measurement level. This paper is intended for any level of SAS® user. This paper is also written to an audience with a background in theoretical and applied statistics, though the information within will be presented in such a way that any level of statistics/mathematical knowledge will be able to understand the content.

INTRODUCTION

We can be certain that all parametric tests in a statistical analysis assume some certain characteristics (or assumptions) about the data. Depending on the parametric analysis, the assumptions vary. A violation of any of these assumptions changes the conclusion of the research and interpretation of the results. Therefore, all research, whether for a journal, thesis/dissertation, or report, must check and adhere to these assumptions for accurate interpretation and model integrity.

COMMON ASSUMPTIONS

The following assumptions are commonly found in statistical research:

Assumptions of Normality: Most of the parametric tests require that the assumption of normality be met. Normality means that the distribution of the test is normally distributed (or bell-shaped) with 0 mean, with 1 standard deviation and a symmetric bell shaped curve.

Assumptions of Homogeneity of Variance: The assumption of homogeneity of variance is that the variance within each of the populations is equal.

Assumptions of Homogeneity of Variance-Covariance Matrices: The assumption for a multivariate approach is that the vector of the dependent variables follow a multivariate normal distribution, and the variance-covariance matrices are equal across the cells formed by the between-subjects effects.

INTRODUCTION TO THE DATSET

The dataset used for this paper is easily accessible by anyone with access to SAS®. It is a sample dataset titled “lipids”. The background to this sample dataset states that it is from a study to investigate the relationships between various factors and heart disease. In order to explore this relationship, blood lipid

screenings were conducted on a group of patients. Three months after the initial screening, follow-up data was collected from a second screening that included additional information such as gender, age, weight, total cholesterol, and history of heart disease. The outcome variable of interest in this analysis is the reduction of cholesterol level between the initial and 3-month lipid panel or “cholesterolloss”. The predictor variables of interest are age (age of participant), weight (weight at first screening), cholesterol (total cholesterol at first screening), triglycerides (triglycerides level at first screening), HDL (HDL level at first screening), LDL (LDL level at first screening), height (height of participant), skinfold (skinfold measurement), systolicbp (systolic blood pressure) diastolicbp (diastolic blood pressure), exercise (exercise level), and coffee (coffee consumption in cups per day).

DATA CLEANING AND PREPARATION

As a first step in the examination of our research question – do target health outcome variables contribute to the amount of cholesterol lost between baseline and a 3 month follow-up – we must first identify which variables will be used in the analysis, what these variables look like, and how these variables will interact with each other. In short, we must clean and prepare the data for our analysis. This may seem redundant, but it is a worthy note to make considering the type of analysis we are about to conduct. We will begin by identifying the dataset and making sure that it is appropriately imported into the SAS environment. At this time we will also use the CONTENTS procedure to check the structure and types of variables we will be working with:

```
/* Example of Multicollinearity Findings */
libname health
"C:\ProgramFiles\SASHome\SASEnterpriseGuide\7.1\Sample\Data";

data health;
set health.lipid;
run;

proc contents data=health;
title 'Health Dataset with High Multicollinearity';
run;
```

ASSUMPTIONS OF LINEAR REGRESSION

Linear regression is an analysis that assesses whether one or more predictor variables explain the dependent (criterion) variable. This type of regression has five key assumptions.

- Linear relationship
- Multivariate normality
- No or little multicollinearity
- No auto-correlation
- Homoscedasticity

Additionally, it is necessary to make a note about sample size for this type of regression model. In Linear regression the sample size rule of thumb is that the regression analysis requires at least 20 cases per independent variable in the analysis.

ASSUMPTION OF LINEAR RELATIONSHIP

Linear regression needs the relationship between the independent and dependent variables to be linear. It is also important to check for outliers since linear regression is sensitive to outlier effects. One way to test the linearity assumption can be through the examination of scatter plots.

ASSUMPTION OF MULTIVARIATE NORMALITY

Linear regression analyses require all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot. Normality can be checked with a goodness of fit test, such as the Kolmogorov-Smirnov test. When the data is not normally distributed a non-linear transformation (e.g., log-transformation) might fix this issue.

ASSUMPTION OF ABSENCE OF MULTICOLLINEARITY

Linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other.

As stated above, multicollinearity may be tested with three central criterion:

- Correlation matrix: when computing the matrix of Pearson's Bivariate Correlation among all independent variables the correlation coefficients should hopefully be smaller than 0.8.

- Tolerance: the tolerance measures the influence of one independent variable on all other independent variables; the tolerance is calculated with an initial linear regression analysis. Tolerance is defined as $T = 1 – R^2$ for these first step regression analysis. With $T < 0.1$ there might be multicollinearity in the data and with $T < 0.01$ there certainly is.

- Variance Inflation Factor (VIF): the variance inflation factor of the linear regression is defined as $VIF = 1/T$. With $VIF > 10$ there is an indication that multicollinearity may be present; with $VIF > 100$ there is certainly multicollinearity among the variables.

- Condition Index: the condition index is calculated using a factor analysis on the independent variables. Values of 10-30 indicate a mediocre multicollinearity in the linear regression variables, values > 30 indicate strong multicollinearity.

If multicollinearity is found in the data, centering the data (that is deducting the mean of the variable from each score) might help to solve the problem. However, the simplest way to address the problem is to remove the independent variables with high VIF values. Other alternatives to tackle the problems is conducting a factor analysis and rotating the factors to insure independence of the factors in the linear regression analysis, using ridge regression, LASSO, or Elastic Net techniques.

ASSUMPTION OF THE ABSENCE OF AUTOCORRELATION

Linear regression analyses require that there exists little or no autocorrelation in the data. Autocorrelation occurs when the residuals are not independent from each other. In other words when the value of $y(x+1)$ is not independent from the value of $y(x)$.

While a scatterplot allows you to check for autocorrelations, you can test the linear regression model for autocorrelation with the Durbin-Watson test. Durbin-Watson's d tests the null hypothesis that the residuals are not linearly auto-correlated. While d can assume values between 0 and 4, values around 2 indicate no autocorrelation. As a rule of thumb values of $1.5 < d < 2.5$ show that there is no auto-correlation in the data. However, the Durbin-Watson test only analyses linear autocorrelation and only between direct neighbors, which are first order effects.

ASSUMPTION OF HOMOSCEDASTICITY

Lastly, linear regression analyses assume the presence of homoscedasticity. Examination of a scatter plot is good way to check whether the data are homoscedastic (in other words, the residuals are equal across the regression line).

The Goldfeld-Quandt Test can also be used to test for heteroscedasticity. The test splits the data into two groups and tests to see if the variances of the residuals are similar across the groups. If homoscedasticity is present, a non-linear correction might fix the problem.

ASSUMPTIONS OF LOGISTIC REGRESSION

Logistic regression is quite different than linear regression in that it does not make several of the key assumptions that linear and general linear models (as well as other ordinary least squares algorithm based models) hold so close: (1) logistic regression does not require a linear relationship between the dependent and independent variables, (2) the error terms (residuals) do not need to be normally distributed, (3) homoscedasticity is not required, and (4) the dependent variable in logistic regression is not measured on an interval or ratio scale.

However, logistic regression still shares some assumptions with linear regression, with some additions of its own.

ASSUMPTION OF APPROPRIATE OUTCOME STRUCTURE

To begin, one of the main assumptions of logistic regression is the appropriate structure of the outcome variable. Binary logistic regression requires the dependent variable to be binary and ordinal logistic regression requires the dependent variable to be ordinal.

ASSUMPTION OF OBSERVATION INDEPENDENCE

Logistic regression requires the observations to be independent of each other. In other words, the observations should not come from repeated measurements or matched data.

ASSUMPTION OF THE ABSENCE OF MULTICOLLINEARITY

Logistic regression requires there to be little or no multicollinearity among the independent variables. This means that the independent variables should not be too highly correlated with each other.

ASSUMPTION OF LINEARITY OF INDEPENDENT VARIABLES AND LOG ODDS

Logistic regression assumes linearity of independent variables and log odds. Although this analysis does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds.

ASSUMPTION OF A LARGE SAMPLE SIZE

Finally, logistic regression typically requires a large sample size. A general guideline is that you need at minimum of 10 cases with the least frequent outcome for each independent variable in your model. For example, if you have 5 independent variables and the expected probability of your least frequent outcome is .10, then you would need a minimum sample size of 500 (10*5 / .10).

TESTING THE ASSUMPTIONS

ASSUMPTIONS OF NORMALITY: COMMON TESTS

To test the assumption of normality, the following measures and tests can be applied:

Skewness and Kurtosis: To test the assumption of normal distribution, Skewness should be within the range ±2. Kurtosis values should be within range of ±7.

This test can be completed through use of the Univariate procedure:

```
/* Testing for Normality */
proc univariate data=health plots;
      var age weight cholesterol triglycerides hdl ldl height skinfold
      systolicbp diastolicbp exercise coffee cholesterolloss;
run;
```

The UNIVARIATE Procedure
Variable: Age

Moments			
N	95	Sum Weights	95
Mean	24.3157895	Sum Observations	2310
Std Deviation	3.26901364	Variance	10.6864502
Skewness	2.01555274	Kurtosis	5.34546659
Uncorrected SS	57174	Corrected SS	1004.52632
Coeff Variation	13.4439955	Std Error Mean	0.33539372

Figure 1: Results for Testing Skewness & Kurtosis

Shapiro-Wilk's W test: Most of the researchers use this test to test the assumption of normality. Wilk's test should not be significant to meet the assumption of normality.

Kolmogorov-Smirnov test: In the case of a large sample, most researchers use K-S test to test the assumption of normality. This test should not be significant to meet the assumption of normality.

Q-Q plot: Most researchers use Q-Q plots to test the assumption of normality. In this method, observed value and expected value are plotted on a graph. If the plotted value vary more from a straight line, then the data is not normally distributed. Otherwise data will be normally distributed.

The Shapiro-Wilk's W, Komogorov-Smirnov tests, and Q-Q Plots can be completed through use of the Capability procedure:

```
/* Testing for Normality - Shapiro-Wilk's W, Komogorov-Smirnov tests,
and Q-Q Plots */
proc capability DATA=health NORMAL;
      var age weight cholesterol triglycerides hdl ldl height skinfold
      systolicbp diastolicbp exercise coffee cholesterolloss;
      QQPLOT age weight cholesterol triglycerides hdl ldl height
      skinfold systolicbp diastolicbp exercise coffee cholesterolloss
            /NORMAL(MU=EST SIGMA=EST COLOR=RED L=1);
      PPPLOT age weight cholesterol triglycerides hdl ldl height
      skinfold systolicbp diastolicbp exercise coffee cholesterolloss
            /NORMAL(MU=EST SIGMA=EST COLOR=RED L=1);
      HISTOGRAM /NORMAL(COLOR=MAROON W=4) CFILL = BLUE CFRAME = LIGR;
      INSET MEAN STD /CFILL=BLANK FORMAT=5.2 ;
run;
```

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.771653	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.235291	Pr > D	<0.0100
Cramer-von Mises	W-Sq	1.308639	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	7.082537	Pr > A-Sq	<0.0050

Figure 2: Results for Testing Shapiro-Wilk's W & Komogorov-Smirnov

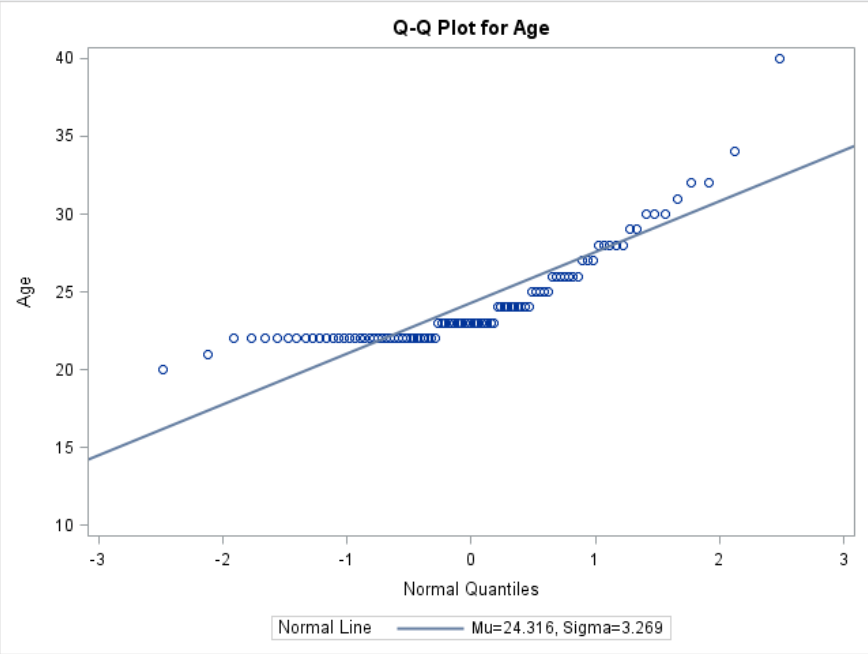

Figure 3: Checking Q-Q Plots

ASSUMPTIONS OF HOMOGENEITY OF VARIANCE (HOMOSCEDASTICITY): COMMON TESTS

Levene's test: To test the assumption of homogeneity of variance, Levene's test is used. Levene's test is used to asses if the groups have equal variances. This test should not be significant to meet the assumption of equality of variances.

Levene's test can be completed through use of the GLM procedure:

```
/* Testing for Homogeneity of Variance - Levene's Test */
proc glm data=health;
   class exercise coffee;
   model cholesterolloss = exercise;
   means exercise / hovtest=levene; /* can specify type=abs|square */
```

```
run;
```

Levene's Test for Homogeneity of CholesterolLoss Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Exercise	4	5979783	1494946	1.44	0.2444
Error	30	31093925	1036464		

Figure 4: Levene's Test for Homogeneity of Variance

Plot Residuals by Predicted Values:

```
/* Plot Residuals by Predicted Values */
proc reg data= health;
      model cholesterolloss = age weight cholesterol triglycerides hdl
      ldl height skinfold systolicbp diastolicbp exercise coffee;
      plot r.*p.;
run;
quit;
```

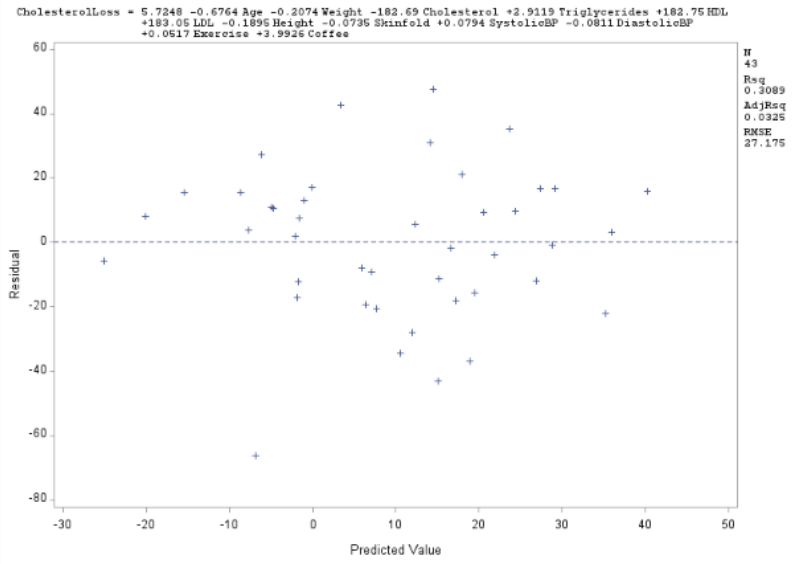

Figure 5: Plot Residuals by Predicted Values

White Test: This statistic is asymptotically distributed as chi-square with k-1 degrees of freedom, where k is the number of regressors, excluding the constant term. Two other tests that can be employed are the Breusch-Pagan Test and Lagrange Multiplier (LM) Test. If you check the P-value of Q statistics and LM tests, a P-value greater than .05 indicates homoscedasticity. If the p-value of White test and Breusch-Pagan test is greater than .05, the homogeneity of variance of the residual has been met (Homoscedasticity).

```
/* Homoscedasticity Test - White and Breusch-Pagan Test */
proc model data= health;
      parms a1 b1 b2 b3;
      cholesterolloss = a1 + b1*age + b2*weight + b3*cholesterol;
      fit cholesterolloss / white pagan=(1 age weight cholesterol)
```

```
      out=resid1 outresid;
run;
quit;
```

The MODEL Procedure

Nonlinear OLS Summary of Residual Errors

Equation	DF Model	DF Error	SSE	MSE	Root MSE	R-Square	Adj R-Sq
CholesterolLoss	4	39	24909.8	638.7	25.2728	0.2230	0.1632

Nonlinear OLS Parameter Estimates

Parameter	Estimate	Approx Std Err	t Value	Approx Pr > |t|
a1	-4.75027	34.7714	-0.14	0.8920
b1	-0.04688	1.3803	-0.03	0.9731
b2	-0.25581	0.1567	-1.63	0.1106
b3	0.301901	0.1174	2.57	0.0140

Number of Observations		Statistics for System	
Used	43	Objective	579.2971
Missing	52	Objective*N	24910

Heteroscedasticity Test

Equation	Test	Statistic	DF	Pr > ChiSq	Variables
CholesterolLoss	White's Test	6.29	9	0.7101	Cross of all vars
	Breusch-Pagan	2.36	3	0.5015	1, Age, Weight, Cholesterol

Figure 6: Homoscedasticity - White's and Breusch-Pagan Tests

There are several ways in which violations of this test can be adjusted. You can employ the Box-Cox transformations of the dependent variable or through use of Weighted Least Squares.

Box-Cox Transformation:

```
/* Box-Cox Transformation as an Adjustment */
proc transreg data=health  test;
      model boxcox(cholesterolloss) = identity(age weight cholesterol);
run;
```

Transformation	Best Lambda
Square	1.5 to 2.5
None	0.75 to 1.5
Square-Root	0.25 to 0.75
Natural Log	-0.25 to 0.25

Inverse Square-Root	-0.75 to –0.25
Reciprocal	-1.5 to -0.75
Inverse Square	-2.5 to -1.5

Weighted Least Squares: if variable transformation does not solve this problem, then we can use weighted least squares. You can construct these weights through the following steps: (1) compute the absolute and squared residuals, (2) find the absolute and squared residuals versus the independent variables to get the estimated standard deviation and variance, and (3) compute the weights using the estimated standard deviations and variance.

```
/* Weighted Least Squares as an Adjustment */
proc reg data=health;
   model cholesterolloss=age weight cholesterol;
   output out=WORK.PRED r=residual;
run;

data work.resid;
  set work.pred;
  absresid=abs(residual);
  sqresid=residual**2;

proc reg data=work.resid;
    model absresid=age weight cholesterol;
    output out=WORK.s_weights p=s_hat;
   model sqresid=age weight cholesterol;
    output out=WORK.v_weights p=v_hat;
run;

** compute the weights using the estimated standard deviations**;
data work.s_weights;
      set work.s_weights;
            s_weight=1/(s_hat**2);
      label s_weight = "weights using absolute residuals";

** compute the weights using the estimated variances**;
data work.v_weights;
      set work.v_weights;
            v_weight=1/v_hat;
      label v_weight = "weights using squared residuals";

** Do the weighted least squares using the weights from the estimated
standard deviation**;
proc reg data=work.s_weights;
      weight s_weight;
      model cholesterolloss=age weight cholesterol;
run;

** Do the weighted least squares using the weights from the estimated
variances**;
proc reg data=work.v_weights;
      weight v_weight;
      model cholesterolloss=age weight cholesterol;
      run;
```

```
quit;
```

Weight: v_weight weights using squared residuals

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	25.27382	8.42461	7.04	0.0007
Error	39	46.64962	1.19614		
Corrected Total	42	71.92344			

Root MSE	1.09368	R-Square	0.3514
Dependent Mean	-0.70855	Adj R-Sq	0.3015
Coeff Var	-154.35553		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-34.15771	20.68932	-1.65	0.1068
Age	1	0.34253	1.06795	0.32	0.7501
Weight	1	-0.14328	0.11784	-1.22	0.2314
Cholesterol	1	0.30981	0.07072	4.38	<.0001

Figure 7: Weighted Least Squares Adjustment for Homoscedasticity

ASSUMPTIONS OF HOMOGENEITY OF VARIANCE-COVARIANCE MATRICES: COMMON TESTS

Box's M test: This test is used to test the multivariate homogeneity of variance-covariance matrices assumption. An insignificant value of Box's M test shows that those groups do not differ from each other and would meet the assumption. It is also worthy to note that Box's M is extremely sensitive to departures from normality. Therefore, if the sample is not normally distributed (an assumption we covered earlier), you should not use the Box's M test. Box's M also has very little power (Cohen, 2008) for data with small sample sizes. If you use the Box's M test on a small sample and your result is not significant, this does not necessarily indicate that the covariance matrices are equal. The test also has issues in the other direction, receiving criticism for being overly sensitive to larger sample sizes. It tends to report a statistically significant result when one doesn't actually exist. To address this particular issue, it is recommended that you use a smaller alpha level (Hahs-Vaughn, 2016).

Additionally, Box's M Test is common in SPSS, but not in SAS. One way to request a Box's M Test is to use the Discrim procedure. Make sure to indicate `pool=test` and `wcov` as options.

```
/* Testing for Homogeneity of Variance-Covariance Matrices - Box's M
Test */
proc discrim data=health method=normal pool=test wcov;
      class exercise;
      var age weight cholesterol triglycerides hdl ldl height skinfold
systolicbp diastolicbp cholesterolloss;
run;
```

SAS also offers Bartlett's test as a variation of Box's M, through use of the GLM procedure

```
/* Testing for Homogeneity of Variance-Covariance Matrices - Bartlett's
Test */
proc glm data=health;
   class exercise coffee;
   model cholesterolloss = exercise;
   means exercise / hovtest=bartlett;
run;
```

Bartlett's Test for Homogeneity of CholesterolLoss Variance			
Source	DF	Chi-Square	Pr > ChiSq
Exercise	6	8.0566	0.2340

Figure 8: Bartlett's Test for Homogeneity of Variance-Covariance Matrices

Randomness: Most of the statistics assume that the sample observations are random. The Wald-Wolfowitz test, also known as the Runs test for randomness, is often used to test this assumption. A run is a set of sequential values that are either all above or below the mean. To simplify computations, the data are first centered about their mean. To carry out the test, the total number of runs is computed along with the number of positive and negative values. A positive run is then a sequence of values greater than zero, and a negative run is a sequence of values less than zero. We can then test if the number of positive and negative runs are distributed equally in time.

The following statements create an example data set using the random number generator RANNOR. The Wald-Wolfowitz test will be performed on the variable age.

```
/* Testing for Randomness - Wald-Wolfowitz Test */
data health;
      drop i;
            do i=1 to 75;
             age=rannor(123);
             output;
      end;
run;
```

The MEAN=0 option in the PROC STANDARD step below centers the variable age about its mean.

```
proc standard data=health out=health_two mean=0;
      var age;
run;
```

The following DATA step computes the total number of runs (RUNS), the number of positive values (NUMPOS), and the number of negative values (NUMNEG).

```
data runcount;
      set health_two nobs=nobs;
            if age=0 then delete;
            if age>0 then n+1;
            if age<0 then m+1;
            retain runs 0 numpos 0 numneg 0;
            previous=lag(age);

            if _n_=1 then do;
                  runs=1;
                  prevpos=.;
                  currpos=.;
```

```
                    prevneg=.;
                    currneg=.;
                end;

                else do;
                    prevpos=( previous > 0 );
                    currpos=( d > 0 );
                    prevneg=( previous < 0 );
                    currneg=( d < 0 );

                    if _n_=2 and (currpos and prevpos) then numpos+1;
                        else if _n_=2 and (currpos and prevneg) then
                        numneg+1;
                        else if _n_=2 and (currneg and prevpos) then
                        numpos+1;
                        else if _n_=2 and (currneg and prevneg) then
                        numneg+1;

                    if currpos and prevneg then do;
                        runs+1;
                        numpos+1;
                    end;

                    if currneg and prevpos then do;
                        runs+1;
                        numneg+1;
                    end;
                end;
        run;

        data runcount;
            set runcount end=last;
            if last;
        run;
```

Finally, these steps compute and display the Wald-Wolfowitz (or Runs) test statistic and its *p*-value.

```
        data waldwolf;
            label z='Wald-Wolfowitz Z'
                    pvalue='Pr > |Z|';
            set runcount;
            mu = ( (2*n*m) / (n + m) ) + 1;
            sigmasq = ( (2*n*m) * (2*n*m-(n+m)) ) / ( ((n+m)**2) * (n+m-1) );
            sigma=sqrt(sigmasq);
            drop sigmasq;

            if N GE 50 then Z = (Runs - mu) / sigma;
                else if Runs-mu LT 0 then Z = (Runs-mu+0.5)/sigma;
                else Z = (Runs-mu-0.5)/sigma;

            pvalue=2*(1-probnorm(abs(Z)));
        run;

        title  'Wald-Wolfowitz Test for Randomness';
        title2 'H0: The data are random';

        proc print data=waldwolf label noobs;
```

```
        var z pvalue;
        format pvalue pvalue.;
run;
```

Wald-Wolfowitz Test for Randomness
H0: The data are random

Wald-Wolfowitz Z	Pr > |Z|
.001550389	0.9988

Figure 9: Wald-Wolfowitz Test for Randomness

Multicollinearity: Multicollinearity means that the variables of interest are highly correlated, and high correlations should not be present among variables of interest. To test the assumption of multicollinearity, VIF and Condition indices can be used, especially in regression analyses. A value of VIF >10 indicates multicollinearity is present and the assumption is violated.

Our first step is to explore the correlation matrix. We can do this through implementation of the CORR procedure:

```
/* Assess Pairwise Correlations of Continuous Variables */
proc corr data=health;
var age weight cholesterol triglycerides hdl ldl height skinfold
systolicbp diastolicbp exercise coffee cholesterolloss;
title 'Health Predictors - Examination of Correlation Matrix';
run;
```

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Age	Weight	Cholesterol	Triglycerides	HDL	LDL	Height	Skinfold	SystolicBP	DiastolicBP	Exercise	Coffee	CholesterolLoss
Age	1.00000 95	0.08935 0.3892 95	0.26282 0.0101 95	0.21167 0.0395 95	0.20310 0.0484 95	0.21588 0.0356 95	-0.02080 0.8414 95	0.10625 0.3055 95	0.02384 0.8186 95	-0.06384 0.5388 95	-0.12193 0.2392 95	0.25069 0.0142 95	0.09914 0.5270 43
Weight	0.08935 0.3892 95	1.00000 95	-0.02188 0.8333 95	0.10757 0.2994 95	-0.27555 0.0069 95	0.05743 0.5804 95	0.69794 <.0001 95	0.07427 0.4744 95	0.15740 0.1277 95	0.13627 0.1879 95	0.03254 0.7542 95	0.05720 0.5819 95	-0.24221 0.1176 43
Cholesterol	0.26282 0.0101 95	-0.02188 0.8333 95	1.00000 95	0.40081 <.0001 95	0.35246 0.0005 95	0.96170 <.0001 95	-0.07521 0.4688 95	0.07588 0.4649 95	-0.04103 0.6930 95	0.15969 0.1221 95	0.01305 0.9001 95	-0.01157 0.9114 95	0.40318 0.0073 43
Triglycerides	0.21167 0.0395 95	0.10767 0.2994 95	0.40081 <.0001 95	1.00000 95	-0.27838 0.0063 95	0.48904 <.0001 95	0.04071 0.6953 95	0.09292 0.3704 95	0.14545 0.1596 95	0.14073 0.1737 95	-0.11162 0.2815 95	-0.00360 0.9731 95	0.11396 0.4669 43
HDL	0.20310 0.0484 95	-0.27555 0.0069 95	0.35246 0.0005 95	-0.27838 0.0063 95	1.00000 95	0.08340 0.4217 95	-0.24465 0.0169 95	0.11116 0.2835 95	-0.06008 0.5630 95	0.02410 0.8167 95	-0.03055 0.7688 95	0.10955 0.2906 95	0.19099 0.2199 43
LDL	0.21588 0.0356 95	0.05743 0.5804 95	0.96170 <.0001 95	0.48904 <.0001 95	0.08340 0.4217 95	1.00000 95	-0.00777 0.9404 95	0.04547 0.6617 95	-0.03028 0.7708 95	0.16118 0.1187 95	0.02672 0.7972 95	-0.04585 0.6591 95	0.37389 0.0136 43
Height	-0.02080 0.8414 95	0.69794 <.0001 95	-0.07521 0.4688 95	0.04071 0.6953 95	-0.24465 0.0169 95	-0.00777 0.9404 95	1.00000 95	-0.13762 0.1835 95	0.08432 0.4166 95	0.06327 0.5424 95	0.00521 0.9600 95	0.07165 0.4902 95	-0.27042 0.0795 43
Skinfold	0.10625 0.3055 95	0.07427 0.4744 95	0.07588 0.4649 95	0.09292 0.3704 95	0.11116 0.2835 95	0.04547 0.6617 95	-0.13762 0.1835 95	1.00000 95	-0.09901 0.3398 95	-0.03817 0.7134 95	-0.26581 0.0092 95	0.07833 0.4505 95	-0.03538 0.8218 43
SystolicBP	0.02384 0.8186 95	0.15740 0.1277 95	-0.04103 0.6930 95	0.14545 0.1596 95	-0.06008 0.5630 95	-0.03028 0.7708 95	0.08432 0.4166 95	-0.09901 0.3398 95	1.00000 95	0.33476 0.0009 95	-0.05138 0.6209 95	-0.05048 0.6271 95	-0.07917 0.6136 43
DiastolicBP	-0.06384 0.5388 95	0.13627 0.1879 95	0.15969 0.1221 95	0.14073 0.1737 95	0.02410 0.8167 95	0.16118 0.1187 95	0.06327 0.5424 95	-0.03817 0.7134 95	0.33476 0.0009 95	1.00000 95	-0.03647 0.7257 95	0.03908 0.7069 95	0.13192 0.3991 43

Figure 10: Multicollinearity - Pearson Correlation Coefficients

Keep in mind, while reviewing these results we want to check to see if any of the variables included have a high correlation – about 0.8 or higher – with any other variable. As we can see, upon review of this correlation matrix, there seems to be some particularly high correlations between a few of the variables. Some relationships of note would be Cholesterol / LDL (0.96) and Weight / Height (0.70). Next we will examine multicollinearity through the Variance Inflation Factor, Tolerance, and Collinearity Diagnostics. This can be done by specifying the vif, tol, and collin options respectively after the model statement:

```
proc reg data=health;
      model cholesterolloss = age weight cholesterol triglycerides hdl
      ldl height skinfold systolicbp diastolicbp exercise coffee / vif
      tol collin;
      title 'Health Predictors - Multicollinearity Investigation of VIF
      and Tol';
run;
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	5.72484	108.12644	0.05	0.9581	.	0
Age	1	-0.67645	2.20644	-0.31	0.7613	0.32637	3.06405
Weight	1	-0.20743	0.27789	-0.75	0.4612	0.32763	3.05224
Cholesterol	1	-182.68577	170.82886	-1.07	0.2934	4.326797E-7	2311178
Triglycerides	1	2.91187	2.73231	1.07	0.2951	0.00034921	2863.60930
HDL	1	182.75031	170.71293	1.07	0.2929	0.00000516	193966
LDL	1	183.05303	170.82561	1.07	0.2925	5.113026E-7	1955789
Height	1	-0.18955	1.61295	-0.12	0.9072	0.43651	2.29616
Skinfold	1	-0.07347	0.53443	-0.14	0.8916	0.77820	1.28502
SystolicBP	1	0.07945	0.63738	0.12	0.9016	0.66694	1.49939
DiastolicBP	1	-0.08111	0.43028	-0.19	0.8518	0.66583	1.50190
Exercise	1	0.05167	0.05513	0.94	0.3562	0.77863	1.28430
Coffee	1	3.99259	3.68202	1.08	0.2868	0.44992	2.22261

Figure 11: Tolerance and VIF Investigation Results

When considering tolerance, we want to make sure that no values fall below 0.1. In reviewing our results, we can see several variables – namely cholesterol, triglycerides, HDL, and LDL – had values well below our 0.1 cutoff value. As for variance inflation, the magic number to look out for is anything above the value of 10. This finding is echoed in review of the Variance Inflation results, where these same variables reveal values far larger than our 10 cutoff for this column. Next, we will look at the collinearity diagnostics for an eigensystem analysis of covariance comparison:

Collinearity Diagnostics														
			Proportion of Variation											
Number	Eigenvalue	Condition Index	Intercept	Age	Weight	Cholesterol	Triglycerides	HDL	LDL	Height	Skinfold	SystolicBP	DiastolicBP	Exercise	Coffee
1	11.29489	1.00000	0.00001138	0.00004637	0.00006086	1.18985E-10	6.589162E-7	2.117859E-9	2.001E-10	0.00001648	0.00096170	0.00002146	0.00011281	0.00154	0.00092480
2	0.68622	4.05704	0.00000331	0.00000701	0.00001442	1.19889E-10	1.143653E-8	2.18204E-10	4.01922E-10	0.00000233	0.00417	0.00000916	0.00000136	0.14262	0.28041
3	0.47052	4.89952	1.797612E-8	0.00000475	0.00006283	1.43701E-10	0.00007230	6.603999E-9	6.81584E-10	7.934101E-7	0.00922	0.00000181	0.00000261	0.35333	0.11353
4	0.27571	6.40063	0.00007350	0.00002441	0.00066563	4.12089E-15	0.00024185	6.082316E-8	3.70204E-10	0.00008619	0.06181	0.00011487	0.00038936	0.17610	0.07292
5	0.14667	8.77543	0.00009651	0.00053911	0.00028359	1.554489E-9	0.00000596	7.219312E-8	1.834574E-9	0.00014692	0.77592	0.00026966	0.00215	0.19373	0.00009012
6	0.06145	13.55776	0.00082045	0.00016295	0.02554	4.483083E-8	0.00000217	0.00000126	5.61196E-8	0.00162	0.01235	0.00304	0.00411	0.00601	0.00073618
7	0.02723	20.36602	0.00093349	0.00170	0.04781	4.702702E-8	0.00025889	0.00000293	2.825254E-7	0.00015620	0.00302	0.00381	0.02703	0.00766	0.08364
8	0.02089	23.25002	0.00003049	0.04483	0.02385	4.595332E-9	0.00004015	0.00000125	4.865568E-8	0.00044312	0.00851	0.00011118	0.44175	0.01589	0.00722
9	0.00826	36.97981	0.03667	0.00022313	0.32353	7.986649E-9	0.00012325	0.00000321	1.621936E-7	0.00897	0.00304	0.04829	0.21593	0.07217	0.04171
10	0.00535	45.53079	0.00836	0.74848	0.10288	1.801143E-8	0.00013411	0.00000190	9.625344E-9	0.02801	0.01629	0.00809	0.13539	0.00271	0.23169
11	0.00195	76.05944	0.07125	0.17866	0.11829	2.005126E-8	5.868795E-8	6.652202E-7	2.868043E-8	0.13625	0.00141	0.85602	0.14837	0.00064741	0.16488
12	0.00085064	115.23088	0.87069	0.01200	0.27559	5.802692E-9	0.00002858	1.490124E-7	2.525126E-8	0.70634	0.10713	0.06251	0.00076513	0.02725	0.00001376
13	9.448677E-9	34574	0.01106	0.01333	0.08142	1.00000	0.99909	0.99999	1.00000	0.12396	0.00617	0.01772	0.02460	0.00133	0.00234

Figure 12: Collinearity Investigation Results

ASSUMPTIONS OF THE ABSENCE OF AUTOCORRELATION: COMMON TESTS

Another common assumption is the need for independence of error terms. It states that the errors associated with one observation are not correlated with the errors of any other observation. It is a problem when you use time series data. Autocorrelation inflates significance results of coefficients by underestimating the standard errors of the coefficients. Hypothesis testing will therefore lead to incorrect conclusions.

Durbin Watson Test: PROC REG tests for first-order autocorrelations using the Durbin-Watson coefficient (DW). The null hypothesis is no autocorrelation. A DW value between 1.5 and 2.5 confirms the absence of first-order autocorrelation. If DW value less than 1.5, it indicates positive autocorrelation. If DW value greater than 2.5, it indicates negative autocorrelation

```
/* Test for AutoCorrelation - Durbin-Watson */
proc reg data = health;
      model cholesterolloss = age weight cholesterol triglycerides hdl
      ldl height skinfold systolicbp diastolicbp / dw;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: CholesterolLoss

Durbin-Watson D	1.622
Number of Observations	43
1st Order Autocorrelation	0.187

Figure 13: Durbin-Watson D Test for Autocorrelation Results

Lagrange Multiplier Test: It can be used for more than one order of auto correlation. It consists of several steps. First, regress Y on Xs to get residuals. Compute lag value of residuals up to pth order. Replace missing values for lagged residuals with zeros. Rerun regression model including lagged residual variable as an independent variable.

```
/* Test for AutoCorrelation - Lagrange Multiplier Test */
proc autoreg data = health;
      model cholesterolloss = age weight cholesterol triglycerides hdl
      ldl height skinfold systolicbp diastolicbp / dwprob godfrey;
run;
```

Correction of Autocorrelation: (1) add lagged transforms (lag value) of the dependent value, and (2) use the Autoreg procedure.

ASSUMPTIONS OF A LINEAR RELATIONSHIP: COMMON TESTS

[Testing Outliers] Box Plot Method: If a value is higher than the 1.5*IQR above the upper quartile (Q3), the value will be considered as outlier. Similarly, if a value is lower than the 1.5*IQR below the lower quartile (Q1), the value will be considered as outlier. In SAS, the plots option in the Univariate procedure tells SAS to generate Box Plot graph.

See this website for information on a macro that uses the Box Plot Method.

[Testing Outliers] Studentized Residuals: Residuals are the difference between the observed value and the predicted value. Standardized Residuals are the residuals divided by the standard error of estimate. Lastly, studentized Residuals are the residuals divided by the standard error of the residual with that case deleted. If an absolute value of studentized residual is greater than 3, the observation is considered as an outlier.

```
/* Studentized residuals - Check Outliers*/
ods graphics on;
proc reg data=health;
      model cholesterolloss = age weight cholesterol triglycerides hdl
      ldl height skinfold systolicbp diastolicbp exercise coffee / stb
      clb;
      output out=stdres p= predict r = resid rstudent=r h=lev
      cookd=cookd dffits=dffit;
run;
quit;
ods graphics off;

/* Print only those observations having absolute value of studentized
residual greater than 3*/
proc print data=stdres;
      var age weight cholesterol triglycerides hdl ldl height skinfold
      systolicbp diastolicbp exercise coffee cholesterolloss;
      where abs(r)>=3;
run;
```

Obs	Age	Weight	Cholesterol	Triglycerides	HDL	LDL	Height	Skinfold	SystolicBP	DiastolicBP	Exercise	Coffee	CholesterolLoss
23	23	182	189	47	50	138.2	75.5	10	124	73	60	1	-73

Figure 14: Outlier Results for Studentized Residuals

[Testing Outliers] Cook's D: Cook's D can also be used to test for outliers. The general rule of thumb, is the higher the value of Cook's D, the more influential the point is. Additionally, if the Cook's D value is greater than 4/(number of observations), the value is considered an outlier.

```
/* Cook's D - Check Outliers */
proc print data=stdres;
      where cookd > (4/51);
      var age weight cholesterol triglycerides hdl ldl height skinfold
systolicbp diastolicbp exercise coffee cholesterolloss;
run;
```

Obs	Age	Weight	Cholesterol	Triglycerides	HDL	LDL	Height	Skinfold	SystolicBP	DiastolicBP	Exercise	Coffee	CholesterolLoss
9	23	178	234	307	28	201.1	73.5	5	124	82	60	1	59
11	26	188	258	299	30	223.2	73	19	130	86	0	1	-18
12	22	150	212	52	69	142.2	64.25	15	120	74	0	0	-16
13	22	123	137	158	29	105.5	64.25	21	120	74	0	4	-28
23	23	182	189	47	50	138.2	75.5	10	124	73	60	1	-73
29	28	150	228	480	29	191.3	66	22	138	82	120	0	4
34	40	217	277	240	71	202.2	75	30	128	80	0	5	34

Figure 15: Outlier Results for Cook's D

[Testing Linearity] Scatter Plot: You can use a scatter plot to test for linearity. Does the scatter plot show a linear pattern in the data?

```
/* Scatter Plot for Testing Linearity */
ods graphics on;
proc reg data=health;
      model cholesterolloss = age weight cholesterol triglycerides hdl
      ldl height skinfold systolicbp diastolicbp exercise coffee /
      partial;
run;
quit;
ods graphics off;
```

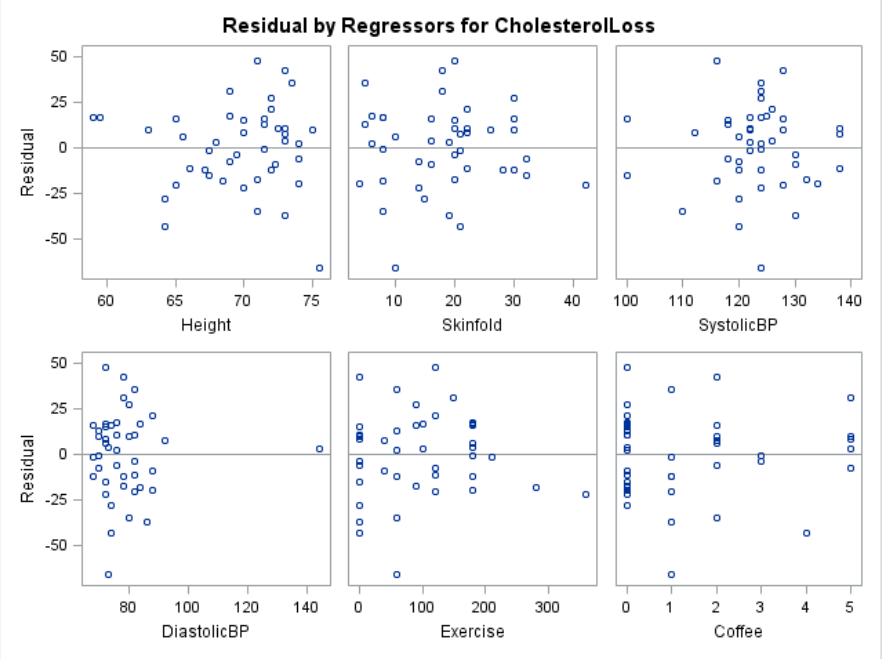

Figure 16: Scotter Plot Results for Testing Linearity

[Testing Linearity] Correlation Between Independent and Dependent Variable: There should be a moderate and SIGNIFICANT correlation score between dependent variable and independent variable.

```
/* Testing Correlation between Dependent and Independent Variables */
proc corr data=health;
```

```
        var age weight cholesterol triglycerides hdl ldl height skinfold
        systolicbp diastolicbp exercise coffee;
        with cholesterolloss;
run;
```

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	Age	Weight	Cholesterol	Triglycerides	HDL	LDL	Height	Skinfold	SystolicBP	DiastolicBP	Exercise	Coffee
CholesterolLoss	0.09914 0.5270 43	-0.24221 0.1176 43	0.40318 0.0073 43	0.11396 0.4669 43	0.19099 0.2199 43	0.37389 0.0135 43	-0.27042 0.0795 43	-0.03538 0.8218 43	-0.07917 0.6138 43	0.13192 0.3991 43	0.22724 0.1428 43	0.08732 0.5777 43

Figure 17: Correlation Coefficients for Independent vs Dependent Variables

NOTES ON TRANSFORMING VARIABLES TO MEET AN ASSUMPTION

Transforming variables is often done to correct for outliers and assumption failures (normality, linearity, and homoscedasticity/homogeneity); however, interpretation is then limited to the transformed scores. Examples of different transformations are: taking the square root of the variable(s); taking the natural logarithm; multiplicative inverse; for skewed variables, and reflecting the variable before applying the desired transformation.

- Violations of homogeneity usually can be corrected by transforming the DV. If you can not transform the DV, then you can use a more stringent alpha level for the untransformed DV

- Ensure that the transformed variable(s) meets the assumptions (such as normality, little to no outliers, etc…). Often, you are not sure what transformation would work best to meet the assumptions; trial and error.

- Usually, if some variables are skewed and others are not, the transformations provide an improvement; however, that is not always the case.

- To transform for normality: According to some research, taking the inverse of the scores is the best of several alternatives for skewed (or J-shaped) distributions. However, according to Tabachnick & Fidell (2007), this alternative may not render the distribution normal.

- When the error variance appears to be constant (Homoscedasticity), only X needs be transformed to linearize the relationship. Transform independent variable to Log10(X), Inverse(X), Square root(X), Square(X), Exp(X), 1/X, Exp(-X). When the error variance does not appear constant it may be necessary to transform Y or both X and Y. You can then run Box-Cox Transformations for Dependent Variable to control for any additional violations.

Additionally, examining the means for untransformed scores is the same as examining the medians for transformed scores; the transformation affects the mean but not the median because the median only depends on rank order. Therefore, the means of transformed variables is the same as the median of untransformed variables.

According to Tabachnick & Fidell (2007), to reflect a variable, find the largest score in the distribution and then add 1 to it; this forms a constant that is larger than any other score in the distribution. Create a new variable by subtracting each score from the constant. Interpret this reflected variable appropriately: reverse the direction of the interpretation or re-reflect the variable after transforming it; or, keep in mind that if smaller scores represented negative units before the transformation, then after the transformation the smaller scores will represent positive units.

CONCLUSION

In order to ensure that your model is appropriately interpreted, it is important to make sure that all assumptions have been tested and any violations have been corrected. This may seem daunting, but the processes to do this are easy and the corrections are not painful. It is always worth it to make sure that the results you are reporting are correct!

REFERENCES & RECOMMENDED READING

Beck, N. & Jackman, S. (1998). Beyond Linearity by Default: Generalized Additive Models. American Journal of Political Science. 42 (2): p 596-627.

Beef, A., le Cessie, S., & Dekkers, O. M. (2015). Models with Transformed Variables: Interpretation and Software. Epidemiology. 26(2): e16-e17.

Bhalla, D. (Accessed August 2018). Identify and Remove Outliers With SAS. Available at: https://www.listendata.com/2014/10/identify-and-remove-outliers-with-sas.html

Bhalla, D. (Accessed August 2018). Detect Non-Linear and Non-Monotonic Relationship Between Variables. Available at: https://www.listendata.com/2015/03/detect-non-linear-and-non-monotonic.html

Box, G. E. P., 1949. A general distribution theory for a class of likelihood criteria. Biometrika, 36: 317–346.

Cohen, B. (2008). Explaining Psychological Statistics. John Wiley & Sons.

Cornell University (2012). Interpreting Coefficients in Regression with Log-Transformed Variables. StatNews #83. Available at: https://www.cscu.cornell.edu/news/statnews/stnews83.pdf

Eckel, S. (2008). Lecture 14: Interpreting logistic regression models. [Lecture]

Elswick, R. K., Schwartz, P. F., & Welsh, J. (1997). Interpretation of the odds ratio from logistic regression after a transformation of the covariate vector. Statistics in Medicine. 16: p 1695-1703.

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., & Tu, X. M. (2014). Log-transformation and its implications for data analysis. Biostatistics in psychiatry. 26(2): 105-109.

Hahs-Vaughn, D. (2016). Applied Multivariate Statistical Concepts. Taylor & Francis.

Hocking, R. (2003). Methods and applications of linear models. 2nd Edition, Wiley.

Huang, C. (2012). Detecting and Responding to Violations of Regression Assumptions. [Lecture].

Layard, M. (1974). A Monte Carlo Comparison tests for equality of covariance matrices. Biometrika. 16, 461-465.

Lund, B. (2015). Selection and Transformation of Continuous Predictors for Logistic Regression. Paper 2687-2015.

MathWorks. (2018). Nonlinear Logistic Regression. Available at: https://www.mathworks.com/help/stats/examples/nonlinear-logistic-regression.html

Statistics Solutions. (2018). Assumptions of Linear Regression. Available at www.statisticssolutions.com/assumptions-of-linear-regression/

Statistics Solutions. (2018). Assumptions of Logistic Regression. Available at www.statisticssolutions.com/assumptions-of-logistic-regression/

Statistics Solutions. (2018). Transforming variables to meet an assumption. Available at: http://www.statisticssolutions.com/transforming-variables-to-meet-an-assumption/

Stoltzfus, J. C. (2011). Logistic Regression: A Brief Primer. Society for Academic Emergency Medicine. Doi: 10.1111/j.1553-2712.2011.01185.x

Myers, R. (1990). Classical and modern regression with applications. 2nd Edition, Duxbury.

Politzer-Ables, S. (2016). Interpretation of coefficients in logistic regression. Available at: http://www.mypolyuweb.hk/~sjpolit/logisticregression.html

Ravishanker, N. & Dey, D. (2002). A First course in linear model theory, Chapman & Hall / CRC, Boca Raton.

Tabachnick, B.G. and Fidell, L.S. (2007). Using Multivariate Statistics. 5th ed. Boston: Allynand Bacon.

Teknomo, K.(2018). Nonlinear Transformation. Available at: http://people.revoledu.com/kardi/

Sarlija, N., Bilandzic, A., and Stanic, M. (2017). Logistic regression modelling: procedures and pitfalls in developing and interpreting prediction models. Croation Operational Research Review. 8: 631-652.

UCLA Statistics. (2016). FAQ How do I interpret a regression model when some variables are log transformed? Available at: https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqhow-do-i-interpret-a-regression-model-when-some-variables-are-log-transformed/

UCLA Statistics. (2016). FAQ How do I interpret a regression model when some variables are log transformed? Available at: https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/

Burok. (Unknown). Dealing with Model Assumption Violations. [Lecture Notes]. Available at: https://academic.macewan.ca/burok/Stat378/notes/remedies.pdf

ACKNOWLEDGMENTS

Thank you to Lisa Shank and Natasha Schvey for all your statistical questions concerning assumption violations. If it wasn't for you, I would not have been motivated to write this paper!

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Deanna N Schreiber-Gregory, MS
Data Analyst / Research Associate
Contractor with Henry M Jackson Foundation for the Advancement of Military Medicine
Department of Internal Medicine
Uniformed Services University of the Health Sciences
E-mail: d.n.schreibergregory@gmail.com

Karlen Bader
Research Assistant
Henry M Jackson Foundation for the Advancement of Military Medicine
Uniformed Services University / Walter Reed Medical Center