SESUG Paper 121-2025

Multiple Logistic Regression Using SAS and SPSS

Amita Patil, Walden University; Manaswi Chigurupati, Assure Specialty Pharmacy;
Rachel TeWinkel, Walden University, Kelcey Sihanourath, Walden University.

ABSTRACT

This paper describes how to conduct a multiple logistic regression using SAS and SPSS. It gives details on a) framing a research question, b) running all assumptions, c) running multiple logistic regression model, d) interpretation, e) reporting results, and f) writing results in APA style for publication.

INTRODUCTION

This paper describes steps for framing a research question, developing null and alternative hypotheses, and checking assumptions and conducting multiple logistic regressions in SAS and SPSS. While developing research question the use of terms like 'predict' or 'likelihood' indicates regression model is used. It is a good practice to list independent variables (predictors) first followed by dependent variable (outcome) while framing a research question. Null hypothesis indicates the independent variables do not predict the outcome whereas alternative hypothesis indicates that independent variables predict outcome. To answer the research question, assumptions are checked, and if met, multiple logistic regression is run. As outcome is dichotomous, we conducted binary logistic regression. If the outcome is ordinal, use ordinal logistic regression, and if outcome categories are more than two then use multinomial logistic regression technique.

Research Question: To what extent do cholesterol, weight status, and sex predict the likelihood of high blood pressure levels?

Null Hypothesis (H_o): Cholesterol, weight status, and sex do not predict the likelihood of high blood pressure level.

Alternative Hypothesis (H_a): Cholesterol, weight status, and sex do predict the likelihood of high blood pressure level.

METHODS

Multiple logistic regression steps using SAS and SPSS are described below. Steps are conducted in SAS 'On Demand for Academics' and SPSS v30. Data is from SASHELP library, Heart dataset.

Assumptions

1) Dependent variable should be measured at the nominal level (dichotomous here)
2) Independent variables can be continuous, or categorical
3) Independence of observations and the dependent variable should have mutually exclusive and exhaustive categories
4) No multicollinearity among independent variables (if two independent continuous variables)
5) Linear relationship between any continuous independent variables and the logit transformation of the dependent variable. A continuous independent variable is "linear on the log-odds scale" if the relationship between it and the natural log of the odds (also known as the "logit") of the dependent variable is linear.
6) No extreme outliers

MULTIPLE LOGISTIC REGRESSION USING SAS

Assumption Testing in SAS

The first assumption fulfills as outcome variable High_BP is nominal variable with two categories. Hence, we will use binary logistic regression. We have more than one predictor variable, so the second assumption is met as we have two categorical independent variables i.e., Weight_status and Sex, and one continuous independent variable Cholesterol (refer Table 1). The outcome is categorized as '1: Yes' and '0: No' and are mutually exclusive and exhaustive, used from SASHELP Heart dataset

If independent variable is categorical then cell frequency should be greater than 5. After running the SAS code, we get that expected count of predictors Sex with High_BP >5 for all cells and Weight_status with High_BP >5 for all cells (see Figure 1 and Figure 2). Multicollinearity is checked when we have two continuous independent variables using Pearsons's correlation test not described in paper as we do not have two independent continuous variables.

For assumption #6, visual inspection of histogram of Cholesterol indicates normal distribution and box plot shows outliers (see Figure 3). The skewness for Cholesterol variable is 0.8 and kurtosis is 2.1, which are in normal limits (see Figure 4). Outliers to be removed from data before proceeding to final analysis.

The fifth assumption is linearity on the log-odds scale. This would mean that the natural log of the odds of High_BP has a linear association with the independent variable, Cholesterol. Most of the time solution is to categorize continuous types of variables so that we no longer must worry about the assumption of linearity on the log-odds scale. Instead, we conducted the Box-Tidwell test where an insignificant *p*-value would indicate the assumption is met. We created a Box-Tidwell variable named BT_Chol by multiplying Cholesterol variable and with its log variable i.e., log (Cholesterol). Then run logistic regression using Cholesterol and Box-Tidwell variable, including all other variables. The p-value is 0.45, hence assumption is met **(**see Figure 5).

As all our assumptions were met, we can proceed with multiple logistic regression. The variance and model fit is explained in the output of multiple logistic regression is shown in Figure 6. Results remain same for SAS and SPSS, hence we have reported towards end and are explained in APA style in 'RESULT' section Figure 7 for SAS results.

Table 1

SAS Variable Name	Type	Categories
High_BP	Character	1: Yes 0: No
Sex	Character	Male, Female
Weight_Status	Character	Underweight, Normal, Overweight
Cholesterol	Numeric	Numeric

SAS CODE

```
*Run contents to understand the type of variable;
PROC CONTENTS DATA=BPdataset;
RUN;

*Run a chi-sq test for each categorical predictor with outcome to check for expected count assumption, if expected counts >5 then assumption is met ;
PROC FREQ DATA=BPdataset;
TABLES SEX*HIGH_BP/CHISQ EXPECTED;
RUN;
PROC FREQ DATA= BPdataset;
TABLES WEIGHT_STATUS*HIGH_BP/CHISQ EXPECTED;
RUN;

*Run a distribution check for continuous predictors assumptions to see if it is normal and check for outliers in box plot to check for assumption #6;
PROC UNIVARIATE DATA= BPdataset;
VAR CHOLESTEROL;
HISTOGRAM;
RUN;

PROC SGPLOT DATA= BPdataset;
VBOX CHOLESTEROL;
RUN;

*Run assumptions #5 check, linear relationship between continuous independent variables and the logit transformation of the dependent variable;
*Create a Box-Tidwell variable;
DATA BPdataset;
SET BPdataset;
log_cholesterol=log(CHOLESTEROL);
BT_Chol= CHOLESTEROL* log_cholesterol;
RUN;
```

```
*Run regression using continuous variable and Box Tidwell variable and all variables for assumption check;;
PROC LOGISTIC DATA= BPdataset;
Model High_BP= CHOLESTEROL BT_Chol SEX WEIGHT_STATUS;
RUN;

*All assumptions are met so we will proceed with multiple logistic regression analysis, you can use descending function if your event is coded 1, as by default SAS predicts probability of 0. We used event="1). The  lackfit will run Hosmer-Lemeshow test and RSQ will produce r-squared values;
PROC LOGISTIC DATA= BPdataset;
        CLASS SEX (REF="Female") WEIGHT_STATUS (REF="Normal");
        MODEL HIGH_BP (event="1") = CHOLESTEROL SEX WEIGHT_STATUS/ lackfit rsq;
RUN;
```

Figure 1

Assumption Check for Categorical Predictor Sex using Chi-sq Test in SAS

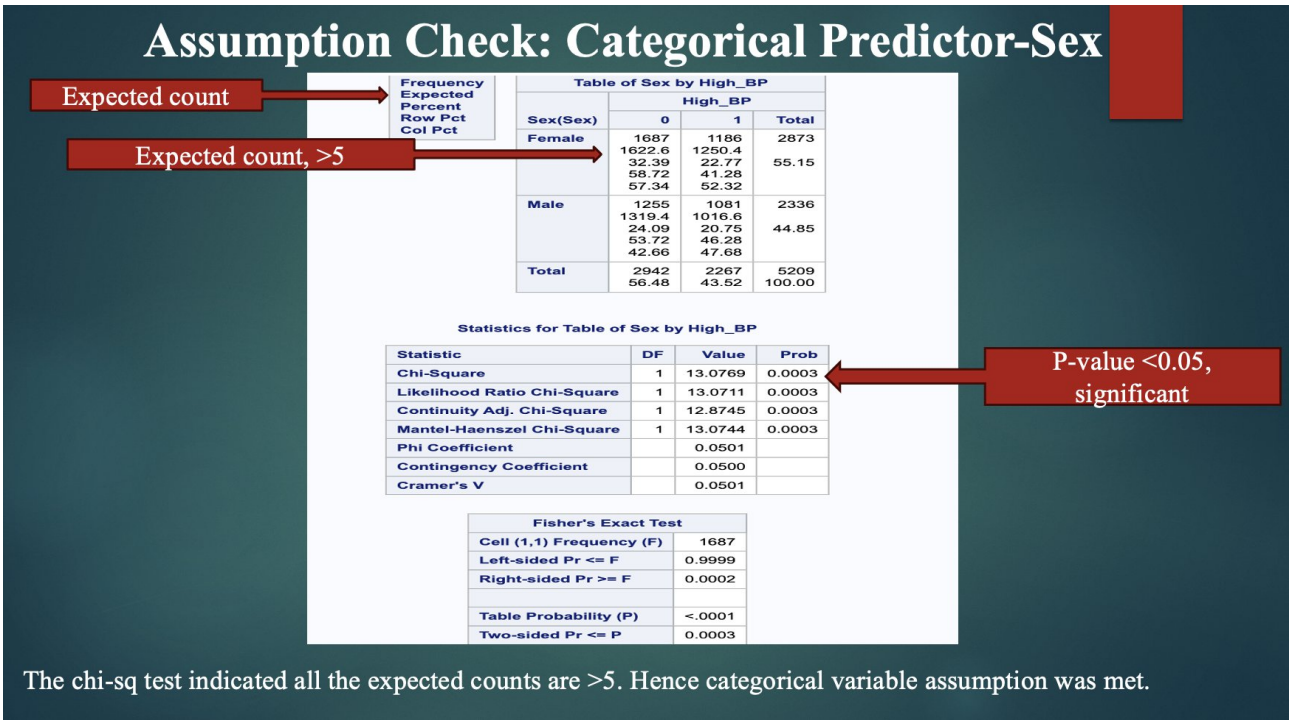


Table of Sex by High_BP

Frequency / Expected / Percent / Row Pct / Col Pct	High_BP 0	High_BP 1	Total
Female	1687 1622.6 32.39 58.72 57.34	1186 1250.4 22.77 41.28 52.32	2873 55.15
Male	1255 1319.4 24.09 53.72 42.66	1081 1016.6 20.75 46.28 47.68	2336 44.85
Total	2942 56.48	2267 43.52	5209 100.00

Statistics for Table of Sex by High_BP

Statistic	DF	Value	Prob
Chi-Square	1	13.0769	0.0003
Likelihood Ratio Chi-Square	1	13.0711	0.0003
Continuity Adj. Chi-Square	1	12.8745	0.0003
Mantel-Haenszel Chi-Square	1	13.0744	0.0003
Phi Coefficient		0.0501	
Contingency Coefficient		0.0500	
Cramer's V		0.0501	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	1687
Left-sided Pr <= F	0.9999
Right-sided Pr >= F	0.0002
Table Probability (P)	<.0001
Two-sided Pr <= P	0.0003

Figure 2

Assumption Check for Categorical Predictor Weight Status using Chi-sq Test in SAS

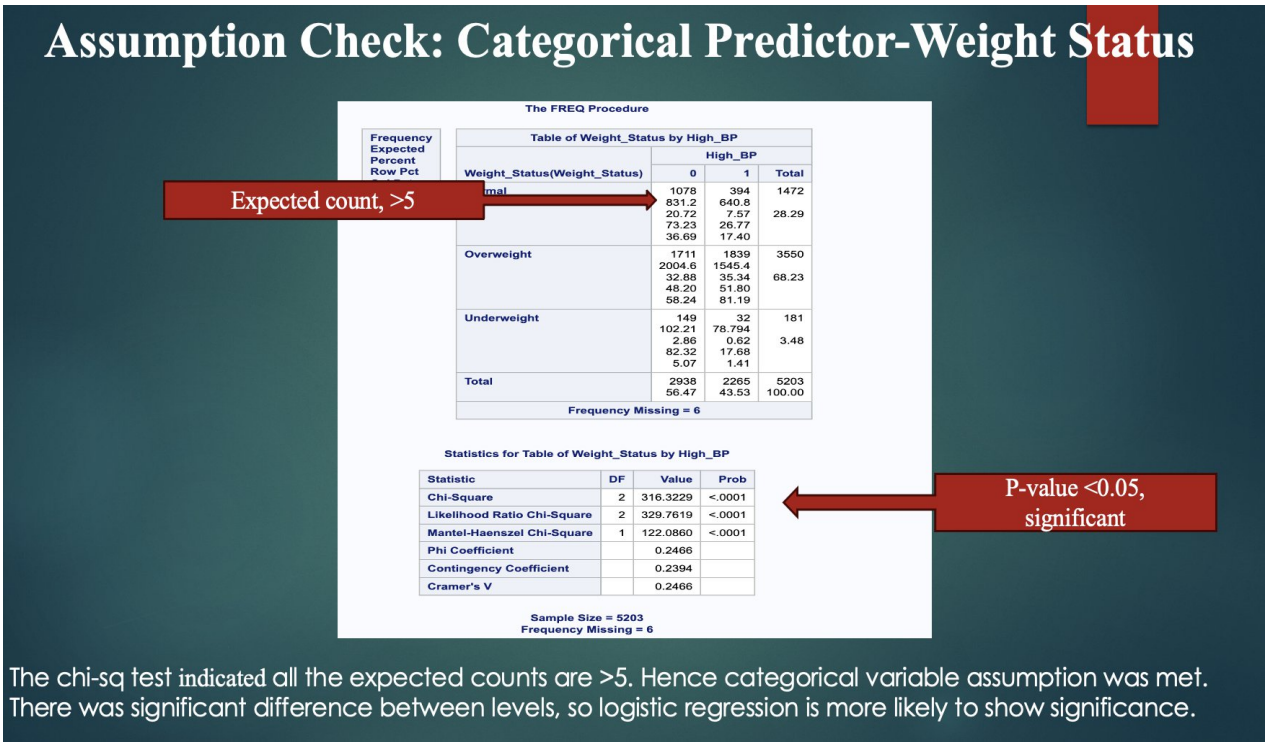


Figure 3

Assumption Check for Continuous Predictor Cholesterol using Histogram and Box Plot in SAS

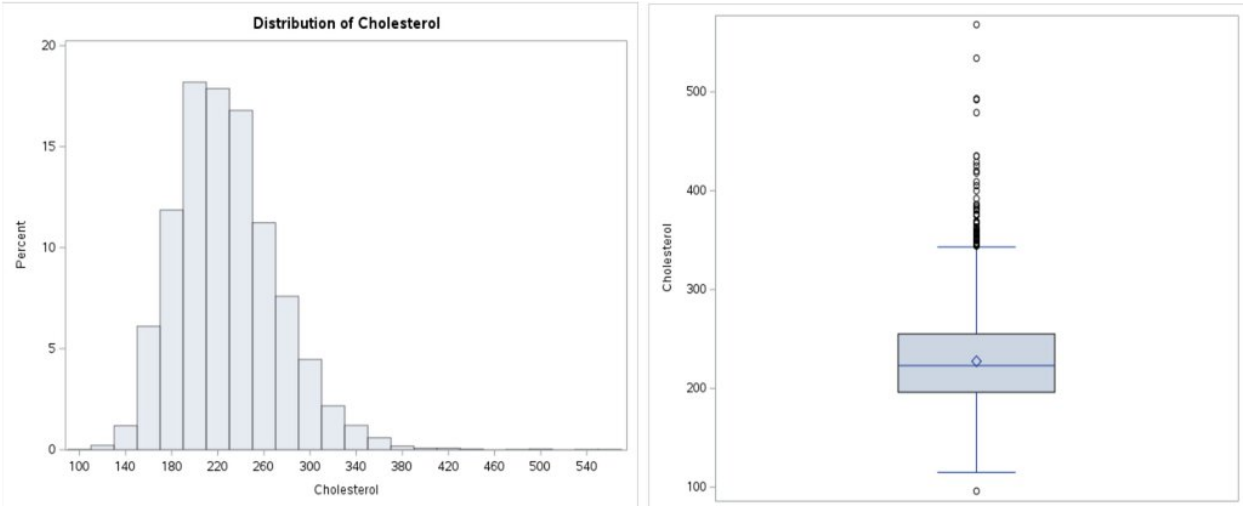


Figure 4

Assumption Check for Continuous Predictor Cholesterol: Skewness and Kurtosis in SAS

Moments			
N	5057	Sum Weights	5057
Mean	227.417441	Sum Observations	1150050
Std Deviation	44.9355238	Variance	2019.2013
Skewness	0.81634421	Kurtosis	2.10376843
Uncorrected SS	271750510	Corrected SS	10209081.8
Coeff Variation	19.7590491	Std Error Mean	0.63189269

Basic Statistical Measures			
Location		Variability	
Mean	227.4174	Std Deviation	44.93552
Median	223.0000	Variance	2019
Mode	200.0000	Range	472.00000
		Interquartile Range	59.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	359.8988	Pr > |t|	<.0001
Sign	M	2528.5	Pr >= |M|	<.0001
Signed Rank	S	6394577	Pr >= |S|	<.0001

Figure 5

Assumption Check for **Linear Relationship Between Continuous Independent Variable Cholesterol and the Logit Transformation of the Dependent Variable** in SAS

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	3.1498	1.0772	8.5507	0.0035
Cholesterol		1	-0.0290	0.0293	0.9768	0.3230
BT_Chol		1	0.00339	0.00452	0.5629	0.4531
Sex	Female	1	0.0908	0.0298	9.2868	0.0023
Weight_Status	Normal	1	0.1761	0.0784	5.0443	0.0247
Weight_Status	Overweight	1	-0.8120	0.0733	122.5998	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Cholesterol	0.971	0.917	1.029
BT_Chol	1.003	0.995	1.012
Sex Female vs Male	1.199	1.067	1.348
Weight_Status Normal vs Underweight	0.631	0.420	0.950
Weight_Status Overweight vs Underweight	0.235	0.158	0.350

Figure 6

R-square and Hosmer-Lemeshow Goodness-of-Fit Test in SAS

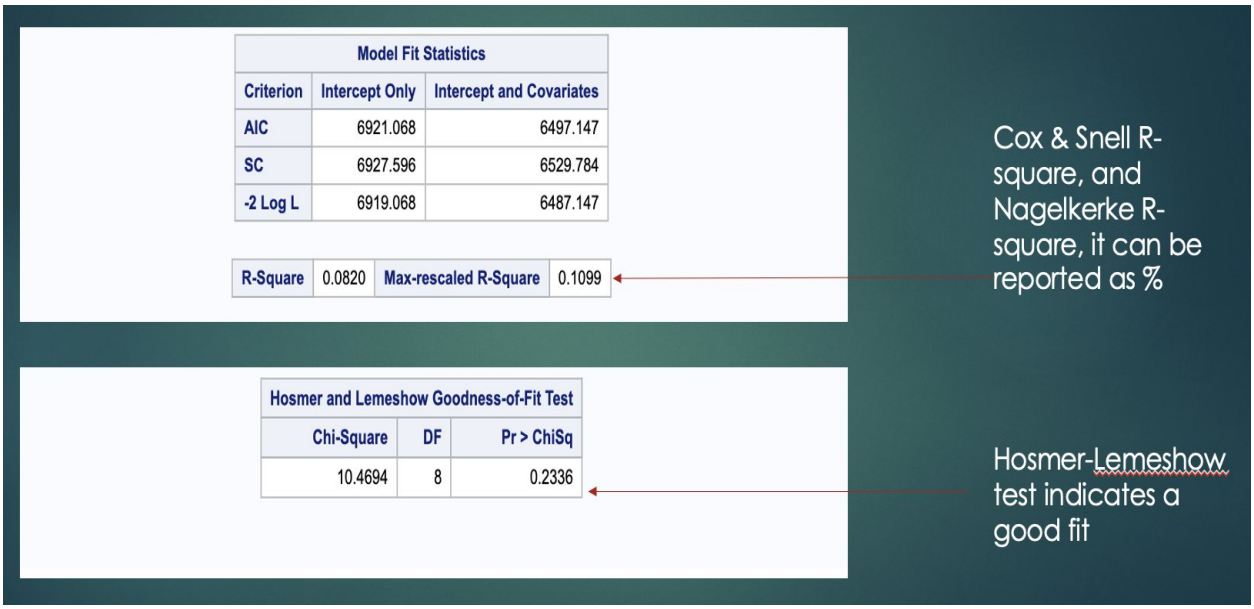
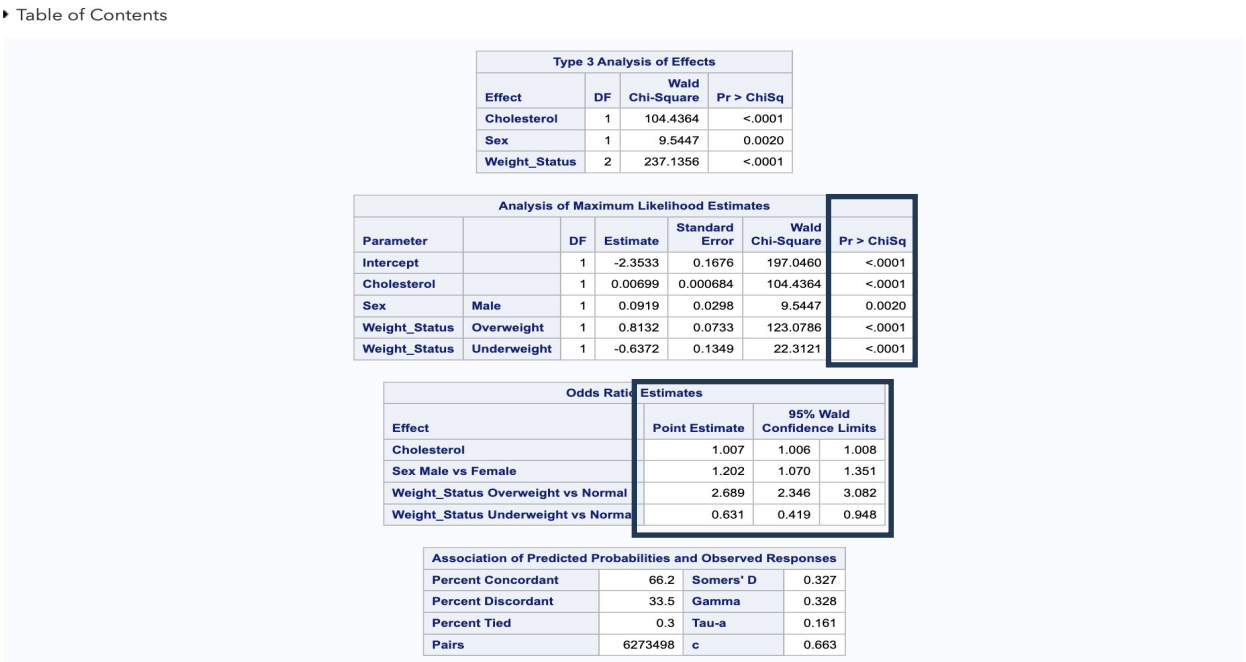


Figure 7

Multiple Logistic Regression Results in SAS



MULTIPLE LOGISTIC REGRESSION USING SPSS

Assumption Testing in SPSS

For categorical predictors run chi-sq test, we see expected counts >5 for Sex and Weight_status (see Figure 8 and Figure 9)

1) Analyze →Descriptive Statistics → Crosstabs…
2) In “Column(s)” place the outcome variable and in “Row(s)” place the predictor
3) Click “Statistics” and check “Chi-Square”. Click Continue
4) Click “Cells” and check “Observed”, “Expected” and request Percentages for “Rows”
5) Click OK to run analysis

Figure 8

Assumption Check for Categorical Predictor using Chi-sq/Crosstab in SPSS

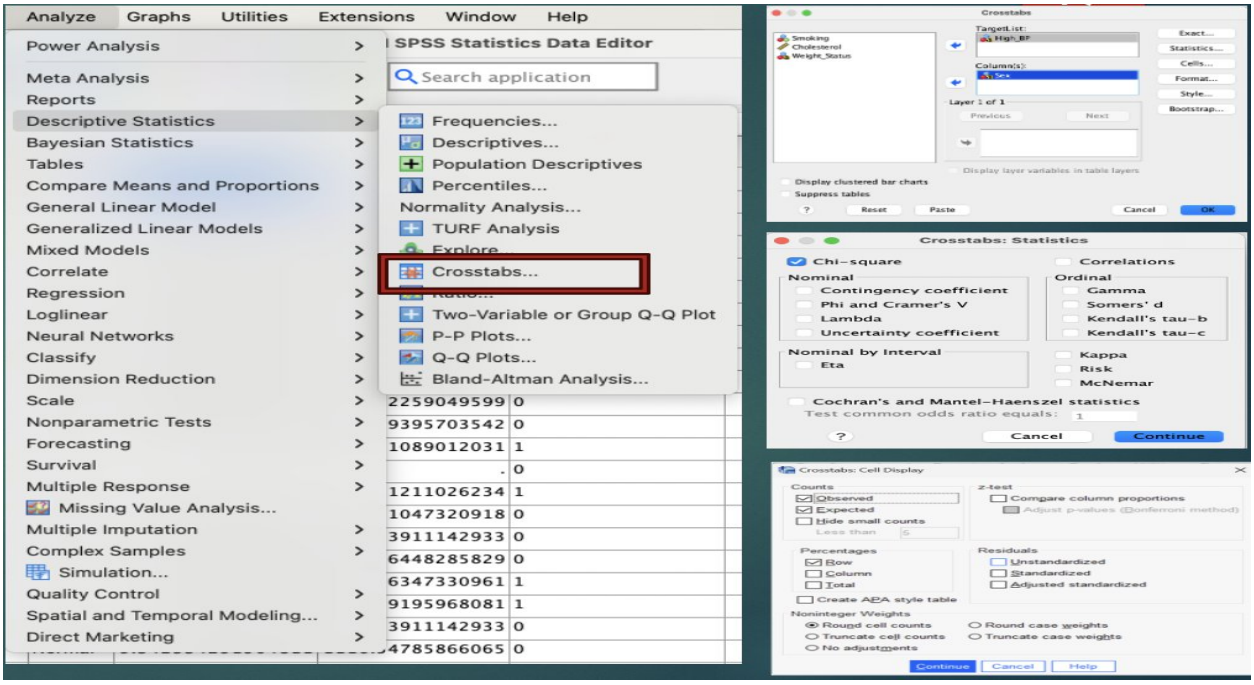
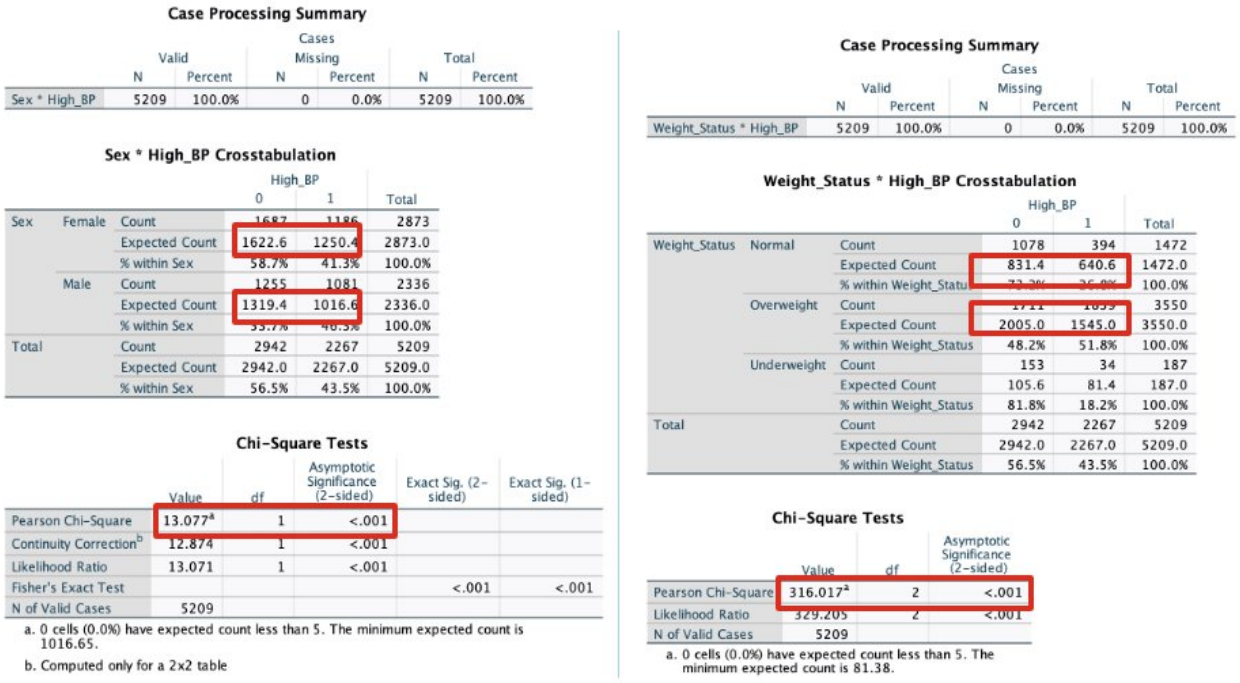


Figure 9

Assumption Check for Categorical Predictor Sex and Weight_status in SPSS



Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Sex * High_BP	5209	100.0%	0	0.0%	5209	100.0%

Sex * High_BP Crosstabulation

			High_BP		
			0	1	Total
Sex	Female	Count	1687	1186	2873
		Expected Count	1622.6	1250.4	2873.0
		% within Sex	58.7%	41.3%	100.0%
	Male	Count	1255	1081	2336
		Expected Count	1319.4	1016.6	2336.0
		% within Sex	53.7%	46.3%	100.0%
Total		Count	2942	2267	5209
		Expected Count	2942.0	2267.0	5209.0
		% within Sex	56.5%	43.5%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	13.077[a]	1	<.001		
Continuity Correction[b]	12.874	1	<.001		
Likelihood Ratio	13.071	1	<.001		
Fisher's Exact Test				<.001	<.001
N of Valid Cases	5209				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 1016.65.

b. Computed only for a 2x2 table

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Weight_Status * High_BP	5209	100.0%	0	0.0%	5209	100.0%

Weight_Status * High_BP Crosstabulation

			High_BP		
			0	1	Total
Weight_Status	Normal	Count	1078	394	1472
		Expected Count	831.4	640.6	1472.0
		% within Weight_Status	73.2%	26.8%	100.0%
	Overweight	Count	1711	1839	3550
		Expected Count	2005.0	1545.0	3550.0
		% within Weight_Status	48.2%	51.8%	100.0%
	Underweight	Count	153	34	187
		Expected Count	105.6	81.4	187.0
		% within Weight_Status	81.8%	18.2%	100.0%
Total		Count	2942	2267	5209
		Expected Count	2942.0	2267.0	5209.0
		% within Weight_Status	56.5%	43.5%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	316.017[a]	2	<.001
Likelihood Ratio	329.205	2	<.001
N of Valid Cases	5209		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 81.38.

For checking assumption, linear relationship with logit-odds, we transform variable. To assess this, use the Box-Tidwell Test (see Figure 10).

Steps:

1) Transform →Compute Variable
2) In “Numeric Expression” type Ln(predictor)*predictor
3) In “Target Variable” name the new variable BT_predictor

Using Box-Tidwell variable run logistic regression (see Figure 11).

1) Analyze → Regression → Binary logistic
2) In “Dependent” place the outcome
3) In “Covariates” place the BT_predictor, the original predictor, and all other variables in the model
4) Click OK to run analysis

Figure 10

Continuous Predictor Linearity with Logit-odds Assumption Check in SPSS

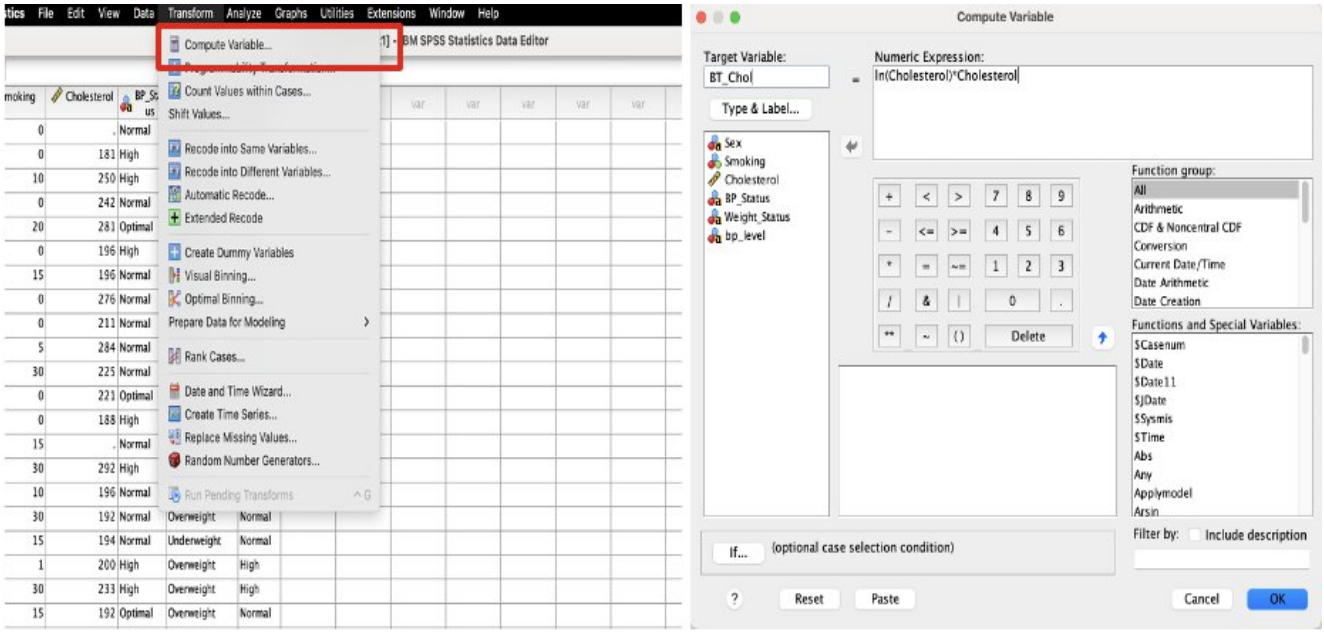
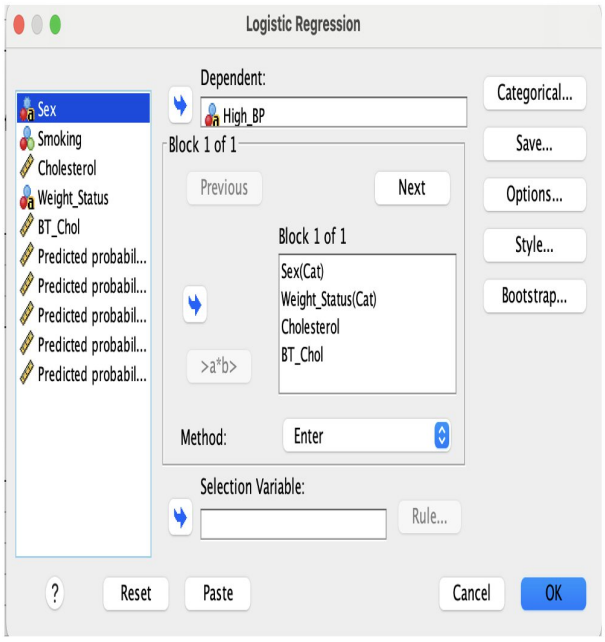


Figure 11

Run Logistic Regression for Box-Tidwell Variable in SPSS



Classification Table[a]

Observed			Predicted High_BP 0	Predicted High_BP 1	Percentage Correct
Step 1	High_BP	0	1996	857	70.0
		1	1049	1155	52.4
	Overall Percentage				62.3

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Sex(1)	.186	.060	9.745	1	.002	1.204
	Weight_Status			236.587	2	<.001	
	Weight_Status(1)	.432	.203	4.524	1	.033	1.540
	Weight_Status(2)	1.420	.197	51.822	1	<.001	4.136
	Cholesterol	.029	.029	.984	1	.321	1.029
	BT_Chol	-.003	.005	.567	1	.451	.997
	Constant	-3.859	1.089	12.560	1	<.001	.021

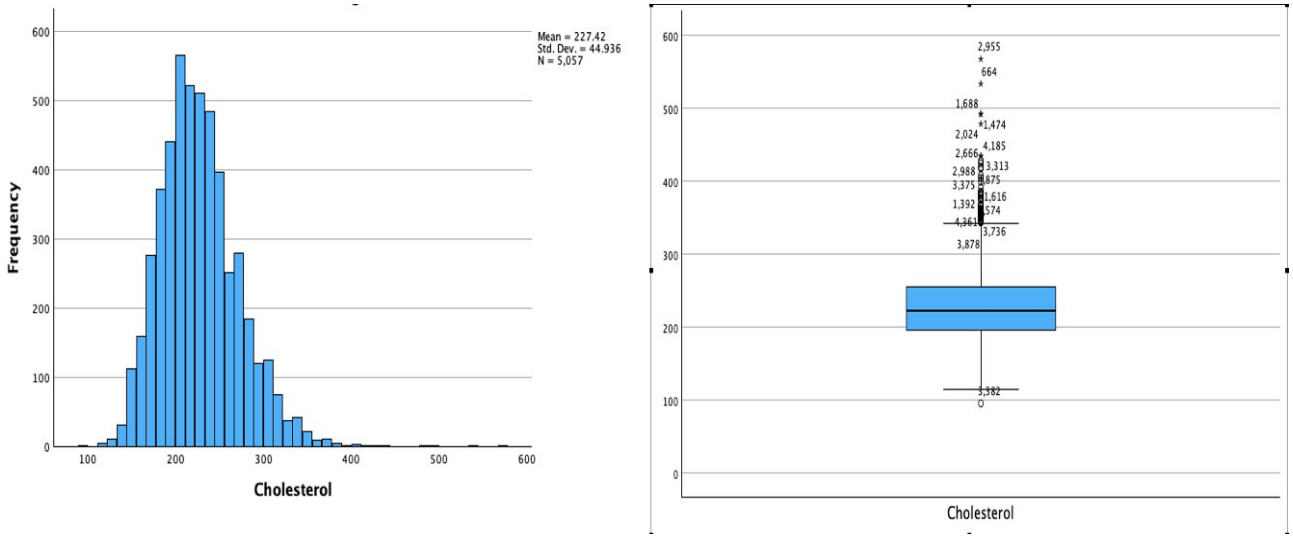
a. Variable(s) entered on step 1: Sex, Weight_Status, Cholesterol, BT_Chol.

Last assumption #6 check for continuous predictor (see Figure 12).

1. Analyze → Descriptive Statistics → Explore
2. In “Dependent List” place all continuous variables
3. Click OK to run analysis

Figure 12

Histogram and Boxplot of Cholesterol in SPSS



Conducting Multiple Logistic Regression (see Figure 13)

1. Analyze → Regression → Binary Logistic
2. In “Dependent” place the outcome and in “Covariates” place all the predictors.
3. In “Save” check “Probabilities”
4. In “Options” check “CI for exp(B),” and “Hosmer-Lemeshow goodness of fit”

IF USING CATEGORICAL PREDICTOR

5. In “Categorical” move predictor to “Categorical Covariates” and select a reference category
6. “First” means the smaller number is the reference category

   (If 1 = *males* and 0 = *females*; then *females* is reference group)
7. Click OK to run analysis

Figure 13

Multiple Logistic Regression Steps in SPSS

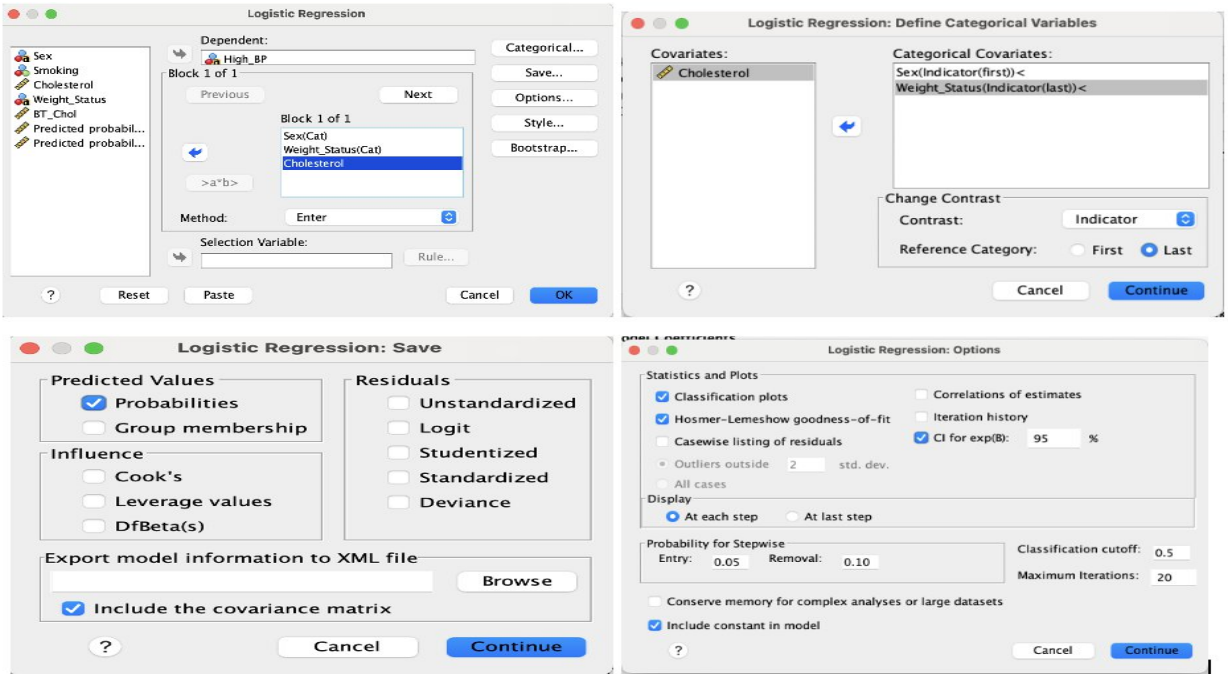


Figure 14

Multiple Logistic Regression Model Fit in SPSS

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	432.725	4	<.001
	Block	432.725	4	<.001
	Model	432.725	4	<.001

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	6494.245[a]	.082	.110

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Figure 15

Multiple Logistic Regression Results in SPSS

Classification Table[a]

Observed			Predicted High_BP 0	Predicted High_BP 1	Percentage Correct
Step 1	High_BP	0	2005	848	70.3
		1	1070	1134	51.5
	Overall Percentage				62.1

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
Step 1[a]	Sex(1)	.188	.059	10.011	1	.002	1.207	1.074	1.356
	Weight_Status			237.279	2	<.001			
	Weight_Status(1)	.433	.203	4.560	1	.033	1.542	1.036	2.296
	Weight_Status(2)	1.422	.197	52.064	1	<.001	4.146	2.818	6.102
	Cholesterol	.007	.001	105.341	1	<.001	1.007	1.006	1.008
	Constant	-3.062	.245	156.324	1	<.001	.047		

a. Variable(s) entered on step 1: Sex, Weight_Status, Cholesterol.

*Note: SPSS result for weight status is different from SAS, as weight categories can be referenced either Indicator 'first' or 'last'. If you want to keep refence group specific recode the variable using Transform then recode into different variable and then use that one, so that you can use Indicator function.

RESULT

The logistic regression results were significant, $\chi^2(4, 5209) = 432.72$, $p = .001$. The model explained between 8.2% (Cox and Snell R^2) and 11.0% (Nagelkerke R^2) of the variance in High BP level. The Hosmer-Lemeshow test indicated the model was a good fit at $\chi^2(8) = 10.57$, $p = .233$ (see

Figure 6, Figure 14). Males are more likely to have high blood pressure level compared to females, Wald (1) = 10.01, $p = .002$, (OR 1.207, 95% CI[1.074, 1.356]) controlling for weight status, and cholesterol (see Figure 7,Figure 15). Participants having higher cholesterol were more likely to have high BP level, Wald (1) = 105.341, $p < .0001$, (OR=1.007, 95% CI [1.006, 1.008]), controlling for sex and weight status. Being overweight had 2.68 times more odds of having high blood pressure when compared normal weight, $p < .0001$, 95% CI [2.34, 3.08]) whereas underweight were less likely to have high BP when compared normal weight, $p < .0001$, (OR=0.631, 95% CI [0.0419, 0.948]), refer to Figure 7.

CONCLUSION

It is essential to run assumption check before running logistic regression model. If assumptions are not met there are techniques like data transformation, removing outliers, recoding variables, if cell size smaller than combining the categories with justification and literature reference.

REFERENCES

Grubber, Janet. 2019. “The Thorn in My Side!! Logistic Regression Continuous Variables that Violate the Assumption of Linearity on the Log-odd” SESUG Paper 233-2019: Lexjansen. Available at https://www.lexjansen.com/sesug/2019/SESUG2019_Paper-233_Final_PDF.pdf

Lund Research Ltd. (2018). “Binomial logistic regression using SPSS statistics.” Laerd Statistics. https://statistics.laerd.com/spss-tutorials/binomial-logistic-regression-using-spss-statistics.php

Wagner, W. E., III. (2019-04-17). Using IBM® SPSS® Statistics for Research Methods and Social Science Statistics, 7th Edition. [VitalSource Bookshelf 10.5.3]. Retrieved from vbk://9781506388991

ACKNOWLEDGMENTS

Special thanks to Henna Patani for reviewing the paper.

RECOMMENDED READING

Base SAS Programing Guide

Chapter 9: Using IBM® SPSS® Statistics for Research Methods and Social Science Statistics 7th Edition

CONTACT INFORMATION

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

Amita Patil
Walden University
Amita.patil705@gmail.com