Data Screening for Multiple Regression

Sparky Flame

School of Behavioral Sciences, Liberty University

Author Note

Sparky Flame

I have no known conflict of interest to disclose.

Correspondence concerning this article should be addressed to

Sparky Flame

Email: sparky.flame@liberty.edu

Data Screening of Categorical and Continuous Variables

Data screening was accomplished for the variables gender, a dichotomous (categorical) variable with two groups of male and female, Stress, as measured by DASS-Stress, a continuous-interval level of measurement variable, and Depression, as measured by DASS-Depression, a continuous-interval level of measurement variable, from the EDCO 745 course dataset in preparation to conduct a simple moderator analysis using Hayes Model 1. A frequency table was created for gender (see Tables 1). Results of the frequency tables indicated slightly more male (N = 704) than female (N = 596) participants (Table 1).

Table 1

Frequency for Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	704	54.1	54.2	54.2
	Female	596	45.8	45.8	100.0
	Total	1300	99.9	100.0	
Missing		2	.1		
Total		1302	100.0		

Tests of assumptions were conducted for multiple regression.

1. There must be one criterion/outcome variable (Y) that is measured at the continuous level (i.e., the interval or ratio level). The criterion variable, Depression, is continuous (interval level of measurement) by research design.
2. There must be two or more predictor variables (X_i) that are measured either at the continuous or nominal level. The predictor variables are gender and Stress. Gender, the

moderator variable (*W*), which is a type of predictor, is a nominal level of measurement and Stress is continuous (interval level of measurement) by research design.

3. There must be independence of observations (i.e., independence of residuals). Independence of observations (autocorrelation) is tested using the Durbin-Watson statistic. The Durbin-Watson statistic is developed when one conducts the regression as part of the output. Values of the Durbin-Watson statistic close to 2 indicate no autocorrelation (independence of observations). Values of 1 to 3 satisfy this requirement. The Durbin-Watson statistic for the regression is 2.076, demonstrating independence of observations (see Table 2).

Table 2

Model Summary

					Change Statistics					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change	Durbin–Watson
1	.857[a]	.734	.734	6.32596	.734	1704.711	2	1233	.000	2.076

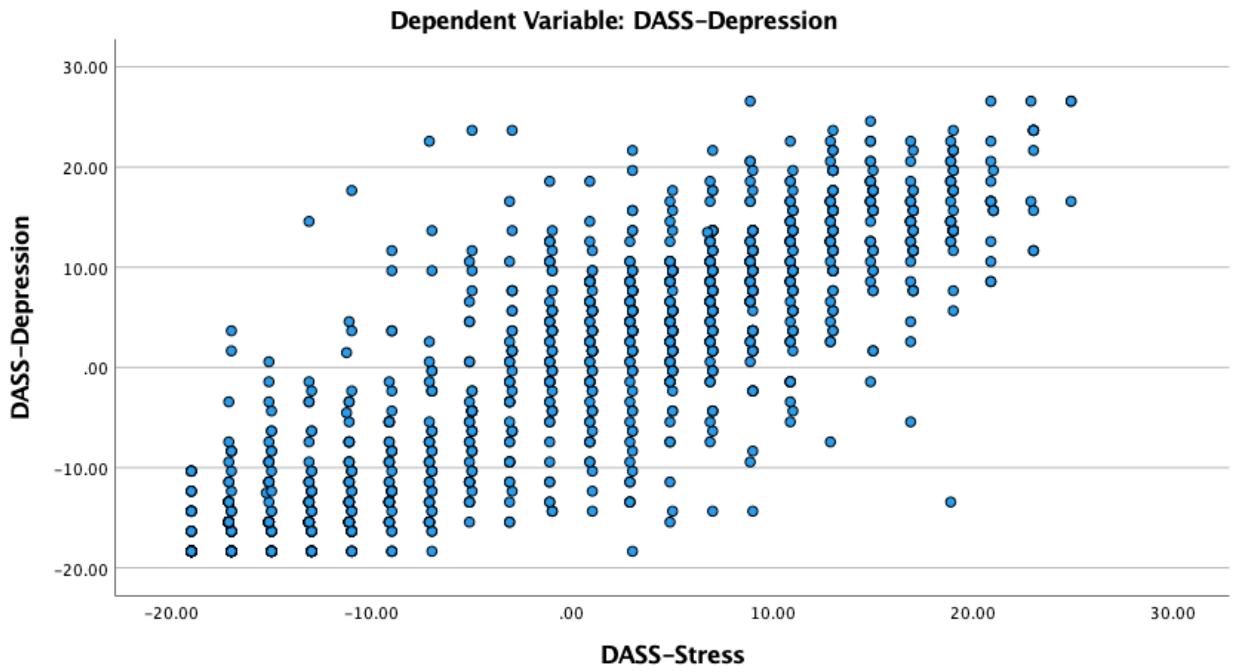
a. Predictors: (Constant), DASS–Stress, Do you identify as:

b. Dependent Variable: DASS–Depression

4. There must be a linear relationship between (a) the criterion/outcome and each of the predictor variables, and (b) the criterion and predictor variables collectively. The linear relationship between variables may be tested by scatterplot for each pairing with the criterion, as well as by an examination of the plot of the residuals. This is a visual test. Based upon a scatterplot between Stress and Depression, there is a linear relationship between the two variables (see Figure 1).

Figure 1

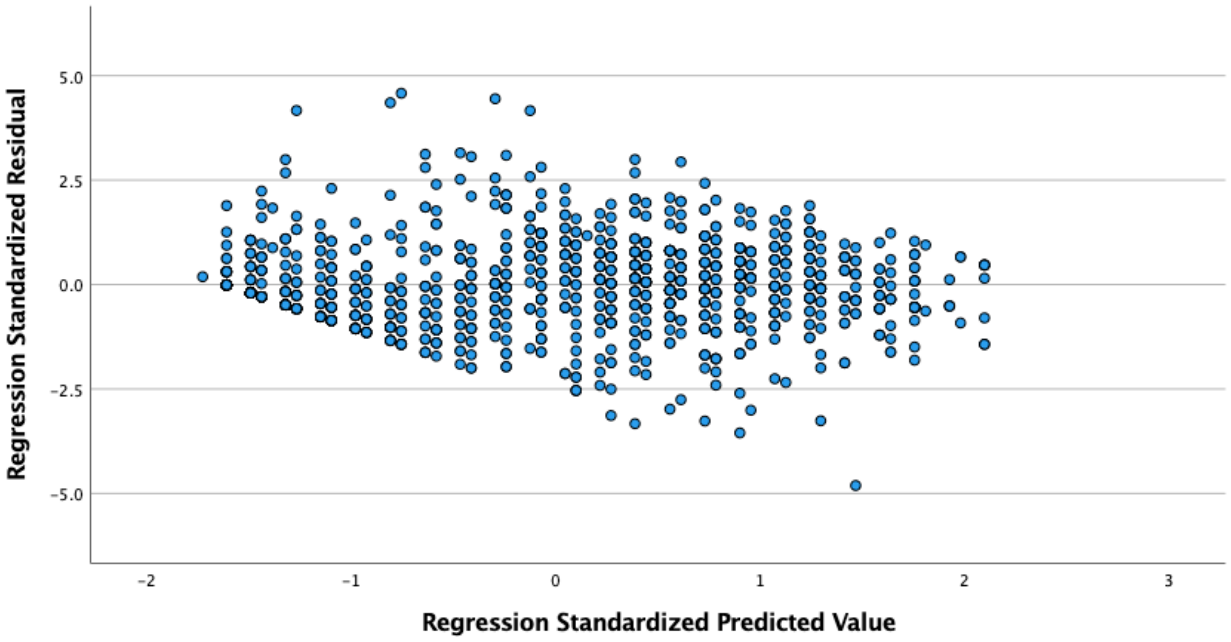
Scatterplot of Stress and Depression



Because gender is a nominal variable, a scatterplot cannot be developed for relationships with this variable. However, one may examine the overall linearity of the model, which is the combination of gender and Stress in relation to Depression, which is completed through partial regression plots. Based upon the result of the partial regression plot, there is a linear relationship (See Figure 2).

Figure 2

Partial Regression Plot of Predictors Gender and Stress Against Depression



5. There must be homoscedasticity of residuals (equal error variances). The assumption is tested using a visual test. One examines the plot of residual (error) variances (Figure 2) to determine if the residuals are relatively equal as indicated by a box shape across the figure. Instances in which the residuals are cone-shaped indicate a lack of homoscedasticity. Figure 2 indicates a slightly diamond shape with narrowing at each end, indicating homoscedasticity of residuals is questionable.
6. There must be no multicollinearity. Tested using the variance inflation factor (VIF), a score of 4 or less indicates no multicollinearity. Multicollinearity is the phenomenon when the predictor variables approximately measure the same construct. If multicollinearity is present, it may be eliminated by removing one of the variables from the analysis. The VIF for the present analysis is 1.007, indicating no multicollinearity, and is presented in Table 3 in the far-right column.

Table 3

Regression Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	2.502	.641		3.905	<.001					
	Do you identify as:	-1.232	.357	-.051	-3.449	<.001	-.120	-.098	-.051	.993	1.007
	DASS-Stress	.899	.016	.851	57.814	.000	.855	.855	.849	.993	1.007

a. Dependent Variable: DASS-Depression

7. There must be no significant outliers, high leverage points, or highly influential points. The assumption is tested using casewise diagnostics, which identify these three phenomena. Instances of these points should be removed from the dataset and the dataset reevaluated. Casewise diagnostics were completed for the regression, revealing 16 records with extreme violations, as shown in Table 4. These records will be removed prior to completing the regression.

Table 4

Casewise Diagnostics

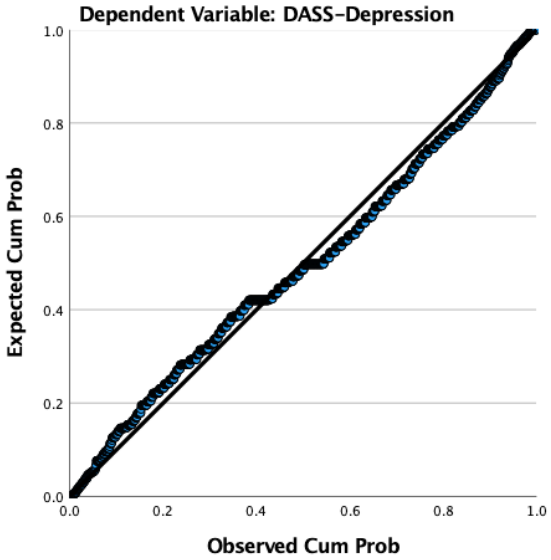
Case Number	Std. Residual	DASS-Depression	Predicted Value	Residual
215	3.120	30.00	10.2615	19.73848
279	-3.006	8.00	27.0132	-19.01322
281	3.152	32.00	12.0599	19.94010
285	-4.807	2.00	32.4084	-30.40837
362	-3.548	4.00	26.4470	-22.44696
597	-3.264	4.00	24.6486	-20.64858
682	4.449	42.00	13.8583	28.14171
998	-3.328	.00	21.0518	-21.05181
999	4.164	42.00	15.6567	26.34333
1023	-3.133	.00	19.8197	-19.81969
1059	-3.258	10.00	30.6100	-20.60998
1062	3.063	32.00	12.6262	19.37384
1261	3.094	34.00	14.4245	19.57545
1280	4.353	36.00	8.4631	27.53686
1294	4.580	38.00	9.0294	28.97060
1303	4.168	30.00	3.6343	26.36575

a. Dependent Variable: DASS-Depression

8. The residuals (errors) must be approximately normally distributed. The test of residual normality is tested using a normal P-P plot. Normality is indicated when the points of the scatterplot fall along or near the 45-degree line. The normal P-P plot for the regression indicates an approximately normal distribution of the residuals. See Figure 3.

Figure 3

Normal P-P Plot of Regression Standardized Residuals



The data met the tests of assumptions, with homogeneity of variances left as a questionable result. A moderator analysis may now be conducted.