Interpreting the Basic Outputs (SPSS) of Multiple Linear Regression

Chuda Prasad Dhakal, PhD

Tribhuvan University, Institute of Agriculture and Animal Sciences, Rampur Campus, Chitwan, Nepal

Abstract: ***Regression analysis is one of the important tools to the researchers, except the complex, cumbersome and the expensive undertaking of it; especially in obtaining the estimates correctly and interpreting them plentifully. We perceive a need for more inclusive and thoughtful interpretation of (in this example) multiple regression results generated through SPSS. The objective of this study is to comprehend and demonstrate the in-depth interpretation of basic multiple regression outputs simulating an example from social science sector. In this paper we have mentioned the procedure (steps) to obtain multiple regression output via (SPSS Vs.20) and hence the detailed interpretation of the produced outputs has been demonstrated. We have illustrated the interpretation of the coefficient from the output, Model Summary table (R^2, Adj. R^2, and SE); Statistical significance of the model from ANOVA table, and the statistical significance of the independent variables from coefficients table. An expansive and attentive interpretation of multiple regression outputs has been explained untiringly. Both statistical and the substantive significance of the derived multiple regression model are explained. Every single care has been taken in the explanation of the results throughout the study to make it a competent template to the researcher for any real-life data they will use. Because every effort has been made to clearly interpret the basic multiple regression outputs from SPSS, any researcher should be eased and benefited in their fields when they use multiple regression for better prediction of their outcome variable.***

Keywords: Multiple regression,Regression outputs, R squared, Adj. R Square, Standard error, Multicollinearity

1. Introduction

Regression analysis technique is built on many statistical concepts including sampling, probability, correlation, distributions, central limit theorem, confidence intervals, z-scores, t-scores, hypothesis testing and more (Interpreting regression output, without all the statistics theory, n.d).Interpretation of the results catches up the issues of 1) analysing the correlation and directionality of the data, 2) estimating the model, i.e., fitting the line, and 3) evaluating the validity and usefulness of the model. Hence interpreting its output generally is bulky. One example that suits this issue (interpreting of the regression results) is, Nathans et al.(2012) reveals that, there is no single right way to interpret regression results, and although reliance on beta weights may feel right because it is normative practice, it provides very limited information. Also, the authors [(Martin, 2018); (Klees, 2016); (Armstrong, 2011); (Dion, 2008); (Guthery & Bingham, 2007); (Seva et al., 2010) and (Miler, n.d)]; have shed light on the importance and the power of regression and its challenge if the models investigated were valid at all.

To obtain credible and valid estimates of regression passing across frequent cumbersome steps that may derail at any time of the analysis, and interpreting them (the regression results) is always a challenge. In their papers the authors mentioned above have emphasized on, the caution any researcher has to take, s/he has to seek to reach their correct results and interpretation, about the completeness and the comprehensive level of interpretation. The dimension the interpretation has to cover for any balanced presentation of the regression results (both statistical and the substantive significance) when writing an application of the regression (especially multiple regression) results.

Keeping this view, this paper is intended to be a quick and easy-to-follow summary of the interpreting of regression analysis outputs. However, the scope of this paper is limited to shed light only on basic insights the regression output gives, based on the multiple regression output that looks like in SPSS software.

In this paper we want to open researchers' eyes wide through which multiple regression output can be viewed successfully. For instance, as Sweet and Karen (2012) initiate the discussion as, 'interpretation of the statistics in multiple regression is, the same as in bivariate regression, other than in multiple regression the effects of multiple independent variables often overlap in their association with the dependent variable.

This paper is structured by (a) defining data (b) considering the specified no of basic multiple regression output (c) describing how each output are interpreted, and (d) summarizing the discussion. In conclusion, we will demonstrate a data-driven example and a results and discussion section that researchers can use as a template for interpreting and reporting multiple regression outputs.

2. Materials and Methods

For twelve families, hypothetical data (Table 1): dependent variable (y) = *hours per week* husband spends in house work, and the two independent variables, x_1 = *no of children* in the families,x_2 = *wife's years of education*and x_3=husband's *years of education*, is considered for the study.

Table 1: Hypothesized data

Family	A	B	C	D	E	F	G	H	I	J	K	L
Y	4	1	3	5	3	2	4	5	4	4	4	5
x_1	3	0	3	2	2	1	3	4	4	3	2	3
x_2	16	10	14	18	14	14	12	14	16	12	16	16
x_3	18	18	16	12	16	18	12	12	14	12	16	16

Any fit of a multiple regression model is valid, if and only if satisfied are the underlying assumptions,1) dependent variable should be measured on a continuous scale (i.e., it is either an interval or ratio variable). 2) there are two or more independent variables, which can be either continuous (i.e., an interval or ratio variable) or categorical (i.e., an ordinal or nominal variable). 3) independence of observations (i.e., independence of residuals), 4) linear relationship between (a) the dependent variable and each of the independent variables, and (b) the dependent variable and the independent variables collectively. 5) homoscedasticity 6) data must not show multicollinearity, 7) there should be no significant outliers, high leverage points or highly influential points and 8) the residuals (errors) are approximately normally distributed.

Considering none of the eight assumptions mentioned earlier, have been violated, regression output to the given data was generated through the following steps conducted in SPSS Vs. 20.

Click‘Analyse’,‘Regression’, ‘Linear’. Select ‘*hours per week*’ in the dependent variable box and ‘*no of children*’ and ‘*years of education*’ in the independent variable box. Select ‘enter’ as the as the method [The default method for the multiple linear regression analysis]. Click ‘Statistics’, select [‘Model fit’ and ‘Estimates’ are default selections] ‘R squared change’, ‘Confidence Intervals’, ‘Part and partial correlations’ and ‘Collinearity diagnostics’ and click ‘continue’. Outputs (Model summary table, Anova and Coefficients) generated through the command mentioned afore are discussed and interpreted systematically in the following result and discussion section. At times while discussing, same output table has been replicated as per the ease to see the results close by.

3. Results and Discussion

Our research question for the multiple linear regression is: Can we explain the outcome variable, *hours per week* that a husband spends at house work with the given independent variables *no of children*, *wife’s year of education* and *husband’s years of education*?

Determining how well the model fits

The first table of interest is the *model summary* (Table 2). This table provides the *R*, R^2, adjusted R^2, and the standard error of the estimate, which can be used to determine how well a regression model fits the data:

Table 2: Model summary

Model summary				
Model	R	R Square	Adjusted R Square	Std. Error of the estimate
1	.925[a]	.856	.803	.547

a. Predictors: (Constant), husband’s years of education, wife’s years of education, no of children

The "R" column represents the value of *R*, the *multiple correlation coefficient*. *R* can be considered to be one measure of the quality of the prediction of the dependent variable; in this case, *hours per week*. A value of .925 in this example, indicates a good level of prediction. The "R Square" column represents the R^2 value (also called the coefficient of determination), which is the proportion of variance in the dependent variable that can be explained by the independent variables.

You can see from our value of .856 that our independent variables explain 85.6 % of the variability of our dependent variable, *hours per week.* And 14.4% (100%-85.6%) of the variation is caused by factors other than the predictors included in this model. At first glance, R-squared seems like an easy to understand statistic that indicates how well a regression model fits a data set. However, it doesn’t tell us the entire story. To get the full picture, one must consider R^2 value in combination with residual plots, other statistics, and in-depth knowledge of the subject area.

According to Frost (2017) caveats about R^2 is: small R-squared values are not always a problem, and high R-squared values are not necessarily good. For instance, for an outcome variable like human behaviour which is very hard to predict, a high value of R-squared is almost impossible. And, this does not mean any predicted model to such case is always useless. A good model can have a low R^2 value. On the other hand, a biased model can have a high R^2 value! A variety of other circumstances can artificially inflate R^2.

To accurately report the data interpretation of "Adjusted R Square" (*adj.* R^2) is another important factor. A value of .803 (coefficients table) in this example indicates true 80.3% of variation in the outcome variable is explained by the predictors which are to keep in the model. High discrepancy between the values of R-squared and Adjusted R Squareindicates a poor fit of the model. Any addition of useless variable to a model causes a decrease in adjusted r-squared. But, for any useful variable added, adjusted r-squared will increase. Adjusted R^2 will always be less than or equal to R^2. Adjusted R^2 therefore, adjusts for the number of terms in a model. As R^2 always increases and never decreases, it can appear to be a better fit with the more terms added to the model and the adjusted R^2 penalizes one from being completely misleading.

Stephanie (2018) cautions about how to differentiate between R^2 and adjusted R^2. R^2 Shows how well data points fit a regression line assuming every single variable explains the variation in the dependent variable which is not true. Whereas, adjusted R^2 tells how well the data points fit a regression line showing the percentage of variation explained only by the independent variables that actually affect the dependent variable. In addition, example of interpreting and applying a multiple regression model (n.d.)reveals that the "adjusted R^2" is intended to "control for" overestimates of the population R^2 resulting from small samples, high collinearity or small subject/variable ratios. Its perceived utility varies greatly across research areas and time.

The standard error (in this example .55)of a model fit is a measure of the precision of the model. It is the standard deviation of the residuals. It shows how wrong one could be if s/he used the regression model to make predictions or to estimate the dependent variable or variable of interest. As R^2 increases the standard error will decrease. On average, our estimates of hours per week with this model will be wrong by .55which is not an ignorable amount given the scale of hours per week. And hence, the standard error is wished to be as small as possible. The standard error is used to get a confidence interval for the predicted values.

Correlated predictors (multicollinearity) may cause large standard error of the estimate of the regression coefficient. However, even with the presence of multicollinearity the regression can still be precise if the "magnified" standard error is still small enough.

Statistical significance of the model

The F-ratio in the ANOVA (Table 3) tests whether the overall regression model is a good fit for the data. The table shows that the independent variables statistically significantly predict the dependent variable, *F* (3, 8) = 15.907, p (.001) < .05 (i.e., the regression model is a good fit of the data).

Table 3: ANOVA

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	14.274	3	4.758	15.907	.001[b]
	Residual	2.393	8	.299		
	Total	16.667	11			

a. Dependent Variable: hours per week

b. Predictors: (Constant), husband's years of education, wife's years of education, no. of children

Statistical significance of the independent variables

Statistical significance of each of the independent variables tests whether the unstandardized (or standardized) coefficients are equal to 0 (zero) in the population(i.e. for each of the coefficients, H_0: β = 0 versus Ha: β ≠ 0 is conducted). If *p* < .05, the coefficients are statistically significantly different to 0 (zero). The usefulness of these tests of significance are to investigate if each explanatory variable needs to be in the model, given that the others are already there.

Table 4: Coefficients

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
	B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
(Constant)	2.021	1.681		1.203	.263					
No. of children	.367	.185	.348	1.984	.082	.759	.574	.266	.584	1.711
Wife's year of education	.271	.080	.491	3.386	.010	.641	.767	.454	.853	1.173
Husband's years of education	-.211	.081	-.425	-2.584	.032	-.653	-.675	-.346	.663	1.509

a. Dependent Variable: hours per week

Given that, the *t*-value and corresponding *p*-value are in the "t" and "Sig." columns (Table 4), respectively, in this example, the tests tell us that *wife's years of education p(.010)<0.05* and *husband's years of education p(.032)<0.05* are significant, *but no of children is not significant P(.082)>0.05*. This means that the explanatory variable *no of children* is no more useful in the model, when the other two variables are already in the model. In other words, with *wife's years of education and husband's years of education* in the model, *no of children* no more adds a substantial contribution to explaining *hours per week*.

Like the standard error of model fit discussed above, the standard error of the coefficients in regression output are also wished to be as small as possible. It reflect show wrong you could be, while estimating its value. For instance, in this example relative to the coefficient .271 of *wife's years of education* its standard error .080 is small.

Estimated model coefficients

The general form of the equation to predict *hours per week* from no of children, *wife's years of education*, and *husband's years of education*, is:

Predicted *hours per week* = 2.021 + 0.367 (*no of children*) + 0.271(*wife's year of education*) – 0.211 (*husband's years of education*)

This is obtained from the (Table 5) below:

Table 5: Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
	B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
(Constant)	2.021	1.681		1.203	.263					
No. of children	.367	.185	.348	1.984	.082	.759	.574	.266	.584	1.711
Wife's year of education	.271	.080	.491	3.386	.010	.641	.767	.454	.853	1.173
Husband's years of education	-.211	.081	-.425	-2.584	.032	-.653	-.675	-.346	.663	1.509

a. Dependent Variable: hours per week

Constant 2.021, is the predicted value for the dependent variable (in this example)*hours per week* if all independent variables, *no of children* = 0, *wife's years of education* = 0 and *husband's years of education*=0. That is, we would expect an average *hour per week* of 2.021 husband spends for house work when all predictor variables take the value 0. For this reason, this is only a meaningful interpretation if it is reasonable that the predictors can take the value 0 in practice. Besides, Karen (2018) reveals that the data set should include values for all predictors that were near 0. Therefore, if both of these conditions are not true, the constant term (the y-intercept) in the regression line really has no meaningful interpretation.

However, as it (the y-intercept) places the regression line in the right place, this is always kept in there while presenting the regression model. In this example, it is easy to see that in the data set, *no of children* sometimes is 0, but *both wife's years of education* and *husband's years of education*, are not close to 0, then our intercept has no real interpretation.

Unstandardized coefficients indicate how much the dependent variable varies with an independent variable when all other independent variables are held constant. The regression coefficient provides the expected change in the dependent variable (here: *hours per week*)for a one-unit increase in the independent variable. Referring to the coefficients (Table 5) above the unstandardized coefficient for *no of children* is 0.367. This means for every unit increase (one child increase) in *no of children*, there is 0.367 hours increase in *hours per week*. But each one-year increase in *husband's years of education* causes reduction (the negative sign of the coefficient) in *hours per week* by 0.211hours.

Accordingly, standardized coefficients are called beta weights, given in the "beta" column. The beta weight measure how much the outcome variable increases (in standard deviations) when the predictor variable is increased by one standard deviation assuming other variables in the model are held constant. These are useful measures to rank the predictor variables based on their contribution (irrespective of sign) in explaining the outcome variable.

Hence in this case, *wife's years of education* is the highest contributing (.491) predictor to explain *hours per week*, and the next is *husband's years of education* (-.425). However, only when the model is specified perfectly and there is no multicollinearity among the predictors, Stephanie (2018) explains.

Zero order partial and part correlation

Table 6: Coefficients

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
	B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
(Constant)	2.021	1.681		1.203	.263					
No. of children	.367	.185	.348	1.984	.082	.759	.574	.266	.584	1.711
Wife's year of education	.271	.080	.491	3.386	.010	.641	.767	.454	.853	1.173
Husband's years of education	-.211	.081	-.425	-2.584	.032	-.653	-.675	-.346	.663	1.509

a. Dependent Variable: hours per week

Zero order correlation are the bivariate correlation between the predictors and the dependent variable. Hence .759 in this example is the direct effect of *no of children* on *hours per week*, this ignores the effect of other two predictor variables that may/may-not be influencing the dependent variable.

But when the effect of the other two independent variables are accounted (but kept constant) to *no of children* and *hours per week*, the correlation changes to be less strong.574, which is partial correlation. And, For the same case, the part correlation .266 is the correlation between *no of children* and *hours per week* where the effect of the other two independent variables are completely excluded out.

From the causal perspective, this means (in this example) if we change *no of children*, we change the other variables, too. Now, when we model the TOTAL effects from *no of children* on *hours per week*, we have to account for the direct effect (which appears to be strong) and the indirect effect of *no of children* influencing the other variables which in turn influence *hours per week*. When we combine the strong direct impact with the indirect effects, we end up with an overall "weaker" impact.

Because part correlations are the correlations that presume the effect of the other predictors have been excluded out, these are helpful to identify if the multiple regression used was beneficial. I.e. to estimate gain in predictive ability (how much gain had been there in the predictive ability due to the combination of the predictors in the model) of the model. In this example, *part coefficients of determination* (SPSS does not produce this) are $(.266)^2$, $(.454)^2$ and $(-.346)^2$ respectively for the *predictors no of children, wife's years of education*, and *husband's years of education*. These unique contributions of the predictors when added up, approximates (7.1+20.6+12) = 39.7% of the variation in the outcome variable. And this percentage of variance in the response variable is different from the R-squared value (85.6 %) in the model. Meaning that (85.6-39.7=45.9) 46% overlapping predictive work was done by the predictors. Which is not that bad. This proves the combination of the variables had been quite good.

The information in the (Table 6) above also allows us to check for multicollinearity. A common rule of thumb: for any predictor VIF > 10 should be examined for possible multicollinearity problem (Dhakal, 2016). In our multiple linear regression model. VIF should be < 10 (or Tolerance > 0.1) for all variables, which they are.

4. Summary

Putting the above all together we could write up the results as follows:

A multiple regression was run to predict *hours per week* a husband spends at house work, from *no of children, wife's years of education* and *husband's years of education*. The model statistically significantly predicted *hours per week* F (3, 8) = 15.907, p(.001) < .05, $R^2 = 0.856$. Out of three only two variables *wife's years of education p (.010)< .05* and *husband's years of education p (.032)< .05* added statistically significantly to the prediction. The highest contributing predictor is *wife's years of education* (.491) and, and the next is *husband's years of education* (-.425) to explain *hours per week*. Multicollinearity problem does not exist in the model as VIF for all variables is < 10 (or Tolerance > 0.1). And, 46 % overlapping predictive work was done by the predictors. This proves the combination of the variables had been quite good.

No of children is not significant P (.082)>0.05, has therefore no substantial contribution in explaining *hours per week*, when the other two significant predictors are already in the model. Dare to answer the questions new experiment will pose, when the study is replicated with only the two significant predictors?

5. Conclusion

The demonstration of interpreting of multiple regression output obtained through SPSS is descriptive and intuitive. Henceforth, it can be used by the researchers, students, and the related faculties as a template while each one of the related would be using real data for problem solving researches and the studies.

References

[1] Armstrong, J. S. (2011). Illusions in Regression Analysis. *ScholarlyCommons*. Retrieved from http://repository.upenn.edu/marketing_papers/173Accessed on 13 June 2018.

[2] Dhakal, C.P. (2016). *Optimizing multiple regression model for rice production forecasting in Nepal*.(Doctoral thesis, Central Department of Statistics, Tribhuvan University Nepal).

[3] Dion, P.A. (2008). Interpreting structural equation modeling results: A reply to Martin and Cullen. *Journal of Business Ethics*, *83*(3), 365–368. Springer, Stable URL http://www.jstor.org/stable/25482382 Retrieved from https://www.jstor.org/stable/25482382?seq=1#page_scan_tab_contents Accessed on 15 June 2018.

[4] Example of interpreting and applying a multiple regression model. (n.d). Retrieved from http://psych.unl.edu/psycrs/statpage/full_eg.pdfAccessed on 11 June 2018.

[5] Frost, J. (2017). How to interpret R-squared in regression analysis. Retrieved fromhttp://statisticsbyjim.com/regression/interpret-r-squared-regression/ Accessed on 02 June 2018.

[6] Grace, B. J., &Bollen, A. K. (2005). Interpreting the results from multiple regression and structural equation models. Bulletin of the Ecological Society of America, *86*(4), 283 – 295. ISSN:0012-9623, EISSN:2327-6096, doi: 10.1890/0012-9623(2005)86[283:ITRFMR]2.0.CO;2

[7] Guthery, F.S., & Bingham, R. (2007). A primer on interpreting regressionmodels. *Journal of Wildlife Management*, *71*(3) 684 – 692. ISSN:0022-541X, EISSN:1937-2817. The Wildlife Society doi:10.2193/2006-285

[8] Interpreting regression output (Without all the statistics theory). (n.d). *GraduateTutor.com*. Retrieved from http://www.graduatetutor.com/statistics-tutor/interpreting-regression-output/ Accessed on 29 May 2018.

[9] Klees, J. S. (2016). Inferences from regression analysis: Are they valid? *University of Maryland, USA*. Retrieved from http://www.paecon.net/PAEReview/issue74/Klees74.pdfAccessed on 13 June 2018

[10] Martin, K. G. (2018). Interpreting Regression Coefficients. *The Analysis Factor*. Retrieved from https://www.theanalysisfactor.com/interpreting-regression-coefficients/Accessed on 13 June 2018.

[11] McCabe, G.P. (1980). The interpretation of regression analysis results in sex and race discrimination problems. *The American Statistician 34*(4) 212-215. ISSN:0003-1305 EISSN:1537-2731, American Statistical Association,doi: 10.1080/00031305.1980.10483030.

[12] Miler, J.E. (n.d). Interpreting the substantive significance of multivariable regression coefficients., *Rutgers University*. Retrieved from http://www.statlit.org/pdf/2008MillerASA.pdfAccessed on 13 June 2018.

[13] Nathans, L., Oswald, F. L., &Nimon, K. (2012). Interpreting multiple linear regression: A guidebook of variable importance. *Practical Assessment, Research & Evaluation*, *17*(9). Retrieved fromhttp://pareonline.net/getvn.asp?v=17&n=9Accessed on 29 May 2018

[14] Seva, U. L., Ferrando, P.J., and Chico, E. (2010). Two SPS program for interpreting multiple regression results. *Behaviour Research Methods42*(1)29–35. Springer, Retrieved from https://link.springer.com/article/10.3758/BRM.42.1.29 15 June 2018

[15] Stephanie. (2018).Adjusted R-Squared: What is it used for?*How to Stat*. Retrieved from http://www.statisticshowto.com/adjusted-r2/ Accessed on 02 June 2018

[16] Stephanie. (2018). Beta Weight: Definition, uses. Retrieved from http://www.statisticshowto.com/beta-weight/ Accessed on 02 June 2018

[17] Sweet, S. & Martin, K.G. (2012). Data analysis with SPSS: A first course in applied statistics (4th ed.). *Pearson Education, Inc., Publishing as Allyn& Bacon, 75 Arlington, Boston USA*.