A Learning Aid for Ushering Logistic Regression Early in Introductory Analytics Courses

Niki Kunene
kunenek@easternct.edu
Dept. Accounting and Business Information Systems
Eastern Connecticut State University
Willimantic, 06226, USA

Katarzyna Toskin
toskink1@southernct.edu
Business Information Systems
Southern Connecticut State University
New Haven, 06515, USA

Abstract

Logistic regression (LR) is a foundational supervised machine learning algorithm and yet, unlike linear regression, appears rarely taught early on where analogy to linear regression is an advantage. In this paper, we investigate prevalence of LR in the learning outcomes and topic coverage of undergraduate statistics courses that are part of analytics curricula. A sample of 50 syllabi from undergraduate business statistics courses shows only two percent of the courses included LR. Conceivable reasons for this dearth of LR content is likely related to topic complexity, time constraints, and varying degrees of tool ease of use and support. These constraints represent an opportunity to fill-in-the-gaps using easy to understand and non-intimidating techniques. We propose a logistic regression flow diagramming visual aid that can be used by instructors and novice students learning regression leveraging the introduction proximity to linear regression to readily boost their understanding of a foundational technique.

Keywords: Logistic Regression, Flow Diagram, Predictive Analytics, Data Analytics, Flow Chart, Pedagogical Aid

1. INTRODUCTION

Logistic regression is a classical model in statistics used for estimating conditional probabilities (Berkson, 1944). Logistic regression is also foundational to predictive analytics in multiple ways: 1. Logistic regression is a supervised classification (machine learning) algorithm used in many problem classes that seek to predict the probability of a target variable. 2. Among competing machine learning classification algorithms, e.g., support vector machine (SVM), and random forest, logistic regression is relatively simpler, and it is aided by having a (familiar) analogy to linear regression. 3. Because it is relatively simpler, good enough and easy to implement, it typically serves as a benchmark model when performing analysis for comparison to other algorithms. 4. Lastly, logistic regression is a gateway to learning neural networks (in that, in neural network representation, each neuron can be conceived as a small regression classifier). For these reasons, it is not surprising that logistic regression is widely used and taught for predictive analytics.

However, logistic regression appears to be rarely taught in the foundational statistics courses that are part of analytics curricula. The underlying reasons are likely manifold. But, it is not

unreasonable to surmise that the topic is deemed too complex for the timing of foundational statistical courses on a college campus for undergraduate students, or that time does not allow, or that the technical tools we use in these courses may be a hurdle that potentially magnifies time constraints.

We surmise that not teaching logistic regression is a lost opportunity for: broadening the introduction to analytics to a larger audience and offering students an opportunity to experience a realistic introduction to data analytics, using one of the most widely used algorithms, early in students' campus careers; as well as giving a wider audience (i.e., non-analytics students) access to a technique that is applicable to a wide range of real-world problems and is used to read or access research publications in the social sciences (Linneman 2021).

The purpose of this paper is to: first, we present evidence showing the relative dearth of logistic regression instruction/content in undergraduate business statistics courses (notwithstanding its importance to reliant business analytics and data science curricula). Second, we identify conceivable reasons and limitations for why logistic regression, in contrast to linear regression, is rarely included in introductory courses. Lastly, as a mechanism to work around these reasons and limitations, we propose a teaching visual aid for logistic regression. The proposed teaching aid may be used to supplement teaching activity and support students either, by reinforcing instruction or can be used subsequently when reviewing the topic.

2. BACKGROUND

Teaching Logistic Regression

Logistic regression is, broadly, like multiple regression but, where the outcome variable is a categorical variable and predictor variables may be continuous or categorical. In its simplest form, it allows us to predict which two categories a person or thing is likely to belong to, given other (additional) data. Albeit, the principles underlying logistic regression have a few parallels to ordinary least squares regression (OLS), logistic regression is rarely taught in foundational classes even though its analogy and instructional (time) proximity to OLS would be pedagogically advantageous. We see evidence of this absence in prior literature outlining content maps for analytics.

Sircar (2009) maps the analytics curriculum which includes a course on "Applied Regression Analysis in Business" that does not cover logistic regression. Similarly, mapping content topics to student prior experience for each topic in a "Big Data Analytics" course development and roll out, Hill & Kline (2014) show linear regression, both simple and multiple, are "partially covered in previous statistics courses", while logistic regression is a "new topic for *most* students" (as opposed to *new for all*). The work of Kline & Hill and Sircar suggests that both linear and logistic regression, to lesser and greater extents respectively, are candidates for review work if not full-on reinstruction or, even a "filling-the-gap" approach suggested by (Bauman & Tuzhilin (2018) where instructors can pre-identify a library of learning materials, related to the structure of the course, to be recommended to students as remedial learning materials to close knowledge gaps.

To the best our knowledge the literature exploring why logistic regression is scarcely taught in business statistics courses is missing. However, the same issue has been explored in the social sciences, e.g., sociology (Linneman, 2021; Lottes, DeMaris, & Adler, 1996; Walsh, 1987) and psychology. Several explanations have been advanced: while statistics courses have grown and may also be taught, strictly-speaking, by non-statisticians (Utts, 2015); for instance, the analogy used in sociology is, statistics is taught within sociology departments (Linneman, 2021). In data analytics, analytics courses that require logistic regression may be taught within information systems, business or economics departments rather than by statisticians from math and statistics departments.

At the same time, logistic regression (LR) has also grown in popularity in other social sciences like sociology and psychology (which matters for data analytics minor programs looking to attract students from other majors), Linneman (2021) argues that because of LR's widespread use in scientific literature (we would add, and in data analytics), student understanding of logistic regression would contribute to students' quantitative literacy. Linneman emphasizes a contention also made by Walsh (1987): "a grasp of logistic regression will not only assist students in their own research efforts, but it will also enable them to intelligently read and evaluate current research in their field" (p. 178). Student understanding of logistic regression, in data analytics, would expand the universe of problems and their contexts with which students are able to engage, a pedagogically approach is superior to amassing repetitions in limited contexts (Schmidt & Bjork, 1992). Further, introducing logistic

ISSN 2473-4901
v7 n5598

regression following linear regression is consistent with pedagogy spacing strategies that are designed to optimize short and long-term retention of knowledge (Lyle, Bego, Hopkins, Hieb, & Ralston, 2020); in this particular case, a spacing strategy would leverage student understanding of both probability and linear regression, topics that are widely taught in introductory statistics courses. In this paper, we assert that the absence of LR in content coverage early in students' exposure to related analytics content is a missed opportunity.

3. INSTRUCTIONAL GAP: DATA SUPPORT

To verify what we intuited experientially, we investigated the extent of the coverage of logistic regression in business statistics undergraduate, courses. We searched for publicly available business statistics syllabi, online, across the United States, randomly selected 50 sample syllabi and analyzed their learning outcomes and course outlines. We found, that only 2% of the sample business statistics courses covered logistic regression, compared to the 88% that cover simple linear regression and 76% cover multiple linear regression. (See Figure 1 below). We see that the business statistics courses we rely on as prerequisites of our data analytics curricula, on the main, do not introduce logistic regression.

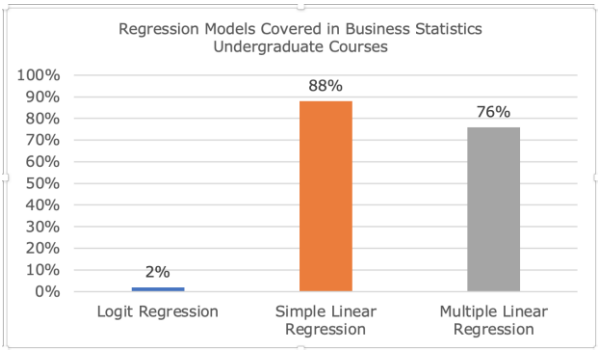


Figure 1. Sample of Business Statistics Courses: Comparative Frequency of Regression Model Coverage

We note that the schools in our sample happen to be located in 25 distinct states. They were primarily 4-year, AACSB accredited public institutions across the nation. Table 1, below, breaks down the characteristics of these institutions.

Institutional Characteristic		**%**
Program duration		
	4-year	88
	2-year	12
School Type		
	Public	68
	Private	32
AACSB Accredited		76
Modality		
	On-ground	94
	Online	6

Table 1. Institutional characteristics from the sample

Table 1, seems to, at least in part, support our prior experiential observations that there is an instructional gap for logistic regression in the statistics courses that are integral to many analytics curricula though, the reasons for this gap we can only surmise at this stage.

4. THE INSTRUCTIONAL GAP

The reasons for scarcity of logistic regression coverage in introductory business statistic courses could be a result of several factors from time constraints, relative complexity (perceived and/or real), student preparedness as well tool efficacy and availability.

Time and complexity. Hill and Kline (2014) caution that one of the main challenges facing analytics instructors, is the tension between the need to cover or review underlying knowledge or content and available course-time is, "*teaching tasks may take longer than expected. The instructor should be prepared to allocate additional class time or provide significant time for guidance outside of class*" (Hill & Kline, 2014, p. 6).

Additionally, though somewhat analogous, forming and interpreting the logit function is relatively more complex for logistic regression compared to simple and multiple linear regression. This complexity may be intimidating. Studies show that affective reasons are one of the reasons contributing to students' difficulty with interpreting and communicating the results of their analyses (Ashaari, Judi, Mohamed, & Wook, 2011; Reid & Petocz, 2002; Toskin & Kunene, 2020). There have been suggestions in the literature that (though unlikely in our belief) instructors who are not themselves statisticians may not feel necessarily comfortable with their own abilities to efficiently/effectively unpack the

complexity for students (Linneman, 2021), especially when coupled with time constraints, as in the case when introductory statistics courses are brought into the domain or department (e.g., in business, sociology, and psychology) .

Tool efficacy and availability. Compared to simple and linear regression, tool support for logistic regression is generally not as user friendly or necessarily easy or cheap to access. For example, both MS Excel and R, which are easily available to most students, run good linear regression models generating comprehensive output. However, running a logistic model in MS Excel involves either multiple manual calculation steps together with the use of an Excel Analysis Toolpak that includes Solver; and/or teaching students how to use the open-source add-in, RegressIt for Excel or, the XLMiner Analysis Toolpak. Minitab which is available to faculty and students at a large discount also supports logistic regression from a GUI interface (though the latest version no longer runs natively on MacOS). R on the other hand requires several lines of commands to create comprehensive output. Other software, like Stata and SAS, also require some use of the command line but, unlike R they are relatively expensive licensed-programs; this often means they are not easily available to students. IBM's SPSS, which may not be easily available to students, is easy to use and generates comprehensive output for logistic analyses.

Below, we propose an aid for logistic regression that helps students readily draw on their prior knowledge, i.e., as analogy with linear regression, and gives both instructors and students a visual aid that supports a continuity of instruction. The proposed artifact uses flow diagramming, an established pedagogical aid that makes complexity readily accessible to a novice. (Toskin & Kunene, 2020). The proposed artifact is not, however, intended as a substitute for initial instructional time, but rather as tool to support or facilitate instruction, especially where time constraints are a factor, and/or for student review.

5. PROPOSED ARTIFACT

Following, Toskin & Kunene (2020) and others, we propose a visual aid for logistic regression because logistic regression is foundational to predictive analysis and, in the social sciences, there are many problems that are and can be expressed as binary predictive problems. And yet, in our experience, with evidence from the work of Hill & Kline (2014), logistic regression is rarely taught in introductory statistics courses even though it has analogy and some similarities with linear or ordinary least squares (OLS) regression.

The proposed aid can be used to supplement instruction or, can be assigned by analytics faculty as remedial supplemental material to close knowledge gaps and help students catch-up in an accessible and time saving manner. This logistics regression aid assumes prior knowledge in linear regression (see Toskin & Kunene, 2020) as well as some knowledge of probability. The proposed aid follows Toskin & Kunene (2020) and uses flow diagramming which is an easily understandable and widely used pedagogical aid for graphically depicting, for the novice student, the prior knowledge necessary to successfully interpret and communicate a regression model. In this context, this aid and that in Toskin & Kunene (2020) could well serve as part of a library of supplemental aids that instructors of introductory analytics courses can use as part of readily enabling needed fill-in-the-gap approaches (Bauman & Tuzhilin, 2018) in the teaching of analytics.

In the following section we describe the proposed flow diagram. The diagram is included in Appendix A. In general, the artifact can be used independently of tool. For reasons related ease of use and recognizing instructor time constraints, we designed the aid around SPSS. We elaborate of the reasons for this choice in section 7, tool selection.

Background. Logistic regression is the regression model we fit when the target or response variable is categorical, namely dichotomous (in its simplest form), multiple (more than 2 levels), and ordinal. Logistic regression differs from multiple regression because it is intended to predict the probability of an event occurring or group membership using a maximum likelihood estimation method. Additionally, the dependent variable can only take on two values 0 or 1, thus the probability must fall within this range. As a result, logistic regression uses a logistic curve rather than a linear relationship of regression to model the relationship between dependent and independent variables (Hair et al., 2009).

Practically, in analytics a logistic regression is used as an algorithm for solving classification problems. Thus functionally, they are an opportunity to leverage what students already know to introduce them to an important class of problems in data analytics. In effect, the response or target variable serves as a classifier. It is easier

for students to start with binary response problems.

There are additional key differences:
1. The coefficients are converted to log odds.
2. Model fit cannot technically be assessed using R-squared. Thus, pseudo R-square values can be used, with some caution, to assess model fit (there are several competing pseudo R-square calculations),
3. Logistic regression also introduces a classification table to evaluate the predictive accuracy of the (classification) model.

These differences may seem a little harder for students to grasp and interpret initially. The intent is to leverage what students already know from linear regression and probability to guide them through key aspects of interpreting and communicating binary logistic analysis. The multiple linear regression teaching aid (Toskin & Kunene, 2020) focused on five key elements: significance F (p-value for the F statistic), the *intercept* or constant; *coefficients* of the hypothesized independent variables and their respective p-values. For logistic regression, we draw students' attention to concepts with near analogy. It is important for instructors to add caution where the analogy is not precise or complete, as in the interpretation of any pseudo R-squared values and the equation where the logit of the mean of y is a linear function of the predictors.

This teaching aid also employs a simple, easy to understand visual tool using flow diagramming. Flow diagrams have an well-established track record for similar tasks (Toskin & Kunene, 2020).

The logistic regression flow diagram

The proposed flow diagram (see Appendix A) focuses on five key steps. (As discussed below, SPSS is the easiest and most comprehensive tool for the task):
1. Interpret significance of Chi-square statistic (p-value),
2. Interpret the *intercept* or constant.
3. Sequentially locate and interpret *coefficients* of the hypothesized independent variables and their respective p-values.
4. Evaluate common pseudo *R-square measures* for model fit (e.g., Cox & Snell R-squared, Nagelkerke) (not directly analogous with R-squared in OLS, interpret with some caution)
5. *Understand the "hit ratio" in the classification table to assess predictive accuracy of the model*. This step could also be undertaken earlier in the process, as a first or second step. It broadly answers the question, how accurately does the model classify (unseen) data.

When students use SPSS for logistic regression for the first time, instructors should draw attention to the fact that SPSS output generates a "null" or baseline model (with only a constant and no independent variables) followed by an estimated model with the chosen predictors (see Appendix B). The null model typically appears under the section named "Block 0" and an estimated model under "Block 1" (assuming no stepwise method is used). Additionally, Block 1 includes chi-square statistic and its p-value, two pseudo R-square measures, i.e., Cox & Snell and Nagelkerke, beta coefficients along with their statistical significance based on the Wald test, and exponentiated beta value (i.e., Exp(B)). The output also includes a classification table that specifies the "hit ratio" and the overall percentage of cases correctly classified to the appropriate dependent group.

In the flow diagram, first we bring students' attention to the Chi-square value that measures the difference in change (the reduction) of log likelihood value between the base/null model which contains only an intercept, and the proposed model that includes specified independent variables. If the p-value of the Chi-square test is statistically significant students are directed to the next step, otherwise they are encouraged to re-evaluate the chosen independent variables in the model.

Step 2: we help students understand the value and meaning of the intercept or constant and its position in the logistic regression equation.

Step 3: students are directed to locate the regression coefficients for each independent variable, one at a time, and assess each p-value for statistical significance. If the regression coefficient is not statistically significant, the dependent variable and may be removed from the model. Otherwise, if it is statistically significant, students are routed to the next step which focuses on the interpretation of each coefficient value (for continuous and categorical variables)

A logit equation is also provided at that step to help students understand how each coefficient contributes to the overall prediction of the dependent variable, and subsequently to use the model for estimation or prediction. Lastly, we introduced the antilog value, Exp(B), to help students interpret the magnitude of the coefficients.

In cases where independent variables are not significant, we assume that with guidance from faculty or prior knowledge, the regression model will be rerun either in a stepwise fashion or by selecting new variables, and the process of interpretation will begin from the beginning.

Step 4: students are directed to examine pseudo R-square values i.e., Cox & Snell and Nagelkerke R-square used in SPSS to broadly assess model fit and interpret their meaning emphasizing that, in general, a higher percentage of variance explained indicates a better model fit. We note here that instructors may want to however point out that these pseudo R-square values are to be used with caution. Examining the model's classification accuracy is functionally more useful.

In the last step, we highlight that this is a classification problem by asking the student to determine the predictive accuracy of the model by examining the "hit ratio", i.e., the percentage correctly classified using the classification table. The higher the percentage of correctly classified cases, the stronger the predictive accuracy of the model.

The proposed flow diagram can be reused by students each time a new model is generated irrespective of the number or type of independent variables used. And can be used by students in various subsequent courses where logistic regression as tool is used.

Technically, although logistic regression might seem more intricate than linear regression, it has the advantage of not needing to meet the strict assumptions of normality, independence, and constant variance of error terms; for students this means no background effort to test for these requisite assumptions is needed. "Logistic regression … is much more robust when these assumptions are not met" (Hair, Black, Anderson, & Tatham, 2009). For novices, this is an advantage in favor of learning the algorithm.

6. DISCUSSION

In this paper, we use flow diagraming, a proven pedagogical aid for unlocking complexity for a novice, to propose a logistic regression remediation artifact. The purpose of the artifact is to aid in strengthening students' capacity to interpret and communicate analysis for binary logistic regression models. Our approach is intentionally designed to be limited to a specific competency (interpreting logistic regression analyses) and be accessible and not intimidating to the student.

While the aid is not intended to be self-instructing, in other words that students will receive prior instruction from faculty, it is intended to aid both students and instructors as a resource for students who need to (quickly) review how to interpret the results of their logistic regression output without supplemental instruction.

This artifact is designed in a manner informed by and consistent with the design proposed and tested in Toskin & Kunene (2020). We believe the logistic regression aid proposed here similarly makes it easier for both instructors and students to quickly "brush up" on necessary knowledge for understanding logistic regression output.

A secondary objective of our study is to contribute to a potential **library** of accessible and supplemental materials for introductory data analytics courses that instructors can assign or recommend to their students as a mechanism to close knowledge gaps in practicable ways that recognize **time** constraints and the need to support a range of students in our analytics classes. We believe each aid is most useful and accessible where it is focused on a specific purpose and designed to minimize **complexity** for its audience. Flow diagramming is proven in this sense (complexity reduction and therefore time-savings new learners).

Third, the proposed visual aid is consistent with proven teaching strategies in this area, by leveraging student familiarity with prior content, i.e.., linear regression, and drawing analogies where possible (Lottes et al., 1996).

Fourth, we wanted to design a logistic regression artifact that lends itself to use in any task requiring students to run logistic regression analyses, interpret and communicate the analysis. This would enable a student, even in their earlier years on campuses, to use the tool across multiple courses even in settings seeking to advance quantitative numeracy across the curricula. This is a strength inherited from the simplicity and accessibility of flow diagramming.

Tool selection. instructors use various tools in their introductory statistics courses, from Microsoft Excel, to Minitab, R, SPSS, and Stata potentially. It is possible that some even introduce Python. When designing an artifact that supports general conceptual knowledge, the objective is to maximize understanding of the concept while (in this and similar cases) leveraging supporting tool ease of use. To this end, we investigated which would be the easiest

tool to use for logistic regression tool for novices; a tool that would also substantially capture key concepts.

We looked at the following tools, first on ease of use, then on substance. In other words, if a tool was not easy to use, we ruled it out by default then examined the remaining tools for capturing substance.

Tool	Ease of Use	Concepts
Excel, Analytic Solver	Low	
SAS	Low	
Python	Low	
R	Low-Med	Med
Excel, XLMiner	High	Low-Med
Minitab	High	Med
SPSS	High	High

Table 2. Potential supporting technical tools

Any tool that required students to write any form of code to run a logistic regression, we ruled out as too high an ease-of-use bar to cross for an introductory course. Therefore, we discounted Python, SAS and then R. R requires students to know additional commands for displaying key parameters of a logistic regression.

Excel with Analytic Solver: while Excel is easy to use, and arguably Analytic Solver not terribly difficult either, the steps required to perform a logistic regression in Excel are multifold and the extent of the output is restricted to estimating the coefficients of the equation. Additional work is necessary to generate goodness of fit information, and a classification table would not be included. Using Excel with XLMiner was easy but severely lacking in output. It too is best as generating coefficients and their p-values. It also generates a model chi-square, however without associated a p-values.

In the end, the remaining choices were Minitab and SPSS. The two products are both easy to use. However, we found a SPSS produced richer output that is also easier to make sense of. We therefore would recommend the use of SPSS for introductory courses. In cases where students do not have access to SPSS, faculty could generate output for an assigned task and have students focus on the interpretive components of the task, for tool use itself is relatively trivial.

7. CONCLUSION

Data analytics is a growing area for employment and career development. Data analysts and data scientists are in high demand with average salaries that remain in good health for both junior and senior analysts. And thus, the growth in analytics programs offers opportunities for students to enter a fruitful and financially rewarding field upon graduation. And logistic regression is a gateway to many applications of classification problems, yet it is largely untaught in introductory statistics courses.

Teaching logistic regression soon after linear regression would leverage students' immediate understanding of the latter. In other words, the introduction to logistic regression would not be a semester or more after the fact. It would also give an opportunity to a larger body of students an opportunity to sample an important analytics tool. For analytics programs, this matters if we want to expose as many students as possible to the types of problems students will be encountering in the field.

The proposed flow diagramming tool can help students recap or supplement instruction in ways that feel accessible and non-intimidating, and by being strategically focused on key knowledge concepts and any subsequent gaps. Flow diagrams have a proven track record for easily and quickly fostering conceptual understanding; they can serve as a good enough alternative to lecturing or re-lecturing. The related gains in time saved can be directed towards building deeper analytics domain knowledge and stronger skills.

The contribution of this paper is, first, we show from data that there is evidence that logistic regression is indeed excluded from instruction in the undergraduate business statistics courses that analytics curricula rely on (despite its relevance to both business analytics and data science curricula). We then offer conceivable reasons and limitations for why logistic regression, in contrast to linear regression, is rarely included in these introductory courses. Lastly, we put forward a mechanism to subvert these reasons/limitations by proposing teaching visual aid for logistic regression.

We support the design of creative mechanisms to enable students to readily access content and unlock complexity that helps students early in their academic careers. The development of

similar aids for other topics on the analytics content map can help serve as part of a library of supplemental aids to be used as part of fill-in-the-gap approaches (Bauman & Tuzhilin, 2018) in the teaching of analytics.

8. LIMITATIONS AND FUTURE RESEARCH

Limitations of our proposed artifact are the following: In general, it is arguably designed with the SPSS tool output in "mind, meaning faculty may have to clarify that intercept" and "constant" are interchangeable. However, conceptually, it is usable with output generated from other tools. We assume, our audience are novices with limited technical wherewithal and therefore may need help generating output in cases where the tool assumes broader technical competencies. Second, though designed it in a manner consisted with similar prior artifacts (Toskin & Kunene, 2020). The proposed artifact has not yet been tested empirically. Future research can empirically test and validate the artifact in use by students.

9. REFERENCES

Ashaari, N. S., Judi, H. M., Mohamed, H., & Wook, M. T. (2011). Student's attitude towards statistics course. *Procedia-Social and Behavioral Sciences, 18*, 287-294.

Bauman, K., & Tuzhilin, A. (2018). Recommending remedial learning materials to students by filling their knowledge gaps. *MIS Quarterly, 42*(1), 313-332.

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American statistical association, 39*(227), 357-365.

Hair, J. F., Black, W. C., Anderson, R. E., & Tatham, R. L. (2009). *Multivariate Data Analysis* (7th ed.): Macmillan Publishing Co., Inc

Hill, S. E., & Kline, D. M. (2014). *Teaching "big data" in a business school: Insights from an undergraduate course in big data analytics.* Paper presented at the Proceedings of the Information Systems Educators Conference ISSN.

Linneman, T. J. (2021). From Measures of Association to Multilevel Models: Sociology Journals and the Quantitative Literacy Gap. *Teaching Sociology, 49*(1), 45-57.

Lottes, I. L., DeMaris, A., & Adler, M. A. (1996). Using and interpreting logistic regression: A guide for teachers and students. *Teaching Sociology*, 284-298.

Lyle, K. B., Bego, C. R., Hopkins, R. F., Hieb, J. L., & Ralston, P. A. (2020). How the amount and spacing of retrieval practice affect the short-and long-term retention of mathematics knowledge. *Educational Psychology Review, 32*(1), 277-295.

Reid, A., & Petocz, P. (2002). Students' conceptions of statistics: A phenomenographic study. *Journal of Statistics Education, 10*(2).

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological science, 3*(4), 207-218.

Sircar, S. (2009). Business intelligence in the business curriculum. *Communications of the Association for Information Systems, 24*(1), 17.

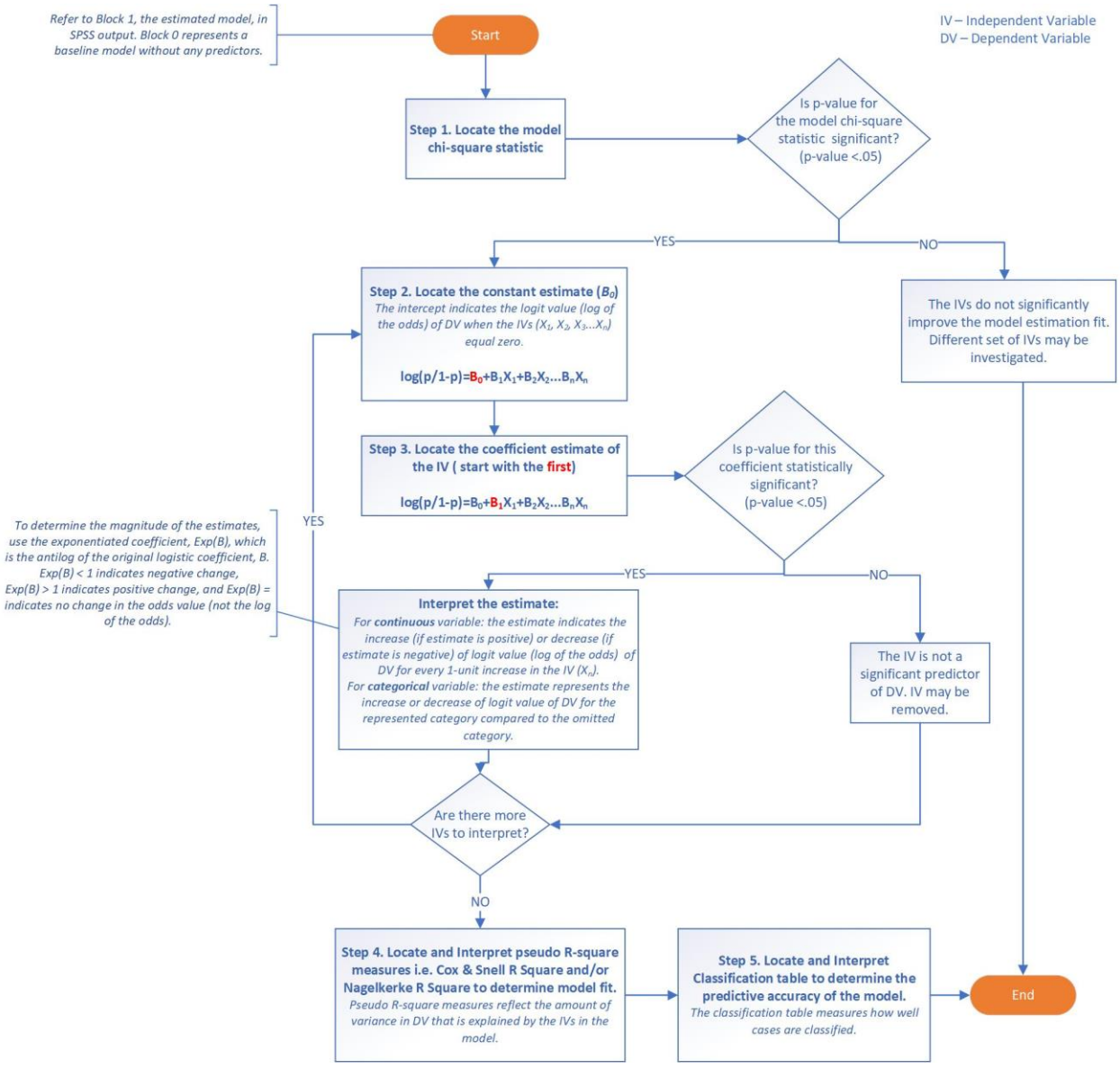
Toskin, K., & Kunene, N. (2020). *A Guide to Interpreting and Communicating Regression Analyses for Data Analytics Students.* Paper presented at the Proceedings of the EDSIG Conference ISSN.

Utts, J. (2015). The many facets of statistics education: 175 years of common themes. *The American Statistician, 69*(2), 100-107.

Walsh, A. (1987). Teaching understanding and interpretation of logit regression. *Teaching Sociology*, 178-183.

ISSN 2473-4901
v7 n5598

Appendix A
LOGISTIC REGRESSION FLOW DIAGRAM



ISSN 2473-4901
v7 n5598

Appendix B

EXAMPLE OF SPSS OUTPUT FOR LOGISTIC REGRESSION ANALYSIS

Logistic Regression

Logistic Regression - Case Processing Summary - July 10, 2021

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	392	51.0
	Missing Cases	376	49.0
	Total	768	100.0
Unselected Cases		0	.0
Total		768	100.0

a. If weight is in effect, see classification table for the total number of cases.

Logistic Regression

Logistic Regression - Dependent Variable Encoding - July 10, 2021

Dependent Variable Encoding

Original Value	Internal Value
0	0
1	1

Block 0: Beginning Block

Block 0: Beginning Block - Classification Table - July 10, 2021

Classification Table[a,b]

Observed			Predicted		
			DIABETES		Percentage Correct
			0	1	
Step 0	DIABETES	0	262	0	100.0
		1	130	0	.0
	Overall Percentage				66.8

a. Constant is included in the model.

b. The cut value is .500

Block 0: Beginning Block

Block 0: Beginning Block - Variables in the Equation - July 10, 2021

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-.701	.107	42.674	1	.000	.496

ISSN 2473-4901
v7 n5598

Block 0: Beginning Block

Block 0: Beginning Block - Variables not in the Equation - July 10, 2021

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	pregnant	25.804	1	.000
		glucose	104.252	1	.000
		pressure	14.552	1	.000
		triceps	25.677	1	.000
		insulin	35.617	1	.000
		mass	28.602	1	.000
		pedigree	17.177	1	.000
		age	48.241	1	.000
	Overall Statistics		135.543	8	.000

Block 1: Method = Enter

Block 1: Method = Enter - Omnibus Tests of Model Coefficients - July 10, 2021

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	154.077	8	.000
	Block	154.077	8	.000
	Model	154.077	8	.000

Block 1: Method = Enter

Block 1: Method = Enter - Model Summary - July 10, 2021

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	344.021[a]	.325	.452

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Block 1: Method = Enter

Block 1: Method = Enter - Classification Table - July 10, 2021

Classification Table[a]

			Predicted		
			DIABETES		Percentage Correct
	Observed		0	1	
Step 1	DIABETES	0	233	29	88.9
		1	56	74	56.9
	Overall Percentage				78.3

a. The cut value is .500