An Approach 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 (LoR) is a foundational supervised machine learning algorithm and yet, unlike linear regression, appears rarely taught early on, where analogy and proximity to linear regression would be an advantage. A random sample of 50 syllabi from undergraduate business statistics courses shows only two percent of the courses included LoR. Conceivable reasons for this dearth of LoR content is likely related to topic complexity, time constraints, and varying degrees of tool ease of use and support. We propose that these constraints can be countered by: [1] introducing logistic regression early, [2] informed tool selection prioritizing ease of use with comprehensive output, and [3] using/developing innovative, accessible, and easy to understand concept learning aids. This approach would leverage the proximity to linear regression and probability readily embed distributed practice for student 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 that is 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 analyses 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 in predictive analytics. We argue that, for pedagogical reasons, logistic regression should be introduced early because, long-term retention has been shown, repeatedly, in the psychological sciences to be positively impacted by *distributed practice* (Dunlosky, Rawson, Marsh, Nathan, &

Willingham, 2013). Specifically, distributed practice includes both *spacing effects* and *lag effects* where spacing outperforms massing (Benjamin & Tullis, 2010); and spacing with longer lags has advantage over spacing with shorter lags (Delaney, Verkoeijen, & Spirgel, 2010; Dunlosky et al., 2013). Thus, we argue, in the same way that probability and linear regression concepts are introduced early and repeated in subsequent analytics courses, introducing logistic regression would yield similar retention benefits for students.

However, logistic regression appears to be rarely taught in the foundational statistics courses that are part of analytics curricula. We argue that not teaching logistic regression in introductory statistics courses is a lost opportunity for leveraging the benefits of distributed practice that would be afforded by the teaching of linear regression in these courses.

The purpose of this paper is to: first, investigate the inclusion of logistic regression instruction and content in undergraduate business statistics courses. Second, we identify conceivable reasons and limitations why logistic regression, in contrast to linear regression, is rarely included in introductory courses. Thirdly, we propose workarounds to overcome these reasons and limitations. The proposed workarounds include, introducing logistic regression early, judicious tool selection that takes into account ease of use without compromising adequate concept coverage, as well as the development of innovative teaching aids to support instruction, student reviewing and distributed practice.

2. BACKGROUND

Logistic regression (LoR) 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. Although, 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 to OLS, and the instructional (time) proximity to both OLS and conditional probability on which it relies would be pedagogically advantageous.

We also see evidence of this absence of logistic regression in prior literature that outlines content maps for analytics programs. Sircar (2009) maps the analytics curriculum which includes a course on "Applied Regression Analysis in Business" that does not cover logistic regression. Similarly, Hill & Kline (2014) map content topics to student prior experience for each topic in a "Big Data Analytics" course development and roll out, show linear regression, both simple and multiple, are "partially covered in previous statistics courses", while logistic regression is a "new topic for *most* students".

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, the use of logistic regression (LoR) has also grown in popularity in other social sciences like sociology and psychology (which matters for data analytics minor programs looking to expand appeal to students from other majors), Linneman (2021) argues that because of LoR'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 (2021) 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) students are able to engage with; this is an approach that is superior to amassing repetitions in limited contexts (Schmidt & Bjork, 1992). Further, introducing logistic regression following linear regression is consistent with spacing strategies in pedagogy 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 maintain that the absence of LoR in content coverage early in students' exposure to related analytics content is a missed opportunity.

3. INSTRUCTIONAL GAP: DATA FINDINGS

We hypothesized that the teaching of logistic regression in introductory business statistics courses would be found in at least 10% of courses.

We then 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. Schools in our sample happen to be located in 25 distinct states. The sample consists of Research 1 schools, teaching-oriented schools as well as liberal arts colleges. They were primarily 4-year institutions with a large majority being AACSB accredited and, 68% our sample were public institutions.

Table 1 below, breaks down the broad characteristics of these institutions.

Institutional Characteristic	%
Program length:	
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

From the sample, we found that only 2% of the business statistics courses covered logistic regression.

The 2% offerings for LoR compare to 88% of the same courses that offer simple linear regression and 76% that offer multiple linear regression. See Figure 1 below.

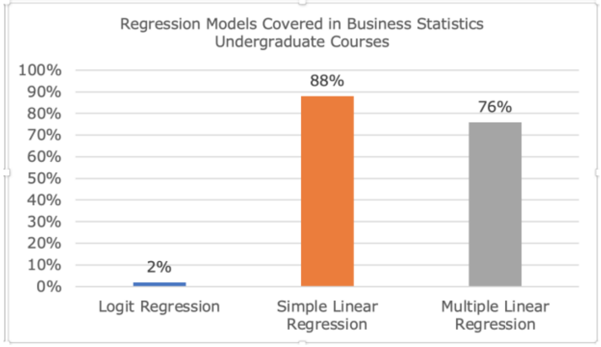


Figure 1. Sample of Business Statistics Courses: Comparative Frequency of Regression Model Coverage

We also found that there is a base of shared introductory topics within these offerings; primarily, these are: descriptive statistics; probability and probability distributions; sampling and estimation; and sampling distribution; [often included in this set is hypothesis testing and statistical inference, although this is not consistently offered. We will call this set, *Level 1* topics in an introductory business statistics course. Please note, correlation is frequently listed as a distinct topic from descriptive statistics in our sample syllabi.

In addition to *Level 1* topics, course offerings may also offer simple linear regression, multiple linear regression and (n)one to all of the following: Correlation, ANOVA, Chi-square (test for independence) and forecasting. We will call this set, *Level 2* topics.. Table 2 shows that simple (88%) and multiple (76%) linear regression are offered by many of the courses followed by correlation with 56% of the courses offering it. LoR is at the tail of the list at 2%.

Topic	Coverage %
Simple Linear Regression	88%
Multiple Linear Regression	76%
Correlation	56%
ANOVA	30%
Chi-square	22%
Forecasting	16%
Logit Regression	2%

Table 2. Level 2 Topic Coverage

In our sample, it is interesting to observe that among the Level 2 topics, while a large majority of courses cover linear regression (simple and multiple), in many cases, instructors/programs make decision choices between *Correlation*, *ANOVA*, *Chi-square*, *Forecasting* and *Logistic Regression* (see Figure 2 below).

Figure 2 below shows the frequency with which combinations of Level 2 topics are offered. At the top, 74% of the courses offered both simple and multiple Linear regression (together, simply referred to as Linear regression in the chart). When we ask: In addition to Linear regression (the most frequently offered Level 2 topic), which additional topic(s) is/are offered? We found that only about half the courses also cover Correlation (48%) as a distinct topic, whereas 26% add ANOVA. 18% cover Chi-square in addition to Linear regression. 14% cover Linear regression and ANOVA and Chi-square. 12% cover Linear regression and Correlation and Chi-square. 10% cover Linear regression, ANOVA, and Chi-square.

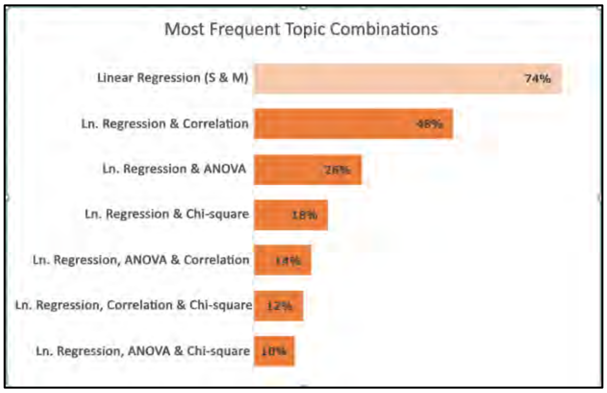


Figure 2. Most Frequent Topic Combinations

Lastly, we also found that there is an additional tier of topics offered within these courses. For example, we found syllabi that include *derivatives* and *optimization*; *capacity planning decisions*; *decision trees*; and index numbers (as in consumer and producer price indices). We will call this set, *Level 3* topics. There is less of a pattern to these offerings; they may be related to program/school or instructor specific priorities.

The data we gathered seems to, at least in part, support our prior experiential supposition that there is an instructional gap for logistic regression in introductory business statistics courses. While the specific reasons for this gap are subject to further empirical research, we do see evidence that the statistics courses we rely on as prerequisites of our data analytics curricula, on the main, do not introduce logistic regression (LoR). Below we explore and discuss this gap and propose some solutions.

4. DISCUSSION

Instructional gap: factors

The reasons for scarcity of logistic regression coverage in introductory business statistic courses may arise for several reasons. Reasons include, but are not limited to, the following: time constraints; the relative complexity of LoR, tool efficacy and easy access to the tool; and program/instructor priorities.

Time. 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/content and available course-time, that "*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). Introducing LoR is likewise cast in this context.

Relative complexity. Though somewhat analogous to simple and multiple linear regression, forming and interpreting the logit function is relatively more complex for logistic 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, 2021). There have been suggestions in the literature that (though unlikely in our belief) instructors may not feel necessarily comfortable to efficiently/effectively unpack the complexity for students (Linneman, 2021). This may also be attributed to instructors' own limited experience with (teaching) logistic regression.

Specifically, reasons why logistic regression is relatively more complex are: [1] unlike linear regression, we fit a regression model when the target or response variable is categorical (dichotomous in its simplest form), and ordinal. [2] 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 regression relationship to model the relationship between the dependent and independent variables (Hair, Black, Anderson, & Tatham, 2009). The response or target variable therefore serves as a *classifier* in many problem applications.

There are additional key differences:

1. The coefficients are converted to log odds. Most people have difficulty thinking in terms of log-odds (Lottes et al., 1996)
2. Model fit cannot technically be assessed using R-squared. Pseudo R-square values may be used, with some caution, to assess model fit. There are multiple Psuedo R-squares measures to choose from.
3. Logistic regression introduces a classification table to evaluate the predictive accuracy of the (classification) model.

These differences may seem a bit harder for students to grasp and interpret initially.

Tool efficacy and easy access. 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 and relatively cheaply 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 add-ins, e.g., 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, or the LINDO add-in, *What's Best!*. These add-on tools add a layer of complexity. *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* also require some use of the command line and, unlike R, are licensed programs. This often means, in general, they are not necessarily 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.

Instructional gap: proposed solutions

In this subsection we discuss how strategic tool selection, and use of innovative teaching aids, may coalesce to afford time-savings that instructors can leverage to introduce logistic regression early.

Tool selection. Instructors use various tools in their introductory statistics courses, from Microsoft Excel, to Minitab, R, SPSS, SAS and Stata potentially. It is possible that some even introduce Python. A tool's ease of use while preserving key concept coverage are important considerations. We investigated which would be the easiest tool for novices to use for logistic regression; 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, with add-ins	Low	
Python	Low	
SAS Programming	Low	
R	Low-Med	Med
Excel, XLMiner	High	Low-Med
Minitab	High	Med
SPSS	High	High

Table 3. 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 steep an "ease-of-use" bar to cross for an introductory course. Therefore, we discounted Python, SAS programming. R requires students to know additional commands for displaying key parameters of a logistic regression; we rated this as an ease-of-use hurdle. We did not assess SAS Enterprise Miner (EM) for data mining which is GUI based and includes a logistic regression algorithm, but the interface is geared towards data mining and machine learning which we believe is not suitable for an introductory statistics course.

Excel with Analytic Solver: while Excel is easy to use, and arguably Analytic Solver is 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 requiring additional add-ins like *What's Best!*. Using Excel with XLMiner was easy but severely lacking in output. It, too, is best for generating coefficients and their p-values. It also generates a model chi-square, however without the associated model p-values.

In the end, the remaining choices were Minitab and SPSS. The two products are both very easy to use. However, we found, SPSS produced richer output that is also easier to make sense of. We would recommend the use of SPSS for introductory courses. In cases where students do not have access to SPSS, faculty could generate logistic regression output for an assigned task and have students focus on the interpretive components of the task. Selecting a tool that is easy to use (while preserving important concept coverage) is an important time-saving consideration.

Innovative teaching aids. To reduce the complexity associated with teaching logistic regression, a possible solution is to develop and/or utilize existing innovative teaching aids. A *visual artifact* for introducing students to logistic analysis is an example of a teaching aid. Following, Toskin & Kunene (2021), we created a an example of a visual aid for logistic regression that uses flow diagramming for interpreting logistic regression output for SPSS, the tool we found easier to use. In the following section we describe the flow diagram. The diagram is included in Appendix A.

The flow diagram (see Appendix A) focuses on five key steps:

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. In other words, not to be interpreted as explaining variance)
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". 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) which is easier to interpret). 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 (i.e., odds of success), 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 and/or by selecting new variables, and the process of interpretation will start 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 or value 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, they do not explain variance as in linear regression.

Examining the model's classification accuracy is functionally more useful.

In *Step 5*, 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.

To help students transition from linear regression to logistic regression, this visual aid draws on similarities between the two techniques. For instance, the multiple linear regression teaching aid (Toskin & Kunene, 2021) 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. Here, for logistic regression, we draw students' attention to logistic regression concepts with near analogy to take advantage of what we can assume is, recent, student prior knowledge. This would alleviate the need to review or (re)introduce analogous regression concepts thus serving as time-saving affordance.

Program/instructor priorities, time.
In Figure 2, we show we found that only 2% of our sample courses covered logistic regression, while 74% covered linear regression (Level 2 topics). We also observe that even among Level 2 topics, there is some level of choice-making by programs or instructors (Figure 2). In other words, only some of Level 2 topics are found in an introductory business statistics course. Priorities are chiefly about purpose and time or, preference.

In this paper we argue for the inclusion of logistic regression early as a priority based on proven pedagogical strategies for long-term retention and the importance of logistic regression in analytics. Recognizing that there are time constraints, we have proposed a couple of affordances for time-savings. Nevertheless, if the case for logistic regression is made, what then could instructors trade off from available list of tradeoffs? It will depend on priorities. We would suggest leveraging concept proximity especially for those concepts that students are more likely to find complex. This would imply, if two topics compete for time and importance equally, and one has near analogy to a covered topic or concept, and the other not, prioritize the one with analogy and cover it with proximity to its analogy. We would expect introductory courses that restrict themselves to Level 1 topics only (i.e., also exclude linear regression) are necessarily exempt from this discussion in part because they may be taught as 2-credit courses. However, for courses that do offer Level 2 and 3 topics, introduction to logistic regression may be included using the approach proposed in this paper. It may or would require the exclusion of one extant topic from Level 2 or 3, e.g., ANOVA or decision trees, with discretion. Furthermore, correlation and chi-square can also be introduced in proximity with their concept relations, e.g., correlation with descriptive statistics.

The benefit of introducing LoR resides in offering undergraduate students the advantage of learning logistic regression, an important foundation in analytics, through strategic proximity and strategic repetitions, a pedagogical approach associated with better long-term retention.

5. CONCLUSION

In this paper, we establish that 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, based on our investigation logistic regression is rarely taught in introductory business statistics courses. Our findings are consistent with evidence from the work of Hill & Kline (2014) that logistic regression is (also) rarely taught in introductory statistics courses we rely on for analytics curricula.

We explored three challenges as likely contributing to the absence of logistic regression introductory courses. These challenges are time constraints, relative complexity, and tool efficacy and easy access. In our experience these challenges are not mutually exclusive and are interrelated. Therefore, reducing at least one of the challenges helps alleviate the remaining two constraints. Furthermore, taking advantage of the (time) proximity to linear regression (plus analogy) and proximity to probability concepts, teaching logistic regression early would embed *distributed practice* along the students' introductory analytics courses, with a better likelihood of increasing retention (Benjamin & Tullis, 2010; Delaney et al., 2010; Dunlosky et al., 2013) and therefore reducing perceived complexity for the student and *reducing the time needed for reinstruction*.

The appropriate choice of tool (i.e., its ease of use and relative comprehensiveness) is an important decision when attempting to address the above challenges. A tool that is easier to use and can generate comprehensive output (without requiring advanced skills) would reduce the amount of time needed to produce meaningful output from the analysis, ultimately reducing some of the complexity in performing analyses. Thus, freeing instructional time for work on comprehension and interpretation

Similarly developing or utilizing aids that are easier to understand and assist students in unpacking concept density, for example visual aids, has an effect on time and complexity constraints. In this paper we provide an example using flow diagraming, a proven pedagogical aid for unlocking complexity for novices. Such an aid can be used to strengthen students' capacity to interpret and communicate analysis for binary logistic regression models. Furthermore, the 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 it can be used by students in subsequent courses where logistic models are used. In that sense, it also lends itself to repetition and spacing strategies.

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. The growth in undergraduate analytics programs offers opportunities for students to enter a fruitful and financially rewarding field upon graduation. Logistic regression is a base competency in analytics, a gateway to many applied classification problems, undergraduate analytics students should be strategically afforded repeat encounters with it to gain competency.

Teaching logistic regression soon after linear regression would not only take advantage of students' immediate understanding of linear regression, it would also give an opportunity to a larger body of students to experiment with an important analytics technique using problem examples they are likely to recognize. For analytics programs, this matters if we want to expose as many students as possible to the types of problems students encounter in the field.

The contribution of this paper is, first, we show from empirical 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). It is conceivable that this influences its lack of coverage in introductory analytics courses identified in Hill & Kline (2014). We then explore conceivable reasons and limitations for why logistic regression, in contrast to linear regression, is rarely included in these introductory courses. Our findings also show that programs/instructors already make choice decisions in deciding which topics to prioritize particularly with respect to Level 2 and 3 topics. Lastly, we have proposed some mechanisms for instructors to subvert these constraints. Our proposed approach seeks to leverage both pedagogical strategies (proximity to analogy and repetition) and innovative aids to reduce the amount of time it would otherwise take to introduce logistic regression.

Finally, we support the design of creative mechanisms to enable students to readily access content and unlock complexity that would help students early in their academic careers. The development of similar aids for other topics on the *analytics content map* may help serve as part of a library of supplemental aids to be used in fill-in-the-gap approaches (Bauman & Tuzhilin, 2018) for the teaching of analytics. The example flow-diagramming aid introduced in this paper as well as those in Toskin & Kunene (2021) may serve as part of 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 for closing 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 effective 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, for complexity reduction and therefore as a time-saving affordance for new learners.

6. LIMITATIONS AND FUTURE RESEARCH

Though cognizant of time constraints, that adding some topic (i.e., logistic regression) implies forsaking another, we have not stipulated which topic may or should be excluded to accommodate logistic regression. Specifically, the answer depends on context and program objectives but, even for the general case, it is a question we will explore in future research."

7. 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.

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive psychology, 61*(3), 228-247.

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

Delaney, P. F., Verkoeijen, P. P., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. *Psychology of learning and motivation, 53*, 63-147.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4-58.

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.

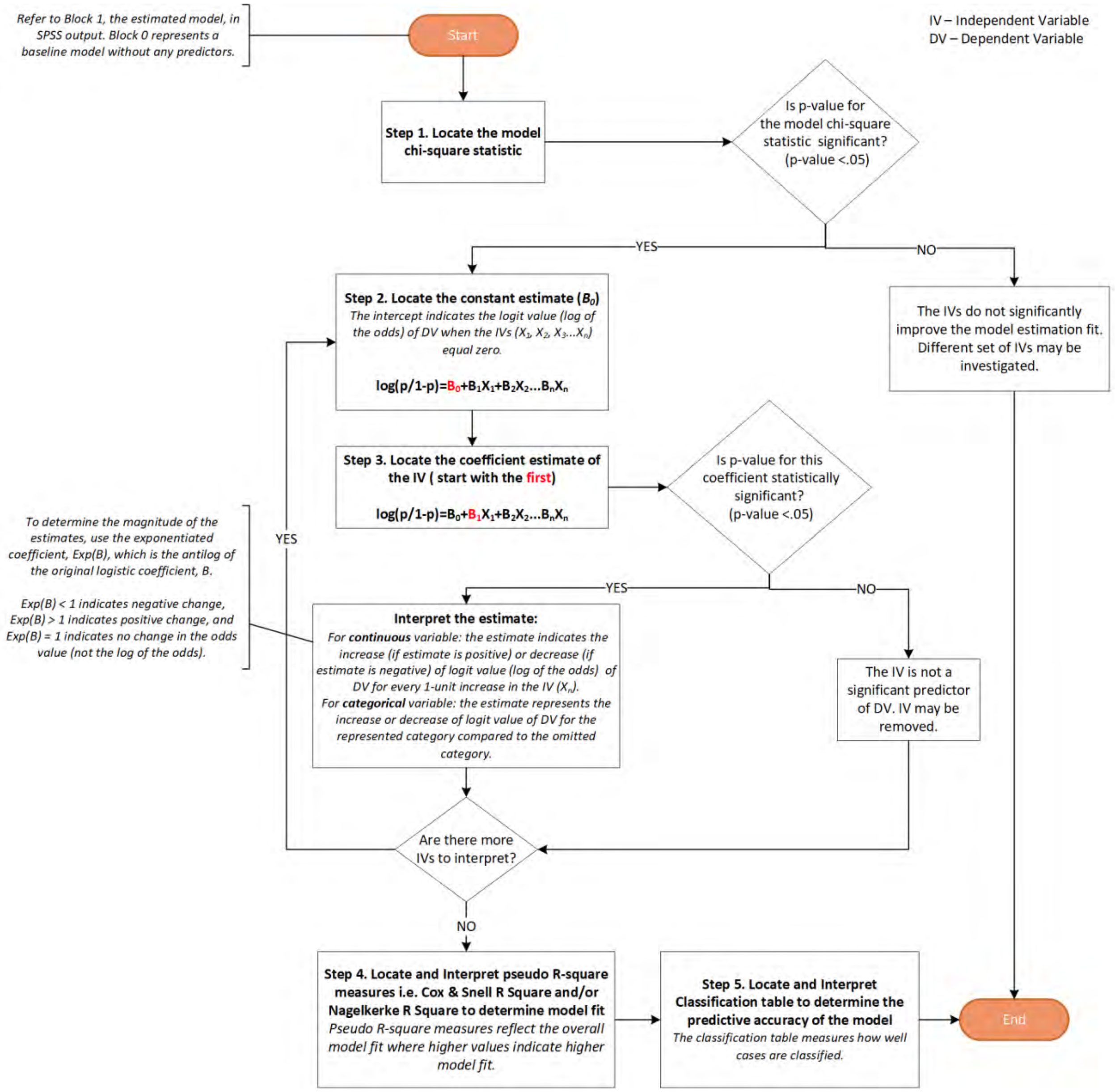
Toskin, K., & Kunene, N. (2021). Flow Through Regression: A Guide to Interpreting and Communicating Regression Models for Data Analytics Students. *Information Systems Education Journal, 19*(5), 5.

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.

Appendix A

LOGISTIC REGRESSION FLOW DIAGRAM



Refer to Block 1, the estimated model, in SPSS output. Block 0 represents a baseline model without any predictors.

IV – Independent Variable
DV – Dependent Variable

Step 1. Locate the model chi-square statistic

Is p-value for the model chi-square statistic significant? (p-value <.05)

The IVs do not significantly improve the model estimation fit. Different set of IVs may be investigated.

Step 2. Locate the constant estimate (B_0)
The intercept indicates the logit value (log of the odds) of DV when the IVs (X_1, X_2, X_3...X_n) equal zero.

$\log(p/1-p)=B_0+B_1X_1+B_2X_2...B_nX_n$

Step 3. Locate the coefficient estimate of the IV (start with the first)

$\log(p/1-p)=B_0+B_1X_1+B_2X_2...B_nX_n$

Is p-value for this coefficient statistically significant? (p-value <.05)

To determine the magnitude of the estimates, use the exponentiated coefficient, Exp(B), which is the antilog of the original logistic coefficient, B.

Exp(B) < 1 indicates negative change, Exp(B) > 1 indicates positive change, and Exp(B) = 1 indicates no change in the odds value (not the log of the odds).

Interpret the estimate:
*For **continuous** variable: the estimate indicates the increase (if estimate is positive) or decrease (if estimate is negative) of logit value (log of the odds) of DV for every 1-unit increase in the IV (X_n).*
*For **categorical** variable: the estimate represents the increase or decrease of logit value of DV for the represented category compared to the omitted category.*

The IV is not a significant predictor of DV. IV may be removed.

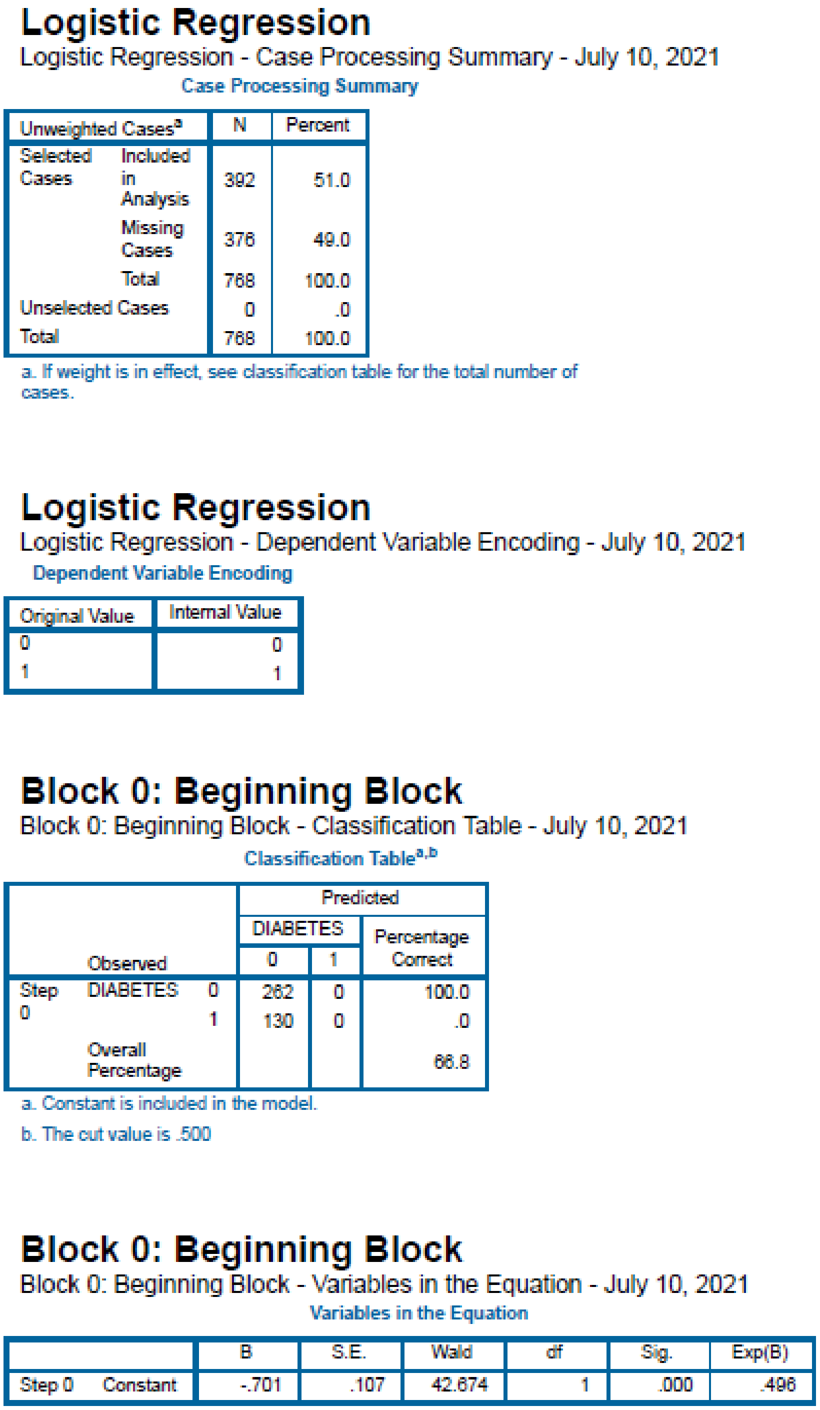
Are there more IVs to interpret?

Step 4. Locate and Interpret pseudo R-square measures i.e. Cox & Snell R Square and/or Nagelkerke R Square to determine model fit
Pseudo R-square measures reflect the overall model fit where higher values indicate higher model fit.

Step 5. Locate and Interpret Classification table to determine the predictive accuracy of the model
The classification table measures how well cases are classified.

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]

			Predicted		
			DIABETES		Percentage Correct
Observed			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.

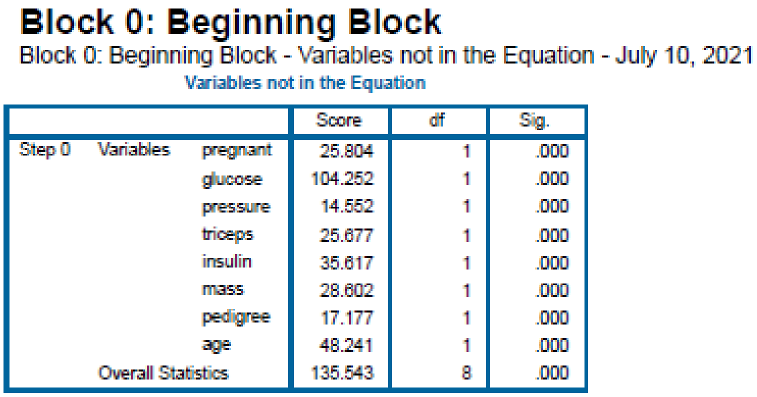
b. The cut value is .500

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

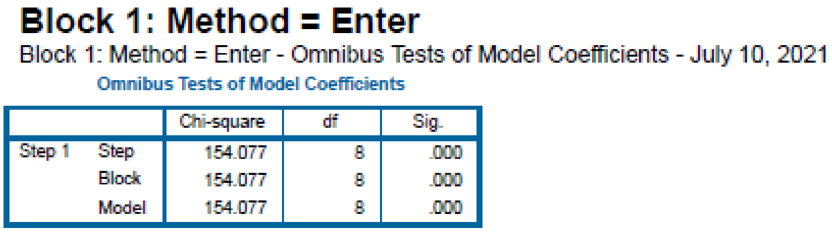


Block 0: Beginning Block

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

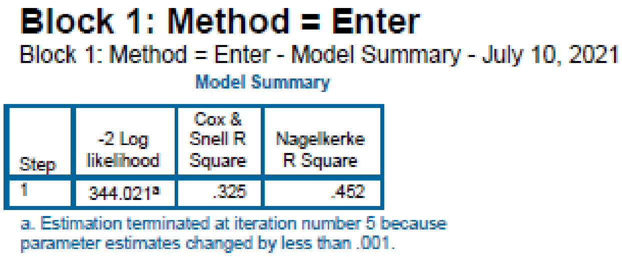


Block 1: Method = Enter

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



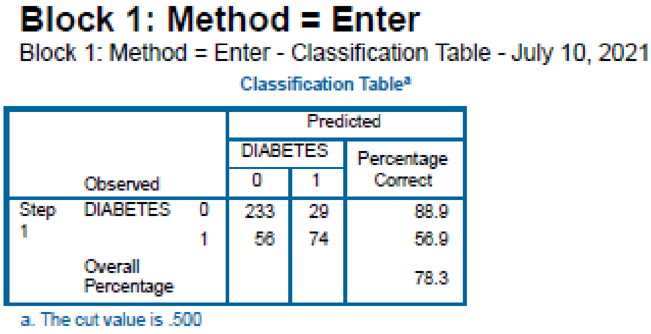
Block 1: Method = Enter

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