Hoang, Viet Ngu & Watson, Justin
(2022)
Teaching binary logistic regression modeling in an introductory business analytics course.
Decision Sciences Journal of Innovative Education, *20*(4), pp. 201-211.

https://doi.org/10.1111/dsji.12274

Received: 13 August 2021 | Accepted: 27 June 2022

DOI: 10.1111/dsji.12274

DECISION SCIENCES
JOURNAL OF INNOVATIVE EDUCATION

TEACHING BRIEF

Teaching binary logistic regression modeling in an introductory business analytics course

Viet-Ngu Hoang[1,2] | **Justin Watson**[2]

[1]Faculty of Business and Law, Queensland University of Technology, Brisbane, Queensland, Australia

[2]Adjunct Professor, International School, Vietnam National University, Hanoi, Vietnam

Correspondence
Viet-Ngu Hoang, Queensland University of Technology, 2 George Street, Brisbane, Queensland, 4000, Australia.
Email: vincent.hoang@qut.edu.au

Abstract

There is an increasing demand to introduce Introductory Business Analytics (IBA) courses into undergraduate business education. Many real-world business contexts require predictive analytics to understand the determinants of a dichotomous outcome; hence, IBA courses should include binary logistic regression analysis. This article provides our reflective discussions on the design of learning activities and assessments to assist business students in learning binary logistic regression in an IBA course. Data on student engagement and learning outcomes are used to shed light on the impacts of teaching logistic regression on student learning and experience. Notably, students opt to focus their assessment work more on logistic regression than on multiple regression analysis, showing the potential attraction of students toward binary logistic regression analysis. We also observed several challenges, mainly related to the use of Excel, that require special attention from instructors.

KEYWORDS

business analytics, classification modeling, binary logistic regression, linear regression

1 | INTRODUCTION

In response to the rapid development of the digital economy and data-based decision-making in the business world, many universities have introduced applied business analytics into undergraduate business curricula (Al-Haddad et al., 2019; Mamonov et al., 2015; Phelps & Szabat, 2017; Yazici, 2020). A significant portion of business decision-making requires dichotomously framing the determinates of outcomes. Examples include whether consumers purchase or consume a product, whether they are satisfied with a product, repeat purchasing, and whether they engage with marketing campaigns (Lawson & Montgomery, 2006). In these circumstances, traditional descriptive analytics is less informative. Students must move toward predictive analytics, including classification-modeling frameworks such as logistic regression analysis, to gain deeper insights that add value in business contexts. Therefore, Introductory Business Analytics (IBA) courses should include binary classification modeling using logistic regression.

While introducing binary logistic modeling into IBA courses creates more real-world learning opportunities for students, teaching it in undergraduate IBA courses faces significant challenges. First, given limited contact hours for lectures and tutorials within a single course, introducing new topics must be at the expense of other topics. Second, logistic modeling is more advanced than linear regression analysis and requires more resources for effective teaching and learning. This is because logistic regression uses maximum likelihood estimation, which requires software and expertise that are less accessible to students with limited prior statistical knowledge. Such expertise includes understanding concepts such as likelihood functions, optimization, odds ratios, sigmoid functions, and logarithms and exponentials. In our case, our student base consists primarily of first-year undergraduate students who have limited or no prior exposure to spreadsheeting and statistical modeling.

Literature on best practices for designing and improving analytics curricula has been growing in recent decades (Cegielski & Jones-Farmers, 2016; Johnson et al., 2022).

However, there is limited literature on teaching binary logistic regression in IBA or statistics courses. One exception is the study by Li et al. (2018), which highlights the inclusion of logistic regression in an introductory business statistics course for Master of Business Administration (MBA) students. To fill this gap in the literature, our article demonstrates how binary logistic regression can be taught in introductory analytics courses to first-year undergraduate business students. We focus on learning activities designed using an experiential learning pedagogy that addresses the challenges we face and justifies our approach with an analysis of student learning and engagement. We also provide anecdotal insights on the effectiveness of using binary logistic regression in promoting the analytical thinking of undergraduate business students.

Undoubtedly, teaching binary logistic regression requires computational software. From the software tools reviewed in Johnson and Berenson (2019), we decided to use Microsoft Excel because it has proven to be an effective teaching tool in IBA courses (Mamonov et al., 2015; Warner & Meehan, 2001), important for the employability of business graduates (Rienzo & Chen, 2018), and essential for entry-level jobs in business analytics (Cegielski & Jones-Farmer, 2016). However, Excel does not have an in-built tool that allows students to run logistic regression. For this, we used a freely available Excel add-in called RegressItLogistic. While this add-in is useful for teaching modeling, it posed several challenges that we address below.

2 | LITERATURE REVIEW

Despite the growing literature for designing and improving analytics curricula (Cegielski & Jones-Farmers, 2016; Johnson et al., 2022), there is still limited discussion around teaching binary logistic regressions in IBA courses, especially for undergraduate students. This section aims to provide a brief literature review to justify the inclusion of binary logistic regression as part of predictive analytics in IBA courses for undergraduate business students.

2.1 | Predictive analytics and binary logistic regression

Standard IBA courses often introduce fundamental descriptive and predictive analytics to business students. Predictive analytics uses modeling techniques to predict the future by examining historical data, detecting patterns or relationships, and extrapolating these patterns or relationships to situations that exist outside of available data (Barton & Court, 2012; Davenport & Patil, 2012; Evans, 2020). In business contexts, predictive analytics helps organizations proactively allocate resources toward future situations, a significant component of an organization's competitive advantage and strategic objectives (Davenport, 2006).

Predictive analytics has applications in many areas of business, particularly in budgeting, managerial accounting, consumer behaviors and relations, risk management, and supply chain management (Waller & Fawcett, 2013). Studies have linked the effective use of predictive analytics to revenue generation, cost minimization, operational efficiency, and enhanced customer service (Dey & Kumar, 2010). Thus, business graduates need knowledge and skills in predictive analytics, especially on how predictive analytics can be used to generate insights that enhance business decision-making in a variety of contexts (Cegielski & Jones-Farmer, 2016; Johnson et al., 2020; Waller & Fawcett, 2013).

Basic analytics techniques such as simple and multiple linear regression are fundamental to predictive analytics and are covered in most IBA courses in undergraduate business degrees. However, linear regression models are more suitable for modeling continuous outcome variables, while many important real-world business situations lend themselves toward modeling discrete outcome variables. In these cases, logistic regression analysis is more appropriate (Lawson & Montgomery, 2006).

Literature has shown that logistic regression is among the most popular methodological tools in predictive analytics (Brusco, 2022). Given that more introductory IBA and statistics books are including logistic regression (Morrell & Auer, 2007), one could expect that logistic regression has been taught widely in undergraduate IBA courses. However, our literature review finds limited discussion on curriculum design for those courses. Most studies provide insights on teaching logistic regression in advanced undergraduate courses or post-graduate courses. For example, Li et al. (2018) highlight the inclusion of logistic regression in an introductory business statistics course for post-graduate students, and Brusco (2022) describes the use of Excel templates to enhance conceptual understanding of binary logistic regression for graduate students. Morrell and Auer (2007) illustrate classroom activities related to logistic regression in an advanced applied statistics course.

Importantly, these studies show that teaching logistic regression helps reinforce linear regression concepts (Morrell & Auer, 2007) and enhance critical and analytical thinking (Brusco, 2022; Li et al., 2018). In contrast to linear regression, binary logistic regression offers unique opportunities for students to learn about out-of-sample testing and how to evaluate models using classification performance metrics (accuracy, sensitivity, and specificity). Asking students to use these metrics to evaluate models also reinforces prior learning on errors in prediction. These learning opportunities have the potential to improve critical- and analytical- thinking skills in students, which are essential for business graduates (Braun, 2020; Calma & Davies, 2021).

Conversely, there are many reasons to support the exclusion of logistic regressions in IBA courses. First, time constraints require instructors to choose certain topics at the expense of others. Second, logistic regression requires different statistical considerations than linear regression (Lawson & Montgomery, 2006). Logistic regression uses maximum likelihood estimation, which requires software and expertise that are less accessible to students with limited prior statistical knowledge. Such expertise includes understanding concepts such as likelihood functions, optimization, odds ratios, sigmoid functions, and logarithms and exponentials; thus, logistic regression can be perceived as being too advanced for undergraduate IBA courses. Finally, as we discuss shortly, there is a lack of accessible and freely available computational software for students to seamlessly run logistic regression and perform necessary model diagnostics.

We believe that the absence of binary logistic regression in IBA courses could lead to a significant loss of learning opportunities for undergraduates. At least at our institution, this represents a considerable problem because many business students do not pursue further studies in business analytics. Thus, there are limited opportunities for them to learn about modeling dichotomous situations and develop critical- and analytical-thinking skills in business contexts. This could lead to business graduates being ill-prepared to tackle real-life business analytics projects (Dykes, 2016). This is concerning, given the increasing demand from businesses for graduates that can make data-driven decisions (Davenport & Harris, 2007; Hazen et al., 2016).

2.2 | The use of Microsoft Excel

The choice of computational software tools is essential to the success of an IBA course (Johnson & Berenson, 2019; Johnson et al., 2020; Rienzo & Chen, 2018). The Guidelines for Assessment and Instructions in Statistics Education College Report identifies the use of computational software as one of six recommendations to enhance the pedagogical delivery of statistics courses (Carver et al., 2016).

From the software tools surveyed in Johnson and Berenson (2019), Microsoft Excel supports topics covered in most introductory statistics courses. Excel has become one of the most used software packages for IBA and statistics courses (Haskin & Krehbiel, 2012; Warner & Meehan, 2001), and many IBA textbooks provide resources for using Excel (Camm et al., 2020; Evans, 2020). Using Excel to teach analytics also improves student employability as competency in Excel is expected of business graduates (Rienzo & Chen, 2018) and essential for entry-level jobs in business analytics (Cegielski & Jones-Farmer, 2016). Interestingly, Al-Haddad et al. (2019) report that students who learned Excel alongside IBA performed better than students learning only Excel skills in an information technology course. Also, literature shows that purpose-built Excel templates or workbooks can play a vital role in teaching logistic regression or Bayesian methods in more advanced courses (Brusco, 2022; Johnson & Berenson, 2019). Guided by these studies, we decided to use Excel in our IBA courses.

However, Excel has many limitations that inhibit its use in a project-orientated business analytics course (Johnson & Berenson, 2019). In particular, Excel does not have an in-built capability to perform logistic regression. Many commercial Excel add-ins enable logistic regression analysis, but using them adds costs to the course. At the time of writing this article, there are two freely available Excel adds-in that enable logistic regression with model diagnostic tools: RegressItLogistic (available at https://regressit.com/index.html) and Real-Statistics (available at https://www.real-statistics.com). Alternatively, Google Sheets with the free XLMiner extension (available at https://workspace.google.com/marketplace/app/xlminer_analysis_toolpak/600284989882) allows for logistic regression analysis, although the free version does not allow users to perform model diagnostics such as classification tables and out-of-sample testing.

3 | COURSE BACKGROUND

Our IBA course was offered for the first time in 2021 as an optional core unit to all students who enrolled in newly designed undergraduate business programs at an international university. Of the 80 students enrolled initially, 54 remained by the end of the semester. Students came from diverse disciplinary backgrounds, including economics, finance, financial planning, accounting, marketing, public relations, management, and international business. Enrollment in the following semesters increased to over 160 students, indicating increasing interest in this course.

The course syllabus describes the course as an introduction to a practical framework for data collection, aggregation, processing, and modeling to transform data into business insights. More specifically, students will gain knowledge of key business analytics methods and approaches while building Excel skills to manage and analyze datasets that improve decision-making in contemporary business environments across all business disciplines. Student learning objectives on completion of the course include being able to: (1) apply a selection of methods to collect, aggregate, and process business data from multiple sources; (2) present data and analyze results in effective forms to assist business decision-making in a variety of industrial and organizational contexts; (3) critically analyze data using a variety of methods and approaches to generate business insights and inform business decision-making; and (4) apply key principles and processes of business analytics for the creation of value from data.

4 | OVERVIEW OF LEARNING ACTIVITIES

The course has 11 weeks of learning plus a week of revision. Each week has a 2-h lecture and 1.5-h tutorial conducted in computer rooms (details of weekly learning materials can be found in the Supplementary Material). The content of the course has three main components:

(1) Business problem conceptualization: The first 2 weeks are designed to help students learn how to structure

business problems and translate these problems into analytics problems using the CRoss Industry Standard Process for Data Mining (CRISP-DM) framework. This framework has been widely regarded as the most relevant and comprehensive framework for executing university and real-world analytics projects (Abbasi et al., 2016; Jaggia et al., 2020).

(2) Descriptive analytics: Over 4 weeks, a variety of topics relating to types of data, data visualization, descriptive statistics, and hypothesis testing are taught in lectures and practiced in tutorials using Excel's Data Analysis ToolPak.

(3) Predictive analytics: Causality is taught prior to predictive analytics to highlight the importance of solid theoretical foundations in implementing predictive analytics. Simple and multiple linear regressions were introduced over 2 following weeks. The last 2 weeks focused on logistic regression and forecasting using time-series data.

Tutorials following each lecture worked through several real-world examples. All tutorials were conducted in computer labs, but students could also use their own laptops. Most tutorial questions required students to use Excel and work on a variety of datasets. Particularly, Excel's Analysis ToolPak was introduced during Week 4 and extensively used until Week 9.

The course was designed around our institution's real-world learning framework, where guest lecturers and real-world case studies are utilized in lectures and tutorials. This unit includes three assessment criteria:

1. Assessment 1 (20%)—four online quizzes to help students review the main concepts taught in the first five weeks. Table 1 lists key core concepts that were quizzed in these online tests.
2. Assessment 2 (35%)—a business analytics plan where students demonstrate their skills in framing business problems and using descriptive analytics tools. Details of this assessment, including marking criteria, are in the Supplementary Material.
3. Assessment 3 (45%)—a final report where students demonstrate their skills in linear and logistic regression analysis using data from Assessment 2.

In Assessments 2 and 3, we adopted the consumer-choice data used in Jaggia et al. (2020), which is available from the authors of that article on request. We modified the data, which was originally designed to be analyzed using R, by removing complex categorical variables that are difficult for Excel to analyze. This was also done to better cater to our student demographic and to maintain focus on explaining and predicting phenomena, consistent with our learning objectives. The dataset contains data on the consumption of frozen yogurt and frozen meals for 3189 households, as well as characteristics such as home ownership, the number of individuals within a household, the number of hours worked by the male and female heads of household, and the education of the heads of household. Key variables are defined in Table 2.

5 | OVERVIEW OF PEDAGOGICAL APPROACH AND SCAFFOLDING ACTIVITIES PRIOR TO TEACHING LOGISTIC REGRESSION

We adopted a constructive-alignment approach to teaching for this course (Biggs, 1996; McCann, 2017). Due to the diverse backgrounds of students in our course, we focused on developing a high level of critical thinking around how to use statistical modeling to "explain" and "predict" phenomena (Shmueli, 2010). We introduced a systematic approach to modeling (Figure 1) when teaching simple and multiple linear regression for the purpose of "explaining." We used data on cross-country happiness from Helliwell et al. (2020) and data on the income of business graduates from Evans (2020) during the lecture and designed tutorial activities in the following weeks to allow students to replicate the analysis. The focus during these 2 weeks was on how to use multiple regression to "explain" the phenomena. Also, students were given opportunities to utilize prior knowledge of causal analysis and hypothesis testing to interpret results from multiple regression analysis.

Students were exposed to binary data prior to binary logistic regression in the form of indicator variables. In tutorials, students accessed exercises and materials on descriptive statistics for binary variables, hypothesis testing for population proportions, and including indicator variables in multiple regression. Students had gained an understanding of binary data by Week 10 when we introduced binary logistic regression in the lecture. In Weeks 11 and 12, we dedicated computing tutorials implementing logistic regression analysis on the dataset for Assessment 3. Details of these activities are in the next section.

6 | EXCEL ADD-IN—REGRESSITLOGISTIC

In our own assessment, RegressItLogistic provides a user-friendly presentation of summaries for different model specifications, which is helpful for evaluating the performance of their models. Therefore, we decided to use RegressItLogistic in our first semester 2021 course offering.[1] A trade-off is that this add-in is not compatible with some Macintosh computers. To mitigate this issue, we encouraged Mac users to use university desktops or the virtual desktop offered by the university for tutorial sections and assessments.[2]

[1] We used the Real-Statistics add-in in 2022 to overcome the fact that RegressItLogistic had issues on some Mac computers.

[2] During the first semester of 2021, amid the global Coronavirus disease (COVID-19) pandemic, most of students enrolled in this course were local students who were able to attend classes on campus and use the university computers during the period that logistic regression topics were taught. However, we noticed a reduction in the face-to-face attendance during Weeks 9–12.

TABLE 1 Core concepts examined in four online quizzes

Quiz number	Core concepts examined	Number of questions
1	Elements of business analytics projects	2
1	CRISP-DM framework	3
1	Types of data	5
1	Types of analytics	3
1	Proper uses of data visualization	2
2	Measure of central tendency	4
2	Measures of dispersion, skewness, and outliers	5
2	Measures of association	3
2	Probabilities, frequency and frequency distributions	3
3	Characteristics of normal distributions	3
3	Characteristics of Bernoulli distributions	3
3	Characteristics of sampling distributions	3
3	Types of random sampling	4
3	Differences between point and interval estimates	2
4	Characteristics of *t*-distribution	3
4	Applications of confidence interval estimates for population proportions	4
4	Applications of confidence interval estimates for population mean (knowing population standard deviation)	4
4	Applications of confidence interval estimates for population mean (not knowing population standard deviation)	4

TABLE 2 Descriptions of key variables for the dataset used in Assessments 2 and 3

Variable	Description
HH_ID	The household's identification number
HHOwn	Whether a household owns their home; 1 if yes and 0 otherwise
HHNbr	The number of members in the household
MWrkHrs	The average hours worked each week by the male head of household
MEdu	Whether the male head of household has an education equivalent to a college degree or above; 1 if yes and 0 otherwise
FWrkHrs	The average hours worked each week by the female head of household
FEdu	Whether the female head of household has an education equivalent to a college degree or above; 1 if yes and 0 otherwise
YogExp	A household's frozen yogurt expenditures per annum (in $)
DinExp	A household's frozen meal expenditures per annum (in $)
TotalExp	A household's total frozen yogurt and frozen meals expenditures per annum (in $)
Yog	Whether a household has purchased frozen yogurt in a given year; 1 if yes and 0 otherwise
Din	Whether a household has purchased frozen meals in a given year; 1 if yes and 0 otherwise
YogOrDin	Whether a household has purchased frozen yogurt or frozen meals in a given year; 1 if yes and 0 otherwise

One issue with using RegressItLogistic (and other freely available add-ins) is that there are limited video resources available for both instructors and students. Therefore, we decided to demonstrate its use in lectures. In addition, we recorded a separate video clip that uses another dataset to demonstrate how to install the add-in, prepare a dataset, run binary logistic regression, diagnose models, and interpret key results. This video can be made available on request. In addition, we created a page of useful Excel and RegressItLogistic resources on the blackboard site for the unit and showed the page to students during lectures and tutorials.

One advantage of using RegressItLogistic is its ability to report Model Summaries, which list estimated results of key parameters used to compare across different model specifications. Figure 2 presents an example that we demonstrated in our lecture during Week 11. Using these Model Summaries sheets, we provided instructions on how students could use rates of true positives and true negatives to justify which model should be used for predictive purposes. The models

1. Descriptive analysis & checking out for outliers in both Y and X variables
2. Consider the use of causal graphs.
3. Correlation matrix of all available variables (this helps to detect collinearity or multicollinearity issues). If high correlation coefficients, then we need to decide what variables better than others.
4. Construct a model with all "good" candidates of available independent variables. (Note: some prefers to use another approach: starting from a very simple model then adding more variables into it).
5. Examine the coefficients and p-values for each independent variables.
   - Do relationships follow some theories or norms or expectations?
   - If p-values > 10%, consider to remove and run step 4 again, then check adjusted R-square.
6. Once majority (or all) x variables are statistically significant and the signs of coefficients are consistent with expectations, then you are closer to a good model.
7. Check any violation of assumptions (for example, using residual plots).

FIGURE 1 The systematic model-building approach

Summary of Regression Model Results

Logistic Model For Approved	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Run Time	10/11/21 2:17 PM	10/11/21 2:51 PM	10/11/21 3:34 PM	10/11/21 3:35 PM	10/11/21 3:36 PM	10/11/21 3:36 PM	10/11/21 3:37 PM
# Fitted	45	45	45	45	45	45	45
Mean	0.467	0.467	0.467	0.467	0.467	0.467	0.467
Standard Deviation	0.499	0.499	0.499	0.499	0.499	0.499	0.499
# Variables	5	4	3	2	2	2	4
RMSE	0.000	0.220	0.235	0.262	0.262	0.262	0.000
R-squared	1.000	0.772	0.743	0.701	0.701	0.701	1.000
Adjusted R-squared	0.807	0.611	0.615	0.604	0.604	0.604	0.839
Maximum VIF	4.121	2.091	1.338	1.093	1.093	1.093	2.421
AIC	12.005	24.199	23.964	24.596	24.596	24.596	10.003
Area Under ROC Curve	1.000	0.986	0.982	0.976	0.976	0.976	1.000
Cutoff Level	0.500	0.500	0.500		0.500	0.500	0.500
Percent Correct	100.0%	93.3%	91.1%		88.9%	88.9%	100.0%
True Positive Rate	100.0%	95.2%	95.2%		90.5%	90.5%	100.0%
True Negative Rate	100.0%	91.7%	87.5%		87.5%	87.5%	100.0%
Test Variable	Test_Sample	Test_Sample	Test_Sample	Test_Sample	Test_Sample	Test_Sample	Test_Sample
Percent Correct	100.0%	100.0%	80.0%	80.0%	80.0%	80.0%	100.0%
True Positive Rate	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
True Negative Rate	100.0%	100.0%	66.7%	66.7%	66.7%	66.7%	100.0%
Coefficients	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Constant	-112.897 (0.663)	-34.044 (0.012)	-30.252 (0.007)	-34.654 (0.007)	-34.654 (0.007)	-34.654 (0.007)	163.353 (0.898)
Credit_Score	0.18 (0.625)	0.045 (0.012)	0.04 (0.007)	0.045 (0.007)	0.045 (0.007)	0.045 (0.007)	
Homeowner	8.113 (0.941)	3.205 (0.124)	5.133 (0.013)				74.783 (0.951)
Revolving_Balance	0.001 (0.873)	0 (0.917)	0 (0.723)				0.007 (0.711)
Revolving_Utilization	-66.922 (0.644)						-613.132 (0.630)
Years_of_Credit_History	0.279 (0.947)	0.267 (0.252)		0.428 (0.025)	0.428 (0.025)	0.428 (0.025)	-3.412 (0.729)

... | Descriptive Stat | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 | **Model Summaries**

FIGURE 2 A screenshot of "Model Summaries" shown in lecture during Week 11

presented in Figure 2 were estimated from a small dataset with 45 observations in a training set, five observations in a testing set, and six observations in a holdout/prediction set. We highlighted to the students that due to this specific small dataset, the model seems to be overfitted and that several models yielded 100% prediction accuracy. To further enhance learning, we asked students to run two binary regression models on data for Assessment 3 using RegressItLogistic during tutorial classes in Week 12. We spent around 15 min during the class discussing the results reported in the Model Summaries sheet as shown in Figure 3. Unfortunately, we do not see this function built in other Excel add-ins including RealStatistics or the free XLMiner extension in Google Sheets. If not using the RegressItLogistic add-in, instructors teaching model comparison need to construct templates for the students to use or show students how to do comparisons.

7 | ACTIVITIES TO ASSIST STUDENTS IN LEARNING BINARY LOGISTIC REGRESSION

In the Week 10 lecture, students were introduced to dichotomous classification models using credit-card-application approval data from Evans (2020). Specific concepts taught included: (1) differences between logistic and linear regression; (2) measuring classification performance using accuracy, sensitivity, and specificity metrics; (3) out-of-sample testing; and (4) predicted probabilities. We are happy to provide access to spreadsheets and lecture notes on request. Also, guided by the constructive-alignment approach, we aligned learning activities in tutorials during Weeks 11 and 12 with tasks in Assessment 3. During these tutorials, activities were designed to show students how to use the

Logistic Model For Q9_4Dummy	Model 1	Model 2
Run Time	10/18/21 1:10 PM	10/22/21 9:51 AM
# Fitted	63	63
Mean	0.730	0.730
Standard Deviation	0.444	0.444
# Variables	9	8
RMSE	0.376	0.376
R-squared	0.229	0.227
Adjusted R-squared	0.000	0.000
Maximum VIF	4.793	1.475
AIC	76.640	74.826
Area Under ROC Curve	0.798	0.790
Cutoff Level	0.500	0.500
Percent Correct	84.1%	84.1%
True Positive Rate	95.7%	95.7%
True Negative Rate	52.9%	52.9%
Test Variable	Test_Sample	
Percent Correct	42.9%	
True Positive Rate	60.0%	
True Negative Rate	0.0%	
Coefficients	Model 1	Model 2
Constant	0.439 (0.921)	1.459 (0.698)
Q10_2Worthwhile	1.495 (0.049)	1.563 (0.036)
Q15_Female	0.266 (0.837)	-0.175 (0.825)
Q15_Male	0.541 (0.667)	
Q16_less20hrs	-0.555 (0.539)	-0.5 (0.578)
Q16_more20hrs	0.169 (0.845)	0.134 (0.875)
Q17_age	-0.08 (0.272)	-0.085 (0.240)
Q20_OPScores	-0.012 (0.725)	-0.018 (0.555)
Q23_Engagement	2.518 (0.044)	2.541 (0.042)
Q7_7Dummy	0.779 (0.281)	0.829 (0.244)

FIGURE 3 A screenshot of "Model Summaries" shown in computing tutorial during Week 12

RegressItLogistic add-in to implement logistic regression analysis for the last assessment.

As noted earlier, students were required to use a relatively large dataset in Assessments 2 and 3. These assessment exercises were modeled around the CRISP-DM framework (Wirth & Hipp, 2000), which had been taught and practiced in lectures and tutorials in the first several weeks of the course and in Assessment 2. Assessment 2 focuses solely on descriptive analytics. Students were advised that the focus of Assessment 3 is on using multiple linear regression and logistic regression to answer two questions: (1) which households are likely to purchase yogurt and frozen dinner products? and (2) how much money is each household likely to spend on each product? Our lectures and tutorials during Weeks 9–12 noted that multiple regression should be used to answer the first question and logistic regression to answer the second.

Students were given the marking rubric as part of the assessment document (see Supplementary Material), and they were required to submit the final report together with Excel spreadsheet files showing their modeling work. The data analysis presented by student work accounts for 15% of the entire mark, the articulation of modeling, evaluation, and deployment aspects of the CRISP-DM framework accounts for 45%, and other aspects account for the remaining 40%.

As mentioned earlier, computing tutorials used Assessment 3 data in the tutorial classes of Weeks 11 and 12, with a focus on binary logistic regression. There were two tutorial classes: one face-to-face class in the computing labs and one online class on Zoom. Attendance in both classes was around 60%–65%. Both classes were recorded, and video recordings were released to students. Specific tasks completed in the tutorial during Week 11 include: choosing relevant dependent and independent variables, cleaning data, transforming categorical data into dummy variables, partitioning data and preparing the final dataset, estimating a benchmark logistic regression model, and understanding the classification table and out-of-sample testing results from this benchmark regression. In addition, students were asked to vary the cut-off values and observe how changes in this cut-off value affected the classification outcomes.

Tasks completed in the Week-12 tutorial were designed to assist students in applying the systematic model-building approach (Figure 1) to make predictions using binary logistic regression. Specifically, instructors in this tutorial class asked students to: (1) vary the set of independent variables and re-run binary regression with these new sets of independent variables, and (2) compare these new models with the benchmark model estimated in the prior week. Instructors demonstrated one example of adding a new variable into the benchmark model. Attention was given to changes in the classification tables as well as the statistical significance of key independent variables. All these activities were illustrated in Excel.

8 | EVIDENCE OF STUDENT LEARNING AND IMPORTANT OBSERVATIONS

Our main quest in this article is to assess if binary logistic regression is effective in promoting higher-order thinking in students enrolled in our course. Most students were able to (1) transform original variables into dummy variables, (2) use logistic regression to model the dichotomous choice of the case study, and (3) interpret the general relationships implied by the logistic regression coefficients and their statistical significance. In particular, many students were able to (1) try to specify many explanatory variables in their modeling and (2) successfully utilize out-of-sample testing to validate and tune their models. These two specific areas are considered higher-order thinking learning outcomes in this course.

Instructor observations also noted that students attending lectures and tutorials were highly engaged with the task and took advantage of RegressItLogistic's interactivity to expand their knowledge and gain deeper insights into the determinants of consumer choice. To provide an example of how RegressItLogistic promoted higher-order thinking, students who ran a model that included all independent variables noticed that the model incorrectly classified all households as being purchasers of frozen goods. Our observations noted that this led to students engaging in extensive constructive dialogue on the following points:

1. Whether it is better to incorrectly classify a household as a purchaser when they would not purchase, or incorrectly classify a household as a non-purchaser when they would purchase.
2. Whether it would be more appropriate to adjust the cut-off rate to find an optimal trade-off between the accuracy, sensitivity, and specificity of the model.
3. What additional variables could potentially improve the predictive power of the model and the practicality of collecting such data.
4. Whether the purpose of modeling the problem is to correctly classify observations or draw inferences about the statistical significance of independent variables on consumer choice or a combination of the two.

Table 3 reports the results for each criterion of Assessment 3. The overall average (median) performance on Assessment 3 for the 54 students who submitted was 65.33% (65.5%), with a standard deviation of 9.29%, suggesting a credit average on the assessment (a credit grade is achieved for scores from 65% to below 75%). Focusing solely on the criteria most relevant to the use of RegressItLogistic, students did relatively well in utilizing the modeling capabilities of RegressItLogistic as assessed through Criterion 2, with an average (median) mark of 71.30% (73.33%) and a standard deviation of 12.70%. Interestingly, students achieved a pass average for the articulation of their modeling, model evaluation, and practical deployment (Criterion 4), with an average (median) mark of 64.24% (64.44%) and a standard deviation of 10.13%.

When interpreting these results, it is important to consider some of the challenges we encountered when implementing RegressItLogistic into classroom and assessment experiences. First, there were compatibility issues for some students using Mac computers. We were aware of this issue and asked students to use Windows computers to do the assessment. However, a small number of submissions did indicate in their report that they had difficulties in using Mac devices and that it was too late for them to access alternative computers. We suspect those students engaged with relevant materials close to the due date.

Students were able to effectively engage with the technical aspects of the task, with only a small proportion of students (5.6%) not using RegressItLogistic or the Data Analysis ToolPak to run linear or logistic regressions to address the research questions. Students were able to effectively develop an understanding of the business context and create justified hypotheses for the relationships between dependent and independent variables, which were then used effectively to justify variable selection. Both strengths are reflected in the results reported for Criteria 2 and 3 in Table 2.

The most significant error we observed in student work was the misinterpretation of logistic regression coefficients. All students who ran logistic regression were able to correctly interpret the sign of the coefficients to imply the direction of the relationship between the dependent and independent variables. However, only two students correctly interpreted the coefficient estimates as log-odds, with the remaining students interpreting logistic regression coefficients in a similar manner to linear regression coefficients. This suggests that instructors need to pay particular attention to how the interpretation of logistic regression coefficients is taught and to explicitly model how linear and logistic regression coefficients have different interpretations. Another significant weakness displayed by students was the evaluation of the practical significance of their results, with students not explicitly connecting their findings back to their expectations and value creation for the organization.

Although students were explicitly required to use multiple linear regression analysis to determine factors affecting household expenditure on frozen products, it was interesting to note that many students did not fully engage with multiple regression as part of the exercise at all. Almost all students (89%) ran some multiple linear regressions as part of their analysis, but few (around 39%) discussed their results, evaluated the model, or discussed the practical implications of their findings in detail. This had a significant negative impact on their mark for Criterion 4. In contrast, logistic

TABLE 3 Summary statistics for student performance on Assessment 3 by criterion (N = 54)

	Criterion				
	1	2	3	4	5
	Analysis using CRISP-DM framework (10 marks)	Data analysis using Excel (15 marks)	Translation of business questions into analytics problems (10 marks)	Articulation of modeling, evaluation and deployment (45 marks)	Professional written communication (20 marks)
Mean	64.26%	71.30%	64.35%	64.24%	64.35%
St. dev.	9.39%	12.70%	11.70%	10.13%	11.11%
Median	60.00%	73.33%	70.00%	64.44%	65.00%

regression analysis was used, discussed, and evaluated extensively by students to classify households as consumers of frozen yogurt, frozen meals, or both. Students also made extensive use of the diagnostic tools to validate and inform their modeling approach. We interpret this as a sign that students were engaged with logistic regression concepts.

There are several possible explanations for this. First, this could be due to differences in the motivation and perception of students regarding the assessment task, even though from our perspective, students were clearly informed that both linear and logistic regression were required in their assessment. The learning focus during class teaching and supporting resources was allocated equally to both modeling strategies; we also created a "Frequently Asked Questions" section that appeared at the top of the Assessment 3 folder on our course's blackboard site, clearly stating that both types of regressions were required to answer the two research questions. Blackboard analytics data showed that, on average, a student viewed this page more than nine times. Thus, despite our best efforts, it is possible that students could have perceived it differently since logistic regression was covered closer to the assessment due date.

Second, it could be that students in our cohort (primarily first-year undergraduate students) are more familiar with the idea of making dichotomous choices than they are with predicting a continuous variable because it is more relevant to their daily lives. Third, it could be that students see the modeling of dichotomous choices as more interesting and appealing. If these two reasons are true, the inclusion of logistic regression could be worthwhile in IBA courses because it encourages a connection between students and the assessment task. This promotes engagement by empowering students with the belief that they can tackle familiar real-world problems (Johnson & Delawsky, 2013). Indeed, informal discussion between students and instructors indicated that students found the concept of logistic regression analysis interesting due to its novelty and practical relevance.

Another important observation is related to those students who did not engage in classes. Online and face-to-face tutorial attendances were 60%–65% during the tutorials dedicated to logistic and multiple regression analysis, and all teaching classes were recorded and released to students on the Echo360 platform. However, only 7% of students viewed the recordings for multiple regression while 28% of students viewed the recordings for logistic regression. This suggests that slightly more than 25% of students were relying on self-directed learning with no engagement with the formal teaching materials. Due to the nature of learning topics, lecture slides and answer guides to tutorial questions did not contain detailed instruction on how to use RegressItLogistic. To support self-directed learners, we provided two YouTube video clips on simple and multiple regression, together with two video clips on logistic regression (one from the add-in publisher and one from the teaching team). Blackboard data analytics showed few views of these resources. However, we did not place these learning resources under the Assessment 3 folder but in a separate folder called "Excel Resources." We suspect this could be a reason why these resources were not used effectively by those students.

Another important observation drawn from the submitted work was that three students (5.6%) did not utilize RegressItLogistic or the Analysis ToolPak to address any of the assessment questions. Although we were not able to link those students with their engagement with classes and teaching materials, it is likely that those students were self-directed learners. This observation highlights challenges associated with the impact of COVID-19 and online learning on student engagement across all universities in our country. Challice et al. (2021) reported that undergraduate student ratings of the quality of their educational experience fell sharply from 78% in 2019 to 69% on average across universities in the country in 2020. Among the six different aspects of student experience surveyed, learner engagement had the biggest drop in overall student satisfaction, falling from 60% in 2019 to a record low of 44% in 2020. Hence, student engagement in learning requires urgent attention in higher-education institutions.

On a final note, an anonymous in-class poll in the last tutorial shows that 77% of students agreed or strongly agreed with the question "In terms of Assessment 3, do you agree that the assessment structure helps you apply the content to real-world problem solving?," providing further support for the learning benefits of RegressItLogistic.

9 | CONCLUSION

In this article, we justify the need to teach binary logistic regression for predictive purposes in IBA courses within undergraduate business curricula. Importantly, we provide our reflections on the design and implementation of several learning strategies and resources specific to binary logistic regression analysis.

We adopt a constructive-alignment approach to teaching and focus on developing a high level of critical thinking around how to use binary logistic regression for predictive purposes. In our design, logistic regression analysis was introduced after students had familiarity with Excel and had learned about using multiple linear regression to explain phenomena. We use several real-world examples to demonstrate how to implement binary logistic regression and classify binary outcomes in Excel using a freely available add-in (RegressItLogistic). These learning activities were conducted in both lectures and computer lab tutorials. We also required students to conduct binary logistic regression analysis in one assessment item, providing several scaffolded Excel-based activities around that task.

Using data from student engagement with learning materials and their assessment performance, we report three key observations. First, students appear to have more interest in learning logistic regression than multiple regression; hence, we argue that including logistic regression in IBA courses could provide more real-world and engaging learning experiences for students. Second, there are very few user-friendly and freely available add-ins that support logistic regression analysis in Excel. Third, students who did not engage with learning materials had trouble implementing binary logistic regression to complete the assessment. Together, these findings support the inclusion of logistic regression analysis into undergraduate business curricula and provide guidance on additional materials to support students, particularly those who show little engagement with learning activities and materials.

ACKNOWLEDGMENTS

Open access publishing facilitated by Queensland University of Technology, as part of the Wiley - Queensland University of Technology agreement via the Council of Australian University Librarians.

ORCID

Viet-Ngu Hoang https://orcid.org/0000-0002-9742-2378

REFERENCES

Abbasi, A., Sarker, S. & Chiang, R.H. (2016) Big data research in information systems: toward an inclusive research agenda. *Journal of the Association for Information Systems*, 17(2), 1–32.

Al-Haddad, S., Thorne, B., Ahmed, V. & Sause, W. (2019) Teaching information technology alongside business analytics: case study. *Journal of Education for Business*, 94(2), 92–100.

Barton, D. & Court, D. (2012) Making advanced analytics work for you. *Harvard Business Review*, 90(10), 78–83.

Biggs, J. (1996) Enhancing teaching through constructive alignment. *Higher Education*, 32(3), 347–364.

Braun, N.M. (2020) Critical thinking in the business curriculum. *Journal of Education for Business*, 79(4), 232–236.

Brusco, M. (2022) Logistic regression via Excel spreadsheets: mechanics, model selection, and relative predictor importance. *INFORMS Transactions on Education. Advance online publication*. https://doi.org/10.1287/ited.2021.0263

Calma, A. & Davies, M. (2021) Critical thinking in business education: current outlook and future prospects. *Studies in Higher Education*, 46(11), 2279–2295.

Camm, J.D., Cochran, J., Fry, M. & Ohlmann, J.W. (2020) *Business analytics*. Boston, MA: Cengage Learning.

Carver, R., Everson, M., Gabrosek, J., Horton, N., Lock, R., Mocko, M., Rossman, A., Roswell, G.H., Velleman, P. & Witmer, J. (2016) *Guidelines for assessment and instruction in statistics education (GAISE) college report 2016*. Accessed 4 August, 2022, Available at: https://commons.erau.edu/publication/1083

Cegielski, C.G. & Jones-Farmer, A. (2016) Knowledge, skills, and abilities for entry-level business analytics positions: a multi-method study. *Decision Sciences Journal of Innovative Education*, 14(1), 91–118.

Challice, G., Bolton, L., Spencer, L., Johnston, B., Kim, C., Smith, D., Tobias, P., Behr, K., Hoque, B., Feng, J. & Pool, K. (2021) 2020 Student experience survey. Accessed 4 August, 2022, Available at: https://www.qilt.edu.au/docs/default-source/resources/ses/2020/2020-ses-national-report.pdf?sfvrsn=d1237953_5

Davenport, T.H. (2006) Competing on analytics. *Harvard Business Review*, 84(1), 98–107.

Davenport, T.H. & Harris, J.G. (2007) *Competing on analytics, the new science of winning*. Boston, MA: Harvard Business School Press.

Davenport, T.H. & Patil, D.J. (2012) Data scientist. *Harvard Business Review*, 90(5), 70–76.

Dey, D. & Kumar, S. (2010) Reassessing data quality for information products. *Management Science*, 56(12), 2316–2322.

Dykes, B. (2016) Data storytelling: the essential data science skill everyone needs. Available at: https://www.forbes.com/sites/brentdykes/2016/03/31/datastorytelling-the-essential-data-science-skill-everyone-needs/ [Accessed 13th June 2021]

Evans, J.R. (2020) *Business analytics: global edition*. Harlow, UK: Pearson Education Limited.

Haskin, H.N. & Krehbiel, T.C. (2012) Business statistics at the top 50 US business programmes. *Teaching Statistics*, 34(3), 92–98.

Hazen, B.T., Skipper, J.B., Ezell, J.D. & Boone, C.A. (2016) Big data and predictive analytics for supply chain sustainability: a theory-driven research agenda. *Computers & Industrial Engineering*, 101(1), 592–598.

Helliwell, J.F., Huang, H., Shun, W. & Norton, M. (2020) *Social environments for world happiness. World Happiness Report 2020*. New York, NY: Sustainable Development Solutions Network.

Jaggia, S., Kelly, A., Lertwachara, K. & Chen, L. (2020) Applying the CRISP-DM framework for teaching business analytics. *Decision Sciences Journal of Innovative Education*, 18(4), 612–634.

Johnson, C.S. & Delawsky, S. (2013) Project-based learning and student engagement. *Academic Research International*, 4(4), 560–570.

Johnson, M.E. & Berenson, M. (2019) Choosing among computational software tools to enhance learning in introductory business statistics. *Decision Sciences Journal of Innovative Education*, 17(3), 214–238.

Johnson, M.E., Albizri, A. & Jain, R. (2020) Exploratory analysis to identify concepts, skills, knowledge, and tools to educate business analytics practitioners. *Decision Sciences Journal of Innovative Education*, 18(1), 90–118.

Johnson, M.E., Misra, R. & Berenson, M. (2022) Teaching Bayesian and Markov methods in business analytics curricula: an integrated approach. *Decision Sciences Journal of Innovative Education*, 20(1), 17–28.

Lawson, C. & Montgomery, D.C. (2006) Logistic regression analysis of customer satisfaction data. *Quality and Reliability Engineering International*, 22(8), 971–984.

Li, M., Mickel, A. & Taylor, S. (2018) Should this loan be approved or denied?": a large dataset with class assignment guidelines. *Journal of Statistics Education*, 26(1), 55–66.

Mamonov, S., Misra, R. & Jain, R. (2015) Business analytics in practice and in education: a competency-based perspective. *Information Systems Education Journal*, 13(1), 4–13.

McCann, M. (2017) Constructive alignment in economics teaching: a reflection on effective implementation. *Teaching in Higher Education*, 22(3), 336–348.

Morrell, C.H. & Auer, R. (2007) Trashball: a logistic regression classroom activity. *Journal of Statistics Education*, 15(1), 1–18.

Phelps, A.L. & Szabat, K.A. (2017) The current landscape of teaching analytics to business students at institutions of higher education: who is teaching what? *The American Statistician*, 71(2), 155–161.

Rienzo, T. & Chen, K. (2018) Planning for low end analytics disruptions in business school curricula. *Decision Sciences Journal of Innovative Education*, 16(1), 23–41.

Shmueli, G. (2010) To explain or to predict? *Statistical Science*, 25(3), 289–310.

Waller, M.A. & Fawcett, S.E. (2013) Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. *Journal of Business Logistics*, 34(2), 77–84.

Warner, B. & Meehan, A. (2001) Microsoft Excel™ as a tool for teaching basic statistics. *Teaching of Psychology*, 28(4), 295–298.

Wirth, R. & Hipp, J. (2000) CRISP-DM: towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 11–13 April, Manchester, UK. Menlo Park, CA: AAAI Press, pp. 29–39.

Yazici, H.J. (2020) Project-based learning for teaching business analytics in the undergraduate curriculum. *Decision Sciences Journal of Innovative Education*, 18(4), 589–611.

AUTHOR BIOGRAPHIES

Dr. Viet-Ngu (Vincent) Hoang is an Associate Professor of Economics at Queensland University of Technology (Faculty of Business and Law). He has developed and has been teaching applied business analytics and cost and benefit analysis. Previously, he taught other units in the field of applied economics and policy analysis and valuation. His main research interest lies in the areas of performance benchmarking, climate change adaption, sustainability, and efficiency and productivity analysis. He published more than 60 articles in these fields of research. Dr. Hoang also holds adjunct/fellow positions at International School (Vietnam National University, Hanoi) and IPAG Business School (Paris, France).

Dr. Justin Watson is a sessional academic at the Queensland University of Technology. He has taught extensively in introductory and advanced courses in applied finance, analytics, and statistics at undergraduate and post-graduate levels for the past 7 years. His current area of interest is developing appropriate pedagogical tools for the effective teaching of statistical concepts that promote positive attitudes toward statistics in secondary and tertiary contexts.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Hoang, V.-N. & Watson, J. (2022) Teaching binary logistic regression modeling in an introductory business analytics course. *Decision Sciences Journal of Innovative Education*, 20, 201–211. https://doi.org/10.1111/dsji.12274