Teaching Logistic Regression using Ordinary Least Squares in Excel

Milo Schield, Augsburg College. Minneapolis, MN.

Abstract: The results of a logistic regression are reported in news stories and journal articles. Logistic regression is a useful tool in modeling data with a binary outcome. If we want students appreciate the value of statistics, we should show them the many tools statistics provides. Yet, logistic regression is seldom – if ever – a part of the intro statistics course. In teaching business statistics, where 70% of classes are taught using Excel, the lack of an Excel Logistic Regression command may seem like a sufficient reason. This paper first reviews how Excel Solver can do multivariate logistic regression using MLE but notes that the process is complicated, time-consuming and non-informative. Although modeling binary outcomes using OLS is not justified theoretically, this paper argues that in many cases the difference is not material. Binary outcomes can be modeled efficiently and effectively using ordinary least squares regression in Excel in three ways. (1) The simplest is to use linear OLS to fit binary outcomes. This approach is quick and simple, but it is limited to those cases where the predicted probabilities of zero and one occur well outside the range of interest. (2) Use OLS to fit a logistic function to grouped data. Grouped data can avoid zeroes and ones that would create infinities in the Log[Odds(pGroup)]. This grouped-data approach introduces students to the logistic function and the need to be careful in interpreting the regression coefficient. Unfortunately, obtaining suitable grouped probabilities may be impossible in multivariate analysis with small samples. (3) This paper introduces the idea of using OLS on the log(odds) of 'nudged data'. Nudging involves replacing the binary values of zero and one with epsilon and one minus epsilon respectively. This new approach gives generally good results for bivariate and multivariate regression. The differences are not generally material in showing students how confounding can influence an association. MLE and Logit-OLS-Nudge are compared in handling a Simpson's reversal with two continuous predictors. This logistic-OLS-nudge approach allows statistical educators to show students how controlling for a confounder can influence – even reverse – an association. Showing this upholds the goal of the 2016 update to the GAISE guidelines: "to give students experience with multivariable thinking" so that they "learn to consider potential confounding factors."

Keywords: Statistical literacy, GAISE 2016, statistical education, multivariate methods

1. The Need and the Problem

Many regression models involve a binary outcome: customer (sell vs. pass), loan recipient (payoff vs. default), medical condition (yes or no) or medical outcome (live or die). These models typically model the association between predictor and binary outcome with a logistic function. Hence they are described as logistic regression.

The results of these logistic models are presented in journal articles and in the news media. They typically involve odds instead of probabilities and the slopes are not linear. These models are common, but are seldom shown in an introductory statistics course. Moore et al (2003a, b) is an exception.

Logistic regression is important. Witmer (2010) presented logistic regression as one of the four key statistical tools for a second course in statistical modelling.

	Predictor(s)	
Outcome	Quantitative	Categorical
Quantitative	1. Regression	2. ANOVA
Categorical	3. Logistic Regress.	4. Chi-square

Figure 1: Classifying Models by Variables Involved

The GAISE (2016) guideline encourages statistical educators to use multivariable data and to show the influence of confounders on an observed association. Introducing logistic regression opens the door to other multivariate ideas such as discriminant analysis and it introduces a tool that users in other disciplines need to use. If students fail to see value in statistics, it may be that they aren't given tools that they will use.

A big reason for the omission of logistic regression (aside from the lack of time) is that it doesn’t use ordinary-least-squares (OLS). With a binary outcome, OLS is not justified theoretically: errors are not normally distributed, and the variance in the errors is not constant over the range of predictor values. And the predicted values of a linear model can go outside the relevant range for probabilities. The left side of Figure 2 illustrates those results.

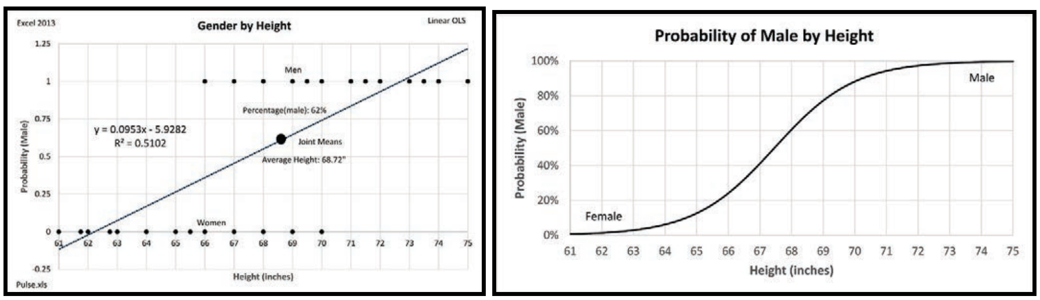


Figure 2: Modelling a binary outcome using linear (left) and logistic (right) models

These two problems (out of range and assumptions) have solutions.

(1) Solving the out-of-range problem. We want a function that respects the lower and upper limits of zero and one, and that gives probabilities for the full range of X values. The right side of Figure 2 illustrates the shape of such a function.

Logistic regression uses an S-shaped curve. Although there are several S-shaped curves, logistic regression involves a particular choice: the log of the odds. Using odds converts the range of a probability from [0, 1] to [0, ∞]. Using Ln(Odds) converts the range of the odds from [0, ∞] to [-∞, +∞].

(2) Solving the assumption violations problem. Maximum Likelihood Estimation (MLE) avoids the assumptions involved in using ordinary least squares (OLS).

Two problems remain: there is no analytic formula for the MLE solution, and Excel has no command to perform a Logistic Regression using MLE.

2. Teaching Logistic Regression

Given the importance of logistic regression and the fact that most students will never take Stat 200, it follows that it should be included – if possible – in an expanded first course.

One way is to use statistical software to generate a logistic function using MLE. Morrell and Auer (2004) showed how they used logistic regression in the classroom to model a student activity they called 'trash-ball'. They used Minitab to obtain the MLE solution.

Since 70% of faculty use Excel in teaching intro business statistics, logistic regression must be teachable using Excel. Excel can be used to generate MLE. Appendix B shows how this can be done. Figure 3 shows the result of modelling gender based on height for college students

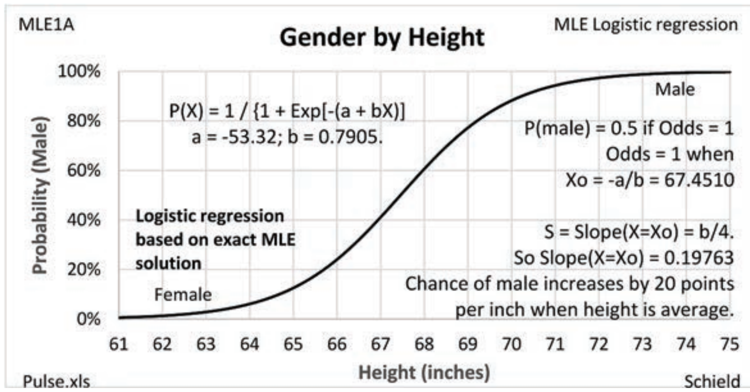


Figure 3: Excel MLE: Final Result Graph

Appendix B show cases where the results of using Excel Solver agree with those using Minitab. Unfortunately using Excel Solver involves so many manual operations that even skilled Excel users lose track of what is happening. Yes, this complexity could be minimized using macros or VBA in Excel, but using these opens up the possibility of a virus and the difficulty of providing student access to these modules.

A third way is to use ordinary least-squares regression (OLSR) in Excel. The phrase "logistic regression using OLS" seems like an oxymoron since OLS is never justified in logistic regression. This paper argues that in many cases the difference is not material and teaching it is justified pedagogically. This paper considers three ways to teach logistic regression using Ordinary Least Squares regression in Excel.

3. Teaching Logistic Regression Using Linear OLS on Individual data

The first way to teach logistic regression using OLS is the simplest: use a linear model. This approach requires no additional complexity; the coefficients are readily understood. Compare the linear OLS results with Logistic MLE in data with increasing correlation.

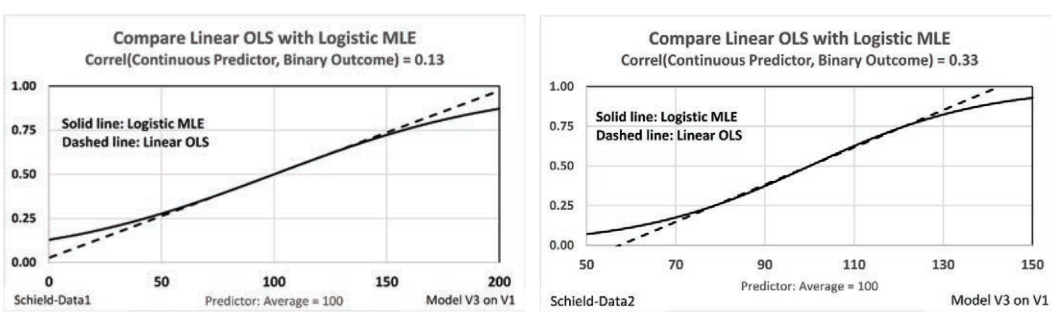


Figure 4: Linear OLS vs. Logistic MLE. Correlation = 0.1 and 0.3

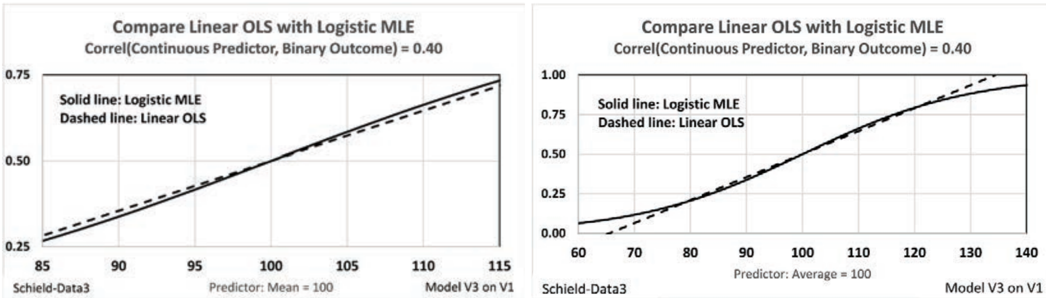


Figure 5: Linear OLS vs. Logistic MLE. Correlation = 0.4

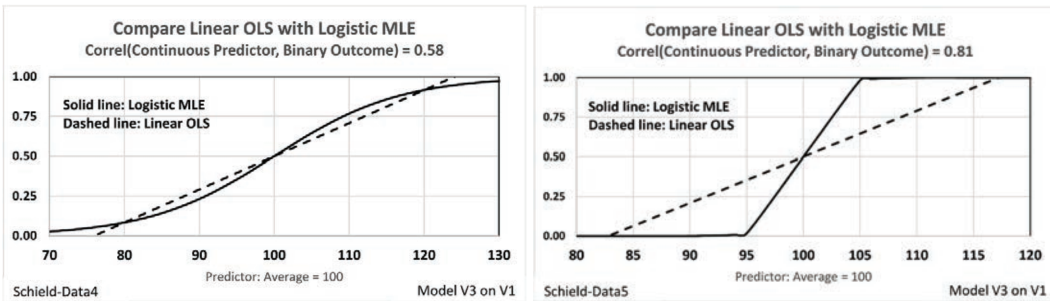


Figure 6: Linear OLS vs. Logistic MLE. Correlation = 0.6 and 0.8

The results of the linear-OLS model and the logistic-MLE model diverge as the predicted result approaches the extremes (zero or one) and as the correlation increases. Schield (2017) contains the associated data, models and graphs.

When the data size is in the hundreds, here are two rules of thumb. Using the linear-OLS model for binary outcomes is generally OK:

- if the predictor range of interest is inside the predictor range associated with the inter-quartile range of the predicted outcome (This guarantees that the predicted probabilities of zero and one have predictor values that are well outside the predictor range of interest.) AND
- if the predictor-outcome correlation is less than 0.4. It is not OK if that correlation is greater than 0.7.

Obviously these rules of thumb depend on the gravity of the decision involved. Using linear-OLS to decide on an advertising campaign is much different from using linear-OLS to approve a new drug.

4. Teaching Logistic Regression Using Logistic-OLS on Grouped Data

A second way to teach logistic regression using Ordinary Least Squares is to fit a logistic function to grouped data.

These groups may have been used in collecting the data or they may arise after collecting the data by grouping continuous predictor values into bins. The goal is to find groups such that the grouped probabilities are never zero or one. Such a grouping eliminates the infinities generated by Odds(P) and Ln(Odds(P)). The log-odd for the various group averages can then be modeled using a linear OLS regression. See Trinfade (1997), Haggstrom (1982), Peck and Devore (2002) and Winner (2017).

Lowry (2017) argues that this grouped-data approach is adequate for two reasons.

> "the first reason, which can be counted as either a high-minded philosophical reservation or a low-minded personal quirk, is that the maximized log likelihood method [MLE] has always impressed me as an exercise in excessive fine-tuning, reminiscent on some occasions of what Alfred North Whitehead identified as *the fallacy of misplaced concreteness*, and on others of what Freud described as *the narcissism of small differences*. The second reason is that in most real-world cases there is little if any practical difference between the results of the two methods [MLE vs OLS of Ln(Odds(Pgrouped))]."

Here are some examples of the Logistic-OLS-Group approach with sample size of 300.

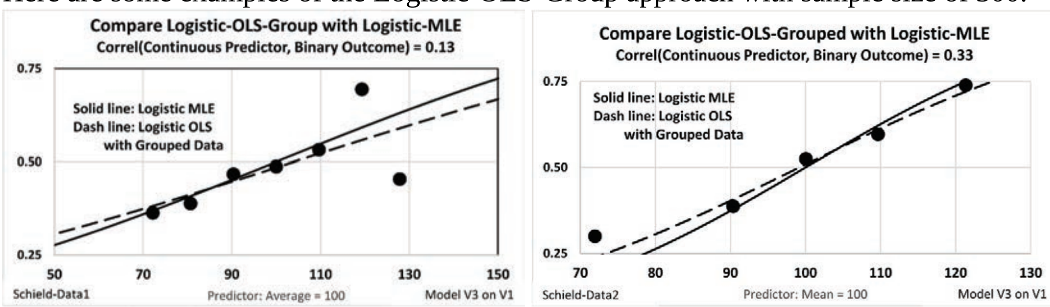


Figure 7: Logistic-OLS-Grouped vs. Logistic-MLE. Correlation = 0.1 and 0.3

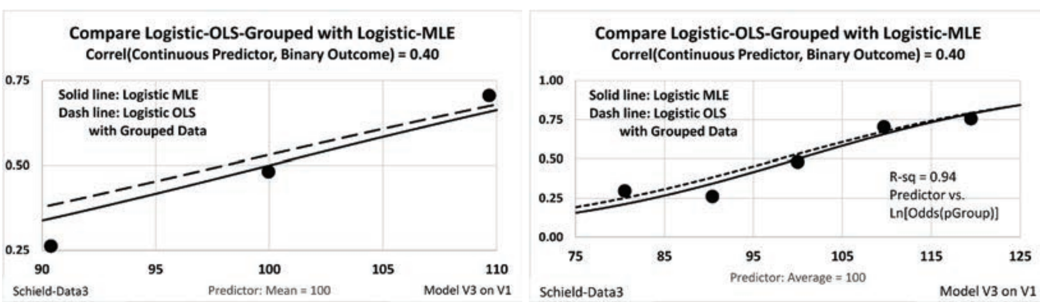


Figure 8: Logistic-OLS-Grouped vs. Logistic-MLE. Correlation = 0.4

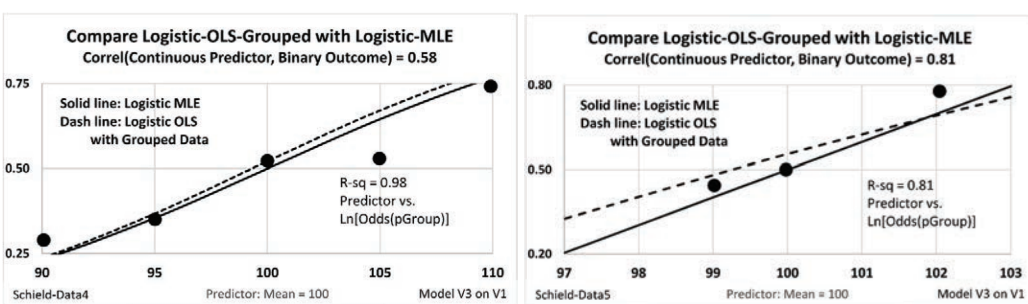


Figure 9: Logistic-OLS-Grouped vs. Logistic-MLE. Correlation = 0.6 and 0.8

Schield (2017) contains the associated data, the models and the graphs.

A big advantage of this OLS-grouped approach is simplicity. Students can see the linear fit in Log-Odds space. They can see how a linear model in Ln-Odds space generates a logistic function in probability space.

There are several difficulties with this OLS-grouped approach.

- Ordinary least squares regression (OLSR) requires that the predictor values involve ratio data. Thus the labels for the predictor groups must preserve the ratio nature of the underlying measurements. This typically requires groups or bins of equal size. If there are outliers in the predictors, the need for equal-sized bins may result in problem bins: empty bins or bins where the grouped probability is either zero or one.
- Avoiding these two problems (empty bins or bins where the average probability is either zero or one) may require increasing or adjusting the bin width. This introduces an additional source of variability. In the worst case, it requires reducing the number of bins, thereby decreasing the accuracy of the model.
- Even if a logistic regression with a single predictor has no problem bins (no empty bins and no bins with average probabilities of zero or one) with bivariate data, problem bins may easily arise when controlling for a confounder forces the data to be multivariate. Thus it may be generally impossible to show confounder influence using OLS on grouped data.

5. Teaching Logistic Regression Using Logistic-OLS on Individual Data

A third way to teach logistic regression using a logistic function with Ordinary Least Squares (OLS) on individual data is to 'nudge' the binary outcomes.

The idea is very simple. Replace binary zero and one with epsilon and one minus epsilon respectively where epsilon is a small positive fraction. This paper uses 0.001. The choice of epsilon can influence the shape of the resulting logistic function. Analyzing this sensitivity is beyond the scope of this paper.

The technical details of this approach are contained in appendices.

Appendix A: Pulse Data
Appendix B: Logistic Regression for Bivariate Data using Excel Solver MLE
Appendix C: Minitab Results using MLE Logistic Regression
Appendix D: Generating Confidence Intervals for Logistic Regression
Appendix E: Slope Statistically Significant
Appendix F: Using Logistic-OLS with Nudge in Excel
Appendix G: Simpson's Paradox with a Continuous Predictor
Appendix H: Using Logistic+OLS+Nudge to show a Simpson's Reversal
Appendix I: Data2 Minitab MLE Before and After Confounding

Schield (2017) contains the data, models and graphs used in Appendix H

6. Compare MLE with Logit-OLS-Nudge using Pulse Data

Figure 10 shows logistic regression of Gender on Height using MLE and OLS with nudge.

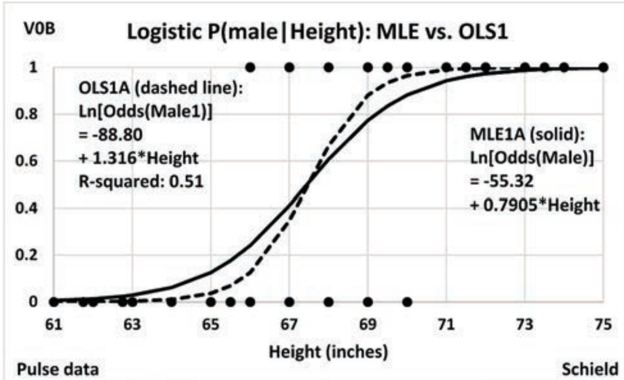


Figure 10: Logistic Regression of Gender by Height: MLE vs OLS with Nudge

Figure 11 shows logistic regression of Gender by Weight using MLE and OLS with nudge.

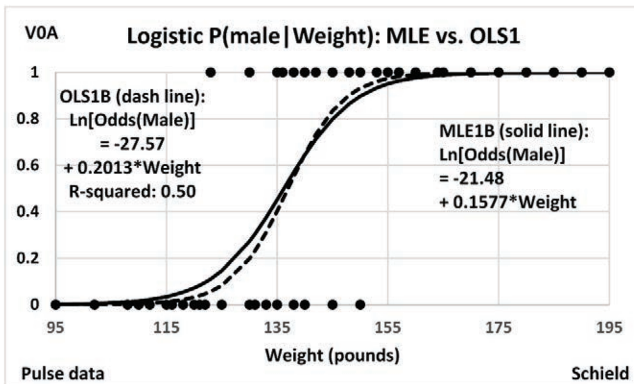


Figure 11: Logistic Regression of Gender by Weight: MLE vs OLS with Nudge

Figure 12 shows logistic regression of Gender by Height & Weight using MLE and OLS..

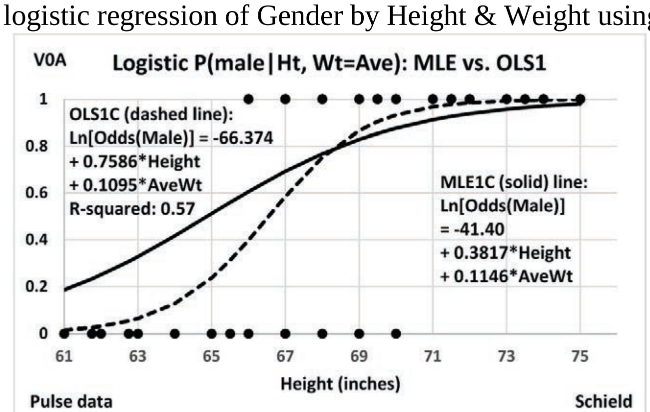


Figure 12: Logistic Regression of Gender by Height & Weight: MLE vs OLS with Nudge

7. Are these Differences Significant?

Are the differences between the MLE models and the Logistic-OLS-nudge models statistically significant? One way to tell is to generate the confidence interval around the MLE line. If it includes the OLS line, then the difference is not statistically significant.

Figure 13 shows both models and the 95% confidence interval around the MLE model when performing a logistic regression of gender on height.

Shows

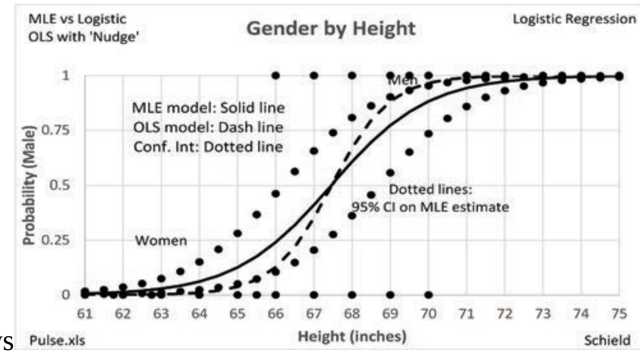


Figure 13: LR of Gender by Height: MLE vs OLS with MLE Confidence Intervals

Figure 14 shows both models and the 95% confidence interval around the MLE model when performing a logistic regression of gender on weight.

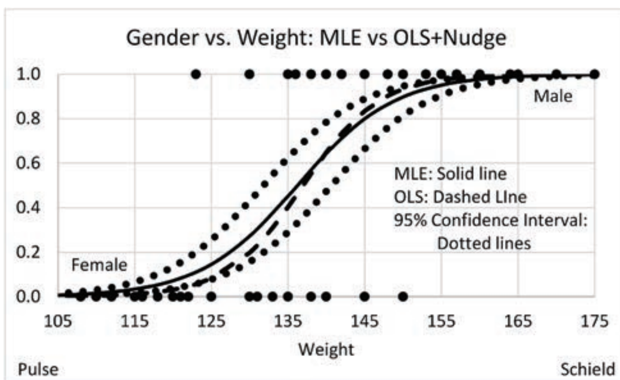


Figure 14: Logistic Regression of Gender by Height & Weight: MLE vs OLS with Nudge

In both of these cases, the OLS logistic-nudge model is generally – but not totally – within the 95% confidence interval around the MLE model.

Second, is the difference between MLE and OLS practically significant? In both cases, the models seem to coincide at the 50% level. The biggest difference is the slope. The MLE model seems flatter than the OLS model. Thus, the OLS model overstates at the higher percentages and understates at the lower percentages.

Just because this Logistic-OLS with nudge approach seems to work for bivariate data is no basis for claiming that it will work with multivariate data. The next section investigates the evidence.

8. Simpson's Reversal using Continuous Predictors and a Binary Outcome

Appendix G discusses the two data sets created to show a Simpson's reversal when the predictor and confounder are both continuous.

Figure 17 shows the results of logistic regression on Data1 before and after controlling for a continuous confounder. The left side uses MLE; the right side uses OLS+Nudge

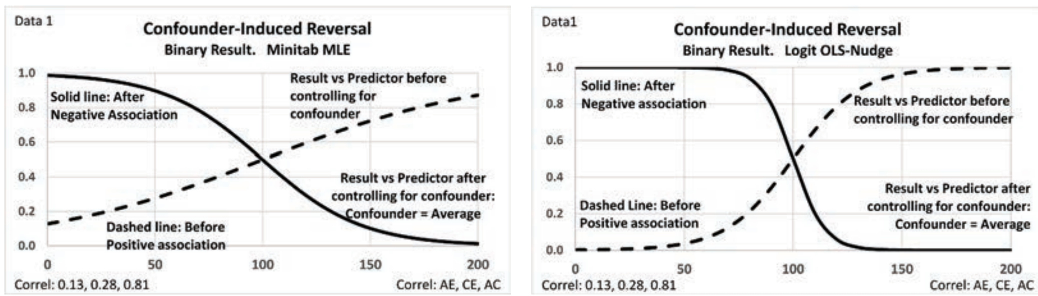


Figure 15: Data1: MLE (left) and OLS+Nudge (right)

Before controlling for the confounder, the slope of the result is positive for both MLE and OLS+Nudge. After controlling for the confounder, the slope is negative for both MLE and OLS+Nudge.

Figure 17 shows the results of logistic regression on Data2 before and after controlling for a continuous confounder. The left side uses MLE; the right side uses OLS+Nudge on the same data.

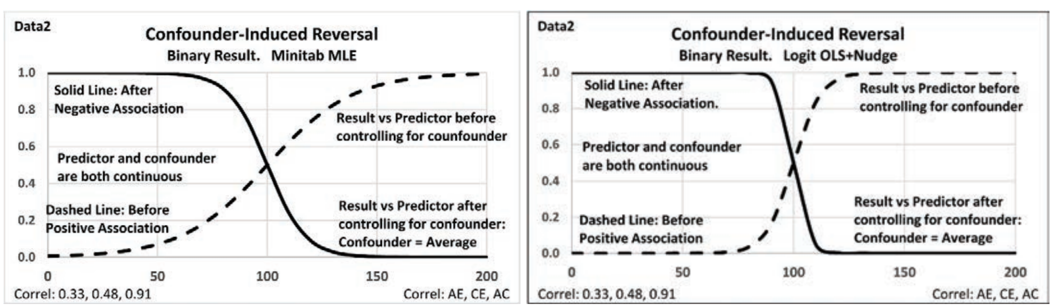


Figure 16: Data #2: MLE (left) and OLS+Nudge (right)

We see the same results. Both methods show the reversal of the association after controlling for the confounder. And in both cases the slope is steeper in the OLS+Nudge than in MLE. But if the goal is teaching, then the OLS-nudge approach clearly satisfies the test of showing the sign reversal.

9. Interpreting the Coefficients and Fit in a Logistic Regression

Interpreting the coefficients for a logistic function is important but not easy. See Miller (2005) for a most excellent discussion. Interpreting the measures of fit is equally important and equally difficult. See Peng et al (2002), Allison (2017), Campbell (1998 and 2006) and Gelder (2012). See Duke (2017) for a discussion of logistic growth.

10. Future Work

Here are some ideas for future work:

- Analyze the difference between the MLE results and the Logit-OLS-nudge results (1) as a function of epsilon, and (2) as a function of the initial correlation.
- See if Logit-OLS can be used to estimate the location and slope of the logistic function at P50 (Odds = 1) which in turn determines the entire logistic function.
- See if this OLS regression can give some estimate of the confidence interval involved by estimating the size (MEo) and location (Xo) of the minimum margin of error and then using these two value to add an error term into the logistic function, ± MEo * [1 + (X-Xo)^2], to generate the models 95% confidence interval.
- Show how this OLS introduction to logistic regression could introduce students (1) to alternative methods of measuring the quality of a model and (2) to other advanced multivariate topics such as discriminant and classification analysis.

11. Conclusion

This paper shows that there several ways to model binary outcomes using Ordinary Least Squares in Excel. More specifically, a pedagogically decent logistic function can be obtained using ordinary least squares regression when the binary outcomes are nudged by a small fraction. Although there can be differences between the logistic-OLS-nudge model and the MLE model, the former seems to be adequate to illustrate logistic regression and the influence of a confounder for teaching purposes. It allows teachers to show a Simpson's reversal using a continuous predictor and a continuous confounder. Teaching logistic regression shows students another tool that statistics provides for analyzing data. Finally, it opens the door to more advanced multivariate topics.

Acknowledgements

To Conrad Carlberg (2012) for showing how to use Excel to do multivariate logistic regression by using Solver for Maximum Likelihood Estimation (MLE). To Cheryl Pammer (2017) at Minitab for giving me the exact Minitab session commands needed to calculate the standard error for the confidence intervals. To Richard Lowry (2017) for noting the lack of practical significance between MLE and OLS with grouped data.

REFERENCES

Allison, Paul (9999). Measures of Fit for Logistic Regression. https://support.sas.com/resources/papers/proceedings14/1485-2014.pdf

Campbell, Michael (1998). Teaching Logistic Regression. ICOTS-5. P. 284-289. https://iase-web.org/documents/papers/icots5/Topic2y.pdf

Campbell, Michael (2006). Teaching Non-Parametric Statistics to Students in Health Sciences. ICOTS-7. https://iase-web.org/documents/papers/icots5/Topic2y.pdf

Carlberg, Conrad (2012). Decision Analytics: Microsoft Excel. Que Publishing.

Duke, Math (2017). Logistic Growth Model. Background: Logistic Modelling. https://services.math.duke.edu/education/ccp/materials/diffeq/logistic/logi1.html

GAISE (2016). Guidelines for Assessment and Instruction in Statistics Education (GAISE) Reports. www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege_Full.pdf

Gelder, Alan B. (2012). Logistic Regression. University of Iowa. http://agelder.weebly.com/uploads/2/3/2/5/23257230/logistic_regression.pdf

Haggstrom, G.W (1982). Logistic Regression and Discriminant Analysis by Ordinary Least Squares. www.dtic.mil/get-tr-doc/pdf?AD=ADA125708

Hardin, (2017). A Statistics course in Logistic Regression taught by James Hardin https://www.statistics.com/logistic-regression/

Harmen, Mark (2017a). Excel Master Series: MBA-Level Statistical Instruction. http://excelmasterseries.com/

Harmen, Mark (2017b). Logistic Regression in 7 Steps in Excel 2010 and Excel 2013. http://excelmasterseries.com/Excel_Statistical_Master/Excel-Logistic-Regression.php

Howell, David (2002). Logistic Regression using SPSS. https://www.uvm.edu/~dhowell/gradstat/psych341/lectures/Logistic%20Regression/LogisticReg1.html

Isaacson, Marc (2008). Using Simulated Surveys to Teach Statistics: A Preliminary Report. ASA Proceedings of the Section on Statistical Education. P. 3124-3130. Copy at http://www.statlit.org/pdf/2008IsaacsonASA.pdf

Lani, James (2017). Logistic Regression. Assumptions, Key Terms and Resources. http://www.statisticssolutions.com/regression-analysis-logistic-regression/

Lowry, Richard (2017). Simple Logistic Regression. http://vassarstats.net/logreg1.html

Miller, Jane (2005). *The Chicago Guide to Writing about Multivariate Analysis*. University of Chicago Press.

Minitab (2017a). Graphs for Simple Binary Logistic Regression. http://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/simple-binary-logistic-regression/interpret-the-results/all-statistics-and-graphs/graphs/

Minitab (2017b). Interpret all statistics for Predict for Binary Logistic. Standard Error (SE) Fit. http://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/predict-for-binary-logistic/interpret-the-results/all-statistics/

Minitab (2017c). Example of Logistic Regression. Odds Ratio of Continuous Predictor http://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/binary-logistic-regression/before-you-start/example/

Minitab (2017d). Find a confidence interval and a prediction interval for the response https://onlinecourses.science.psu.edu/stat501/node/120

Morrell and Auer (2004). Trashball: A Logistic Regression Classroom Activity. Proceedings of the Section on Statistical Education. Copy at www.statlit.org/pdf/2004MorrellAuerASA.pdf

Moore, David S., George P. McCabe, William M. Duckworth, Stanley L. Sclov (2003a). The Practice of Business Statistics. The Companion Volume 17. https://books.google.com/books?id=RG5QAQAAIAAJ

Moore, David S., George P. McCabe, William M. Duckworth, Stanley L. Sclov (2003b). Ch 17: Logistic Regression in *The Practice of Business Statistics: Excel manual.* https://books.google.com/books/about/The_Practice_of_Business_Statistics_Exce.html?id=ZGVMI9xc_dAC

PASSS (2017). Simple Logistic Regression: One Continuous Independent Variable. *Practical Applications of Statistics in the Social Sciences* (PASSS).

www.southampton.ac.uk/passs/full_time_education/multivariate_analysis/simple_logistic_regression_continuous.page Accessed 9/2017.

Peck, R. and J. Devore (2002). Ch 5.5 Logistic Regression in *Exploration & Analysis of Data*. www.cengage.com/c/statistics-the-exploration-analysis-of-data-7e-peck

Phelps, Amy and Kathryn Szabat (2015). The Current Landscape of Business Analytics and Data Science at Higher Education Institutions: Who is Teaching What? http://www.statlit.org/pdf/2015-Phelps-Szabat-ASA-Slides.pdf

Ritter, Brent (2017). YouTube video: How to do Logistic Regression in Excel. https://www.youtube.com/watch?v=rbKtZcrTlr8

Trinfade, David (1997). Regression Models for Binary Response Using Excel and JMP. Statistical Methods Symposium. www.trindade.com/Logistic%20Regression.pdf

Schield, Milo (2016a). Offering STAT 102: Social Statistics for Decision Makers. Invited presentation: IASE Roundtable, Berlin. www.statlit.org/pdf/2016-Schield-IASE.pdf

Schield, Milo (2017). Modeling Binary Outcomes using Ordinary Least Squares in Excel: The Data. www.statlit.org/Excel/2017-Schield-ASA.xlsx

Chao-Ying Joanne Peng, Tak-Shing Harry So, Frances K. Stage and Edward P. St. John (2002). The Use and Interpretation of Logistic Regression in Higher Education Journals: 1988–1999. *Research in Higher Education.* Vol 43, Issue 3, pp 259–293. Available for purchase at https://link.springer.com/article/10.1023/A:1014858517172

Wasserman, Larry (2013). Simpson's Paradox Explained. Normal Deviate at https://normaldeviate.wordpress.com/2013/06/20/simpsons-paradox-explained/

Winner, Lawerence Hummer (2017). Logistic Regression with “Grouped” Data. URL Lobster_Logit.xlsx Accessed 9/2017.

Witmer, Jeffrey (2010). Stats: An Applied Statistics Modelling Course. ICOTS 5. http://icots.info/8/cd/pdfs/invited/ICOTS8_4D4_WITMER.pdf

Zaiontz, Charles (2017). Logistic Regression Using Excel: Real Statistics Resource Pack www.youtube.com/watch?v=EKRjDurXau0 Website: www.real-statistics.com

SCHIELD EXCEL DEMOS

Schield, Milo (2015a). Logistic Regression using MLE in Excel 2013: Case 1A: Gender vs. Height. www.statlit.org/pdf/2015-Schield-Logistic-MLE1A-Excel2013-Slides.pdf

Schield, Milo (2015b). Logistic Regression using MLE in Excel 2013: Case 1C: Gender vs. Height & Weight. Copy at http://www.statlit.org/pdf/2015-Schield-Logistic-MLE1C-Excel2013-Slides.pdf

Schield, Milo (2015c). Logistic Regression using OLS with 'Nudge' in Excel 2013: Case 1A: Gender vs. Height. http://www.statlit.org/pdf/2015-Schield-Logistic-OLS1A-Excel2013-Slides.pdf

Schield, Milo (2015d). Logistic Regression using OLS with 'Nudge' in Excel 2013: Case 1B: Gender vs. Smoking and Height. http://www.statlit.org/pdf/2015-Schield-Logistic-OLS1B-Excel2013-Slides.pdf

Schield, Milo (2015e). Logistic Regression using OLS with 'Nudge' in Excel 2013: Case 1C: Gender vs. Weight. http://www.statlit.org/pdf/2015-Schield-Logistic-OLS1C-Excel2013-Slides.pdf

Schield, Milo (2015f). Logistic Regression using OLS with 'Nudge' in Excel 2013: Case 1D: Gender vs Height and Weight. http://www.statlit.org/pdf/2015-Schield-Logistic-OLS1D-Excel2013-Slides.pdf

Schield, Milo (2015g). Logistic Regression: Comparing MLE with nudged OLS in Excel 2013. www.statlit.org/pdf/2015-Schield-Logistic-MLE-OLS1-Excel2013-Slides.pdf

Schield, Milo (2016b). Logistic Regression using Minitab and Pulse dataset. www.statlit.org/pdf/2016-Minitab-MLE1-Test1.pdf

Appendix A: Pulse Data

The results in Appendices B-G are obtained from a single Minitab dataset called PULSE. It is used in this paper with permission by Minitab. The data was obtained from 92 college students on spring break in Florida in the 1970s. Figure 17 presents the column headings and some data.

Pulse1	Pulse2	Height	Weight	Activity	Run	Smokes	Male
48	54	68	150	1	0	1	1
54	56	69	145	2	0	1	1
54	50	69	160	2	0	0	1
58	70	72	145	2	1	0	1
58	58	66	135	3	0	0	1
58	56	67	125	2	0	0	0
60	76	71	170	3	1	0	1
60	62	71	155	2	0	0	1
60	70	71.5	164	2	0	1	1
60	66	62	120	2	0	0	0

Figure 17 Pulse Data Sample

Pulse1 is a resting pulse. Pulse2 is a second measure of pulse. For those with Run=0, this a repeated measure of rest pulse. For those with Run=1, this second pulse is taken after a brief run in place. The assignment of students to run groups was done randomly.

Figure 18 shows the linear correlation coefficients between all eight variables. Note that Male has a slightly higher linear correlation with height (0.714) than with weight (0.709).

	Pulse1	*Pulse2*	*Height*	*Weight*	*Activity*	*Run*	*Smokes*	*Male*
Pulse1	1							
Pulse2	0.616	1						
Height	-0.212	-0.143	1					
Weight	-0.202	-0.169	0.785	1				
Activity	-0.109	-0.178	0.079	0.004	1			
Run	0.052	0.577	0.228	0.218	-0.007	1		
Smokes	0.129	0.046	0.056	0.200	-0.098	0.066	1	
Male	-0.285	-0.309	0.714	0.709	0.126	0.107	0.129	1

Figure 18 Pulse Data Correlation Matrix

Note the high correlation (0.785) between height and weight among these students. Schield (2016a) argues that using 2/sqrt(n) is sufficient for statistical significance. With n = 92, any randomly obtained |correlation| greater than 0.21 is statistically significant.

Figure 19 shows the ordinal averages and the binary percentages.

	Activity	Run	Smokes	Male
Average	2.120	38.04%	30.43%	61.96%

Figure 19 Pulse Data: Averages and Percentages of Categorical Variables

Appendix B: Logistic Regression for Bivariate Data using Excel Solver MLE

There are two distinct ways of performing logistic regression using Excel and Maximum Likelihood Estimation (MLE).

One is to use an Excel Add-In. See Zaiontz (2017) for a Youtube video on using the Real Statistics add in.

A second way is to use Excel's Solver. See the Ritter (2017) and Harmon (2017a, b) videos. Harmon (2002) and Carlberg (2012 Ch 2, pages 21-52) show step-by-step how to use Excel Solver to perform maximum likelihood estimation (MLE) and generate the best logistic regression.

The demonstration in this paper starts with two variables: height and gender. Gender is coded with 1 for male and 0 for female. Using zero-one coding means that the average of this variable gives the percentage who are male.

Figure 20 shows the initial model. It uses the Ln(Odds) of the average value of Y as the starting point. This choice eliminates some steps. Copying the value from E22 to the clipboard and then doing a special-paste of the value to D3 is the first manual step.

19	D	E	F
20	GENDER & INTERCEPT #1		
21	Male-Pctg	0.62	=AVERAGE(B3:B94)
22	Intercept#1	0.4877	=LN(E21/(1-E21))

Row	D	E
2	Intercept	Slope
3	0.4877	0.0000

Figure 20: Excel MLE: Initial Value

Figure 21 illustrates the worksheet setup with the functions in the first row: G3:K3.

Row	D	E	F	G	H	I	J	K
2	Intercept	Slope		Logit	Odds	Prob Y=1	Prob OK	Ln-LH-OK
3	0.4877	0.0000		0.49	1.63	0.62	0.62	-0.48
4								
5	Sum LnLk	-0.48	=SUM(K3:K94)					
6	Sum LnLk1		Sum #1: Manual					
7	Sum Ln Lk2		Sum #2: Solver MLE					
8	Chi-Sq	0.00	=-2*(E6-E7)					
9	P-Value	1	=CHISQ.DIST.RT(E8,1)					
10								
11	FORMULAS & TEXT: Enter, Copy Down							
12	Logit	G3	=D$3+E$3*A3					
13	Odds	H3	=EXP(G3)					
14	Prob Y=1	I3	=H3/(1+H3)					
15	Prob OK	J3	=IF(B3=1,I3,1-I3)					
16	Ln-LH-OK	K3	=LN(J3)					

Figure 21: Logistic Regression Excel MLE Setup

Figure 22 shows the full worksheet with the functions and the new result in cell E5.

Row	D	E	F	G	H	I	J	K
2	Intercept	Slope		Logit	Odds	Prob Y=1	Prob OK	Ln-LH-OK
3	0.4877	0.0000		0.49	1.63	0.62	0.62	-0.48
4				0.49	1.63	0.62	0.62	-0.48
5	Sum LnLk	-61.11	=SUM(K3:K94)	0.49	1.63	0.62	0.62	-0.48
6	Sum LnLk1		Sum #1: Manual	0.49	1.63	0.62	0.62	-0.48
7	Sum Ln Lk2		Sum #2: Solver MLE	0.49	1.63	0.62	0.62	-0.48
8	Chi-Sq	0.00	=-2*(E6-E7)	0.49	1.63	0.62	0.38	-0.97
9	P-Value	1	=CHISQ.DIST.RT(E8,1)	0.49	1.63	0.62	0.62	-0.48
10				0.49	1.63	0.62	0.62	-0.48
11	FORMULAS & TEXT: Enter, Copy Down			0.49	1.63	0.62	0.62	-0.48
12	Logit	G3	=D$3+E$3*A3	0.49	1.63	0.62	0.38	-0.97
13	Odds	H3	=EXP(G3)	0.49	1.63	0.62	0.38	-0.97
14	Prob Y=1	I3	=H3/(1+H3)	0.49	1.63	0.62	0.62	-0.48
15	Prob OK	J3	=IF(B3=1,I3,1-I3)	0.49	1.63	0.62	0.62	-0.48
16	Ln-LH-OK	K3	=LN(J3)	0.49	1.63	0.62	0.62	-0.48

Figure 22: Logistic Regression Excel MLE: Full spreadsheet

The second manual step is to copy the sum of the log-link from E5 and paste special the value into E6. Figure 23 shows the results of the first solver iteration.

Row	D	E	F
2	Intercept	Slope	
3	0.4877	0.0000	
4			
5	Sum LnLk	-61.11	=SUM(K3:K94)
6	Sum LnLk1	-61.11	Sum #1: Manual
7	Sum Ln Lk2		Sum #2: Solver MLE
8	Chi-Sq	122.23	=-2*(E6-E7)
9	P-Value	2E-28	=CHISQ.DIST.RT(E8,1)

Figure 23: Logistic Regression Excel MLE: First Iteration

The next step is to use Solver. Figure 24 illustrates the startup of Excel's Solver.

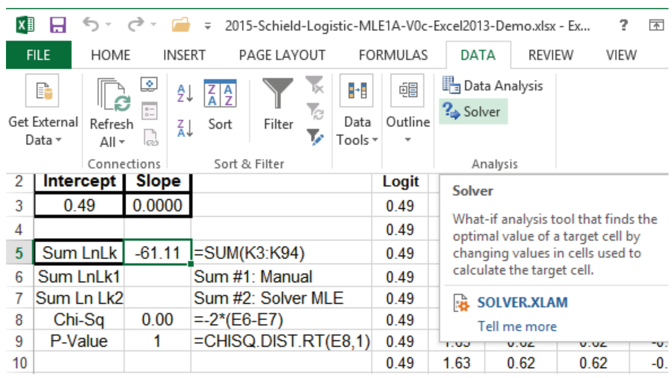


Figure 24: Logistic Regression Excel MLE: Solver Startup

The left side of Figure 25 shows the next manual operation. Select the objective cell (E5) and the variable cells (D3:E3). Select GRC Non-Linear. Select Run/OK.

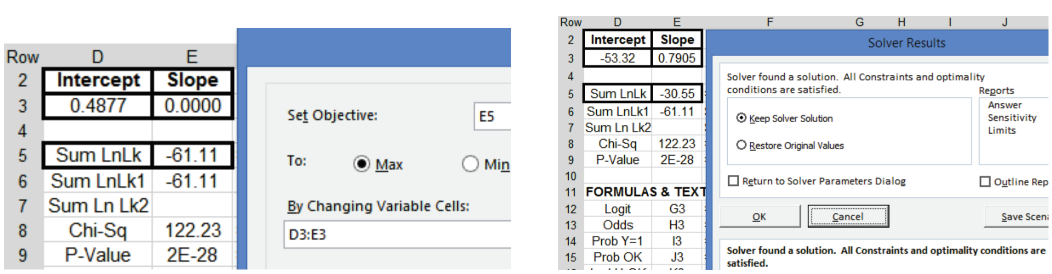


Figure 25: Logistic Regression Excel MLE: Solver Values

The right side of Figure 25 shows Solver's results. Press OK (another manual operation).

After Solver completes satisfactorily, the next manual operation is to copy the Sum of the LogLink from E5 to the clipboard. Paste Special Values into E7. Figure 26 illustrates the final results on the worksheet.

Row	D	E	F
2	Intercept	Slope	
3	-53.32	0.7905	
4			
5	Sum LnLk	-30.55	=SUM(K3:K94)
6	Sum LnLk1	-61.11	Sum #1: Manual
7	Sum Ln Lk2	-30.55	Sum #2: Solver MLE
8	Chi-Sq	61.13	=-2*(E6-E7)
9	P-Value	5E-15	=CHISQ.DIST.RT(E8,1)

Figure 26: Logistic Regression Excel MLE: Final Table

Copying the Sum LnLk to E7 activates the Chi-Squared function which generates a p-value: an incredibly small p-value. So the slope is statistically significant: P-value < 0.05. Note: E-15 means the decimal point is 15 places to the left: 0.000 000 000 000 005.

Since Ln(odds) = Logit, it follows that Odds = Exp(Logit). Since Odds = p/(1-p), it follows that p = Odds / (1+Odds) = Exp(Logit) / [1 + Exp(Logit)] where Logit is a linear function: -53.32 + 0.7905*X. This can be rewritten as p = 1 / [1 + 1/Exp(Logit)] = 1 / [1 + Exp(-Logit)].

Figure 27 shows the final result graphically:

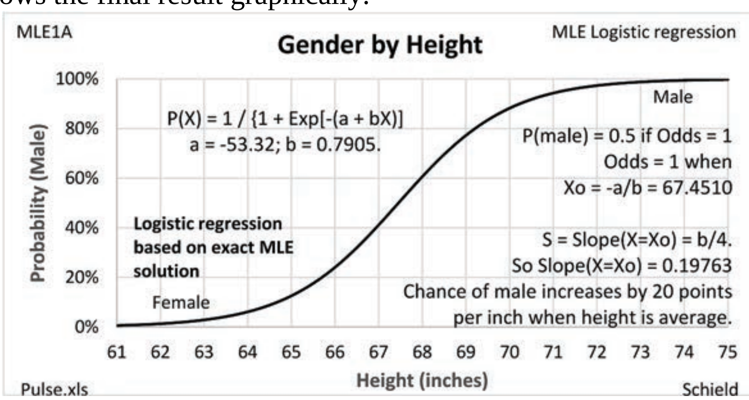


Figure 27: Logistic Regression Excel MLE: Final Graph

The 50% point occurs when X = -a/b. Using calculus, it can be shown that the slope at the 50% point is b/4. Proof:

1. Ln(Odds) = a + bx = Ln[p/(1-p)]
2. Ln(x) = (x-1) - (1/2)(x-1)^2 + (1/3)(x-1)^3 ...
3. Ln(Odds) = a+bx ~ [p/(1-p) -1] = [(2p-1)/(1-p)]
4. Derivative of Ln(Odds) = b = {[2/(1-p)] – [(2p-1)/(1-p)^2]}(dp/dx)
5. b = 4*dp/dx at p = 0.5.

Knowing X50 and the slope specifies the shape of the best fit logistic curve.

Note the number of manual steps. Feedback from skilled Excel users indicated this was not an exercise for a beginner. The number of steps increases with additional predictors.

Schield (2015a and b) presents a more detailed step-by-step presentation.

Appendix C: Minitab Results using MLE Logistic Regression

Minitab provides information on confidence intervals for a continuous predictor. Below is the Minitab output using the same Pulse data (see Appendix A) to model gender using height. See Schield (2016b) for details. As shown in Figure 28 the Minitab and Excel MLE values for the logistic function are identical when rounded at four significant digits.

MLE	Constant	Slope	Log Likelihood
Minitab	-53.3227	0.790517	-30.549
Excel	-53.32	0.7905	-30.55

Figure 28: Minitab and Excel MLE solutions

```
==============================================================
```

Binary Logistic Regression: Male versus Height

```
Link Function: Logit

Response Information
Variable  Value  Count
Male      1         57  (Event)
          0         35
          Total     92

Logistic Regression Table
                                                     Odds     95% CI
Predictor       Coef   SE Coef       Z       P  Ratio  Lower  Upper
Constant   -53.3227   11.4409   -4.66  0.000
Height     0.790517  0.168691    4.69  0.000   2.20   1.58   3.07

Log-Likelihood = -30.549
Test that all slopes are zero: G = 61.129, DF = 1, P-Value = 0.000

Goodness-of-Fit Tests
Method           Chi-Square  DF      P
Pearson             8.00047  19  0.987
Deviance            9.36280  19  0.967
Hosmer-Lemeshow     1.89103   6  0.929

Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)
                         Group
Value     1     2    3    4     5    6    7    8  Total
1
  Obs     0     4    8    7    12    9    9    8     57
  Exp   0.2   2.7  8.9  7.7  11.8  8.7  8.9  8.0
0
  Obs    11    11    9    3     1    0    0    0     35
  Exp  10.8  12.3  8.1  2.3   1.2  0.3  0.1  0.0
Total    11    15   17   10    13    9    9    8     92

Measures of Association:
(Between the Response Variable and Predicted Probabilities)
Pairs        Number  Percent  Summary Measures
Concordant    1801     90.3  Somers' D              0.84
Discordant     116      5.8  Goodman-Kruskal Gamma  0.88
Ties            78      3.9  Kendall's Tau-a        0.40
Total         1995    100.0
```

Binary Logistic Regression: Male versus Weight

Link Function: Logit

Response Information

Variable	Value	Count	
Male?	1	57	(Event)
	0	35	
	Total	92	

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-21.4818	4.48342	-4.79	0.000			
Weight	0.157702	0.0324812	4.86	0.000	1.17	1.10	1.25

Log-Likelihood = -26.126
Test that all slopes are zero: G = 69.974, DF = 1, P-Value = 0.000

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	18.5426	35	0.990
Deviance	17.2106	35	0.995
Hosmer-Lemeshow	3.2500	7	0.861

Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

	Group									
Value	1	2	3	4	5	6	7	8	9	Total
1										
Obs	0	0	3	7	13	11	11	10	2	57
Exp	0.2	0.6	2.2	6.2	14.6	10.4	10.8	10.0	2.0	
0										
Obs	9	9	8	5	4	0	0	0	0	35
Exp	8.8	8.4	8.8	5.8	2.4	0.6	0.2	0.0	0.0	
Total	9	9	11	12	17	11	11	10	2	92

Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs	Number	Percent	Summary Measures	
Concordant	1869	93.7	Somers' D	0.89
Discordant	89	4.5	Goodman-Kruskal Gamma	0.91
Ties	37	1.9	Kendall's Tau-a	0.43
Total	1995	100.0		

Binary Logistic Regression: Male versus Height, Weight

```
Link Function: Logit


Response Information

Variable  Value  Count
Male?     1         57  (Event)
          0         35
          Total     92


Logistic Regression Table

                                                      Odds     95% CI
Predictor       Coef    SE Coef       Z       P  Ratio  Lower  Upper
Constant   -41.3971    11.6260  -3.56  0.000
Height     0.381653   0.187085   2.04  0.041   1.46   1.02   2.11
Weight     0.114585  0.0364816   3.14  0.002   1.12   1.04   1.20


Log-Likelihood = -23.450
Test that all slopes are zero: G = 75.326, DF = 2, P-Value = 0.000


Goodness-of-Fit Tests

Method           Chi-Square  DF      P
Pearson             33.8184  77  1.000
Deviance            33.7175  77  1.000
Hosmer-Lemeshow      3.0383   8  0.932


Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

                             Group
Value    1    2    3    4    5    6    7    8    9    10  Total
1
  Obs    0    0    2    4    7    7    9    9    9    10     57
  Exp  0.1  0.5  1.4  3.4  7.3  8.0  8.5  8.9  9.0  10.0
0
  Obs    9    9    7    5    3    2    0    0    0     0     35
  Exp  8.9  8.5  7.6  5.6  2.7  1.0  0.5  0.1  0.0   0.0
Total    9    9    9    9   10    9    9    9    9    10     92


Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs       Number  Percent  Summary Measures
Concordant    1911     95.8  Somers' D              0.92
Discordant      78      3.9  Goodman-Kruskal Gamma  0.92
Ties             6      0.3  Kendall's Tau-a        0.44
Total         1995    100.0
```

Appendix D: Generating Confidence Intervals for Logistic Regression

Minitab shows the standard error, t-statistic and p-value for each of the model coefficients. Minitab also shows the odds ratio confidence intervals for the predictors.

Figure 29 (left side) shows the data, the logistic model and the associated confidence intervals. Minitab (2017a). The formula for the logistic model is shown. The values of the standard error and the confidence interval are not shown and the values are not readily available. Figure 29 (right side) shows the first menu-based commands to initiate Binary Logistic Regression.

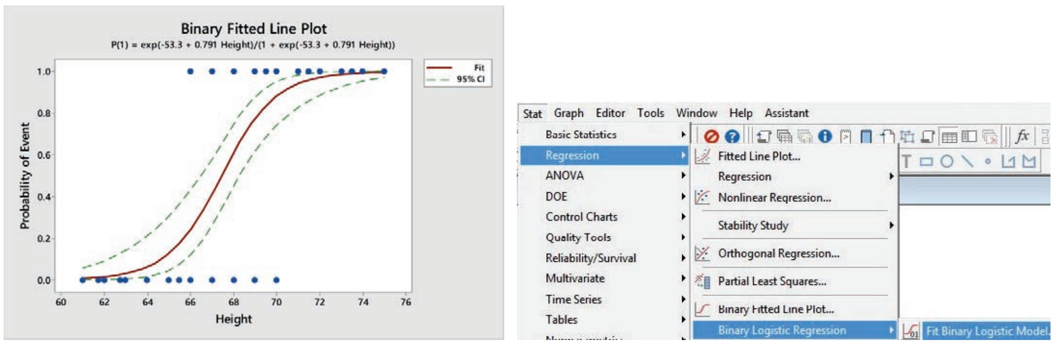


Figure 29: Minitab Logit Model Output and Menu commands

Figure 29 [MS1]shows the menu selections need to generate the MLE coefficients and the variance-covariance matrix. The left side shows the choices for a single predictor: height. The right side shows the choices for two predictors: height and weight.

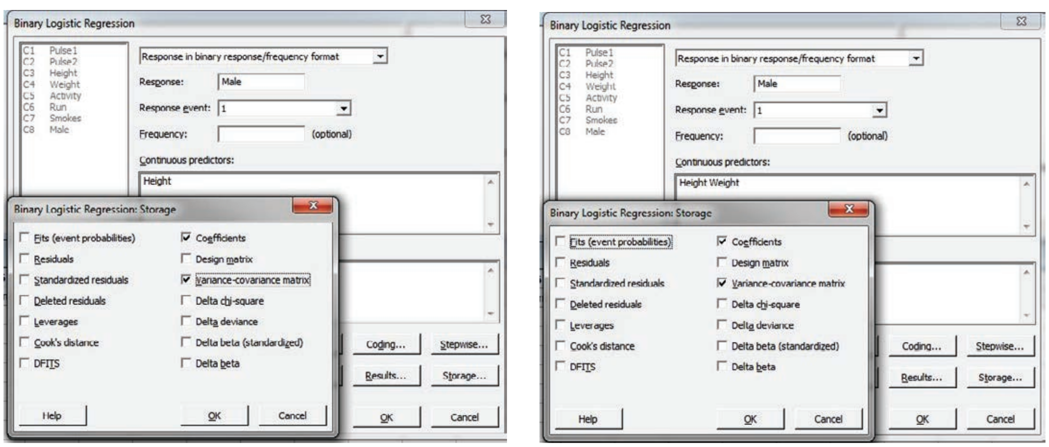


Figure 30: Minitab menus to generate coefficients and variance-covariance matrix.

Some matrix algebra is required to calculate the standard error of the model.

Minitab (2017b) notes that "The standard error of the fit (SE fit) estimates the variation in the estimated probability for the specified variable settings. The calculation of the confidence interval for the prediction uses the standard error of the fit. Standard errors are always non-negative." Minitab (2017c) provides an example with a continuous predictor and a binary predictor. Minitab (2017d) compares a prediction interval from a confidence interval.

Pammer (2017) provided the Minitab commands that generate the standard error and confidence interval for P(male). The left side is for someone of average height. The right side is for someone of average height and average weight.

```
#1 Solve for coefficients and the
#  variance-covariance matrix
Name C9 "COEF" M1 "XPWX".
Gzlm;
   Nodefault;
   REvent 1;
   Response 'Male';
   Continuous 'Height';
   Terms Height;
   Constant;
   Binomial;
     Logit;
   TOdds;
     Increment 1;
   Unstandardized;
   Tmethod;
   Trinfo;
   Tdeviance;
   Tsummary;
   Tcoefficients;
   Tequation;
   Tgoodness;
   TDiagnostics 0;
   Coefficients 'COEF';
   Xpwxinverse 'XPWX'.
#
#2. Calculate P(male|Average height)
Let C10(1) = 1
Let C10(2) = Mean(C3) # Average Height
#
Name C10 "Xh" C11 "LOdd"
Name C12 "OddS" c13 "Prob".
Let C11 = Sum('COEF' * 'Xh')
Let C12(1) = Exp(C11(1))
Let C13(1) = C12(1) / (1 + C12(1))
#
#3. Calculate the standard error
TRANSPOSE C10 M2      #M2 = X'h
MULTIPLY M2 M1 M3     #M3 = X'h(X'WX)^-1
MULTIPLY M3 C10 C14   #C14= X'h(X'WX)^-1X'h
LET C14 = SQRT(C14)   #C14= std(eta_hat)
NAME C14 "StEr".      #C14= Standard Error
#
#4. Confidence interval: P(male|AveHt)
NAME C15 "CNF1"  C16 "CNF2"
LET C15(1) = EXP(C11(1) - 1.96*C14)
Let C15(2) = C15(1) / (1 + C15(1))
LET C16(1) = EXP(C11(1) + 1.96*C14)
Let C16(2) = C16(1) / (1 + C16(1))
```

```
#1 Solve for coefficients and the
#  variance-covariance matrix
Name C9 "COEF" M1 "XPWX".
Gzlm;
   Nodefault;
   REvent 1;
   Response 'Male';
   Continuous 'Height' 'weight';
   Terms Height Weight;
   Constant;
   Binomial;
     Logit;
   TOdds;
     Increment 1 1;    # Two predictors
   Unstandardized;
   Tmethod;
   Trinfo;
   Tdeviance;
   Tsummary;
   Tcoefficients;
   Tequation;
   Tgoodness;
   TDiagnostics 0;
   Coefficients 'COEF';
   Xpwxinverse 'XPWX'.
#
#2. Calculate P(male | Ave Ht+Wt).
Let C10(1) = 1
Let C10(2) = Mean(C3)      # Average Height
LET C10(3) = Mean(C4)      # Average Weight
Name C10 "Xh" C11 "LOdd"
Name C12 "OddS" c13 "Prob".
Let C11 = Sum('COEF' * 'Xh')
Let C12(1) = Exp(C11(1))
Let C13(1) = C12(1) / (1 + C12(1))
#
#3. Calculate the standard error
TRANSPOSE C10 M2       #M2 = X'h
MULTIPLY M2 M1 M3      #M3 = X'h(X'WX)^-1
MULTIPLY M3 C10 C14    #C14= X'h(X'WX)^-1X'h
LET C14 = SQRT(C14)    #C14= std(eta_hat)
NAME C14 "StEr".       #C14= Standard Error

#4. Conf. interval P(male|Ave Ht+Wt)
NAME C15 "CNF1" C16 "CNF2"
LET C15(1) = EXP(C11(1) - 1.96*C14)
Let C15(2) = C15(1) / (1 + C15(1))
LET C16(1) = EXP(C11(1) + 1.96*C14)
Let C16(2) = C16(1) / (1 + C16(1))
```

In both cases, the crucial step is the generation of the standard error that is stored in C14. This matrix algebra mathematics is not for the faint of heart. A short-cut for obtaining an approximate value of the standard error of the model would be appreciated.

Appendix E: Slope Statistically Significant

In doing an OLS regression, a slope coefficient may not be statistically significant. Either the slope in the population is non-existent or the sample size is too small to detect a small non-zero slope in the population.

If the predictor(s) and outcome are continuous, there is no way in OLS regression to look ahead and determine if the resulting slope(s) is statistically significant.

But if the predictor is binary, then the OLS regression will connect the means of the two groups. A simple t-test on the difference in means will determine whether the OLS slope is statistically significant.

If the outcome is binary and the predictors are continuous, it seems that something similar can be done. According to PASSS (2017),

> "we can run a two-sample t test to determine if there is a statistically significant difference in the mean … scores." "This, like all exploratory analysis, can help us determine whether or not it is worth fitting a logistic regression model for these variables. If the difference in mean ... score with respect to [the binary predictor] is insignificant, running a logistic regression wouldn't be the best use of our time, as our results wouldn't be significant."

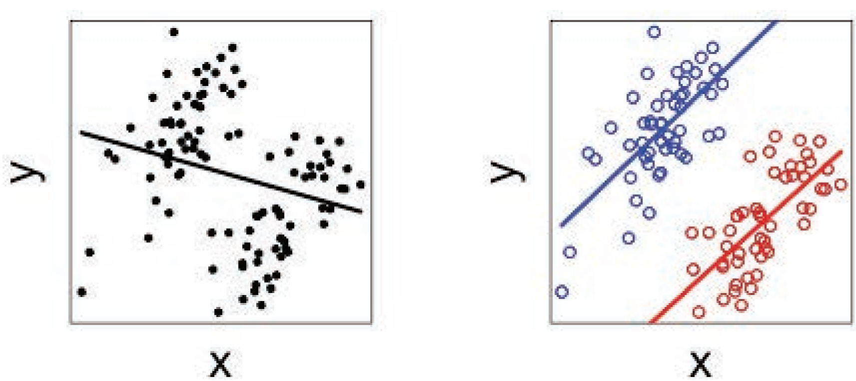
This sounds highly plausible, but the lack of confirmation in other sources is bothersome. For more web background on logistic regression, see Hardin (2017) and Lani (2017).

Appendix F: Using Logistic-OLS with Nudge in Excel

Schield (2015c, d, e, and f) give step by step instructions for doing this. Schield (2015g) compares the results with MLE.

Appendix G: Simpson's Paradox with a Continuous Predictor and Outcome

Simpson's Paradox using categorical predictors and a binary outcome is described as an amalgamation paradox. As such the idea doesn't seem applicable to data using continuous predictors. Wasserman (2013) noted that a two-group continuous version of Simpson's Paradox exists and is sometimes called the ecological fallacy. Wasserman provided Figure 31.



p

Figure 31: Simpson's Paradox: Continuous Predictor & Outcome; Binary Confounder

Appendix H: Using Logistic+OLS+Nudge to show a Simpson's Reversal

Finding data sets involving two continuous predictors and a binary outcome that illustrate a Simpson's Reversal is not easy. The following datasets have been constructed using @Risk simulations in a data-generation program created by Isaacson (2008). The values in the correlation matrix are chosen to clearly illustrate a Simpson's Reversal. Schield (2017) contains the associated data, models and graphs.

The following correlation coefficients in the resulting data were obtained

Variables in	Data #1	Data #2
Predictor-Outcome Correlation	0.131	0.326
Confounder-Outcome Correlation	0.277	0.476
Predictor-Confounder Correlation	0.813	0.907

Data #1 was obtained with @Risk correlation coefficients of 0.2, 0.4 and 0.8 respectively. Data #2 was obtained with @Risk correlation coefficients of 0.4, 0.6 and 0.9. Both predictors had a mean of 100. The binary outcome had an average of 50%.

Data1 and Data2 were modelled using MLE and OLS-Nudge on the Predictor-Outcome association before and after controlling for the confounder.

Using Data1, the following coefficients were obtained using Minitab MLE:

Variables in	Before	After
Intercept or Constant	-1.92	1.52
Predictor Coefficient	0.0192	-0.0435
Confounder Coefficient		0.02825

Using Data1, the following coefficients were obtained using OLS-Nudge:

Variables in	Before	After
Intercept or Constant	0.0266	0.849
Predictor Coefficient	0.0047	-0.00996
Confounder Coefficient		0.006478

Using Data2, the following coefficients were obtained using Minitab MLE:

Variables in	Before	After
Intercept or Constant		
Predictor Coefficient		
Confounder Coefficient		

Using Data2, the following coefficients were obtained using OLS-Nudge:

Variables in	Before	After
Intercept or Constant	-0.675	1.335
Predictor Coefficient	.0117	-.0215
Confounder Coefficient		0.0131

A simple way to check on the validity of the bivariate coefficients is to compute the value of X at which Y = 50%. X50 = -Intercept / Slope. In this data, it should be 100.

This same approach was used to generate the three datasets with higher correlations.

Appendix I: Data2 Minitab MLE Before and After Confounding

Binary Logistic Regression: Result versus Predict

Method

Link function	Logit
Rows used	300

Response Information

Variable	Value	Count	
Result	1	150	(Event)
	0	150	
	Total	300	

Deviance Table

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	1	33.37	33.370	33.37	0.000
Predict	1	33.37	33.370	33.37	0.000
Error	298	382.52	1.284		
Total	299	415.89			

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC
8.02%	7.78%	386.52

Coefficients

Term	Coef	SE Coef	VIF
Constant	-5.152	0.962	
Predict	0.05152	0.00955	1.00

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
Predict	1.0529	(1.0334, 1.0728)

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))
Y' = -5.152 + 0.05152 Predict

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	298	382.52	0.001
Pearson	298	299.04	0.472
Hosmer-Lemeshow	8	2.01	0.981

Binary Logistic Regression: Result versus Predict, Confound

Method

Link function	Logit
Rows used	300

Response Information

Variable	Value	Count	
Result	1	150	(Event)
	0	150	
	Total	300	

Deviance Table

Source	DF	Adj Dev	Adj Mean	Chi-Square	P-Value
Regression	2	99.55	49.774	99.55	0.000
Predict	1	24.22	24.223	24.22	0.000
Confound	1	66.18	66.178	66.18	0.000
Error	297	316.34	1.065		
Total	299	415.89			

Model Summary

Deviance R-Sq	Deviance R-Sq(adj)	AIC
23.94%	23.46%	322.34

Coefficients

Term	Coef	SE Coef	VIF
Constant	4.60	1.68	
Predict	-0.1168	0.0254	5.69
Confound	0.0708	0.0102	5.69

Odds Ratios for Continuous Predictors

	Odds Ratio	95% CI
Predict	0.8897	(0.8465, 0.9352)
Confound	1.0734	(1.0522, 1.0950)

Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))
Y' = 4.60 - 0.1168 Predict + 0.0708 Confound

Goodness-of-Fit Tests

Test	DF	Chi-Square	P-Value
Deviance	297	316.34	0.211
Pearson	297	287.66	0.641
Hosmer-Lemeshow	8	9.55	0.298

Version 4b