

# A Guide to Interpreting and Communicating Regression Analyses for Data Analytics Students

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

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

## Abstract

Faculty teaching data analytics at undergraduate level are often faced with the tension created by student under-preparedness vis a vis underlying and assumed statistical prior knowledge and demands of the course and time constraints. How do faculty close this gap given time constraints and a desire and need to keep the field open to the very student that is likely to be under-prepared? In this paper, we propose the use of flow diagramming as an accessible method for aiding the student recap and upskill by dechunking unfamiliar processes, e.g. interpreting regression analyses, in ways that are time efficient and also not alienating to the student. In preliminary data collected for this pilot study, we found that the use of flow diagrams to close prior knowledge gaps in regression analyses significantly improved student understanding without additional remedial instruction. Time saved can be directed at core learning objectives of the analytics. And yet, accessible remedial tools like flow diagrams may help retain underprepared students in analytics courses.

**Keywords:** Data Analytics, Regression Modeling, Flow Diagram, Flow Chart, Remedial Activity, Remediation

## 1. INTRODUCTION

The demand for analytics knowledge among a variety of users has in business school programs, and specifically in information systems, marketing and finance, led to the incorporation of analytics into curricula at both graduate and undergraduate levels. Though these programs variously cater to different levels of user expertise, from the casual user of analytics to business analysts and data scientists (Watson, 2013, 2015), what they all have in common is an expectation for a foundational understanding of

statistical concepts. As a result, (Arawjo, 2020) data analytics courses typically stipulate an introductory statistics course as a minimal prerequisite because, for students of analytics, statistical understanding is foundational to explanatory statistical modeling, inferential testing and predictive analytics (Shmueli & Koppius, 2011). Yet statistics are and have often been one of the most difficult subjects for undergraduate and even graduate students to grasp. For many higher education institutions, at undergraduate level particularly, the challenge is made compounded for the thousands of students

who graduate high school academically underprepared for college.

Many analytics courses are taught or delivered from within information systems programs, and therefore taught by information systems (IS) faculty because of the large overlap between business analytics, data analytics, and information systems programs (Ceccucci, Jones, Toskin, & Leonard, 2020). In environments where the student population is diverse, faculty teaching these courses have to manage conflicting forces, including: meeting course objectives and analytics content coverage, against balancing the needs and foundational knowledge gaps of underprepared students or students intimidated by the underlying statistical content notwithstanding the prerequisites we impose for our undergraduate courses. Faced with these constraints, faculty and programs, it seems to us, have one of two options, accept that they have to reteach statistical foundations and yield on some the depth in analytics content; we see this in programs that offer *two* undergraduate courses that aim to cover introductory analytics: descriptive analytics, predictive analytics and prescriptive analytics. While understandable, the downside of this approach is arguable watering standards and increasing the cost of the program. Approaches that require student to take a “pass/fail test” to enter analytics courses may result in systemically excluding students graduating from weaker high schools. Another approach, in the absence or complexity and controversy attending remediation, would be for faculty to develop innovative methods and approaches that readily open access to underlying content even for underprepared and/or students with low confidence about the material. Such approaches would strategically and efficiently assist in reviewing underlying concepts to bring students up to some speed while leaving time for coverage of analytics course objectives.

The purpose of this study is to propose an effective pedagogical, visual artifact that increases student efficacy in “relearning” an important underlying statistical concept for analytics without explicit remedial instruction. In this pilot study, we chose the topic of regression analysis: namely, interpreting and communicating one of the most prominent and commonly used statistical modeling techniques, simple and multiple linear regression. The proposed artifact can be applied using a wide range of tools (i.e., SPSS, R, Excel, or Python). We use flow diagramming, an easily understandable and widely used pedagogical aid,

to graphically depict, for the student, prior knowledge (i.e. process or steps) required to successfully interpret and communicate a regression model.

## 2. BACKGROUND

In the experience of the authors, students taking an introductory analytics course often arrive, notwithstanding the typical statistics prerequisite, with little or no recollection or understanding of foundational statistical concepts. This lack of understanding extends both theoretically and in an applied sense. Specifically, our analysis from two data analytics undergraduate courses taught in the Fall 2019, at two regional universities, with coverage of multiple linear regression revealed the following challenges: 80% of students struggled with interpreting and articulating the regression coefficients; 60% of students had a hard time explaining the role of R-squared; 30% of students made incorrect conclusions about the model fit.

These difficulties were not limited to regression analysis, they extended to statistical inference, for example. Undergraduate data analytics students, in our experience, have difficulty interpreting and communicating the results of performed analyses. In reviewing prior literature in statistics education, it shows that there are three types of reasons for these difficulties: *affective* (Ashaari, Judi, Mohamed, & Wook, 2011; Reid & Petocz, 2002), *cognitive* (Chiesi & Primi, 2010), and *pedagogical* (Ramsey, 1999) reasons. Weinberg & Abramowitz (2000, p. 1), researchers in statistics education, concluded that “*our challenge is to find ways of presenting information to our students so that it is accessible, relevant, applicable and even vital to their own areas of interest*”.

The truism that students learn by doing survives, but, from statistics education Ramsey (1999) argued, more thought needs to be put into designing exercises to meet specific objectives. Ramsey distinguishes between three areas of competency as students learn statistics: [1] the opportunity to “gain mechanical skills” in manipulating data and calculating (i.e. practicing calculating means, variances, correlation coefficients, Student *t*'s and F ratios); [2] “Exploring the tools” (exploring the characteristics and properties of the statistics, examining the behavior of the statistics and discovering their limitations to gain a sense of what the calculations are providing (e.g. students should be able to guess-estimate the effect on

variance of adding larger, smaller, equally sized or zero sized data items. [3] The third task, "application", to be completed after [1] and [2] above, is about the interpretation and use of the calculated statistics. In introductory data analytics courses, we may outsource the formal learning of [1] and [2] above to statistics prerequisite courses; [3] falls in the remit of our primary analytics course objectives.

On the other hand, the introduction of technology in statistics education shifted the approach to teaching statistics in ways that are instructional to data analytics. In particular, technology encouraged a shift away from emphasis on computations, formulas, and procedures to an emphasis on "statistical reasoning and the ability to *interpret*, evaluate, and *flexibly apply* statistical ideas" (Ben-Zvi, 2000, p. 130) [emphasis added]. Arguably, this shift presaged the widespread adoption and use of data analytics where ability to interpret, evaluate and apply to a variety of contexts is essential and primary and therefore, the latter as skills we value for analytics students.

In addition, for analytics educators there is the added challenge of managing the tension between giving students greater capacity to interpret, evaluate and apply analytics ideas with an understanding of the bigger picture or problem context as well as the adequate grounding in foundational statistical knowledge. The semi-automated analytics tools we use are a mixed blessing in this regard, especially for the underprepared student. There are useful lessons from the research on statistical education: i.e., foster active learning through alternatives to lecturing (Ben-Zvi, 2000; Cobb, 1992); emphasize oral and written communication (Cobb, 1992); incorporate more data and concepts (Ben-Zvi, 2000); and use *software* designed to foster conceptual understanding (Cobb, 1992). In information systems education, substituting "artifact" (Lee, Thomas, & Baskerville, 2015) for "software" would be entirely consistent with the array of IT/IS artifacts [e.g., constructs, models, methods and instantiations (Hevner, March, Park, & Ram, 2004; March & Smith, 1995) integral to our field and its teaching.

### 3. FLOW DIAGRAMMING USE FOR PEDAGOGY

Flow diagrams, artifacts in computer science and information systems were first introduced into computing by John von Neuman in the 1940s;

they were introduced as a visual representation of the logical structure of a computer program (Ensmenger, 2016). At the time of the ENIAC project, it is understood that flow diagramming was chosen as a form of representation that was readily accessible to the diverse members of the team with different levels of prior knowledge. Flow diagramming was also seen as "superior to introducing a more radical departure in (logical) notation" that some members would have been familiar with (Arawjo, 2020). Priestley's (2018) historical treatment of Von Neuman's work retells how flow diagramming was used in the planning and coding reports of the project to broaden access to understanding of the work by a diverse team:

*"[W]e have acquired a conviction that this programming is best accomplished with the help of some graphical representation of the problem. We have attempted ... to standardize upon a graphical notation ... in the hope that [it] would be sufficiently explicit to make quite clear to a relatively unskilled operator the general outline of the procedure. We further hope that from such a block-diagram the operator will be able with ease to carry out a complete coding of a problem" (p. 59).*

Flow diagramming use increased, primarily because of Von Neuman's fame and it eventually grew to have a commanding influence on software engineering and programming for decades to come (Arawjo, 2020). Flow diagrams have since been used in a variety of contexts, namely: in modeling production processes; in aviation for training and process management (Yazgan & Yilmaz, 2018); in the accounting field to teach CPAs to communicate complex plans (Lehman, 2000); as a quality improvement tool for documenting, understanding, analyzing, and improving business processes (Nesbitt, 1993); to aid reading comprehension in the teaching of the law (Zacharias, 1986). Flow diagram use endures in aiding the teaching of introductory programming and systems analysis and design courses. Although, they have been criticized specifically for their modeling accuracy in programming (Hosch, 1977), they have lasted as both process documentation and teaching aids that make complexity readily accessible to a novice. For information systems and computer science students, they are familiar and useful aids.

In this paper, we propose a method that uses flow diagramming as an accessible teaching aid for reinforcing and developing students' capacity to interpret and communicate their analysis. We use

the instances of both linear and multiple regression models as examples. Our method is also follows a learning theory belief that managing or spacing repetitions in various contexts is better than merely amassing repetitions (Schmidt & Bjork, 1992). So that carefully designed sequences of activities using appropriate tools can help students improve reasoning and understanding over substantial periods of time (Ben-Zvi, 2000; Garfield & Ben-Zvi, 2007). We believe that similar flow diagramming can be used for multiple areas of underlying statistical and modeling concepts, not just regression analysis.

#### 4. METHOD

To illustrate this use of flow diagramming as teaching aid in analytics, we designed two flow diagrams, one for simple linear and multiple regression respectively. We emphasize that the focus is on aiding and strengthening students' capacity for interpretation and communication of analyses (implying understanding). The use of flow diagramming is a deliberate calculus aimed at creating easy access to understanding for the underprepared student and novice. In other words, affect considerations are also at play.

##### Description of Flow diagramming Artifacts

In this section we describe the proposed flow diagramming artifacts: Appendix A to support simple linear regression interpretation, and Appendix E, to support multiple linear regression interpretation. We refer here to regression as a statistical method that seeks to estimate the relationship between an outcome variable and a predictor variable, or set of predictors. In its simplest form, the relationship is expressed as a simple linear equation.

##### Simple Linear Regression

This flow diagram (Appendix A) focuses on four key elements as well as the need to articulate the regression equation: the significance F or *p-value for the F statistic*, the *intercept* or constant; the *slope* or coefficient of the independent variable (DV) and its p-value; and the interpretation of *R-squared* for model fit. We note that for simple linear regression, the instructor may want to remind students that the significance of F statistic (p-value) is equal to the p-value of the coefficient of the independent variable (IV) or slope.

##### Multiple Linear Regression

This flow diagram (Appendix E) focuses on five key elements as well as the need to articulate the multiple regression equation. The process is not dissimilar to simple linear regression. However,

with multiple regression we assume students start with all hypothesized independent predictors included in the regression model. The five key elements are: significance F (p-value for the F statistic), the *intercept* or constant; *coefficients* of the hypothesized independent variables and their respective p-values (we also assume students can, in stepwise fashion, remove non-significant variables, then re-run the model, i.e. return to Step 1 of the flow diagram); the interpretation of *R-squared* for model fit; and *adjusted R-squared*. The final step is intended to help nudge students to use *adjusted R-squared* for model fit for multiple regression. This, to reinforce understanding that adjusted R-squared penalizes the addition of independent variables that do not aid in predicting the dependent variable where R-squared increases with every additional variable regardless of its effect on the dependent variable.

We note that to use both flow diagrams, we make rudimentary assumptions related to prior instruction. For instance, for multiple regression (Appendix E - Step 5), we assume prior instruction did cover that adjusted the R-squared is based on R-squared adjusted for the number of predictors and sample size. We use regression equations, in both cases, without the error term. We also assume students are previously instructed on the fundamentals underlying regression analysis including the checking of regression assumptions: linearity, normality of errors, homoscedasticity, and independence of errors and the role residuals in assessing regression assumptions.

Lastly, the proposed pedagogical flow diagrams are intended to be used over time, with other related and ideally concept repeating assignments. In other words, they can be used again for predictive analytics based on regression models. Below we provide sample assignments and their rubrics (Appendix B, C, and D for simple linear; and Appendix F, G, and H for multiple regression). It is not our intention that these particular flow diagrams be used for instructing regression analysis from scratch per se. Rather we propose that instructors use them as a remedial mechanism, to close the gap of the forgotten or previously ill or misunderstood, and for review. The purpose is to aid students in gaining proficiency on how to *interpret* and *communicate* regression analysis results by focusing on essential information.

## 5. EXPERIMENTAL CONTEXT

To test the efficacy of the proposed pedagogical aids, students enrolled in a business analytics undergraduate course, in the Fall of 2020, at a regional public university were given an opportunity to use the flow diagrams as a mechanism for reviewing assumed prior knowledge and provide feedback for this study. Students participating in the study are required to have previously completed an introductory statistics course as well as an introductory information systems course that includes an introduction to databases and introduction to data analysis and functions in Excel.

### Study Design

Our experimental design used a pre-test/post-test approach (Campbell & Stanley, 2015) to test the effect of using pedagogical aid on student understanding with respect to interpreting and communicating regression analysis results.

Students were presented with four different problems: two for simple regression and two multiple linear regression problems. Each problem included a hypothetical scenario describing the student's role and the problem being investigated, model output (generated using Microsoft Excel data analysis tools), followed by seven questions to test students' understanding of the statistical model presented in the problem.

Step 1: students received a pre-test for simple regression model (Appendix B), followed by a post-test (Appendix C). Although both simple linear regression models used the same data set, the variables used in each model are different. During the post-test, students were asked to use the simple linear regression flow diagram as an aid to formulating response questions for the assigned problem (Appendix A).

Step 2: for multiple linear regression, the process was similar to Step 1 above except a new data set was used to generate the models. Students were presented with a pre-test (Appendix F), followed by a post-test (Appendix G). Variables in the post-test model were changed; and only the post-test included the flow diagram (Appendix E).

To evaluate student responses to the assigned problems consistently and using the same criteria and scoring schema, grading rubrics were used. A grading rubric was designed to evaluate students' efforts interpreting simple linear regression (Appendix D) and multiple linear regression (Appendix H) problems.

### Data Collection – Participants and Procedure

Each student in the course was asked to complete the four problems. Responses were recorded using a Qualtrics survey where each problem was presented in a single screen and students were not permitted to go back to a previous screen. This ensured students could not change or correct their answers in the pretest while completing the flow-diagram aided post-test, for example. No review of linear regression was conducted in class, students had to rely on (assumed) prior knowledge. Out of nineteen students invited to complete the survey, fourteen students participated in the study (73% response rate).

## 6. RESULTS

The data was analyzed using paired sample t-tests. The results presented in Table 1 show statistically significant differences in the mean test scores between the pre-test and post-test for both simple and multiple linear regression indicating that both flow diagrams had a positive effect on student understanding and interpretation of the statistical models presented in the problems. In addition, when asked "How useful did you find the flowchart aid in interpreting the data?", almost 90% of students responded that they found the flow diagram useful ranging from slightly to extremely useful. Furthermore, 93% of students said, they were somewhat likely to extremely likely to use similar flow diagramming aids for other topics in business analytics field.

Outcome	Pre-test		Post-test		t
	Min	Max	Min	Max	
SLR	0	79	0	95	
MLR	0	77	0	96	
Outcome	Pre-test		Post-test		t
	M	SD	M	SD	
SLR	36	27	45	34	-3.20**
MLR	34	29	45	34	-3.94**

Note. SLR – Simple Linear Regression. MLR – Multiple Linear Regression, \*\* p < .01, n = 14

**Table 1. Descriptive Statistics & t-test Results**

## 7. DISCUSSION AND CONCLUSION

In this study, we propose a method for addressing an important tension facing instructors of introductory analytics courses, closing the gap in assumed statistical prior knowledge. In this study, we use examples from simple and multiple regression analysis which are foundational to

predictive analytics. The reasons for the knowledge gaps may be broad, from merely forgetting work from prior semesters (or years as in the case of older returning students), having a poor experience with an instructor, mismatched transfer articulations between community college and a four-year institution, as well as underlying and systemic under-preparedness for college. The net result is, students taking undergraduate analytics courses unprepared for the rigors of the field. To aid students upskill or recap in ways that feel accessible and non-intimidating, we propose the use of strategically designed flow diagrams focused on specific knowledge gap areas. Flow diagrams are tools that have proven to in use to readily foster conceptual understanding; they are a good alternative to lecturing; they are both time effective and time efficient. The gains in time saved can be directed towards analytics content. From a student perspective, flow diagrams are easy to understand and perhaps even familiar for students in information systems and computer science; ease of use and familiarity are precursors to favorable affective assessment. We believe creating mechanism for underprepared students to quickly feel more comfortable or less intimidated by the demands and prior knowledge assumptions of analytics courses can avert conditions where students prematurely drop out of the course out of early frustrations. Retaining underprepared students of all stripes can contribute to opening the field to a diversity of students.

This pilot study shows test scores were higher and statistically significantly when using flow diagrams. Similar methods may not only help student understanding of individual concepts but may in the collective serve as important tools for managing remedial work that attends analytics classes taught at undergraduate level, and therefore retaining underprepared students at relatively little cost.

Finally, information systems students, in particular, could be encouraged to create their own flow diagrams for other analytical processes they find individually challenging, thus to unlocking complexity for themselves. e.g., systematically checking regression assumptions, hypothesis testing from problem statement, data analysis to drawing correct conclusion.

## 8. LIMITATIONS AND FUTURE RESEARCH

The proposed pedagogical flow diagrams do not make a claim comprehensively covering all issues related to regression analysis. For instance, an iterative part of the analysis includes examining

linear regression assumptions inclusive of examining and interpreting residuals; these flow diagram aids do not incorporate that. To minimize the complexity and maintain the effectiveness of the aid, we believe it would require a different but similar flow diagramming aid. Such an aid could, "walk" a student through how to use/interpret the diagnostic features and charts for assessing residuals generated by most statistical tools like R, SPSS and Excel. For example, another flow diagram could be used to aid a student needing remedial activity on how to run the output to examine residuals, remove outliers, or log-transform the data and re-run the regression model before interpretation. Likewise, we do not explicitly model or review steps for performing stepwise regression for multiple regression. For remedial activities, instructors can design such aids. As a field, we could design and share these such aids.

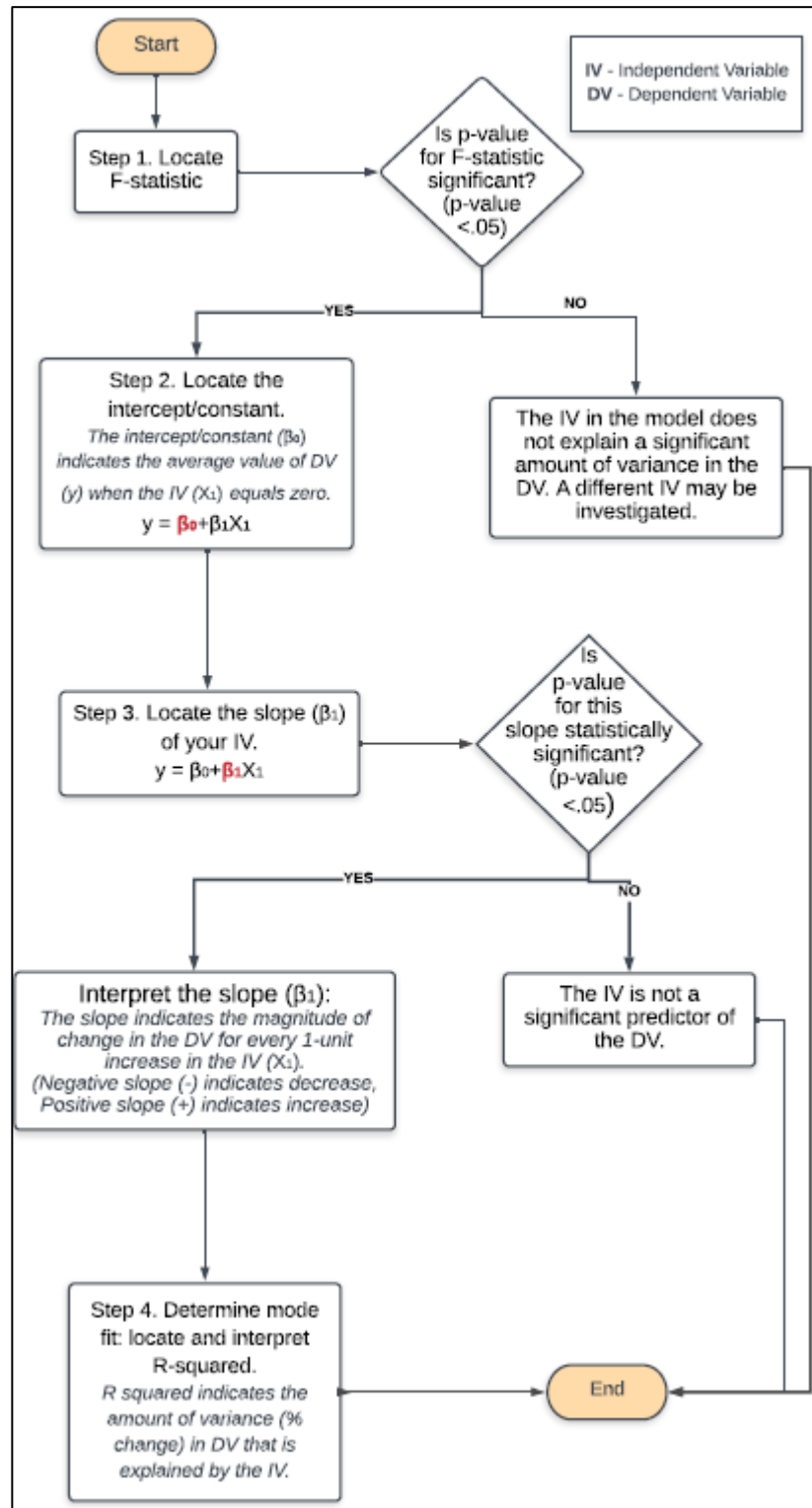
Finally, this study is limited by a small sample size. Though we found statistical significance with a small sample further investigation with a larger sample is warranted to establish generalizability.

## 9. REFERENCES

- Arawjo, I. (2020). *To Write Code: The Cultural Fabrication of Programming Notation and Practice*. Paper presented at the Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.
- Ashaari, N. S., Judi, H. M., Mohamed, H., & Wook, M. T. (2011). Student's attitude towards statistics course. *Procedia-Social and Behavioral Sciences*, 18, 287-294.
- Bahr, P. R. (2008). Does mathematics remediation work?: A comparative analysis of academic attainment among community college students. *Research in Higher Education*, 49(5), 420-450.
- Bahr, P. R. (2010). Preparing the underprepared: An analysis of racial disparities in postsecondary mathematics remediation. *The Journal of Higher Education*, 81(2), 209-237.
- Ben-Zvi, D. (2000). Toward understanding the role of technological tools in statistical learning. *Mathematical thinking and learning*, 2(1-2), 127-155.
- Bettinger, E. P., & Long, B. T. (2009). Addressing the needs of underprepared students in higher education does college remediation

- work? *Journal of Human resources*, 44(3), 736-771.
- Ceccucci, W., Jones, K., Toskin, K., & Leonard, L. (2020). Undergraduate Business Analytics and the Overlap with Information Systems Programs. *Information Systems Education Journal*, 18(4), 22-32.
- Chiesi, F., & Primi, C. (2010). Cognitive and non-cognitive factors related to students' statistics achievement. *Statistics Education Research Journal*, 9(1), 6-26.
- Cobb, G. (1992). Teaching statistics. *Heeding the call for change: Suggestions for curricular action*, 22, 3-43.
- Ensmenger, N. (2016). The multiple meanings of a flowchart. *Information & Culture*, 51(3), 321-351.
- Garfield, J., & Ben-Zvi, D. (2007). How students learn statistics revisited: A current review of research on teaching and learning statistics. *International statistical review*, 75(3), 372-396.
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS quarterly*, 75-105.
- Hosch, F. A. (1977). Whither flowcharting? *ACM SIGCSE Bulletin*, 9(3), 66-73.
- Lee, A. S., Thomas, M., & Baskerville, R. L. (2015). Going back to basics in design science: from the information technology artifact to the information systems artifact. *Information Systems Journal*, 25(1), 5-21.
- Lehman, M. W. (2000). Flowcharting made simple. *Journal of Accountancy*, 190(4), 77.
- March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision support systems*, 15(4), 251-266.
- Nesbitt, T. E. (1993). Flowcharting business processes. *Quality*, 32(3), 34.
- Priestley, M. (2018). *Routines of Substitution: John Von Neumann's Work on Software Development, 1945-1948*: Springer.
- Ramsey, J. B. (1999). Why do students find statistics so difficult. *Proceedings of the 52th Session of the ISI. Helsinki*, 10-18.
- Reid, A., & Petocz, P. (2002). Students' conceptions of statistics: A phenomenographic study. *Journal of Statistics Education*, 10(2).
- Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological science*, 3(4), 207-218.
- Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS quarterly*, 553-572.
- Watson, H. J. (2013). The business case for analytics. *BizEd*, 12(3), 48-54.
- Watson, H. J. (2015). Should you pursue a career in BI/analytics. *Student Edition of the Business Intelligence Journal*.
- Weinberg, S., & Abramowitz, S. (2000). Making general principles come alive in the classroom through an active case studies approach. *Journal of Statistics Education*, 8(2), 15.
- Yazgan, E., & Yilmaz, A. K. (2018). Designing flow chart for aircraft type training in aviation training process management. *Aircraft Engineering and Aerospace Technology*.
- Zacharias, F. C. (1986). Flowcharting the First Amendment. *Cornell L. Rev.*, 72, 936.

### Appendix A: Simple Linear Regression Flow Diagram





### Appendix B: Simple Linear Regression Assignment – Pre-test

**Instructions:** You are a research analyst for an automotive industry. Your boss has asked you to investigate if the engine displacement size (in liters) affects the highway fuel economy (miles per gallon) based on a sample of 234 vehicles. Given the following input from a regression analysis, please interpret the results by answering the questions listed below.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.766020021							
R Square	0.586786672							
Adjusted R Square	0.58500558							
Standard Error	3.835984908							
Observations	234							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1	4847.833384	4847.833384	329.4533329	2.03897E-46			
Residual	232	3413.829009	14.71478021					
Total	233	8261.662393						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	35.69765105	0.720367619	49.55476915	2.1235E-125	34.27835256	37.11694954	34.27835256	37.11694954
Displacement	-3.530588806	0.194513697	-18.15084937	2.03897E-46	-3.913827848	-3.147349764	-3.913827848	-3.147349764

**Questions:**

1. Interpret the overall significance of the regression model. Report the statistic you used to answer this question and its value. (10 points)
2. Report and Interpret the intercept/constant. (10 points)
3. Report and Interpret the slope/regression coefficient of the independent variable (10 points)
4. Report and Interpret the model fit (10 points)
5. Formulate and Interpret the regression equation for this model (10 points)
6. Use the model to predict the value of Y given X1=3.7. (10 points)
7. Write a precise summary of your findings to your boss. Remember that your boss is not a data analyst so make sure to articulate your findings in an easy to understand language. (40 points)

**Notes:** Please see below, the description of the variables and a preview of the first 10 records of the data set.

Manufacturer	Model	Displacement	Year	Cylinders	Transmission	Drive	City	Highway	Fuel	Class
audi	a4	1.8	1999	4	auto(l5)	f	18	29	p	compact
audi	a4	1.8	1999	4	manual(m5)	f	21	29	p	compact
audi	a4	2	2008	4	manual(m6)	f	20	31	p	compact
audi	a4	2	2008	4	auto(av)	f	21	30	p	compact
audi	a4	2.8	1999	6	auto(l5)	f	16	26	p	compact
audi	a4	2.8	1999	6	manual(m5)	f	18	26	p	compact
audi	a4	3.1	2008	6	auto(av)	f	18	27	p	compact
audi	a4 quattro	1.8	1999	4	manual(m5)	4	18	26	p	compact
audi	a4 quattro	1.8	1999	4	auto(l5)	4	16	25	p	compact
audi	a4 quattro	2	2008	4	manual(m6)	4	20	28	p	compact

- Manufacturer – name of the manufacturer of the car
- Model – model name
- Displacement – engine displacement (size) in liters
- Year – year of car manufacture
- Cylinders – number of cylinders
- Transmission – type of transmission such as automatic or manual
- Drive – type of drive such as f = front wheel drive, r = rear wheel drive, 4 = four-wheel drive
- City – city miles per gallon
- Highway – highway miles per gallon
- Fuel – fuel type such as p = premium, r = regular
- Class – type of car

### Appendix C: Simple Linear Regression Assignments – Post-test

**Instructions:** You are a research analyst for an automotive industry. Your boss has asked you to investigate if the number of cylinders affects the city fuel economy (miles per gallon) based on a sample of 234 vehicles. Given the following input from a regression analysis, as well as using the flow chart included at the end of this assignment, please interpret the results by answering the questions listed below.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.805771408							
R Square	0.649267562							
Adjusted R Square	0.647755785							
Standard Error	2.525910807							
Observations	234							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1	2740.13386	2740.13386	429.4728925	1.06838E-54			
Residual	232	1480.212294	6.380225406					
Total	233	4220.346154						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	29.39043476	0.626831425	46.88730267	2.4998E-120	28.1554252	30.62544432	28.1554252	30.62544432
Cylinders	-2.127983841	0.102683449	-20.72372777	1.06838E-54	-2.330295079	-1.925672604	-2.330295079	-1.925672604

**Questions:**

1. Interpret the overall significance of the regression model. Report the statistic you used to answer this question and its value. (10 points)
2. Report and Interpret the intercept/constant. (10 points)
3. Report and Interpret the slope/regression coefficient of the independent variable (10 points)
4. Report and Interpret the model fit (10 points)
5. Formulate and Interpret the regression equation for this model (10 points)
6. Use the model to predict the value of Y given X1=8. (10 points)
7. Write a precise summary of your findings to your boss. Remember that your boss is not a data analyst so make sure to articulate your findings in an easy to understand language. (40 points)

**Notes:** Please see below, the description of the variables and a preview of the first 10 records of the data set.

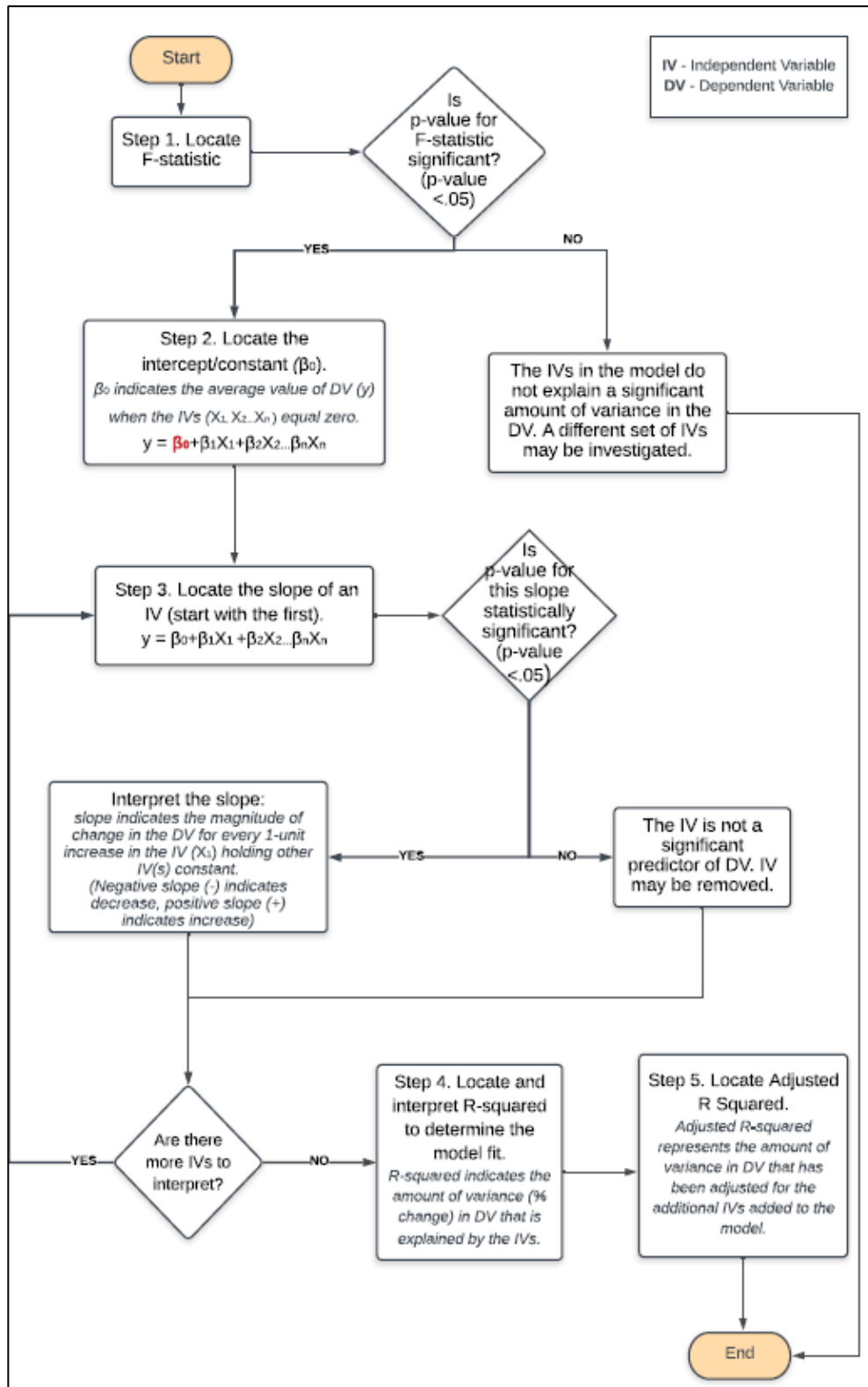
Manufacturer	Model	Displacement	Year	Cylinders	Transmission	Drive	City	Highway	Fuel	Class
audi	a4	1.8	1999	4	auto(l5)	f	18	29	p	compact
audi	a4	1.8	1999	4	manual(m5)	f	21	29	p	compact
audi	a4	2	2008	4	manual(m6)	f	20	31	p	compact
audi	a4	2	2008	4	auto(av)	f	21	30	p	compact
audi	a4	2.8	1999	6	auto(l5)	f	16	26	p	compact
audi	a4	2.8	1999	6	manual(m5)	f	18	26	p	compact
audi	a4	3.1	2008	6	auto(av)	f	18	27	p	compact
audi	a4 quattro	1.8	1999	4	manual(m5)	4	18	26	p	compact
audi	a4 quattro	1.8	1999	4	auto(l5)	4	16	25	p	compact
audi	a4 quattro	2	2008	4	manual(m6)	4	20	28	p	compact

- Manufacturer – name of the manufacturer of the car
- Model – model name
- Displacement – engine displacement (size) in liters
- Year – year of car manufacture
- Cylinders – number of cylinders
- Transmission – type of transmission such as automatic or manual
- Drive – type of drive such as f = front wheel drive, r = rear wheel drive, 4 = four-wheel drive
- City – city miles per gallon
- Highway – highway miles per gallon
- Fuel – fuel type such as p = premium, r = regular
- Class – type of car

**Appendix D:  
Rubric for Grading Simple Linear Regression Pre-test and Post-test**

Grading Rubric: Simple Linear Regression Assignment							
Student Name:							
	Needs Improvement	Satisfactory	Excellent				
	0-1 points	2-3 points	4-5 points	Weight	Points Possible	Points Earned	Weighted Points
<i>Interpretation of overall significance of the model: F statistic/Null Hypothesis</i>	The significance of the model is not reported using the correct statistic. The interpretation is missing or completely incorrect.	The significance of the model is reported correctly and interpretation is partially correct.	The significance of the regression model is reported correctly and interpreted accurately.	0.1	5		0
<i>Interpretation of the intercept/constant: B0</i>	Intercept/constant of regression line is not reported correctly. The interpretation is missing or completely incorrect for the constant/slope.	Intercept/constant is reported correctly and interpretation is partially correct.	Intercept/constant is reported and interpreted accurately.	0.1	5		0
<i>Interpretation of the slope/regression coefficient: B1</i>	Slope/regression coefficient is not reported correctly. The interpretation is missing or completely incorrect for the slope/regression coefficient.	Slope/regression coefficient is reported correctly and interpretation is partially correct.	Slope/regression coefficient is reported and interpreted accurately.	0.1	5		0
<i>Interpretation of model fit: R-squared</i>	R-squared is not reported correctly. The interpretation is missing or completely incorrect for R squared.	R-squared is reported correctly and interpretation is partially correct.	R-squared is reported and interpreted accurately.	0.1	5		0
<i>Formulation and Interpretation of regression equation <math>\hat{Y} = B_0 + B_1X_1</math></i>	The equation is not formulated correctly. The interpretation of the equation is missing or completely incorrect.	The equation is formulated correctly and interpretation is partially correct.	The equation is formulated and interpreted accurately.	0.1	5		0
<i>Calculate the predicted value of Y using a specific value of X1</i>	The predicted value of Y is not calculated correctly.	The predicted value of Y is partially correct.	The predicted value of Y is calculated accurately.	0.1	5		0
<i>Articulate your findings: Write a summary of your findings.</i>	The articulation is incorrect or missing.	The articulation is partially complete and partially correct.	The articulation is complete and accurate.	0.4	5		0
						<b>Total Score</b>	<b>0</b>

### Appendix E: Multiple Linear Regression Flow diagram



**Appendix F:  
Multiple Linear Regression Assignments – Pre-test**

**Instructions:** Department of Public Safety (DOS) has hired you as a consultant to help them explore historical data collected from 50 states of the United States of America. Their first priority is to study if Population size (in millions) and Income per capita (in thousands) affect murder rate (%). Given the following output from a regression analysis, please interpret the results by answering the questions listed below. |

SUMMARY OUTPUT								
<b>Regression Statistics</b>								
Multiple R	0.461732393							
R Square	0.213196803							
Adjusted R Square	0.179715816							
Standard Error	3.343410573							
Observations	50							
<b>ANOVA</b>								
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>			
Regression	2	142.3612699	71.18063494	6.367697658	0.003571525			
Residual	47	525.3845301	11.17839426					
Total	49	667.7458						
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	14.34337892	3.490659875	4.109073767	0.000157786	7.321077034	21.36568081	7.321077034	21.36568081
Population	0.338434824	0.109381813	3.09406851	0.00332034	0.118386999	0.558482648	0.118386999	0.558482648
Income	-0.189425027	0.079472426	-2.383531448	0.021239504	-0.349302926	-0.029547127	-0.349302926	-0.029547127

**Questions:**

1. Interpret the overall significance of the regression model. Report the statistic you used to answer this question and its value. (10 points)
2. Report and Interpret the intercept/constant. (10 points)
3. Report and Interpret the slopes/regression coefficients of the independent variables (10 points)
4. Report and Interpret the model fit (10 points)
5. Formulate and Interpret the regression equation for this model (10 points)
6. Use the model to predict the value of Y given X1=10, X2=50. (10 points)
7. Write a precise summary of your findings to DOS. Remember that your boss is not a data analyst so make sure to articulate your findings in an easy to understand language. (40 points)

**Notes:** Please see below, the description of the variables and a preview of the first 10 records of the data set.

	Population	Income	Illiteracy	LifeExp	Murder	HSGrad	Frost	Area
Alabama	3.615	36.24	2.1	69.05	15.1	41.3	20	50708
Alaska	0.365	63.15	1.5	69.31	11.3	66.7	152	566432
Arizona	2.212	45.3	1.8	70.55	7.8	58.1	15	113417
Arkansas	2.11	33.78	1.9	70.66	10.1	39.9	65	51945
California	21.198	51.14	1.1	71.71	10.3	62.6	20	156361
Colorado	2.541	48.84	0.7	72.06	6.8	63.9	166	103766
Connecticut	3.1	53.48	1.1	72.48	3.1	56	139	4862
Delaware	0.579	48.09	0.9	70.06	6.2	54.6	103	1982
Florida	8.277	48.15	1.3	70.66	10.7	52.6	11	54090
Georgia	4.931	40.91	2	68.54	13.9	40.6	60	58073

Population – population estimate in millions as of July 1, 1975

Income – income per capital as of 1974, in thousands

Illiteracy – percent of population who is illiterate as of 1970

LifeExp – life expectancy in years from 1969-1971

Murder – murder and non-negligent manslaughter rate per 100,000 population as of 1970

HSGrad – Percent of high-school graduates as of 1970

Frost – average days with minimum temperature below freezing between 1930-1960

Area – land area in square miles

### Appendix G: Multiple Linear Regression Assignments – Post-test

**Instructions:** Department of Public Safety (DOS) has hired you as a consultant to help them explore historical data collected from 50 states of the United States of America. They would like to investigate if Murder rate (%) and High School graduation rate (%) affect Life expectancy (number of years). Given the following output from a regression analysis, **as well as using the flow chart attached on the second page of this assignment**, please interpret the results by answering the questions listed below.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.81415359							
R Square	0.662846068							
Adjusted R Square	0.648499092							
Standard Error	0.795871738							
Observations	50							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	2	58.52864627	29.26432314	46.20110018	8.0161E-12			
Residual	47	29.77035573	0.633411824					
Total	49	88.299002						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	70.29708402	1.015667882	69.21266811	5.90943E-49	68.25382379	72.34034424	68.25382379	72.34034424
Murder	-0.237090095	0.035285233	-6.719244063	2.18051E-08	-0.308074828	-0.166105361	-0.308074828	-0.166105361
HSGrad	0.043887299	0.016126888	2.721374355	0.009088366	0.011444185	0.076330413	0.011444185	0.076330413

**Questions:**

1. Interpret the overall significance of the regression model. Report the statistic you used to answer this question and its value. (10 points)
2. Report and Interpret the intercept/constant. (10 points)
3. Report and Interpret the slopes/regression coefficients of the independent variables (10 points)
4. Report and Interpret the model fit (10 points)
5. Formulate and Interpret the regression equation for this model (10 points)
6. Use the model to predict the value of Y given X1=8, X2=70. (10 points)
7. Write a precise summary of your findings to DOS. Remember that your boss is not a data analyst so make sure to articulate your findings in an easy to understand language. (40 points)

**Notes:** Please see below, the description of the variables and a preview of the first 10 records of the data set.

	Population	Income	Illiteracy	LifeExp	Murder	HSGrad	Frost	Area
Alabama	3.615	36.24	2.1	69.05	15.1	41.3	20	50708
Alaska	0.365	63.15	1.5	69.31	11.3	66.7	152	566432
Arizona	2.212	45.3	1.8	70.55	7.8	58.1	15	113417
Arkansas	2.11	33.78	1.9	70.66	10.1	39.9	65	51945
California	21.198	51.14	1.1	71.71	10.3	62.6	20	156361
Colorado	2.541	48.84	0.7	72.06	6.8	63.9	166	103766
Connecticut	3.1	53.48	1.1	72.48	3.1	56	139	4862
Delaware	0.579	48.09	0.9	70.06	6.2	54.6	103	1982
Florida	8.277	48.15	1.3	70.66	10.7	52.6	11	54090
Georgia	4.931	40.91	2	68.54	13.9	40.6	60	58073

- Population – population estimate in millions as of July 1, 1975
- Income – income per capital as of 1974, in thousands
- Illiteracy – percent of population who is illiterate as of 1970
- LifeExp – life expectancy in years from 1969-1971
- Murder – murder and non-negligent manslaughter rate per 100,000 population as of 1970
- HSGrad – Percent of high-school graduates as of 1970
- Frost – average days with minimum temperature below freezing between 1930-1960
- Area – land area in square miles

**Appendix H:  
Rubric for Grading Multiple Linear Regression Pre-test and Post-test**

Grading Rubric: Multiple Linear Regression Assignment							
Student Name:							
	Needs Improvement	Satisfactory	Excellent				
	0-1 points	2-3 points	4-5 points	Weight	Points Possible	Points Earned	Weighted Points
<i>Interpretation of overall significance of the model: F statistic/Null Hypothesis</i>	The significance of the model is not reported using the correct statistic. The interpretation is missing or completely incorrect.	The significance of the model is reported correctly and interpretation is partially correct.	The significance of the regression model is reported correctly and interpreted accurately.	0.1	5		0
<i>Interpretation of the intercept/constant: B0</i>	Intercept/constant of regression line is not reported correctly. The interpretation is missing or completely incorrect for the constant/slope.	Intercept/constant is reported correctly and interpretation is partially correct.	Intercept/constant is reported and interpreted accurately.	0.1	5		0
<i>Interpretation of the slope/regression coefficient: B1 &amp; B2</i>	Slope/regression coefficients are not reported correctly. The interpretation is missing or completely incorrect for the slope/regression coefficients.	Slope/regression coefficients are reported correctly and interpretation is partially correct.	Slope/regression coefficients are reported and interpreted accurately.	0.1	5		0
<i>Interpretation of model fit: R-squared and Adjusted R-squared</i>	R-squared and Adjusted R-squared are not reported correctly. The interpretation is missing or completely incorrect for R squared.	R-squared and Adjusted R-squared are reported correctly and interpretation is partially correct.	R-squared and Adjusted R-squared are reported and interpreted accurately.	0.1	5		0
<i>Formulation and Interpretation of regression equation <math>\hat{Y} = B_0 + B_1X_1 + B_2X_2</math></i>	The equation is not formulated correctly. The interpretation of the equation is missing or completely incorrect.	The equation is formulated correctly and interpretation is partially correct.	The equation is formulated and interpreted accurately.	0.1	5		0
<i>Calculate the predicted value of Y using a specific value of X1 &amp; X2</i>	The predicted value of Y is not calculated correctly.	The predicted value of Y is partially correct.	The predicted value of Y is calculated accurately.	0.1	5		0
<i>Articulate your findings: Write a summary of your findings.</i>	The articulation is incorrect or missing.	The articulation is partially complete and partially correct.	The articulation is complete and accurate.	0.4	5		0
<b>Total Score</b>							<b>0</b>