# Using visualizations for introducing regression concepts

Benjamin Larson
Troy University

Cali Davis
Troy University

Todd Peachey
Troy University

Jeffrey Bohler
Troy University

## ABSTRACT

This case is designed to introduce linear regression and analytics to business students primarily using visualization to explore assumptions, variable types, and results. Since students may have limited exposure to business entities, this case is based on a fictional company that could be explained by discussing simple employment examples such as a fast-food restaurant. The case includes exploring linear regression assumptions using visual representations, results interpretation, and moving into more advanced concepts such as model evaluation to transition to machine learning.

Keywords: Linear Regression, Business Analytics, Statistical Controls, Visualizations, Datasets

Copyright statement: Authors retain the copyright to the manuscripts published in AABRI journals. Please see the AABRI Copyright Policy at http://www.aabri.com/copyright.html

## INTRODUCTION

Linear regression is a method used in data analytics and machine learning. Statistics is a difficult subject for many students and presents challenges in teaching statistics, perhaps using unfamiliar software. We approached this dilemma using platforming and just-in-time Teaching (JiTT) as a pedagogical base to create a series of assignments designed to use visualizations to slowly develop an understanding while providing immediate feedback to the students.

Information from a fictional internal human resources system provides the data for a series of assignments. The case begins with a brief discussion of linear regression, followed by an example fictional story of the dataset, an explanation of the use of SAS software, the potential layouts for assignments, followed by a description of each assignment, the limitations and drawbacks of the case, and finally the potential expansion of the case.

## LINEAR REGRESSION

Most analytics courses dedicate an entire class specifically covering regression (Gorman & Klimberg, 2014). Conceptual understanding is essential; however, most courses use statistics applications such as SAS and SPSS (Gorman & Klimberg, 2014); such applications produce outputs that the analyst needs to interpret rather than calculate. To that end, this case is not meant as a replacement of a statistics course, but a series of exercises to augment understanding and application of regression concepts by using practical examples. As linear regression is a complex topic, the focus is on a simplified evaluation of the assumptions, parameter estimates interpretation, and model assessment. This case assumes that the instructor has a base knowledge in linear regression, however a brief explanation is provided.

Four assumptions are connected to linear repression residuals independence (assumed in our case), linear relationship, homoscedasticity, and the normal distribution of the error terms (Kutner et al., 2005). Additionally, outliers should be explored, and the model assessed to determine if variables are missing normality (Kutner et al., 2005). Linear regression is a versatile model used either as predictive or descriptive model normality (Kutner et al., 2005). Additionally, analysis of outliers may be used in statistical controls (Kutner et al., 2005). While there are formal tests for the assumptions, preliminary evaluations may be made through informal review of residual plots and scatter plots.

Students must comprehend that linear regression generates parameter estimates and corresponding p-values that can test a hypothesis and generate an equation used to predict a dependent variable. There is a trend in a linear statistical relationship plus a constant, plus some error that can be visualized as the "scatter" in a scatterplot. The fitted equation is a modification of our statistical linear equation to find the best fitting line that best "describes" the relationship between the predictor variable(s) (independent variable) and the response variable (dependent variable). Data analysis software like SAS and Excel do this by adjusting the line's position and the slope until the sum of all the squared errors (the difference between predicted and observed responses) has been minimized. Review the following linear equation terms.

Statistical linear equation:
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$y_i$ = the observed response value (DV value)

$\beta_0$ = estimated population regression line constant

$\beta_1$ = estimated population regression line slope

$x_i$ = the predictor variable value (IV value)

$\varepsilon_i$ = error term (difference between $\hat{y}_i$ and $y_i$)

Best Fit linear equation

$\hat{y}_i = b_o + b_1x_i$

$\hat{y}_i$ = the predicted response (or fitted value)

$b_o$ = the estimated Y axis intercept of the best fitting line

$b_1$ = the estimated slope of the best fitting line

$x_i$ = the predictor variable value (IV value)

Thus, the need to determine if $\beta_1$ is not equal to zero ($\beta_1 \neq 0$). Consequently, hypothesis testing is performed on each β by examining if the p-value is less than the accepted α. Given a null hypothesis of $\beta_1 = 0$ and an alternative hypothesis of $\beta_1 \neq 0$, one would reject the null and accept that there is a relationship with $\beta_1$ if the p-value is less than α and the confidence interval does not contain zero. While the accepted α depends upon the problem, a value of 0.05 is used for this case.

It is also crucial for students to understand how to assess the model. To this end, we focus on three terms, which are R-Squared, Adjusted R-Squared, and Root Mean Square Error. R-Squared ($R^2$) or the coefficient of multiple determination measures the variance that the model explains. Adjusted R-Squared (Adjusted $R^2$) adjusts for adding another independent variable so that the model may be evaluated for if adding variable is a significant improvement in the variance explained. Root Mean Square Error (Root MSE) is the standard deviation of the residuals. The individual assignments will list the learning objectives and terms covered.

## THIS CASE

This is a fictitious case. All information contained herein was fabricated by the authors. Any similarity contained herein to actual persons, businesses, events, etc. is purely coincidental and is the responsibility of the authors. Please contact the case authors directly with any concerns.

The dataset was generated using the personal experiences of the authors but is not based upon a specific example. However, it may benefit the students to have the data presented with a narrative. Two narratives that may accompany the case are provided.

### Narrative 1

Your Tech Support provides technical services to residences and small businesses in ten locations across the southeastern United States. Two primary trends led to the formation of this business. There has been a long-term trend of more people working remotely and this trend has increased significantly in 2020. Additionally, significantly more affordable electronics are available to the average homeowner.

There are 400 employees in the company including administrative support and technicians. There are 60 people in the company that provide administration, scheduling, training, and other support services for the technicians. The 340 technicians provide services such as setting up a home network for work, security, entertainment, and convenience. While many people have no problem setting up their home network to meets the needs of their job or just adding conveniences in their home, other people find this challenging. Connecting a home assistant to a security cameras, smart TV, and other appliances is simply beyond the technological ability of many people. These are our customers.

The CEO recently read an article in a business magazine that suggested that height can have a positive influence on salary. There is no reason this should be true in this company. The CEO wants to know if this is the case in Your Tech Support.

**Narrative 2**

Tutors & More, LLC provides academic tutoring and consulting for parents and their high school aged children in ten locations across the southeastern United States. Competition for admission into elite universities has changed in the last few years. Currently, there is more interest from these universities in well-rounded students that offer other qualities such as a record of community service and other activities more so than just academic achievement.

There are 400 employees at Tutors & More, LLC. Tutors and academic consultants make up 340 of the employees. The other 60 employees provide scheduling, training, marketing, and other administrative support. Tutoring services have been available for high school students for many years. However, Tutors & More, LLC offers other services such as admissions application consulting where they provide coaching and mentoring so the student can submit an application that accurately reflects his or her skills and academic capabilities.

The general business model is that tutors meet with parents and students early in their high school years and develop possible plans for the student to pursue. Tutors & More, LLC builds on-going relationships with their clients so they can refine plans for students as their interests change throughout high school.

The CEO recently read an article in a business magazine that suggested that height can have a positive influence on salary. There is no reason this should be true in this company. The CEO wants to know if height and salary are related at Tutors & More, LLC.

**SAS**

SAS Studio is a web-based application. Many data analysts may have programmed using base SAS with an installation on their PC desktop or SAS server. SAS Studio allows the user to write and run SAS code using a web browser, with little to no installation needed. With SAS Studio, users can access datasets and libraries, write new programs and edit or update existing programs. Base SAS software runs behind the scenes of SAS Studio. When commands are entered into SAS Studio, a cloud-based SAS server executes the commands and processes the code. Results are returned to SAS Studio on the local machine’s web browser. This allows for students to utilize the software without installing software. However, it does present version control issues where if the software changes then the instructions may need to change without notification. This case is therefore presented without detailed instructions on running the specific software. Some familiarity such as uploading the dataset to SAS is assumed. Students currently may log in or create an account at https://welcome.oda.sas.com/login.

**ASSIGNMENTS**

JBCA

The assignments are designed using scaffolding and just-in-time teaching as their pedagogical foundations. While correct answers may vary, some general responses to the questions are provided in each assignment.

Scaffolding implies that learning is supported by an external factor such as a mentor, video, or software to allow a learner to solve a problem that they otherwise could not. Scaffolding frameworks have been developed to help to build learning technologies that provide a foundation for pedagogical design and includes the concept of breaking down the problem using three steps, first identifying the various aspects of the problem, second identifying the issues the learners may have with the various aspects, and finally determining how the scaffolding could overcome those issues (Quintana et al., 2004). The goal is to allow students to perform independent research, thus it is important that case assignments served as platforming for the student's research project as well as building upon themselves.

Just-in-time teaching (JiTT) allows instructors to provide information on a concept with students then actively using the concepts and being provided prompt feedback and has been found to be more effective in teaching statistics than other more traditional methods (McGee et al., 2016). Additionally, JiTT also allows for improvements in soft skills such as problem-solving (Turnip et al., 2016). The design principals for JiTT has similar design components to platforming in that its principals contain sequencing, training wheels, and completion strategy (Novak, & Beatty, 2016). Thus, the assignment sequenced proceeded from basic analysis to more complex concepts so that each assignment was manageable and had meaningful results. The assignments initially involved visually exploring the data with scatter plots to familiarize the students with concepts such as outlier and how different datatypes performed in a regression and met general assumption. The exploration of the data was continued by running summary statistics and exploring outliers to determine if the data is appropriate. Next a regression was performed on the given data, and finally additional models were assigned to explore additional model assessment strategies. Additionally, the assignments are designed to be easily graded, facilitating quicker feedback to students, and providing learning assessment that may be incorporated into future lessons.

The dataset was designed using previous real-world business experience. The dataset allows for minor changes to the data to check for plagiarism between student assignments or course sections, as well as flexibility across multiple terms or semesters.

Another important aspect that the authors desired to address was to add training on soft skills into the assignments. Soft skills such as communication of results are vitally important in analytics (Dubey & Gunasekaran, 2015). To this end, students were required to format the results in a professional manner. Open ended questions are also provided to which the students must address with the tables and figures. Finally, instructors do have the option to organize the questions that need to be answered into the form of a memo that may be presented as the final deliverable. The following is a general description of each of the assignments.

## Assignment 1 - Scatter plots

Summary of the assignment: Students will need a SAS user account to access the SAS Studio software as well as to upload the dataset. If they do not have experience with the software, instructions will need to be provided. While the initial assignment is not necessarily scenario based, it can be connected to the initial exploration of if linear regression is appropriate for this dataset.

Learning objectives:

- Use scatter plots as an initial exploration of the data.
- Understanding the linearity assumption.
- Understanding of errors in the data.
- Understanding the impact of outliers:
  - Outlier and variance exploration for feature selection.
  - Outlier exploration as statistical controls.
- Understanding the impact of group classification and continuous variables.
- Understanding of the unequal variance assumption.
- Understanding the normal distribution.

Context for use: Scatter plots are a standard part of the visual displays of data discussed in an introductory statistics or data analytics course. Scatter plots show relationships (or no relationship) between two quantitative variables, expose outlying and unusual data values, and help the student begin to consider assumptions associated with linear regression. The data could be modified for other higher-level courses in regression by imputing data in order to make a case, such as a dataset that requires transformations before linear regression analysis. This assignment should take a student about an hour or less of independent work and takes up about 30 minutes during class.

Description and Teaching Materials: Students do not necessarily have to be assigned any reading material related to scatter plots. They should be provided a basic background of the assumptions of linear regression. Scatter plots are typically covered near the beginning of a statistics or data analytics course and may help students that identify as "visual learners." Students must have access to the data either by having it preloaded into a course or a student can upload the file.

Step 1: Create a series of scatter plots

Once the software is running and the dataset is open, it is time to create the scatter plots. Remember, scatter plots are a useful tool for examining possible relationships between two quantitative variables. This dataset has eight attributes, seven of which are quantitative or numeric. Four scatter plots will be created in this exercise:
Scatter plot 1: X variable = Performance Review; Y variable = $/H.
Scatter plot 2: X variable = Performance Review; Y variable = $/H; filter Position<3.
Scatter plot 3: X variable = Performance Review; Y variable = $/H; filter Position<3; Group variable = Position.
Scatter plot 4: X variable = Performance Review; Y variable = $/H; filter Position=1.
See Figure 1 (Appendix).

Step 2: Analyze the scatter plots

Scatter plots are a great visual for statistics students, especially those in business courses. These four scatter plots offer some insight into the full analysis of the data, which will follow in the form of the Descriptive Statistics, Correlation and Linear Regression sections of this case. The scatter plots are progressive, meaning they represent a thought process based on a visual exploration of the data. Students should recognize the basic concepts of data analysis and graphical displays, as well as the impact of other variables.

Basic Concepts - Start this part of the assignment by pointing out some basic concepts in the first graph (Figure 1 - Scatter plot: Salary and Performance – graph of the overall data for all positions)

- Pattern: Tell the students to look for pattern in the scatter plots. Does the data appear to be positively related – moving upwards from the left to the right? Or negatively related – moving downwards from the left to the right?
- Linear: Are the points following along a general linear pattern? How close are the points to the line? Use the fitted line on the graph to guide this discussion. If not, can another variable help explain this? Should the data be transformed?
- Error: Are any points far away from the line? Does this mean we made an error in the analysis?
- Constant Variance: Does the data seem to have a consistent spread/dispersion across all Performance Reviews? If not, can another variable help explain this?
- Normal Data: Normality in statistics makes everything work out nicely. Are most of the observations in the 95% prediction interval? What is the shape of a normal distribution? Can symmetry be applied to scatter plots?
- Unusual Observations (or Outliers): Why would some of the points be so different? Look at the data point with a high salary and high-performance rating (~$200,000; 19). Which employee could this be? Reexamine the dataset to explore. Remove the outlier, the CEO of the company.
- Continuous and Categorical Variables: Continuous variables are numeric variables where the averages between the values are meaningful and consistent. Categorical variables may be either numeric or alpha numeric and provide a label or name for a record.

Teaching Notes: Using all the scatter plots, ask the students to respond to the following questions. These may become part of a graded assignment for the students.

Question 1. Which graph has the smallest average error? The scatter plot showing only Position 1's Salary and Performance has the smallest average error. While we are not (in this first assignment) given the exact numerical value of the error, students may comment that the amount of variation around the fitted line on the scatter plot is the smallest in this graph.

Question 2. How would you use the graph to describe how a continuous variable (Performance Review) influences the dependent variable (Salary)? The relationship, and therefore the influence, of Performance on Salary can be seen in all four scatter plots. In the first scatter plot, there is a small, positive relationship (graph has slight upward trend from the left to the right) and this relationship only get stronger as the scatter plots become more specific. Even with the y-axis scale change, the scatter plot for only Position 1 shows a strong positive relationship (students may also mention correlation here). This indicates that the Performance Review does influence Salary. While we cannot have a cause and effect relationship stated, the student should mention that the variables do have a strong positive relationship.

Question 3. How would you use the graph to describe how a categorical variable (Position) influences the dependent variable (Salary)? The influence of Position on Salary is best shown in the scatter plot of Grouped by Positions. This scatter plot gives two fitted lines – one fitted line for Position 1 and another fitted line for Position 2. Both fitted lines have a slight positive slope, indicating a small positive relationship. However, the lines are completely difference with respect to their y-intercept value. The mention here of slope and y-intercept will become even more familiar in the Correlation and Linear Regression Assignment.

Question 4. How can an examination of unequal variance influence feature selection? Not only do the scatter plots show differences in features examined, they also provide some textual output commenting on the variance of the y-values across the x-values. Recall from the basic concepts section, that students should have commented that the data do not seem to have a consistent spread/dispersion across all Performance Reviews. This discussion should have also included the fact that some observations with higher performance reviews also appear to have the highest salaries. Leading the student to think about business structure of employees at all levels (Position).

Question 5. How can prediction intervals be used in statistical controls? Prediction intervals are provided on the scatter plots, as bands parallel to the fitted line. The prediction intervals give a visual representation of "being in control" or "being an acceptable distance from the mean" (the fitted line). While none of the scatter plots show all points within the 95% prediction intervals, this is expected.

The case can be made in the scatter plot for Position 1 that all employees do not have the same relationship between salary and performance. Allow students once again to rely on their business acumen to discuss cases of high salary/low performance, low salary/high performance, etc. The value of 95% is a standard evaluation level (α =level of significance = 1-0.95 = 0.05). The students should be aware that the interpretation of these intervals is that of future employees evaluated, 95% would fall inside the prediction intervals, with 5% falling outside the prediction intervals.

Question 6. How can scatter plots be used to evaluate normal distribution? If the normal distribution has already been covered in the class, then mention how the distances from the fitted line (the errors) can be plotted in a histogram to visually assess normality. If the normal distribution has not been covered in the class, then focus on what and how much error should be acceptable.

Since error is simply a measure of how much the data points vary about the fitted line, students should be able to rationalize that most of the data points should vary very little, with a few not varying at all and a few varying more than most. This discussion can create the visual of a bell-shaped distribution where the center (or mean) is at some small value representing little variation and the ends (or tails) represent no variation (0) or a bit larger variation.

Question 7. How can scatter plots be used to evaluate linearity? If the relationship between the x variable and the y variable is linear then there will be no curves or bends in the shape of the data.

## Assignment 2 - Descriptive statistics

Summary of the Assignment: This assignment requires the students to run the basic summary statistics of the data set to include sample size (N), mean, standard deviation, maximum value, minim value, range, skewness and kurtosis. Additionally, the students must create histograms of the $/H variable. If there are outliers, the students must filter the data set to remove them and run the histograms and summary statistics again.

Learning objectives:

- Understanding the role of skewness and kurtosis in understanding normal distribution.
- Understanding the role of min in max in model appropriateness.

- Understanding the impact of outliers.
- Understanding the importance of appropriate filtering.
- Understanding the impact of missing values.
- Understanding need for the number of records.

Context for Use: Descriptive statistics provide information that is required to ensure the dataset meets the assumptions of normality before you can proceed to linear regression. Descriptive statistics also allow filtering of the data to remove outliers that could have an excessive influence on the regression. This assignment should take a student less than two hours or less of independent work and takes up about an hour of time during class.

Description and Teaching Materials. After the students have created the histograms and the data summary tables, they should answer the following questions. These may become part of a graded assignment for the students.

Question 1. How would you use the histograms to describe skewness and kurtosis? A histogram presents the categories on the x axis and the frequencies on the y axis. In the full dataset, the histogram shows many outliers. Most of these outliers are in the right tail which suggests a high positive skewness. A skewness value near 0 is optimal. As skewness values increase or decrease to more than 1 or less than -1, there is evidence of excessive skewness. Additionally, there are a large number of observations in the tails of the distribution that will have a high kurtosis value. An acceptable kurtosis value is near 3 if not standardized and near 0 if standardized. The first distribution shows excessive kurtosis, and the second distribution shows an acceptable value for kurtosis.

Question 2. How would you use the histograms to describe the impact of filtering the dataset on the wage distribution? After filtering the dataset, the histogram shows only a slight positive skewness with a number near 0, and a much lower kurtosis with a small negative number. The removal of the outliers allows us to proceed with the analysis since the dataset now approximates a normal distribution.

Question 3. Location and Position are numeric variables. Are their means and standard deviations meaningful and why/why not? The means and standard deviations of Location and Position are not meaningful. In this data set they are presented as numbers; however, they could have just as easily been presented as text and therefore classified as nominal variables.

Question 4. Why does the line for position 2 in Assignment 1 not continue across the graph? How can we apply this concept using the summary statistics table? The predicted line should only be used for to estimate values that are within the range of the dataset. This underscores the need to provide a summary table that contains a minimum and maximum. See Figure 2 – Salary per hour histograms, Table 3 – Descriptive statistics with outliers, and Table 4 – Descriptive statistics without outliers in the Appendix.

### Assignment 3 - Correlation and linear regression

Summary of the Assignment: For this assignment the scenario is that the student is tasked to build a linear model that may be used as a statistical control for the wages of the company for positions one and two. The students should generate scatter plots for continuous variables as used in assignment 1 and provide summary statistics as used in assignment 2. Students will need to run a correlation analysis and a linear regression model.

Learning Objectives:

- Understanding correlation analysis for feature selection.
- Understanding visually evaluating assumptions based on residual plots.
- Understanding the interpretation of parameter estimates:
    - Hypothesis testing.
    - Best Fit Equation.

Context for Use: Linear Regression is a standard model that may be used in description, prediction, or statistical controls which may be rudimentarily introduced in an introductory statistics or data analytics course. The data is designed for beginning learners but could be modified for other higher-level courses in regression by imputing data in order to make a case, such as a dataset that requires transformations before linear regression analysis. This assignment should take a student about two hours or less of independent work and takes up about an hour of time during class assuming the concepts from Assignments 1 and 2 have been covered. Description and Teaching Materials. Students should be provided a brief presentation or reading which walks them through how parameter estimates relate to the best fit equation. Linear regression may be presented towards the end of an introductory statistics course or at the start of a secondary course after concepts such as normal distribution, statistical significance, and hypothesis testing are covered. As the assignments are meant to build on each other, the students are expected to be able to use materials from prior assignments or to generate new material using skills gained in prior assignments.

Step 1: Create a filtered dataset of position < 3 to be used as the dataset in all subsequent steps (if not already done in assignment 1 or 2).
Step 2: Create scatter plots to explore the data (if not already done in assignment 1 or 2).
Step 3: Create summary statistics (if not already done in assignment 1 or 2).
Step 4: Run a correlation analysis on all continuous variables in the dataset. See Table 4 – Correlation analysis in the Appendix.
Step 5: Analyze the correlation table. Correlation is an important concept in statistics as it provides a measure of the strength of the linear relationship between two quantitative variables.

Basic Concepts - Start this part of the assignment by pointing out the diagonal which indicates the variable's relationship with itself. Explain that the closer the value is to an absolute value of one the stronger the relationship. An absolute value of one represents a perfect correlation. The higher the absolute value of the correlation to the dependent variable the better the variable will be as a predictor and the more likely you want to include the variable in the model. The students should also note issues with multicollinearity and exclude one of the independent variables with high correlations (>|0.9|).

- Variable or Feature Selection: The process that determines which of the variables or features are included within a model.
- Multicollinearity: Multicollinearity is when there is a high correlation between two or more independent or predictor variables. In other words, the variables may measure the same thing. This will not distort the overall accuracy of the model predictions but will distort how the individual parameter estimates appear and may make them counterintuitive so it should be avoided, and a variable excluded if you need to interpret the relationships.

Step 6: Run a linear regression with $/H as the dependent variable, performance and years of employment as the continuous variables and location as a classification or categorical variable.

JBCA

Step 7: Evaluate the model for significance and analyze the predicted model and residual plots for meeting model assumptions.

Basic Concepts – Start by verifying that the students understand the concepts of significance and hypothesis testing. The p-value for the model should be evaluated for significance. If the model is significant, then the assumptions need to be evaluated using knowledge gained from Assignment 1. The students should note that variance for the observations that correspond to Position 2 have larger variance. Students should be encouraged to proceed with the analysis of the model. See Figure 3 – Fit Diagnostics for $/H.

Step 8: Assess the model. See Table 5 – Model Assessment for Positions 1 and 2 in the Appendix.

Basic Concepts – The authors recommend keeping model assessment simple at this point. Focusing on Root MSE and R-Squared. R-Squared represents the amount of the variance that the model explains. The value ranges from zero to one where one represents 100% of the variance explained. The students should be made aware that a model with an $R^2$ of one is inappropriate as it is a deterministic rather than a statistical relationship. The root MSE is the standard deviation of the error terms and students should desire to make the value as small as possible. Further evaluation should be done with the knowledge of the normal distribution and the problem. Roughly 95% of the errors should be within ± two root MSE of the mean.

Step 9: Analyze the parameter estimates. See Table 6 – Parameter estimates for Positions 1 and 2 in the Appendix.

Basic Concepts – Analysis of the parameter estimates occurs in two parts. First the students must evaluate the significance of each $\beta$ to test the null hypothesis that $\beta=0$. If the p-value is less than the assigned $\alpha$ and the confidence limit does not contain zero, then reject the null and accept the alternative that $\beta\neq0$ and therefore there is a relationship between the independent and dependent variable in the model. The student should understand that the parameter estimate is the best estimate of the slope and that it is important to note if the estimate is meaningful. For example, performance review appears to be significant with a p-value <0.0001 and a confidence interval that does not contain zero. Performance review has a parameter estimate of 0.22 which indicates that for each additional point for performance review we would add $0.22 per hour to the estimated wage. By using the minimum and maximum from the summary statistics it can be estimated that the value added by performance ranges from $0.00 to $4.40 and should be considered meaningful in that it would mean a real impact on the expenses of the company.

- Parameter estimate: The most likely value of the relationship between an independent and dependent variable.
- Confidence limits: The high and low values of the parameter estimate in which there is a set probability that the actual parameter estimate will fall in-between.
- P-Value: P-value is a measure of the probability that an observed difference could have occurred just by random chance. If the p-value is less than the α then we reject the null hypothesis.
- Alpha (α): The threshold value needed for significance.

Teaching Notes: Using the correlation table and regression results, ask the students to respond to the following questions. These may become part of a graded assignment for the students.

Question 1. How is the correlation matrix used in feature selection? An independent variable with a high correlation with the dependent variable is likely to be a good predictor and should be included in the model, while independent variables with a correlation closer to 0 likely have little effect and may be left out of the initial analysis. If there are two independent variables that are highly correlated, they may be measuring the same thing and therefore may cause the model to produce counter intuitive parameter estimate and one of the two should be left out of the model.

Question 2. Does height appear to be a good predictor of any of the existing variables? No, height has a low correlation to all the other variables.

Question 3. How can you visually test the assumptions of equal variance, linearity, and normal distribution of error terms and does this model appear to meet assumptions? We would expect the observed versus predicted plot to show that the observed values surround the predicted line. Areas where the observed values are constantly above or below may indicate problems with the assumption of a linear relationship. Areas where the observations become wider or narrower indicate a violation of the assumption of equal variance (homoscedasticity). To test the normal distribution of the error terms, student would expect most of the error terms to fall within two root MSE of the predicted line and that the histogram of the residuals to show a traditional bell shape to the distribution. There appears to be a wider variance with the observations that are associated with position two versus position one. This indicates a potential issue with homoscedasticity.

Question 4. How would the parameter estimates show that a parameter is both significant and meaningful (Chose one parameter as an example and use $\alpha = 0.05$)? Each parameter should be evaluated for significance first by checking if the p-value is <0.05 and that the confidence limits calculated at 95% do not contain 0. For example, performance review appears to be significant with a p-value <0.0001 and a confidence interval that does not contain zero. Performance review has a parameter estimate of $0.22 which indicates that for each additional point for performance review we would add $0.22 per hour to the estimated wage. By using the minimum and maximum from the summary statistics, it can be estimated that the value added by performance ranges from $0.00 to $4.40 and should be considered meaningful in that it would mean a real impact on the expenses of the company.

Question 5. How would the $R^2$ for the model be interpreted and what are the implications? An $R^2$ of 0.91 indicates that 91% of the total variation of the wage is explained. This indicates that the model explains most of the variance. The high value of variance explained indicates that the model can be used in prediction of wages of an employee.

Question 6. How would this model be used in the statistical control process? The distribution of the error terms is approximately normal and 95% of the observations should fall within two root MSE of the predicted value. Any value that is seen outside of this range may be questioned as an outlier and explored.

Question 7. Would one be concerned with the salary of the given employee (Position=1, Performance=10, years of employment = 1, location= 9 and salary = 13)? The estimated salary of the employee would be $13.58 given the parameter estimates and the values for the employee. This indicates that the absolute value of the error for the employee is $0.58 which is within one root MSE and the extent of the error would be expected and therefore would not be an outlier that needs additional exploration.

**Assignment 4 - Modified linear regression and model assessment**

For this assignment the scenario is that the model from Assignment 3 is under scrutiny as the company has been accused of discriminating by height for position 1 employees. The supervisor desires to run the model only for position 1 and determine which is a more appropriate model. The supervisor is interested in if the company discriminates by height.

Learning objectives:

- Model Assessment:
  - $R^2$.
  - Root MSE.
- The relationship between variable significance and model assessment.
- Problem Domain.
- Control Variables.

Context for Use: Linear Regression is flexible in how it is used and is widely used in machine learning. This assignment is useful both in traditional statistics and in transitioning to machine learning to understand the concepts of identifying the data for the problem domain and to demonstrate feature selection using model assessment. This assignment should take a student about two hours or less of independent work and takes up about an hour of time during class assuming the concepts from Assignments 1 and 2 have been covered.

Description and Teaching Materials: Students may be provided a brief presentation on problem solving and identifying the problem domain. The students also should be explained the idea of control variables as the base model before height is added controls for variables that are expected to influence wage. The concepts remain the same as assignment 3 with the additional consideration that the students should be encouraged to maintain their model assessment statistics in a table that will allow for easy comparison.

JBCA

Step 1: Create a filtered dataset by position = 1 to be used as the dataset in all subsequent steps (if not already done in assignment 1 or 2).
Step 2: Create summary statistics (if not already done in assignment 1, 2 or 3).
Step3: Run a linear regression with $/H as the dependent variable, performance and years of employment as continuous variables and location and position as classification or categorical variables.
Step 4: Evaluate the model for significance and analyze the predicted model and residual plots for meeting model assumptions.
Step 5: Assess the model and compare it to the model from assignment 3.

Basic Concepts – The authors recommend first a discussion of the problem domain. By having the scenario written towards only position one, the data should be limited to examining that position. The comparison between the two models should show the effects of unequal variance between positions one and two. The Root MSE declines as does the $R^2$. This is a function of there being less variance to explain in the filtered dataset. The reduction in the Root MSE is large considering the mean of the data. This should stand out and cause the student to consider more than $R^2$ when assessing the model. By comparing the observed versus predicted of both models it should be clear which model better matches the assumptions of running a linear regression model. This is a continuation of the theme of not only feature selection but matching the data to a specific problem domain. See Figure 6 – Visual model assessment in the Appendix.

Step 6: Run a linear regression with $/H as the dependent variable with height, performance and years of employment as continuous variables and location as a classification or categorical variable.
Step 7: Evaluate the model for significance and analyze the predicted model and residual plots for meeting model assumptions.
Step 8: Assess the model and compare it to the model from step 3.

Basic Concepts – A new variable should only be added if it improves the accuracy of the model. It should be evident to the students that root MSE and $R^2$ did not improve. This suggests that the model is not improved by the addition of the variable. This is also evident in adjusted $R^2$ which is the appropriate measure for evaluation as it accounts for the addition of additional variables. The improvement of adjusted $R^2$ would indicate a significant change in the model. It is this kind of evaluation that needs to occur if the analysis moves towards larger dataset and is conducted using machine learning. Since this paper cases uses traditional methods, what will we expect when we look at the parameter estimates for height? See Table 7 – Comparison of models in the Appendix.

Step 9: Analyze the parameter estimate for height. See Table 8 – Parameter estimates with height added to the model.

Basic Concepts – One should expect that height is insignificant. It has a poor correlation as noted in the correlation analysis and did not explain any more variance. In testing the null hypothesis that $\beta=0$, the p-value is 0.37 and that the confidence interval contains zero. This should indicate that in general insignificant variables will not improve the model and should be eliminated. As students may be transitioning to larger dataset and other forms of analysis such as machine learning, it is important to understand the impact on model assessment rather than solely relying on P-values.

JBCA

Teaching Notes. Using the regression results, ask the students to respond to the following questions. These may become part of a graded assignment for the students.

Question 1. Which model is more accurate and why? The model that limits the analysis to position one represents a better model as it better fits the problem domain. While it shows that it explains less variance due to a lower $R^2$, it also has less variance to explain. The lower root MSE indicates a lower variance in the error terms.

Question 2. Does the company discriminate by height? The company does not appear to discriminate by height. The height is insignificant in the model as the p-value of 0.37 is greater than $\alpha$ and the confidence interval contains zero. It also does not improve the amount of variance explained in the model.

Question 3. Is $R^2$ or root MSE more important in the evaluation and why? Both measures are important, but in this case with a relatively high $R^2$ it could be argued that the root MSE is more telling in that it demonstrated that there was a significant reduction in the standard error. Both measures should be evaluated.

Question 4. Why is it important that the $R^2$ is high in exploring concepts such as discrimination? By having a model that controls for a large amount of the variance and controls for known predictors it allows a fair evaluation of a new independent variable.

Question 5. What other analysis could be examined to determine if height is indirectly influencing wages? The correlation analysis can be used to evaluate if it is influencing factors such as years of employment and performance scores. One could also examine the means of the categorical variables to determine if height is different by position or by locations.

## BENEFITS AND LIMITATIONS OF THE CASE

The case is based upon a fictitious dataset that allows the data to be altered to present an infinite number of solutions. While the background stories and data are fictious they are realistic and illustrative. While the values are meant to be representative of a certain employment scenario the trends in the case are by no means universal in industry. In this way, the dataset does lend itself to future modification and for discussion. For example, changes in minimum wage or wage inflation rates may lead to a nonlinear relationship that need to be transformed or made into categories.

## POTENTIAL FOR CASE EXPANSION

There are several ways in which this case may be expanded upon. For example, students could utilize the dataset to perform an ANOVA prior to or after running linear regression. This allows for the exploration of categorical variables such as location or position. This may be paired with exploration of transforming a variable that is nonlinear to better match the data. For, example evaluating if different ranges of performance scores or years of service acted more as categorical variables than continuous variables. Additionally, the concept of interactions may be explored such as between performance and years of service. Concepts such as secondary effects may also be explored examining if insignificant factors in our model may also be indirectly influencing through their relationship to significant factors such height may be insignificant in the final model, but it may influence performance review or length of service. Additional techniques such as stepwise regression may be demonstrated to further confirm results like those in Assignment 4.

JBCA

## REFERENCES

Dubey, R., & Gunasekaran, A. (2015). Education and training for successful career in big data and business analytics. *Industrial and Commercial Training*, 47(4), 174-181.

Gorman, M., & Klimberg, R. (2014). Benchmarking academic programs in business analytics. *Interfaces, 44*(3), 329–341. doi:10.1287/inte.2014.0739

Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2005). Applied linear statistical models (Vol. 5). New York: McGraw-Hill Irwin.

McGee, M., Stokes, L., & Nadolsky, P. (2016). Just-in-time teaching in statistics classrooms. *Journal of Statistics Education*, *24*(1), 16-26.

Novak, G., & Beatty, B. (2016). Designing just-in-time instruction. Instructional-Design Theories and Models, Volume IV: The Learner-Centered Paradigm of Education, 415.

Quintana, C., Reiser, B., Davis, E., Krajcik, J., Fretz, E., Duncan, R., Kyza, E., Edelson, D., & Soloway, E. (2004). A scaffolding design framework for software to support science inquiry. *The Journal of the Learning Sciences*, 13(3), 337-386.

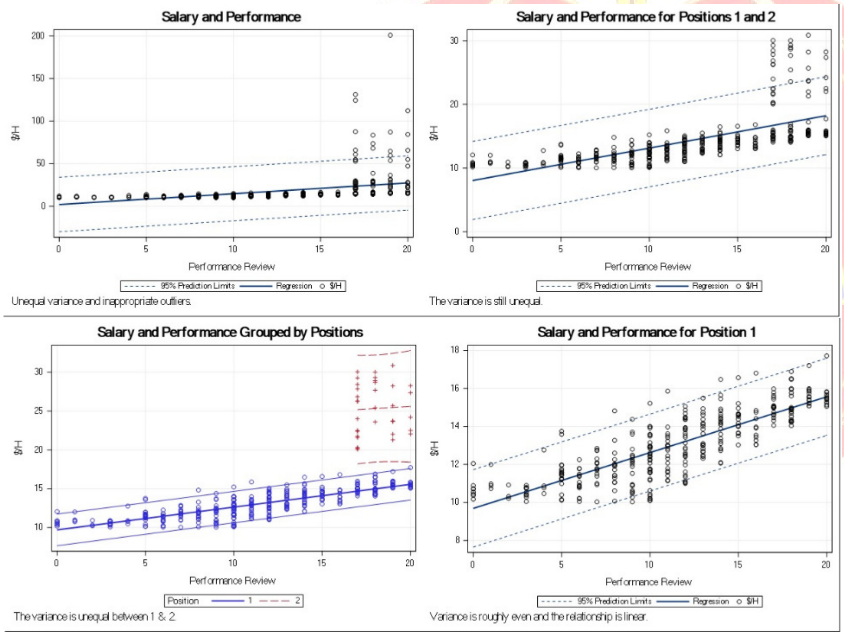
Turnip, B., Wahyuni, I., & Tanjung, Y. I. (2016). The Effect of Inquiry Training Learning Model Based on Just in Time Teaching for Problem Solving Skill. *Journal of Education and Practice*, *7*(15), 177-181.

## APPENDIX

Table 1
*Example of data set*

| Location | Position | $/H | On the Job Errors | Performance Review | Years of Employment | Height | Gender |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 10.77 | 16 | 4 | 1 | 75 | F |
| 4 | 1 | 15.53 | 3 | 17 | 10 | 73 | M |
| 8 | 1 | 12.28 | 10 | 10 | 2 | 63 | M |
| 6 | 1 | 12.07 | 10 | 10 | 3 | 73 | F |
| 5 | 1 | 12.09 | 13 | 7 | 2 | 74 | M |
| 4 | 1 | 12.03 | 12 | 8 | 3 | 63 | F |
| 1 | 1 | 10.99 | 15 | 5 | 1 | 71 | F |
| 9 | 1 | 17.72 | 0 | 20 | 5 | 77 | M |
| 11 | 3 | 54.75 | 3 | 17 | 8 | 72 | F |
| 6 | 1 | 13.64 | 8 | 12 | 5 | 74 | M |

Figure 1
*Scatter plot: Salary and Performance*


Salary and Performance

Unequal variance and inappropriate outliers.

Salary and Performance for Positions 1 and 2

The variance is still unequal.

Salary and Performance Grouped by Positions

The variance is unequal between 1 & 2.

Salary and Performance for Position 1

Variance is roughly even and the relationship is linear.

Figure 2
*Salary per hour histograms*

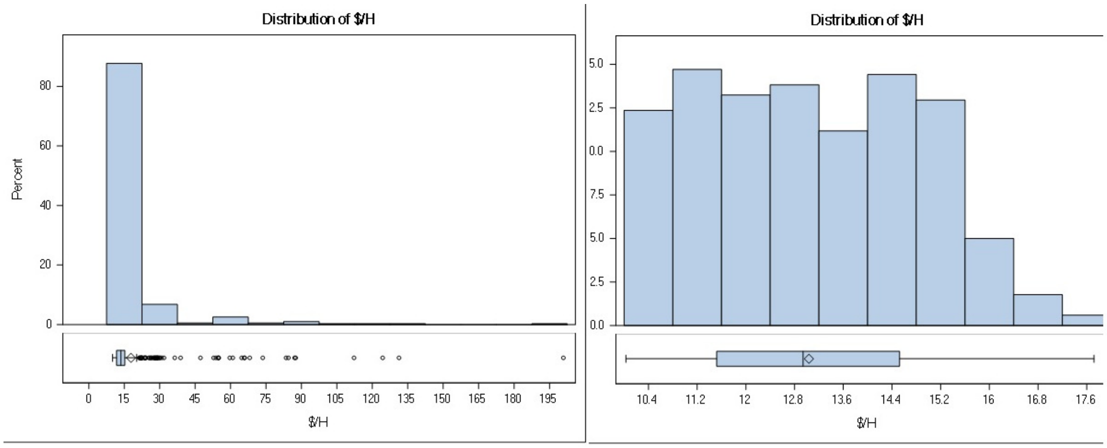


Table 2
*Descriptive statistics all Positions*

| Variable | N | N Miss | Mean | Std Dev | Min | Max | Range | Skew | Kurtosis |
|---|---|---|---|---|---|---|---|---|---|
| $/H | 400 | 0 | 17.85 | 17.41 | 10.03 | 200.88 | 190.85 | 5.69 | 42.11 |
| On the Job Errors | 400 | 0 | 7.58 | 5.22 | 0 | 20.00 | 20.00 | 0.39 | (0.72) |
| Performance Review | 400 | 0 | 12.42 | 5.22 | 0 | 20.00 | 20.00 | (0.39) | (0.72) |
| Years of Employment | 400 | 0 | 4.09 | 2.46 | 1.00 | 10.00 | 9.00 | 0.94 | 0.05 |
| Height | 400 | 0 | 69.21 | 5.42 | 60.00 | 78.00 | 18.00 | (0.04) | (1.15) |

Note: High skewness and kurtosis with outliers

Table 3
*Descriptive statistics only Positions 1 and 2*

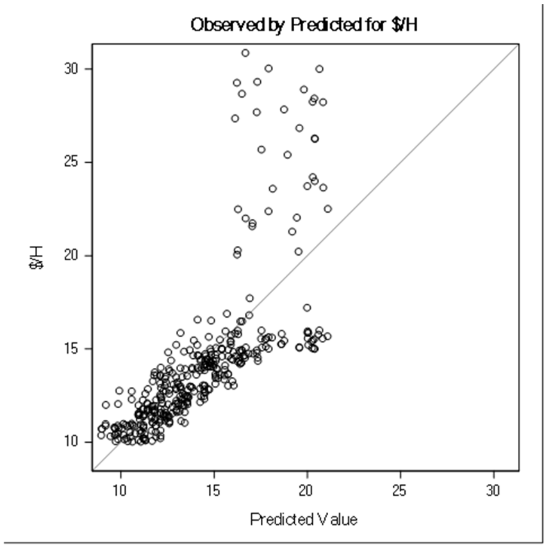
| Variable | N | N Miss | Mean | Std Dev | Min | Max | Range | Skew | Kurtosis |
|---|---|---|---|---|---|---|---|---|---|
| $/H | 340 | 0 | 13.04 | 1.79 | 10.03 | 17.72 | 7.69 | 0.12 | (1.02) |
| On the Job Errors | 340 | 0 | 8.61 | 4.99 | 0 | 20.00 | 20.00 | 0.21 | (0.60) |
| Performance Review | 340 | 0 | 11.39 | 4.99 | 0 | 20.00 | 20.00 | (0.21) | (0.60) |
| Years of Employment | 340 | 0 | 3.51 | 2.04 | 1.00 | 10.00 | 9.00 | 1.28 | 1.57 |
| Height | 340 | 0 | 69.23 | 5.49 | 60.00 | 78.00 | 18.00 | (0.04) | (1.16) |

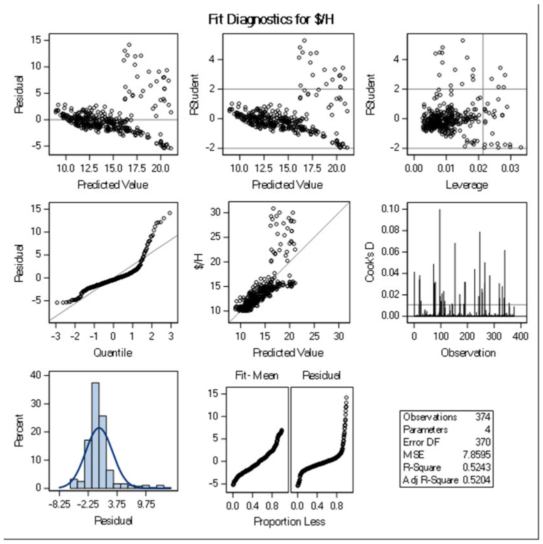
Note: No outliers and more normalized distribution of data

Table 4
*Pearson Correlation Coefficients, N = 340, Positions 1 and 2*

| Variable | $/H | On the Job Errors | Performance Review | Years of Employment | Height |
|---|---|---|---|---|---|
| $/H | 1.00 | (0.65) | 0.65 | 0.70 | (0.06) |
| On the Job Errors | (0.65) | 1.00 | (1.00) | (0.73) | 0.05 |
| Performance Review | 0.65 | (1.00) | 1.00 | 0.73 | (0.05) |
| Years of Employment | 0.70 | (0.73) | 0.73 | 1.00 | (0.07) |
| Height | (0.06) | 0.05 | (0.05) | (0.07) | 1.00 |

Figure 3
*Fit Diagnostics for $/H, Positions 1 and 2*





JBCA

Table 5

*Model Assessment Positions 1 and 2*

| Assessment | Measure |
|---|---|
| Root MSE | 1.23 |
| Dependent Mean | 14.15 |
| R-Squared | 0.91 |
| Adj R-Sq | 0.91 |
| AIC | 544.35 |
| AICC | 545.52 |
| SBC | 219.37 |

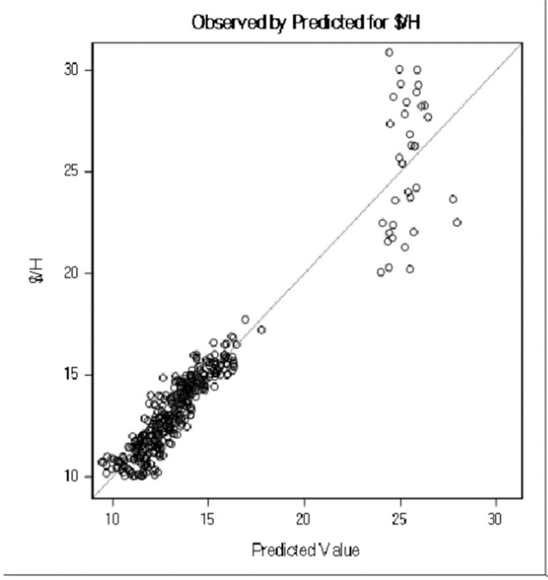
Table 6

*Parameter estimates for Positions 1 and 2*

| Variable | DF | Parameter Estimate | Standard Error | t Value | P Value | 95% Confidence Limits | |
|---|---|---|---|---|---|---|---|
| Intercept | B | 19.28 | 0.39 | 49.47 | <.0001 | 18.51 | 20.04 |
| Performance Review | 1 | 0.22 | 0.02 | 12.06 | <.0001 | 0.19 | 0.26 |
| Years of Employment | 1 | 0.27 | 0.04 | 6.25 | <.0001 | 0.18 | 0.35 |
| Location 1 | B | (0.43) | 0.28 | (1.50) | 0.13 | (0.99) | 0.13 |
| Location 2 | B | (0.19) | 0.28 | (0.66) | 0.51 | (0.75) | 0.37 |
| Location 3 | B | 0.12 | 0.29 | 0.42 | 0.67 | (0.44) | 0.68 |
| Location 4 | B | (0.11) | 0.28 | (0.37) | 0.71 | (0.67) | 0.45 |
| Location 5 | B | (0.01) | 0.29 | (0.02) | 0.99 | (0.57) | 0.56 |
| Location 6 | B | (0.35) | 0.28 | (1.22) | 0.22 | (0.91) | 0.21 |
| Location 7 | B | (0.33) | 0.28 | (1.18) | 0.24 | (0.89) | 0.22 |
| Location 8 | B | (0.43) | 0.28 | (1.53) | 0.13 | (0.99) | 0.12 |
| Location 9 | B | 1.53 | 0.28 | 5.42 | <.0001 | 0.98 | 2.09 |
| Location 10 | 0 | - | . | . | . | . | . |
| Position 1 | B | (9.70) | 0.26 | (37.96) | <.0001 | (10.21) | (9.20) |
| Position 2 | 0 | - | . | . | . | . | . |

Figure 6

*Visual model assessment*



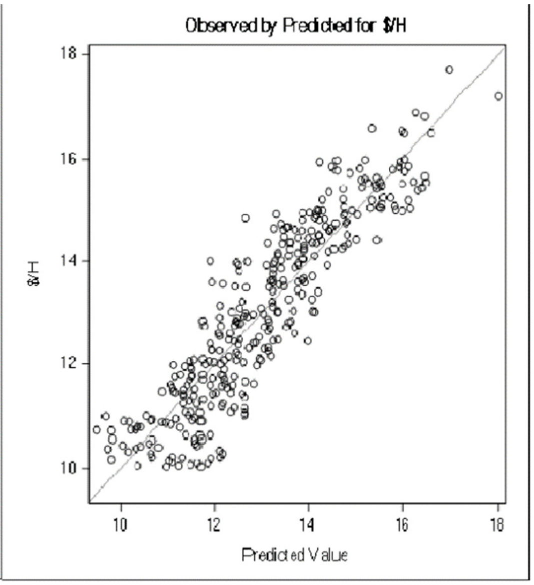


Table 7
*Comparison of models*

| Assessment | Position<br>1 & 2 | Position<br>1 | Position 1<br>w/Height |
|---|---|---|---|
| Root MSE | 1.23 | 0.76 | 0.76 |
| Dependent Mean | 14.15 | 13.04 | 13.04 |
| R-Squared | 0.91 | 0.83 | 0.83 |
| Adjusted R-Squared | 0.91 | 0.82 | 0.82 |
| AIC | 544.35 | 163.38 | 164.55 |
| AICC | 545.52 | 164.49 | 165.85 |
| SBC | 219.37 | -132.67 | -127.67 |

Table 8
*Parameter estimates with height added to the model*

| | | Parameter | Standard | | | 95% Confidence | |
|---|---|---|---|---|---|---|---|
| Variable | DF | Estimate | Error | t Value | P Value | Limits | |
| Intercept | B | 9.86 | 0.55 | 17.86 | <.0001 | 8.78 | 10.95 |
| Performance Review | 1 | 0.21 | 0.01 | 17.80 | <.0001 | 0.19 | 0.23 |
| Years of Employment | 1 | 0.30 | 0.03 | 10.40 | <.0001 | 0.24 | 0.36 |
| Height | 1 | (0.01) | 0.01 | (0.89) | 0.37 | (0.02) | 0.01 |
| Location 1 | B | (0.19) | 0.18 | (1.04) | 0.30 | (0.55) | 0.17 |
| Location 2 | B | 0.07 | 0.18 | 0.40 | 0.69 | (0.29) | 0.43 |
| Location 3 | B | 0.04 | 0.18 | 0.24 | 0.81 | (0.32) | 0.41 |
| Location 4 | B | (0.01) | 0.18 | (0.05) | 0.96 | (0.37) | 0.35 |
| Location 5 | B | 0.26 | 0.18 | 1.40 | 0.16 | (0.10) | 0.62 |
| Location 6 | B | (0.20) | 0.18 | (1.09) | 0.28 | (0.56) | 0.16 |
| Location 7 | B | 0.10 | 0.18 | 0.56 | 0.57 | (0.26) | 0.46 |
| Location 8 | B | (0.16) | 0.18 | (0.89) | 0.38 | (0.53) | 0.20 |
| Location 9 | B | 1.88 | 0.18 | 10.23 | <.0001 | 1.52 | 2.24 |
| Location 10 | 0 | - | . | . | . | . | . |