**Simple Linear Regression: An Example**

We begin with a **research question** defining what we are investigating with regression: Does the time students spend on MindTap help explain (that is to say, predict) their grades?

We will use an significance level for this regression analysis.

This question defines our variables and the relationship that we are investigating. Here the variables are:

* **Dependent variable**:
* **Independent variable**:

The dataset collected is a random sample of students enrolled in BA 3400. For each student, time spent logged on to MindTap (in hours) and overall MindTap score (in points) was recorded. Here is the dataset:

|  |  |  |
| --- | --- | --- |
| **Student #** | **Hours** | **Score** |
| 1 | 20.85 | 88.80 |
| 2 | 10.65 | 85.40 |
| 3 | 25.72 | 99.60 |
| 4 | 7.75 | 68.10 |
| 5 | 18.28 | 78.40 |
| 6 | 11.62 | 75.10 |
| 7 | 17.30 | 78.50 |
| 8 | 15.03 | 84.30 |
| 9 | 9.60 | 77.20 |
| 10 | 10.93 | 90.40 |
| 11 | 14.02 | 82.20 |
| 12 | 15.25 | 91.10 |
| 13 | 17.72 | 98.50 |
| 14 | 9.57 | 71.90 |
| 15 | 14.60 | 86.00 |
| 16 | 16.45 | 85.60 |
| 17 | 6.77 | 63.00 |
| 18 | 12.07 | 85.90 |
| 19 | 13.00 | 83.20 |
| 20 | 22.00 | 97.30 |
| 21 | 4.92 | 81.80 |
| 22 | 14.67 | 86.10 |
| 23 | 22.00 | 87.00 |
| 24 | 24.90 | 90.50 |

After the data collection step, the next thing to do is to graph a scatterplot of Scores vs Time Spent. In the scatterplot, we are chiefly concerned with non-linear patterns – if we see those, then we cannot use linear regression to analyze this data. We may also be able to identify a trend in the data that will give us an idea about the relationship between these two variables. Here is the scatterplot:

Do you see any nonlinear patterns?

Do you see a trend in the data? What sign do you expect on the slope of the regression line?

At this point, we can run the regression in Excel (Demo). See the next page for the results!

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT | | DV: MindTap Score | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |  |  |
| Multiple R | 0.6917 |  |  |  |  |  |  |  |
| R Square | 0.4785 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.4548 |  |  |  |  |  |  |  |
| Standard Error | 6.6202 |  |  |  |  |  |  |  |
| Observations | 24 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |  |  |
| Regression | 1 | 884.5979 | 884.5979 | 20.1839 | 0.0002 |  |  |  |
| Residual | 22 | 964.1917 | 43.8269 |  |  |  |  |  |
| Total | 23 | 1848.79 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 99.0%* | *Upper 99.0%* |
| Intercept | 67.4709 | 3.9186 | 17.2182 | 2.97E-14 | 59.3443 | 75.5976 | 56.4254 | 78.5165 |
| Hours | 1.1151 | 0.2482 | 4.4927 | 0.0002 | 0.6004 | 1.6299 | 0.4155 | 1.8148 |

Using the Regression Output Equations Roadmap, identify and on this output. Then write the Estimated Regression Equation (Remember, the ERE is ):

Before we can use this equation for prediction or interpret the slope, we must confirm that there is a statistically significant relationship between MindTap Score (y) and Time Spent (x) **in the population**. (Remember, if the slope of a line is zero, then there is no relationship between x and y. If the slope of a line is a number other than zero, then there *is* a relationship between x and y.)

Here is the Coefficients Table again:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *Lower 99.0%* | *Upper 99.0%* |
| Intercept | 67.4709 | 3.9186 | 17.2182 | 2.97E-14 | 59.3443 | 75.5976 | 56.4254 | 78.5165 |
| Hours | 1.1151 | 0.2482 | 4.4927 | 0.0002 | 0.6004 | 1.6299 | 0.4155 | 1.8148 |

The sample slope is 1.1151 which is definitely not zero… so that means there is a relationship between Hours and Score, right? Not necessarily! That sample slope just reflects what is going on in this sample of 24 students. We have to use this sample slope to prove that the **population slope**  is not zero. If proven, then that would mean there is a statistically significant relationship between Hours and Score in the **whole population of BA 3400 students**, not just in these 24 students that were in the sample. That is what we need to prove before using the regression to explain or predict anything.

Perform the hypothesis test detailed in **Ch 14: Handout #2** to determine whether there is a statistically significant relationship between Hours and Score in the population:

Now that we have confirmed that Hours and Score are related in the population, we can use our Estimated Regression Equation (ERE) to predict student scores, and we can interpret the slope

First, let’s write down the ERE again:

If 11 hours are spent on MindTap, what is the predicted score? Interpret this value.

Interpret the slope

Report and interpret the confidence interval for the population slope Since we are doing this regression at the significance level, which is a 99% confidence level, we should report and interpret the 99% confidence interval:

**But there is more to regression than that…**

The existence of a significant relationship between y and x is a necessary step – after all, if you cannot confirm a relationship between the IV and the DV, then the regression is no good for anything – but it is not sufficient to stop there. We need a way to judge the quality of the model. How well does it fit the data? Does it make accurate predictions? How much of the variation in y does x explain? Such questions are answered in the Regression Statistics table. But to understand the Regression Statistics table, we need to understand how the regression line (the estimated model) is calculated.

Linear regression calculates the estimated regression equation which minimizes the squared vertical distance between each observed value of in the sample (each ) and the regression line, at every value of in the dataset. This vertical distance between **observed and predicted y** is a very important quantity in regression, called the **Residual** and it is calculated by taking the difference between the observed and predicted values of y for a given observation:

Mathematically speaking, the regression line satisfies the Least Squares Criterion, which is:

So, the regression procedure minimizes the sum of the squared residuals. **Conceptually, the estimated regression equation that linear regression calculates is the line that is as close as possible to all the points in the data set at once: that is what it means to satisfy the Least Squares Criterion.** The regression line that satisfies the Least Squares Criterion is called the best-fit line.

Let’s calculate some residuals for the MindTap data, to gain insight into what regression is doing and thereby understand what the output tells us. Here is part of the MindTap dataset, along with the graph of the data with the regression line:

**where**

|  |  |  |
| --- | --- | --- |
| **Student #** | **Hours** | **Score** |
| 1 | 20.85 | 88.80 |
| 2 | 10.65 | 85.40 |
| 3 | 25.72 | 99.60 |
| 4 | 7.75 | 68.10 |
| 5 | 18.28 | 78.40 |
| … | … | … |

Find Student #4 on the graph. Calculate the residual for this student.

Find Student #3 on the graph. Calculate the residual for this student.

The regression procedure uses the residuals for all 24 students in its placement of the regression line. It takes all the residuals into account at once, squares them (why?), sums them, and places the line where that sum is the smallest it can be. This sum of the squared residuals is extremely important, then! And that is why it is reported in the ANOVA table in the regression output. It is called the **Sum of Squares due to Error** or the **Residual Sum of Squares,** and it is notated as the **SSE.**

Let’s take a look at the ANOVA table from the MindTap regression output. (Refer to the *Regression Output Equations Roadmap* throughout). Where is the SSE and its degrees of freedom?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |
| Regression | 1 | 884.5979 | 884.5979 | 20.1839 | 0.0002 |
| Residual | 22 | 964.1917 | 43.8269 |  |  |
| Total | 23 | 1848.79 |  |  |  |

The other values in the ANOVA table are also important in what we are building towards – that is, assessing the quality of this model.

The **Total Sum of Squares,** is the total amount of prediction error we would make if we used (the mean of ) to predict the observed sample data. Where is the SST and its *df* in the ANOVA table?

* The , and has degrees of freedom

The **Sum of Squares due to Regression,** is how much we will reduce prediction error by using the Estimated Regression Equation instead of the mean to predict the observed sample data. Where is the SSR and its *df* in the ANOVA table?

* The , and has degrees of freedom

The **Mean Square Error,**  is the estimate of the variance of . Remember It is the random error term in the Regression Model: . Where is the MSE in the ANOVA table?

* The

**Putting this all together to explain the ANOVA table in regression:**

* recall: the sample data is our best representation of the underlying population, so if our model is good at predicting the sample values, we can infer it will also be good at predicting population values
* The ANOVA table in the regression output compares two different methods of predicting the sample data in an effort to help us assess the quality of our regression model. We assess the regression by comparing its predictions to the next best alternative
  + Premise: there are two different models you could use to predict . One model includes information about and one does not:
    - First, you could use the Estimated Regression Equation to predict (in other words, predict at each )
    - Second, you could use the next best option, which is to just use the mean of to predict at each
      * This is also called the ‘Intercept only model’ because the equation of does not include any variables, thus being only an intercept, and representing a horizontal line at .
* The ANOVA table reports the total errors made when predicting the actual observed data using each of these models. In this way, we can compare the two possible models and decide whether our regression model is worth using.

So now that we have more understanding of what is in the ANOVA table, we should not be surprised to see all of these values showing up in the Regression Statistics table in the output, which contains the model fit stats we need.

Here is the Regression Statistics table from the *Regression Output Equations Roadmap*:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Regression Statistics*** | |  |  |  |  |  |
| **Multiple R** | or = | → & : Correlation Coefficient | | | | |
| **R Square** |  | → : Coefficient of Determination | | | |  |
| **Adjusted R Square** |  | → = : Adjusted | | | | |
| **Standard Error** |  | →: Standard Error of the Estimate, or Root Mean  Square Error (RMSE) | | | | |
| **Observations** |  | → *n* = # observations | | |  |  |

Here is the Regression Statistics table from the MindTap regression output:

|  |  |
| --- | --- |
| *Regression Statistics* | |
| Multiple R | 0.6917 |
| R Square | 0.4785 |
| Adjusted R Square | 0.4548 |
| Standard Error | 6.6202 |
| Observations | 24 |

blank on purpose

Show the calculation of the Coefficient of Determination, Interpret this value.

Show the calculation of the Standard Error of the Estimate, Interpret this value.

Show the calculation of the Correlation Coefficient, Interpret this value.