**Regression Introduction – 1**

**Regression is estimating one variable conditioned on another variable.**

**. . .**

|  |
| --- |
| **Simple Linear Regression**SummaryIntroductionDefinitions & TerminologyLeast-Squares |

|  |  |  |
| --- | --- | --- |
|  | **Regression Summary** |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Consider the data from five subjects that were asked miles and minutes to arrive at a destination.Let X=Miles and Y=Minutes. **We wish to estimate minutes using miles.**

|  |  |  |  |
| --- | --- | --- | --- |
| Subject | MilesX | MinutesY |  |
| 1 | 1 | 4 |
| 2 | 3 | 6 |
| 3 | 3 | 20 |
| 4 | 5 | 15 |
| 5 | 8 | 20 |

. . . |

Consider Regression.

|  |  |
| --- | --- |
| Regress Y on X to estimate Y using X. |  |
| Regression Equation, Ŷ = (Intercept) + (Slope) \* X |
| Slope = SSXY/SSXX = 57/28 = 2.035714286 ≈ 2.0357Excel: “=Slope(Y-data,X-data) |
| Intercept =Y–(Slope)\*X = (65/5)–(57/28)\*(20/5) = 4.857142857 ≈ 4.857Excel: “=Intercept(Y-data,X-data) |
| Regression Equation, Ŷ = 4.857 + 2.0357 \* X |

Consider Calculations between Miles and Minutes.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Subject | 1 | 2 | 3 | 4 | 5 | Sum |  |  |
| Miles, X | 1 | 3 | 3 | 5 | 8 | 20 | =X |  |
| Minutes,Y | 4 | 6 | 20 | 15 | 20 | 65 | =Y | SS = Sum of Squares of Error |
| X\*X | 1 | 9 | 9 | 25 | 64 | 108 | =X2 | SSXX= ( X –X )2 = X2–(X)\*(X)/n = 108–20\*20/5 = 28 |
| Y\*Y | 16 | 36 | 400 | 225 | 400 | 1077 | =Y2 | SSYY= ( Y –Y )2 = Y2–(Y)\*(Y)/n = 1077–65\*65/5 = 232 |
| X\*Y | 4 | 18 | 60 | 75 | 160 | 317 | =XY | SSXY= ( X –X )\*( Y –Y ) = X\*Y–(X)\*(Y)/n = 317–20\*65/5 = 57 |

|  |  |  |
| --- | --- | --- |
|  | **Regression Introduction** |  |

Let X=Miles and Y=Minutes. **We wish to estimate minutes using miles.**

|  |  |  |  |
| --- | --- | --- | --- |
| Subject | MilesX | MinutesY |  |
| 1 | 1 | 4 |
| 2 | 3 | 6 |
| 3 | 3 | 20 |
| 4 | 5 | 15 |
| 5 | 8 | 20 |

|  |
| --- |
| 1. As ‘Miles’ increases, ‘Minutes’ increases. Thus, there is a positive relationship or positive association between ‘Miles’ and ‘Minutes’. However, the relationship would be negative, if ‘Minutes’ decreased as ‘Miles’ increased.2. Outlier: The data point for subject 3, (X=3,Y=20), might possibly be an outlier. To establish an outlier, an argument needs to be presented based on either strong observation or logic. A strong observation would be when one point is obviously different from all the other points. For example, if we saw a point (X=3, Y=500), we could argue the point is an outlier due to the unreasonable magnitude of the difference between the point and the remaining points. A strong logic would be that subject 3 which generated the data point Y=20 minutes was the only subject that had an accident on his way and that is why the time is so large. An event that the other subjects did not encounter. Also, an event based on a factor not desired in the determination of the relationship between the two variables.3. To represent the relationship between the two variables, consider a linear line with the equation in intercept-slope form, Ŷ=a+b\*X, where a is the intercept and b is the slope. {Example: To illustrate the line through the data, select a=5 and b=2. Then, plot the line Ŷ=5+2\*X through the data.}4. Now the line Ŷ=5+2\*X may be used to determine Y given a value for X.{Example: Using Ŷ=5+2\*X for X=3, Ŷ=5+2\*(3)=11}5. The line, Ŷ=5+2\*X, is called a regression line because it estimates the mean of “Y” given “X” expressed as E[Y|X], or simply, estimates Y given X.6. Since X is used to estimate Y, Y is called the “Dependent Variable” or “Response Variable” and X is called the “Independent Variable” or “Explanatory Variable”.7. Since X is used to estimate Y, the terminology is “Regress Y on X” or “Regress the Dependent Variable on the Independent Variable”. Since X is used to estimate Y, Y is regressed back onto X. |

|  |  |  |
| --- | --- | --- |
|  | **Regression Definitions and Terminology** |  |

Let X=Miles and Y=Minutes. **We wish to estimate minutes using miles.**

|  |  |  |  |
| --- | --- | --- | --- |
| Subject | MilesX | MinutesY |  |
| 1 | 1 | 4 |
| 2 | 3 | 6 |
| 3 | 3 | 20 |
| 4 | 5 | 15 |
| 5 | 8 | 20 |

Consider the data and the line, Ŷ=5+2\*X.

Define the line: Ŷ=b0+b1\*X, where Ŷ is an estimate of Y given X and b0=Intercept, b1=Slope

Therefore, in the regression line Ŷ =5+2\*X, b0=5, b1=2

Consider the terminology:

|  |  |
| --- | --- |
|  | The regression line: Ŷ=5+2\*XThe sample mean of Y isY=Y/n=13The regression equation yields Ŷ =11 for X=3 { Ŷ=b0+b1\*X=5+2\*3=11 }The observed value of Y for X=3 is Y=6 (from table above) |
| Define Regression Model of Fit: Ŷ=b0+b1\*XDefine Regression Model of Data: Y=0+1\*X+Define error,  = (Ŷ – Y) for X=3 between Model (Ŷ =11) and Data (Y=6)Thus, for X=3, =(Ŷ–Y)=(11-6)=5 |
| 0 and 1 are Parameters.  is an error term.b0 and b1 are Estimates of the Parameters 0 and 1E[Y|X] is a Parameter, the population mean of Y|X.Ŷ=b0+b1\*X is the Estimate of the Parameter E[Y|X] |

|  |  |
| --- | --- |
| Model of Data: Y=0+1\*X+(Example: Data sets above.) | Y is the Dependent Variable or Response VariableX is the Independent Variable or Explanatory Variable0 is the “Intercept” parameter1 is the “Slope” parameter.  is the error term, e~N(0,s2) |
| Model of Fit: Ŷ=b0+b1\*X(Example: Ŷ=5+2\*X )(Regression Equation) ( Ŷ = estimate of E[Y|X] ) | Ŷ is the estimate of E[Y|X]E[Y|X] is the population mean of Y conditioned on X b0 is the estimate of 0 (the intercept)b1 is the estimate of 1 (the slope) |
| Mean of Y: isY=Y/n(Example:Y= | Y is the sample mean Y is the estimate of E[Y] not conditioned on X Y is the estimate of the mean of Y independent of X |

|  |  |
| --- | --- |
| Y=0+1\*X+ | Regression Model of Data |
| Ŷ=b0+b1\*X  | Regression Equation Regression Model of the Fit of the Data |

|  |  |  |
| --- | --- | --- |
|  | **Least-Squares Regression** |  |

Model of Data: Y=0+1\*X+

Solve for Error term:  = [ Y – ( 0+1\*X) ]

Substitute Estimates:  = [ Y – ( b0+b1\*X) ]

Square Error term: 2 = [ Y – ( b0+b1\*X) ]2

Sum over all data: 2 = [ Y – ( b0+b1\*X) ]2

Find b0 and b1 that minimizes the sum of the squared error.

|  |
| --- |
| **Min 2 = Min  [ Y – ( b0+b1\*X) ]2****b0,b1 b0,b1** |
| **Results:****b1 = SSXY/SSXX****b0 = Y/n – ( SSXY/SSXX )\*X/n = Y – b1 \*X** |

Using calculus:

|  |
| --- |
| Taking derivative, D(b0)= 2  [ Y – ( b0+b1\*X) ] (–1) = 0Summing through, Y – n\*b0 – b1\*X = 0Collecting terms, b0 = Y/n – b1\*X/nAlternative form, b0 = Y – b1 \*XTaking derivative, D(b1)= 2  [ Y – ( b0+b1\*X) ] (–X) = 0Summing through,  XY – b0\*X – b1\*X2 = 0Substituting for b0,  XY – ( Y/n – b1\*X/n )\*X – b1\*X2 = 0Collecting terms, b1 XY – (X)\*(Y)/n ]/[ X2 – (X)2/n ]Alternative form, b1 = SSXY/SSXX  |

The regression model that minimizes the sum of squared errors is called “Least-squares regression”.

|  |
| --- |
| Model of Data: Y=0+1\*X+Least-squares Regression Model: Ŷ=b0+b1\*X , for b1=SSXY/SSXX and b0=Y – b1 \*X |

Example.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Subject | 1 | 2 | 3 | 4 | 5 | Sum | SS |  |  |
| X=Miles | 1 | 3 | 3 | 5 | 8 | 20 |  |  |  |
| Y=Minutes | 4 | 6 | 20 | 15 | 20 | 65 |  |  | SS=Sum of Squares |
| X\*X | 1 | 9 | 9 | 25 | 64 | 108 | 28 |  | SSXX = 108–20\*20/5 = 28SSYY = 1077–65\*65/5 = 232SSXY = 317–20\*65/5 = 57 |
| Y\*Y | 16 | 36 | 400 | 225 | 400 | 1077 | 232 |  |
| X\*Y | 4 | 18 | 60 | 75 | 160 | 317 | 57 |  |
|  |  |  |  |  |  |  |  |  |  |
| b1=SSXY/SSXX = 57/28 = 2.035714286 ≈ 2.0357b0= Y – b1 \*X = (65/5) – (57/28)\*(20/5) = 4.857142857 ≈ 4.857Least-squares Regression Model: Ŷ=4.857+2.0357\*X |

|  |  |  |
| --- | --- | --- |
|  | **Exercises for Least-Squares Regression** |  |

1.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Index | 1 | 2 | 3 | 4 | 5 |  | b0 | b1 |
| X | 2 | 3 | 4 | 6 | 9 |  | 3.026 | 0.578 |
| Y | 4 | 5 | 5 | 7 | 8 |  |  |  |

2.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Index | 1 | 2 | 3 | 4 | 5 |  | b0 | b1 |
| X | 2 | 3 | 4 | 6 | 9 |  | 19.649 | -1.052 |
| Y | 18 | 17 | 15 | 12 | 11 |  |  |  |

3.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Index | 1 | 2 | 3 | 4 | 5 |  | b0 | b1 |
| X | 2 | 3 | 4 | 6 | 9 |  | -0.870 | 0.890 |
| Y | 1 | 2 | 2 | 5 | 7 |  |  |  |

4.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Index | 1 | 2 | 3 | 4 | 5 |  | b0 | b1 |
| X | 2 | 3 | 4 | 6 | 9 |  | 6.831 | -0.006 |
| Y | 7 | 6 | 8 | 6 | 7 |  |  |  |

5.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Index | 1 | 2 | 3 | 4 | 5 |  | b0 | b1 |
| X | 8 | 4 | 1 | 5 | 9 |  | 8.374 | 2.005 |
| Y | 22 | 17 | 10 | 19 | 28 |  |  |  |

6.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Index | 1 | 2 | 3 | 4 | 5 |  | b0 | b1 |
| X | 8 | 4 | 1 | 5 | 9 |  | 27.942 | -1.767 |
| Y | 16 | 20 | 27 | 18 | 11 |  |  |  |

7.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Index | 1 | 2 | 3 | 4 | 5 |  | b0 | b1 |
| X | 1 | 0 | -3 | 6 | -5 |  | 24.808 | 1.040 |
| Y | 26 | 25 | 21 | 31 | 20 |  |  |  |

8.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Index | 1 | 2 | 3 | 4 | 5 |  | b0 | b1 |
| X | 1 | 0 | -3 | 6 | -5 |  | 25.463 | -0.684 |
| Y | 23 | 26 | 28 | 22 | 29 |  |  |  |