**Regression Introduction – 1**

|  |
| --- |
| **Simple Linear Regression**  Concept  Definitions  Terminology  Least-Squares |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Consider the data from five subjects that were asked miles and minutes to arrive at a destination.  Let X=Miles and Y=Minutes. **Can we estimate minutes using miles?**   |  |  |  |  | | --- | --- | --- | --- | | Subject | Miles  X | Minutes  Y |  | | 1 | 1 | 4 | | 2 | 3 | 6 | | 3 | 3 | 20 | | 4 | 5 | 15 | | 5 | 8 | 20 |   . . . |

Consider Miles and Minutes.

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

SS = Sum of Squares of Error = Sum of Squared Errors

|  |
| --- |
| SSXX= ( X –X )2 = X2–(X)\*(X)/n = 108–20\*20/5 = 28  SSYY= ( Y –Y )2 = Y2–(Y)\*(Y)/n = 1077–65\*65/5 = 232  SSXX= ( X –X )\*( Y –Y ) = X\*Y–(X)\*(Y)/n = 317–20\*65/5 = 57 |

|  |  |  |
| --- | --- | --- |
|  | **Regression Concept** |  |

Let X=Miles and Y=Minutes. **Can we estimate minutes using miles?**

|  |  |  |  |
| --- | --- | --- | --- |
| Subject | Miles  X | Minutes  Y |  |
| 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’. If ‘Minutes’ decreased as ‘Miles’ increased, the relationship would be negative.  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, Y=a+b\*X, where a and b are constants.  {Example: To illustrate the line through the data, visually select  a=5 and b=2. Then, plot the line Y=5+2\*X through the data.}  4. The line Y=5+2\*X can be used to estimate Y given a value of X. Technically, estimate of the population mean of Y given X where the population mean of Y is expressed as E[Y|X].  {For example: if X=3, then Y=11, which estimates E[Y|X=3];  and, if X=7, then Y=19, which estimates E[Y|X=7].}  5. The line, Y=5+2\*X, is called a regression line because it estimates the mean of “Y” given “X”, or simply, estimates Y given X.  6. Since X is used to estimate Y, Y is called the “Dependent Variable” and X is called the “Independent 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”. |

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

Let X=Miles and Y=Minutes. **Can we estimate minutes using miles?**

|  |  |  |  |
| --- | --- | --- | --- |
| Subject | Miles  X | Minutes  Y |  |
| 1 | 1 | 4 |
| 2 | 3 | 6 |
| 3 | 3 | 20 |
| 4 | 5 | 15 |
| 5 | 8 | 20 |

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

|  |  |
| --- | --- |
|  | The regression line: Y=5+2\*X  Now, the sample mean of Y isY=Y/n=13  But the regression values of Y for X=3 are:  Model (Ŷ=11): Ŷ=b0+b1\*X=5+2\*3=11  🡨Error,  = (Ŷ – Y) between Model and Data  Data (Y=6): Y=0+1\*X+  0 and 1 are Parameters.  is an error term.  b0 and b1 are Estimates of the Parameters.  E[Y|X] is the Parameter, population mean of Y|X.  Ŷ=b0+b1\*X is the Estimate of the Parameter. |

|  |  |
| --- | --- |
| Model of Data: Y=0+1\*X+  (Example: Data sets above.) | Y is the Dependent Variable representing Data  X is the Independent Variable representing Data  0 is the “Intercept” parameter  1 is the “Slope” parameter.   is the error term, e~N(m,s2) |
| Model of Fit: Ŷ=b0+b1\*X  (Example: Ŷ=5+2\*X )  (Regression Model)  ( Ŷ = estimate of E[Y|X] ) | Ŷ is the estimate of E[Y|X], the population mean of Y conditioned on X representing the value from a regression equation for a given value of 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  or the estimate of E[Y] not conditioned on X  or 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 squared error.

|  |
| --- |
| **Min 2 = Min  [ Y – ( b0+b1\*X) ]2**  **b0,b1 b0,b1** |

Using calculus:

|  |
| --- |
| Taking derivative, D(b0)= 2  [ Y – ( b0+b1\*X) ] (–1) = 0  Summing through, Y – n\*b0 – b1\*X = 0  Collecting terms, b0 = Y/n – b1\*X/n  Alternative form, b0 = Y – b1 \*X  Taking derivative, D(b1)= 2  [ Y – ( b0+b1\*X) ] (–X) = 0  Summing through,  XY – b0\*X – b1\*X2 = 0  Substituting for b0,  XY – ( Y/n – b1\*X/n )\*X – b1\*X2 = 0  Collecting 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 = 28  SSYY = 1077–65\*65/5 = 232  SSXY = 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 ≈ 2.0357  b0= Y – b1 \*X = (65/5) – (57/28)\*(20/5) = 4.857142857 ≈ 4.857  Least-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 |  |  |  |