**Regression Introduction – 2**

|  |
| --- |
| **Simple Linear Regression**  Sum of Squares  Coefficient of Determination  ANOVA  Test of Hypothesis  Residuals |

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 Measures.

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

|  |  |  |
| --- | --- | --- |
|  | **Sum of Squares** |  |

|  |  |
| --- | --- |
|  | Regression Line: Ŷ=b0+b1\*X  Y:Y =Y/n  Ŷ: Ŷ=b0+b1\*X  Y: Y=0+1\*X+ |

**Definitions.**

|  |
| --- |
| Model of Data: Y=0+1\*X+ |
| Least-squares Regression Model: Ŷ=b0+b1\*X , for b1=SSXY/SSXX and b0=Y – b1 \*X |
| Sample mean of Data :Y = Y/n |
| Total Sum of Squares, TSS = (Y –Y)2 = SSYY = 232 |
| Sum of Squares of Model, SSM = (Ŷ –Y )2 = (SSXY)2/SSXX = (572/28) ≈ 116.04  (Sum of Squares of Regression) |
| Sum of Squares of Error, SSE = (Y – Ŷ )2 = SSYY – (SSXY)2/SSXX ≈ 115.96  (Sum of Squares of Residuals) |
| From Example: SSYY = 232; SSXX = 28; SSXY = 57 |
| **Identities.** |
| Total = Model + Error  TSS = SSM + SSE  (Y –Y)2  = (Ŷ –Y )2 + (Y – Ŷ )2  SSYY  = [ (SSXY)2/SSXX ] + [ SSYY – (SSXY)2/SSXX ] |
| 1. Total variation is explained by either the variation due to the model or variation due to random error.  2. As Ŷ approachesY, SSM approaches zero and SSE approaches TSS.  3. As Ŷ approachesY, the total variation is explained less by the model and more by random error.  4. As Ŷ approachesY, the regression slope approaches zero and the correlation coefficient approaches zero. Consider the estimate of the regression line,  Ŷ=b0+b1\*X = (Y – b1 \*X ) + b1 \* X , as b1 approaches zero, Ŷ approachesY. |

|  |  |  |
| --- | --- | --- |
|  | **R2 – Coefficient of Determination** |  |

|  |
| --- |
| Total = Model + Error  TSS = SSM + SSE  (Y –Y)2  = (Ŷ –Y )2 + (Y – Ŷ )2  SSYY  = [ (SSXY)2/SSXX ] + [ SSYY – (SSXY)2/SSXX ] |
| Divide by TSS: TSS/TSS = SSM/TSS + SSE/TSS  1 = [ (SSXY)2/(SSXX \*SSYY ) ] + [ 1 – (SSXY)2/(SSXX \*SSYY )]  4. Define: Coefficient of Determination, R2= SSM/TSS  Thus: 1 = R2 + [ 1 – R2 ] |
| **Coefficient of Determination**  **R2= SSM/TSS = (SSXY)2/(SSXX\*SSYY)**  **0 ≤ R2 ≤ 1** |
| 1. R2 = SSM/TSS represents the percent of total variation explained by the model.  2. 1–R2 = SSE/TSS represents the percent of total variation explained by random error. |
| **The correlation coefficient, r = sqrt(R2) = SSXY/sqrt( SSXX\*SSYY ), –1 ≤ r ≤ +1** |

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 |  |

|  |
| --- |
| Coefficient of Determination, R2 = SSM/TSS  = (SSXY)2/(SSXX\*SSYY) = (572/(28\*232) = 0.500153941 ≈ 0.5 |
| Total Sum of Squares, TSS = (Y –Y)2 = SSYY = 232 |
| Sum of Squares of Model, SSM = (Ŷ –Y )2 = (SSXY)2/SSXX = (572/28) ≈ 116.04  = TSS\*R2 = [SSYY ]\*[ (SSXY)2/(SSXX\*SSYY) ] = (SSXY)2/SSXX = (572/28) ≈ 116.04 |
| Sum of Squares of Error, SSE = (Y – Ŷ )2 = SSYY – (SSXY)2/SSXX ≈ 115.96  = TSS\*(1–R2) = [SSYY]\*[1 – (SSXY)2/(SSXX\*SSYY) ] = SSYY – (SSXY)2/SSXX ≈ 115.96 |
|  |
| Correlation coefficient, r = sqrt(R2) =  =57/sqrt(28\*232) = 0.707215625 ≈ 0.7 |

|  |  |  |
| --- | --- | --- |
|  | **ANOVA – Analysis of Variance** |  |

Consider the values:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Subject | 1 | 2 | 3 | 4 | 5 | Sum |  | SSXY =57 |  | SSM ≈ 116.04 |
| X=Miles | 1 | 3 | 3 | 5 | 8 | 20 |  | SSXX =28 |  | SSE ≈ 115.96 |
| Y=Minutes | 4 | 6 | 20 | 15 | 20 | 65 |  | SSYY =232 |  | TSS=232 |

**ANOVA Table**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SOURCE | DF | SS | MS | Ft | R2 |
| MODEL  (Regression) | 1 | SSM | MSM=SSM/1 | MSM/MSE | SSM/TSS |
| ERROR  (Residuals) | n-2 | SSE | MSE=SSE/(n-2) |  |  |
| TOTAL  (Corrected Total) | n-1 | TSS |  |  |  |

Using Algebra to combine relationships:

|  |
| --- |
| R2 = SSM/TSS ≈ 0.5 |
| TSS = (Y –Y)2 = SSYY = 232 |
| SSM = (Ŷ –Y )2 = TSS\*R2 ≈ 116.04 |
| SSE = (Y – Ŷ )2 = TSS\*(1–R2) ≈ 115.96 |
| r = sqrt (R2) ≈ 0.7 |

**ANOVA Table**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SOURCE | DF | SS | MS | Ft | R2 |
| MODEL | 1 | SSM=TSS\*R2 | MSM=SSM/1 | MSM/MSE | SSM/TSS |
| ERROR | n-2 | SSE=TSS\*(1–R2) | MSE=SSE/(n-2) |  |  |
| TOTAL | n-1 | TSS=SSYY |  |  |  |

**ANOVA Table for Example**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SOURCE | DF | SS | MS | Ft | R2 |  | F=0.05 | P-value |
| MODEL | 1 | 116.04 | 116.04 | 3.00 | 0.50 |  | 10.13 | 0.182 |
| ERROR | 3 | 115.96 | 38.65 |  |  |  |  |  |
| TOTAL | 4 | 232 |  |  |  |  |  |  |

|  |  |  |
| --- | --- | --- |
|  | **Test of Hypothesis** |  |

Consider the values:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Subject | 1 | 2 | 3 | 4 | 5 | Sum |  | SSXY =57 |  | SSM ≈ 116.04 |
| X=Miles | 1 | 3 | 3 | 5 | 8 | 20 |  | SSXX =28 |  | SSE ≈ 115.96 |
| Y=Minutes | 4 | 6 | 20 | 15 | 20 | 65 |  | SSYY =232 |  | TSS=232 |

**ANOVA Table**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SOURCE | DF | SS | MS | Ft | R2 |
| MODEL | 1 | SSM=TSS\*R2 | MSM=SSM/1 | MSM/MSE | SSM/TSS |
| ERROR | n-2 | SSE=TSS\*(1–R2) | MSE=SSE/(n-2) |  |  |
| TOTAL | n-1 | TSS=SSYY |  |  |  |

**ANOVA Table for Example**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| SOURCE | DF | SS | MS | Ft | R2 |  | F=0.05 | P-value |
| MODEL | 1 | 116.04 | 116.04 | 3.00 | 0.50 |  | 10.13 | 0.182 |
| ERROR | 3 | 115.96 | 38.65 |  |  |  |  |  |
| TOTAL | 4 | 232 |  |  |  |  |  |  |

|  |
| --- |
| Regression Test of Model Significance |
| Model of Data: Y=0+1\*X+  Least-squares Regression Model: Ŷ=b0+b1\*X , for b1=SSXY/SSXX and b0=Y – b1 \*X  Degrees of Freedom from ANOVA Table, (1,3). |
| Ho: 1=0 🡨Regression NOT significant  Ha: 1/0 🡨Regression IS significant    CV: + F(0.05:1,3) = + 10.13 🡨(From F-tables)  RR: > +10.13  TS: Ft=3.00 🡨 (From ANOVA Table)  Inference: Do Not Reject Ho 🡨(Since TS is NOT in RR)   |  |  | | --- | --- | |  |   RR | |  |  | | 0 CV F  10.13  Ft=3.00 | | |
| Regression is not significant at a 5% level of significance. |
| P-value is 0.182 🡨 (From AVOVA Table) (From Excel: “=1–F.dist(X,DF1,DF2,1)”) |

|  |  |  |
| --- | --- | --- |
|  | **Residual Analysis** |  |

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

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Index | X | Y |  |  |  |  |  |  |  |  | SSxy= | 57 |
| 1 | 1 | 4 |  | ANOVA |  |  |  |  |  |  | SSyy= | 232 |
| 2 | 3 | 6 |  | SOURCE | DF | SS | MS | F | P |  | SSxx= | 28 |
| 3 | 3 | 20 |  | MODEL | 1 | 116.04 | 116.04 | 3.00 | 0.182 |  | R2= | 0.5 |
| 4 | 5 | 15 |  | ERROR | 2 | 115.96 | 38.65 |  |  |  | b0 = | 4.86 |
| 5 | 8 | 20 |  | TOTAL | 3 | 232 |  |  |  |  | b1 = | 2.04 |

A ‘residual’ or ‘error’ or ‘deviation’ or ‘’ is defined as ‘Y–Ŷ’ for each sample pair.

Residual Plot

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| X | Y | Ŷ | Y–Ŷ |  | –5 –4 –3 –2 –1 0 +1 +2 +3 +4 +5 +6 +7 +8 +9 +10 | |
| 1 | 4 | 6.89 | -2.89 |  | **+** |  |
| 3 | 6 | 10.96 | -4.96 |  | **+** |  |
| 3 | 20 | 10.96 | 9.04 |  |  | **+** |
| 5 | 15 | 15.04 | -0.04 |  | **+** |  |
| 8 | 20 | 21.14 | -1.14 |  | **+** |  |

|  |
| --- |
| The residual plot should be ‘random’ or normally distributed about zero.  One sample seems to be ‘anomalous’ from the others.  What happens if we remove that sample? |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Index | X | Y |  |  |  |  |  |  |  |  | SSxy= | 65.75 |
| 1 | 1 | 4 |  | ANOVA |  |  |  |  |  |  | SSyy= | 170.75 |
| 2 | 3 | 6 |  | SOURCE | DF | SS | MS | F | P |  | SSxx= | 26.75 |
| 3 | 3 | 20 |  | MODEL | 1 | 161.61 | 161.61 | 35.36 | 0.027 |  | R2= | 0.95 |
| 4 | 5 | 15 |  | ERROR | 2 | 9.14 | 4.57 |  |  |  | b0 = | 0.80 |
| 5 | 8 | 20 |  | TOTAL | 3 | 170.75 |  |  |  |  | b1 = | 2.46 |

Residual Plot

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| X | Y | Ŷ | Y–Ŷ |  | –5 –4 –3 –2 –1 0 +1 +2 +3 +4 +5 +6 +7 +8 +9 +10 | |
| 1 | 4 | 3.26 | 0.74 |  |  | **+** |
| 3 | 6 | 8.18 | -2.18 |  | **+** |  |
| 3 | 20 |  |  |  |  |  |
| 5 | 15 | 13.09 | 1.91 |  |  | **+** |
| 8 | 20 | 20.47 | -0.47 |  | **+** |  |

ANOVA table determines the “Significance” of a regression.

Use RESIDUALS to address the “Validity” of a regression.

|  |  |  |
| --- | --- | --- |
|  | **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 |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA | df | SS | MS | F | P |
| Model | 1 | 10.287 | 10.287 | 60.159 | 0.004 |
| Error | 3 | 0.513 | 0.171 |  |  |

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 |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA | df | SS | MS | F | P |
| Model | 1 | 34.083 | 34.083 | 32.805 | 0.011 |
| Error | 3 | 3.117 | 1.039 |  |  |

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 |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA | df | SS | MS | F | P |
| Model | 1 | 24.375 | 24.375 | 88.672 | 0.003 |
| Error | 3 | 0.825 | 0.275 |  |  |

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 |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA | df | SS | MS | F | P |
| Model | 1 | 0.001 | 0.001 | 0.001 | 0.973 |
| Error | 3 | 2.799 | 0.933 |  |  |