THE LINEAR REGRESSION PROCESS

From Generation to Interpretation

Presented By

Frank Briody
frankbriody@gmail.com

Prospect High School
Mt. Prospect, IL

2022 MMC Conference of Workshops

Photo credit - Vito Palmisano

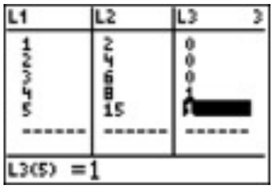
Contents

The Story

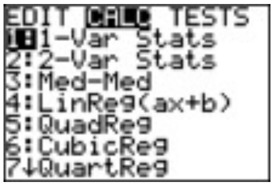
A statistics teacher gives a quiz to a class. The scores were 2, 4, 6, 8, and 15 with one student being absent. Absent student returns the next day...
Student: How am I going to do on the quiz?
Teacher: Well, the class average was...

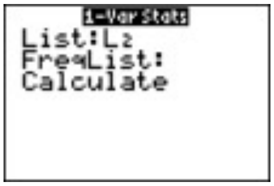
1 Prologue: Standard Deviation

How much variability, on average, is there around the mean?

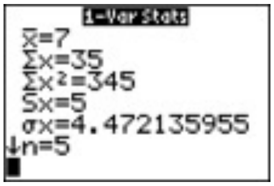


```
L1      L2      L3      3
 1       2       0
 2       4       0
 3       6       0
 4       8       1
 5       15      
------  ------  ------
L3(5) =1
```

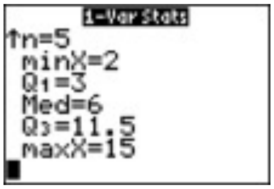




```
1-Var Stats
List:L2
FreqList:
Calculate
```



```
1-Var Stats
x̄=7
Σx=35
Σx²=345
Sx=5
σx=4.472135955
↓n=5
```



```
1-Var Stats
↑n=5
 minX=2
 Q1=3
 Med=6
 Q3=11.5
 maxX=15
```

Score	Deviation	Squared Deviation
x		
2		
4		
6		
8		
15		

The Story Part 2

A statistics teacher gives a quiz to a class. The scores were 2, 4, 6, 8, and 15 with one student being absent. After surveying the class, the teacher knows the hours studied were 1, 2, 3, 4 and 5, respectively. Absent student returns the next day...
Student: How am I going to do on the quiz?
Teacher: That depends - how long did you study?

2 Getting Least Squares Line of Best Fit

Hours (x)	1	2	3	4	5
Score (y)	2	4	6	8	15

2.1 From Summary Statistics

Formulas (given): $\hat{y} = a + bx$ $\quad b = r\frac{s_y}{s_x}$ $\quad a = \bar{y} - b\bar{x}$

Descriptive Statistics: x, y

```
Variable  N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median     Q3  Maximum
x         5   0  3.000    0.707  1.581    1.000  1.500   3.000  4.500    5.000
y         5   0   7.00     2.24   5.00     2.00   3.00    6.00  11.50    15.00
```

Correlations: x, y

```
Pearson correlation of x and y = 0.949
P-Value = 0.014
```

2.2 From Output

Regression Analysis: y versus x

```
The regression equation is
y = - 2.00 + 3.00 x

Predictor     Coef   SE Coef        T       P
Constant    -2.000     1.915    -1.04   0.373
x           3.0000    0.5774     5.20   0.014

s = 1.82574    R-sq = 90.0%    R-Sq(adj) = 86.7%
```

Analysis of Variance

```
Source          DF        SS        MS       F       P
Regression       1    90.000    90.000   27.00   0.014
Residual Error   3    10.000     3.333
Total            4   100.000
```

3 Interpretation

3.1 Slope

Slope represents the **predicted** change in response associated with each unit increase in the explanatory variable, **on average**.

3.2 Y-Intercept

Y-intercept is the predicted value when the explanatory (x) is 0. [Often the y-intercept is useless due to *extrapolation*.]

4 Predicted Values and Residuals

$\hat{y} = -2 + 3x$

Hours	Score	Predicted	Residual
x	y	$\hat{y}$	$y - \hat{y}$
1	2		
2	4		
3	6		
4	8		
5	15		

- Predicted $\hat{y}$: substitute explanatory (x) values into regression equation.
- Residual $y - \hat{y}$ (also called *regression error*) CHECK THIS; actual minus predicted.

The Story Part 3

A statistics teacher gives a quiz to a class. The scores were 2, 4, 6, 8, and 15 with one student being absent. After surveying the class, the teacher knows the hours studied were 1, 2, 3, 4 and 5, respectively. Absent student returns the next day...
Student: How am I going to do on the quiz?
Teacher: That depends - how long did you study?
Student: Does how long I studied really make a difference?

5 The Coefficient of Determination r^2 - Comparing Models

Regression Analysis: y versus x

```
The regression equation is
y = - 2.00 + 3.00 x

Predictor     Coef   SE Coef         T       P
Constant    -2.000     1.915     -1.04   0.373
x           3.0000    0.5774      5.20   0.014

s = 1.82574    R-sq = 90.0%    R-Sq(adj) = 86.7%
```

Analysis of Variance

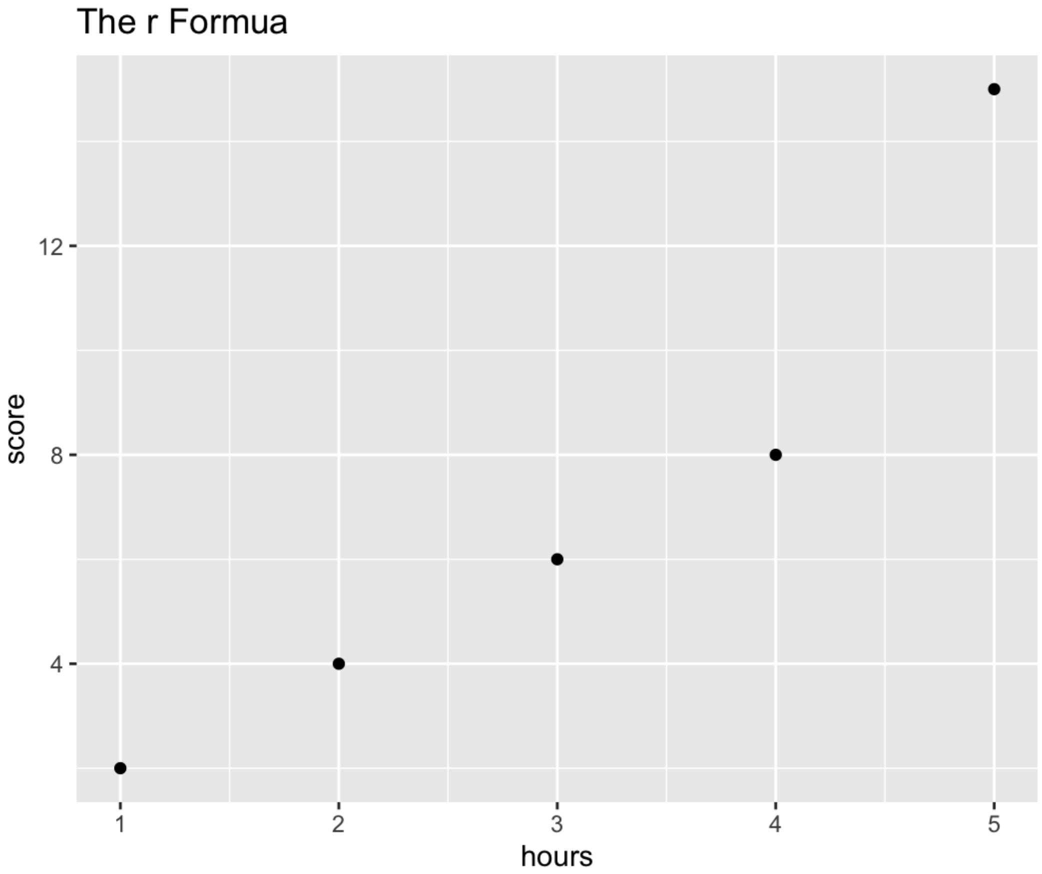
```
Source           DF        SS        MS        F        P
Regression        1    90.000    90.000    27.00    0.014
Residual Error    3    10.000     3.333
Total             4   100.000
```

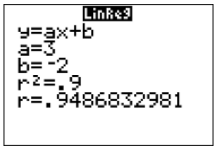
Hours	Score	Predicted	Error	(Error)^2	Residual	(Residual)^2
x	y	$\hat{y}$	$y-\bar{y}$	$(y-\bar{y})^2$	$y-\hat{y}$	$(y-\hat{y})^2$
1	2					
2	4					
3	6					
4	8					
5	15					

6 The Correlation Coefficient r

6.1 Getting r

- $r = \frac{\Sigma z_x \cdot z_y}{n-1}$
- Never calculate by hand; use calculator or computer output.
- Know formula *properties*.





```
LinReg
y=ax+b
a=3
b=-2
r²=.9
r=.9486832981
```

Correlations: x, y

```
Pearson correlation of x and y = 0.949
P-Value = 0.014
```

6.2 Five R Properties

- Examples