# THE LINEAR REGRESSION PROCESS

From Generation to Interpretation

Presented By

Frank Briody
*frankbriody@gmail.com*

*Prospect High School*
*Mt. Prospect, IL*

*Photo credit - Vito Palmisano*

2022 MMC Conference of Workshops

## Contents

Put quote here. There is only one large computer program I have used in which there are to a decent approximation 0 bugs: Don Knuth's TeX

– *Jaap Weel*

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.

## 1 Standard Deviation: The Non-Resistant Measure of Spread

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

| Score $x$ | Deviation | Squared Deviation |
|---|---|---|
| 2 | | |
| 4 | | |
| 6 | | |
| 8 | | |
| 15 | | |

The Story
A statistics teacher gives a quiz to a class. The scores were 2, 4, 6, 8, and 15 with one student absent.

## 2 Getting Least Squares Line of Best Fit

### 2.1 From Summary Statistics

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

**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 | | | |

### 2.3 Interpretation

#### 2.3.1 Slope

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

#### 2.3.2 Y-Intercept

Y-intercept is the predicted value when the explanatory ($x$) is 0. [Often the y-intercept is useless.]

### 2.4 Predicted Values and Residuals

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

| Hours $x$ | Score $y$ | Predicted $\hat{y}$ | Residual $y - \hat{y}$ |
|---|---|---|---|
| 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*); actual minus predicted.

### 2.5 The Coefficient of Determination $r^2$

**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**

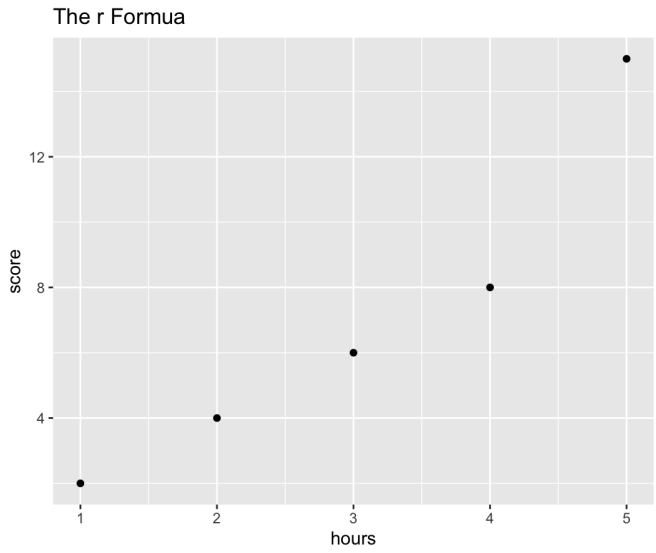
| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Regression | 1 | 90.000 | 90.000 | 27.00 | 0.014 |
| Residual Error | 3 | 10.000 | 3.333 | | |
| Total | 4 | 100.000 | | | |

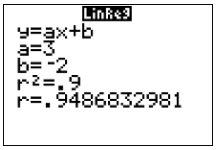
| Hours | Score | Predicted | Error | $(\text{Error})^2$ | Residual | $(\text{Residual})^2$ |
|---|---|---|---|---|---|---|
| $x$ | $y$ | $\hat{y}$ | $y - \bar{y}$ | $(y - \bar{y})^2$ | $y - \hat{y}$ | $(y - \hat{y})^2$ |
| 1 | 2 | | | | | |
| 2 | 4 | | | | | |
| 3 | 6 | | | | | |
| 4 | 8 | | | | | |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

## 3 Summary & Examples

### 3.1 Getting $r$

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




**Correlations: x, y**

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

### 3.2 Five R Properties

- Examples