Lecture 3: Bivariate Data & Linear Regression

1. Introduction
2. Bivariate Data
3. Linear Analysis of Data
   a) Freehand Linear Fit
   b) Least Squares Fit
   c) Interpolation/Extrapolation
4. Correlation

Motivation

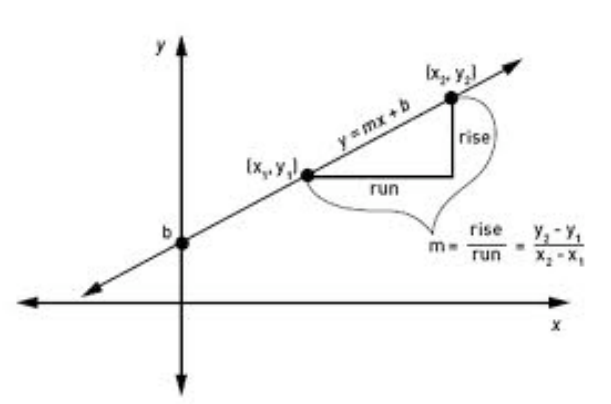
- Visual analysis of data, though useful, is limited
- We use other methods to determine if there is a sufficiently strong relationship between two measurements to possibly eliminate the necessity of making both measurements
  - In many fish, there is a very strong relationship between length and body mass, so that measuring the length alone can provide a very good estimate of the individual's weight
  - Organisms with determinant growth patterns may also have a very strong relationship between their size and age, so that measurements of individual size may be useful in estimating an individual's age
  - In these cases an objective is to determine an equation which **allows you to estimate one measurement from another measurement that might be easier to make**.

Linear Relationships

- A linear relationship between two measurements is the simplest description of a strong relationship
  - Knowing one measurement allows you to find the other by just looking at the associated point on the line
- Linear relationships arise in many biological situations, particularly if the data are *rescaled* (see ch. 4) to account for factors that interact in a non-linear manner
- Recall the slope-intercept form for a line:

$$y = mx + b$$

we will need to go this direction



- Bivariate data are measurements made on two variables per observation
- Bivariate data are often displayed as ordered pairs (x_1, y_1), (x_2, y_2), …, (x_n, y_n) or in a table:

x	x_1	x_2	…	x_n
y	y_1	y_2	…	y_n

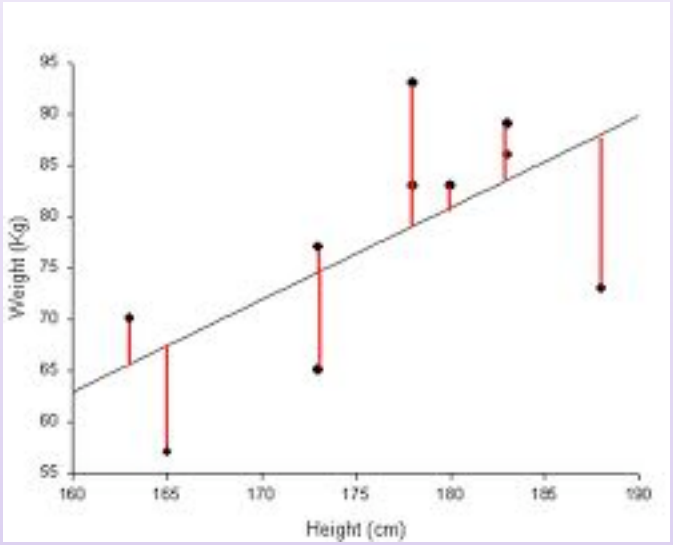
- A **regression** is a formula that describes the relationship between variables. A **linear regression** is one in which the relationship can be expressed in the equation of a line. In this case, we find a line of best fit to describe the relationship.

Freehand Linear Fit

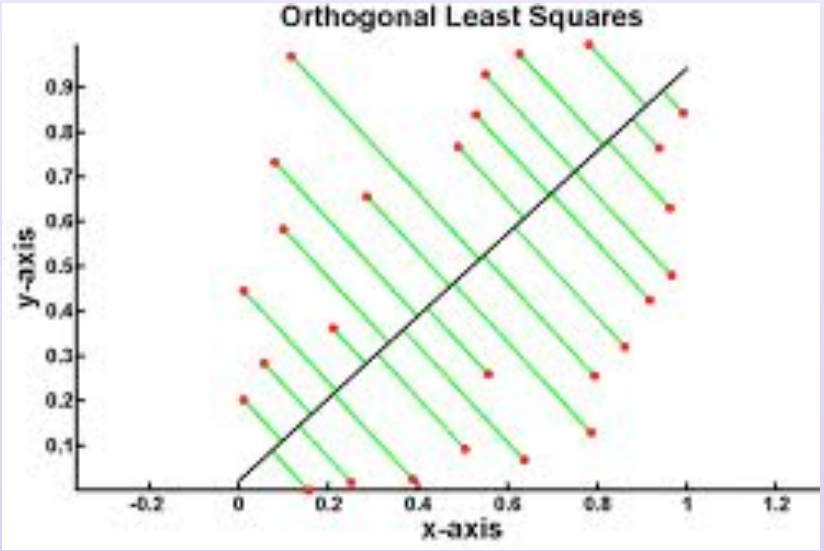
1. See if the data appear to be linearly related
   - Make a scatter plot of the data. Choose your axes so that the range is approximately the **same as that in your data**
   - Which is the horizontal axis and which is the vertical?
     - If one is the dependent (e.g. body weight depends on age, rather than the reverse), then choose the one that is dependent (weight) as the vertical axis and the independent one (age) as the horizontal axis
     - If you do not have any reason to expect that one of the measurements is dependent on the other, then choose one and go with it
   - Eyeball the data and see if there appears to be any relationship. If the scatter plot looks like the points might be described approximately by a line then it is reasonable to proceed with fitting a line

Freehand Linear Fit

2. Draw a freehand line that *estimates* the line of best fit; that is, draw a line through the data, trying to minimize the **vertical** distance between the line and the points:



vertical



orthogonal

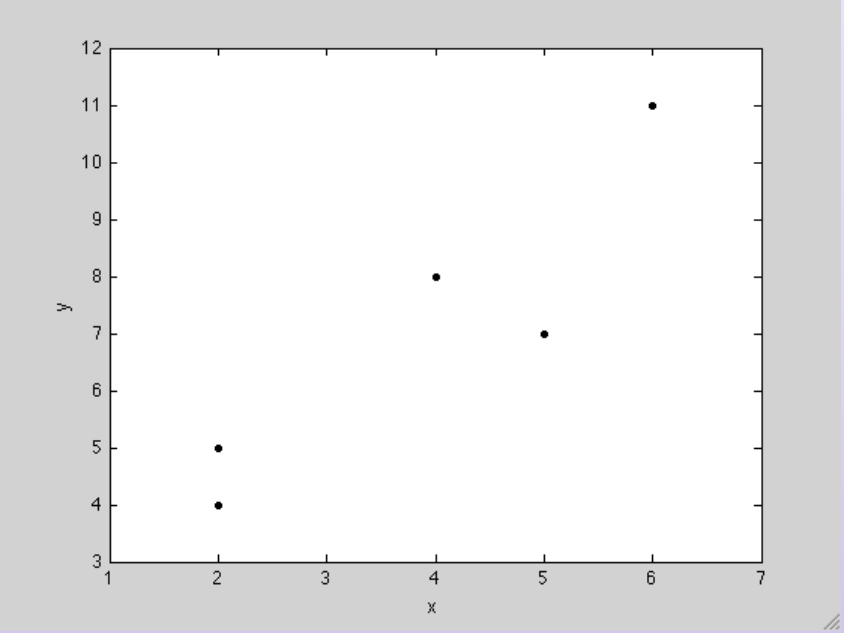
3. Using two points ON THE LINE (not necessarily data points), find the equation of the line.

Example 3.1 (Freehand Linear Fit)

Given the following set of bivariate data, use the freehand method to find a linear equation that fits the data.

x	2	5	2	4	6
y	4	7	5	8	11

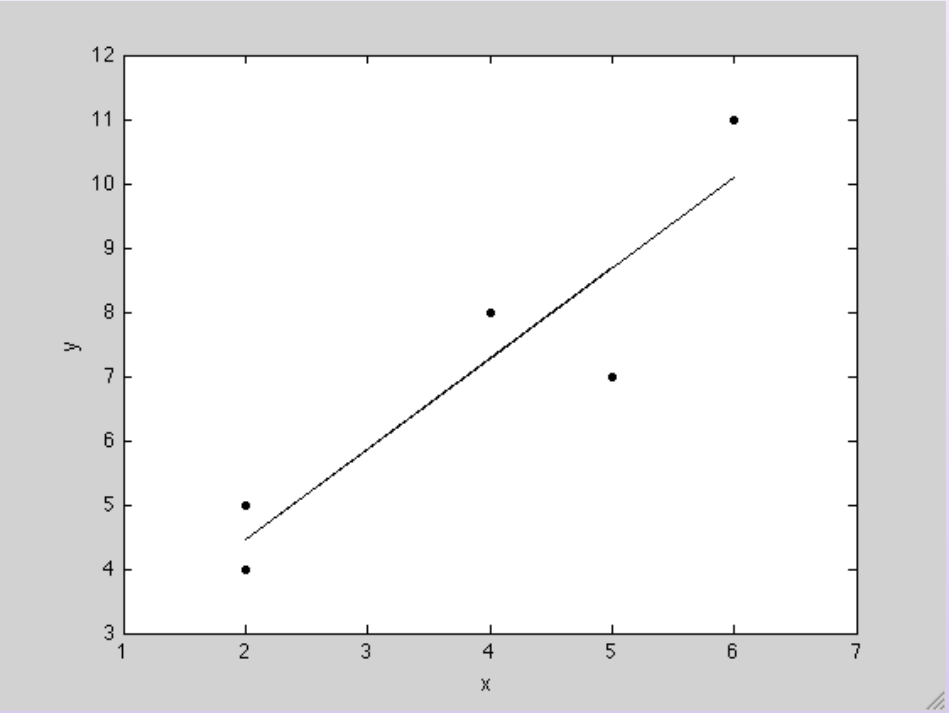
Step 1: Make a scatterplot:



The data seem reasonably linearly related, so we go to the next step.

Example 3.1 (Freehand Linear Fit)

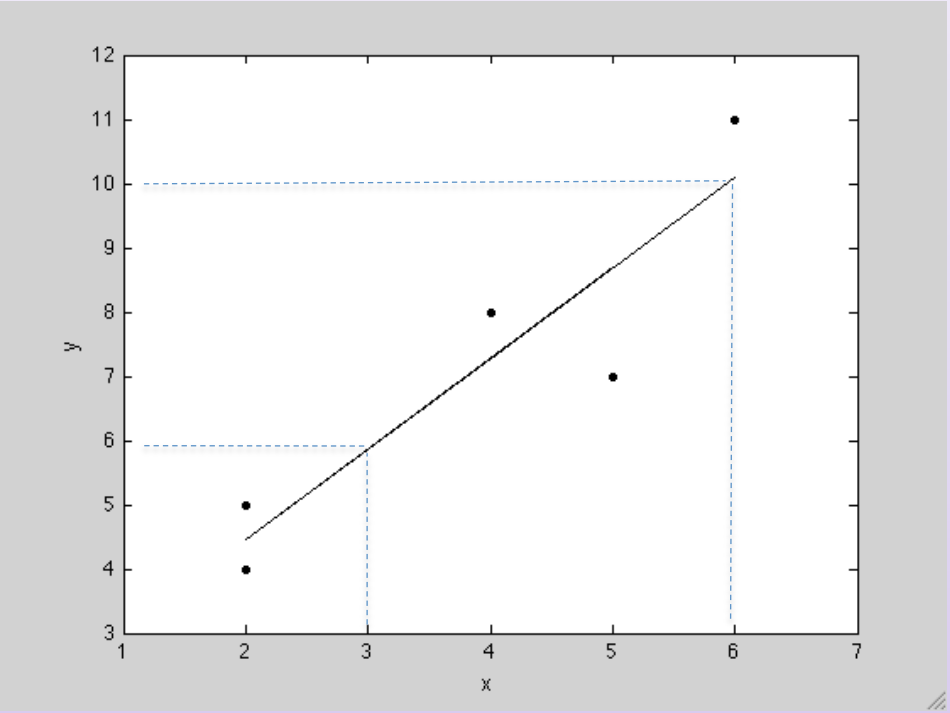
Step 2: Draw a line through the data that attempts to minimize the (vertical) distance between the line and the points:



Notice that our line contains none of the data points!

Example 3.1 (Freehand Linear Fit)

Step 3: Next we pick two points ON THE LINE to determine the equation of the line.



Let's choose (3,6) and (6,10).

Then:

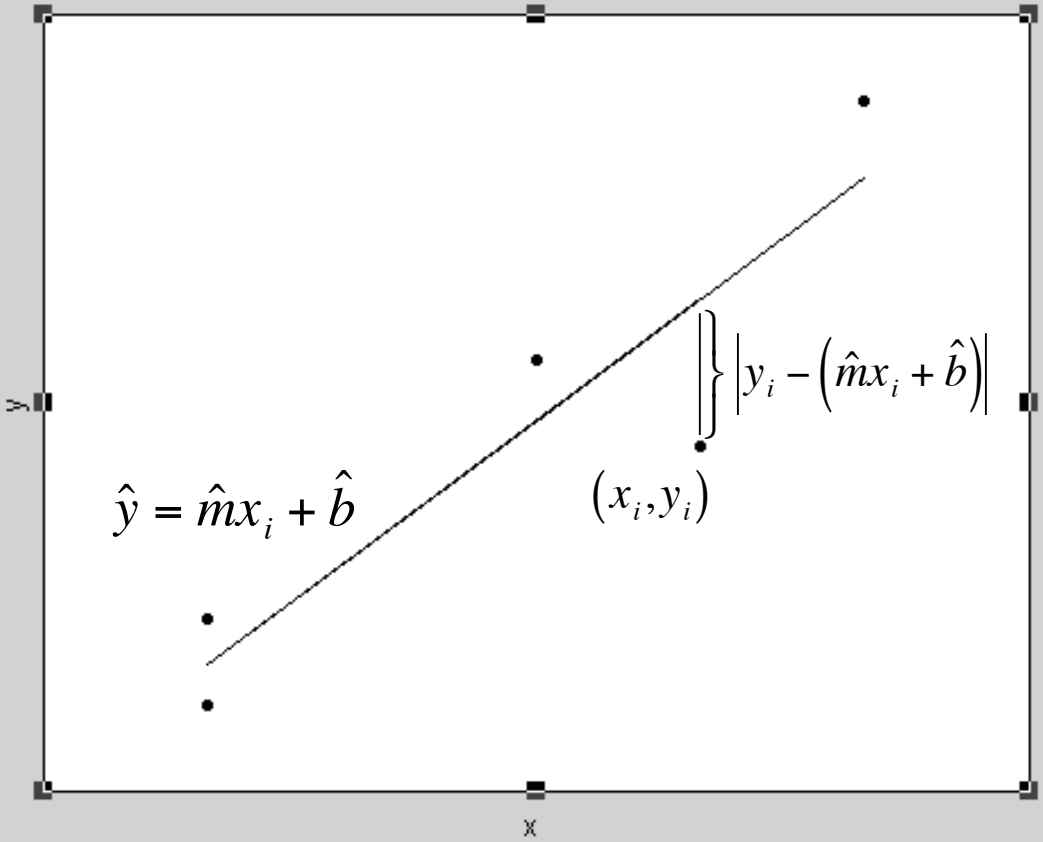
$$m = \frac{10-6}{6-3} = \frac{4}{3}$$

$$y - 6 = \frac{4}{3}(x-3)$$

$$y = \frac{4}{3}x + 2$$

Least Squares Fit

We want to find $\hat{m}$ and $\hat{b}$ to minimize the sum of errors squared:



$$\sum_{i=1}^{n} \left[y_i - \left(\hat{m}x_i + \hat{b} \right) \right]^2$$

Least Squares Fit

Using some simple calculus techniques, one can easily discover that the best such values are given by

$$\hat{m} = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \hat{b} = \bar{y} - \hat{m}\bar{x}$$

where

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \quad \text{and} \quad S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Example 3.2 (Least Squares Linear Regression)

Given the same set of data as before, use the Least Squares method to find a linear equation that fits the data.

x	2	5	2	4	6
y	4	7	5	8	11

Step 1: Find the arithmetic means for x and y:

$$\bar{x} = \frac{2+5+2+4+6}{5} = 3.8 \qquad \bar{y} = \frac{4+7+5+8+11}{5} = 7$$

Step 2: Find S_{xy} and S_{xx}:

$$S_{xy} = \sum_{i=1}^{5}(x_i - 3.8)(y_i - 7) = (\mathbf{2} - 3.8)(\mathbf{4} - 7) + (\mathbf{5} - 3.8)(\mathbf{7} - 7) + (2 - 3.8)(5 - 7) + (4 - 3.8)(8 - 7) + (6 - 3.8)(11 - 7) = 18$$

$$S_{xx} = \sum_{i=1}^{5}(x_i - 3.8)^2 = (\mathbf{2} - 3.8)^2 + (\mathbf{5} - 3.8)^2 + (2 - 3.8)^2 + (4 - 3.8)^2 + (6 - 3.8)^2 = 12.8$$

Example 3.2 (Least Squares Linear Regression)

Step 3: Find $\hat{m}$ and $\hat{b}$:

$$\hat{m} = \frac{S_{xy}}{S_{xx}} = \frac{18}{12.8} = 1.4$$

$$\hat{b} = \bar{y} - \hat{m}\bar{x} = 7 - (1.4)(3.8) = 1.7$$

Step 4: Write down the equation:

$$y = 1.4x + 1.7$$

Compare with our freehand approximation:

$$y = 1.3\overline{3}x + 2$$

Interpolation & Extrapolation

- Our linear least squares fitted line is a simple mathematical model for the dataset. Models can be used to describe relationships and make predictions about some physical system.
- As an example, we can use the linear equation to predict y-values for certain x-values:
  - If the given x-value lies *within* the range of x-values for our dataset, then the predicted y-value is an **interpolation**
  - If the given x-value lies *outside* the range of x-values for our dataset, then the predicted y-value is an **extrapolation**

Interpolation & Extrapolation: Example 3.3

Using the data and best fit line found in Example 3.2, estimate the y value that would correspond with the following x values: (a) x = 5, and (b) x = 10. Which corresponds with performing an interpolation and which with an extrapolation?

Solution:

a) x=5: $y = 1.4(5) + 1.7 = 8.7$ interpolation

b) x=10: $y = 1.4(10) + 1.7 = 15.7$ extrapolation

- We'd like to know if our LSR is good.
- Two sets of measurements are "correlated" if there appears to be some relationship between them
- In this section we present formulae applicable only for *linear* correlation
- Correlated does **not** imply causally related; e.g. leaf length and width

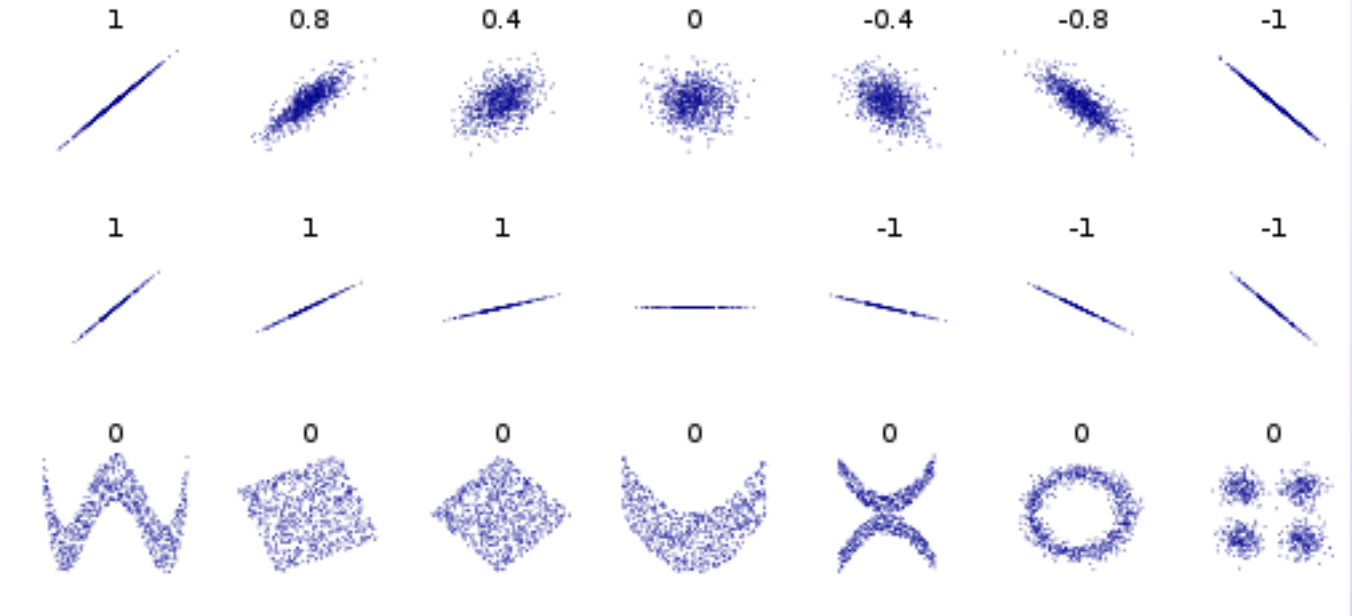
Correlation Coefficient

- The correlation coefficient rho is given by:

$$\rho = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

- This value will always be between -1 and 1
  - If ρ = -1, data are perfectly negatively correlated; if close, highly negatively correlated
  - If ρ = 0, data are uncorrelated; note: this only means they are not linearly related
  - If ρ = 1, data are perfectly positively correlated; if close, highly positively correlated
- ρ^2 x 100 tells us the percent of the variance accounted for by the LSR

Figure 3.1



Various sets of data with their corresponding correlation coefficient shown above each data set. Notice in the bottom row that there can be a correlation in the data, but the value for the correlation coefficient ρ is zero because there is not a linear correlation

Example 3.4

How correlated is the data presented in Example 3.1?

- In order to use the formula, we need to calculate S_{yy}:

$$S_{yy} = \sum_{i=1}^{5} (y_i - 7)^2 = (4-7)^2 + (7-7)^2 + (5-7)^2 + (8-7)^2 + (11-7)^2 = 30$$

- We have, then, the correlation coefficient:

$$\rho = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{18}{\sqrt{(12.8)(30)}} \approx 0.92 \implies \text{highly correlated}$$

- Finally, $\rho^2 = .8464$, so 85% of the variance in the dataset is accounted for by the LSR line

Homework

- Chapter 1: 1.2 - 1.5
- Chapter 2: 2.2–2.4, 2.7, 2.9, 2.10
- Chapter 3: 3.3, 3.4, 3.9

Homework

Exercise 3.4 (d)

The average length and width of various bird eggs are given in the following table.

x	5.8	1.5	2.3	1.0	3.3
y	8.6	1.9	3.1	1.0	5.0

Step 1: Find the arithmetic means for x and y:

$$\bar{x} = \frac{5.8 + 1.5 + 2.3 + 1.0 + 3.3}{5} = 2.78 \qquad \bar{y} = \frac{8.6 + 1.9 + 3.1 + 1.0 + 5.0}{5} = 3.92$$

Step 2: Find S_{xy} and S_{xx}:

$$S_{xy} = \sum_{i=1}^{5} (x_i - 2.78)(y_i - 3.92) = 22.872$$

$$S_{xx} = \sum_{i=1}^{5} (x_i - 2.78)^2 = 14.428$$

Exercise 3.4 (d)

Step 3: Find $\hat{m}$ and $\hat{b}$:

$$\hat{m} = \frac{S_{xy}}{S_{xx}} = \frac{22.872}{14.428} = 1.59$$

$$\hat{b} = \bar{y} - \hat{m}\bar{x} = 3.92 - (1.59)(2.78) = -0.487$$

Step 4: Write down the equation:

$$y = 1.59x - 0.49$$

Homework

Exercise 3.4 (e)

- In order to use the formula, we need to calculate S_{yy}:

$$S_{yy} = \sum_{i=1}^{5} (y_i - 3.92)^2 = ?$$

- We have, then, the correlation coefficient:

$$\rho = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \approx 0.9988 \quad \Rightarrow \text{ highly positively correlated}$$

- Finally, $\rho^2 = .9975$, so 99.75% of the variance in the dataset is accounted for by the LSR line