

STAT 100 Lecture 4: Bivariate Data, Scatter Plots, Linear Regression, and Correlations

Nate Strawn

<http://www.math.umd.edu/~nstrawn/>

1 Lots of definitions

- The Variance: $s^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2$
- The Standard Deviation: $s = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}$
- The Range: $\text{Range} = \max_{1 \leq i \leq N} x_i - \min_{1 \leq i \leq N} x_i$
- The Interquartile Range: $\text{IQR} = Q_3 - Q_1$.

2 Boxplots and Normal Distributions

Today's Agenda

- 1 Visualizing Bivariate Data
- 2 Linear Correlation between Variables
- 3 Linear Regression

Bivariate Data

Bivariate Data is a fancy way to say, “Two-Variable Data.”

Multivariate Data is a fancy way to say, “Many-Variable Data.”

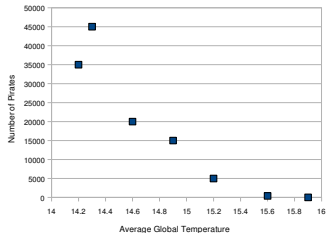
Bivariate Data is a list of numerical pairs

$$(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N).$$

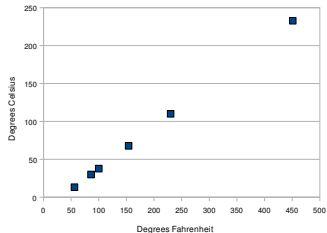
The easiest way to visualize Bivariate Data is through a Scatter Plot.

Scatter Plots

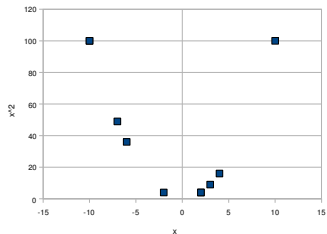
Average Global Temperature vs Number of Pirates



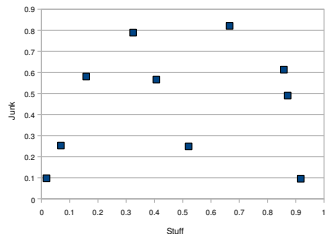
Degrees Fahrenheit vs Degrees Celsius



Integers vs Squares



Stuff vs Junk



Linear Correlation of Bivariate Data

Definition

Given a list of bivariate data $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$, we define the **correlation coefficient** of the x and y variables to be

$$r = \frac{1}{N-1} \sum_{i=1}^N \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right),$$

where s_x and s_y are the standard deviations of the x_1, x_2, \dots, x_N and y_1, y_2, \dots, y_N , respectively.

Formula for the Correlation Coefficient

Formula

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

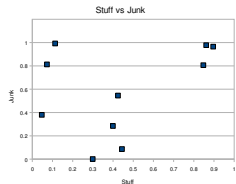
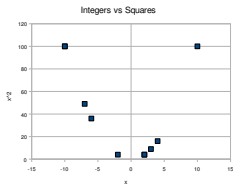
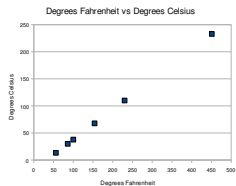
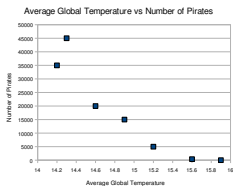
where $S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$, $S_{xx} = \sum(x_i - \bar{x})^2$, and $S_{yy} = \sum(y_i - \bar{y})^2$.

Mean Temp. (x)	Pirates (y)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
14.2	35000	-0.76	17797.57	0.57	316753548	-13475
14.3	45000	-0.66	27797.57	0.43	772704977	-18266
14.6	20000	-0.36	2797.57	0.13	7826405	-999
14.9	15000	-0.06	-2202.43	0	4850691	125
15.2	5000	0.24	-12202.43	0.06	148899263	-2963
15.6	400	0.64	-16802.43	0.41	282321605	-10801
15.9	17	0.94	-17185.43	0.89	295338955	-16203
Tot.=104.7	120417	0	0	2.5	1828695447	-62583
$\bar{x} = 14.96$	$\bar{y} = 17202.43$			S_{xx}	S_{yy}	S_{xy}

We then have that $r = S_{xy}/(\sqrt{S_{xx}}\sqrt{S_{yy}}) \approx -0.93$

Scatter Plots and r-Values

For the following scatter plots, we have $r = -0.93$ and $r = 1.0$ for the first row, and $r = -0.34$ and $r = 0.35$ for the second row.



Standardized Observations of the x -data

Definition

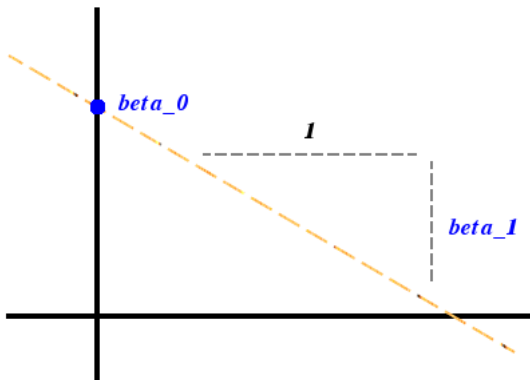
The *Standardized Observations* of the observations x_1, x_2, \dots, x_N are given by

$$\tilde{x}_i = \frac{\text{Observation} - \text{Mean}}{\text{Standard Deviation}} = \frac{x_i - \bar{x}}{s_x}.$$

The conversion to the Standard Observations produces data with mean zero and the standard deviation one. All of the important features of the data are preserved and this conversion makes certain calculations easier.

Quick Review of Linear Equations

Remember that a **line** is both an equation ($y = \beta_0 + \beta_1 x$) and a graph:



Here, β_0 is the y-intercept, and β_1 is the slope of the line.

Definition

The *Least Squares Line* fitting the data

$$(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$$

is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

where

$$\hat{\beta}_1 = S_{xy} / S_{xx} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

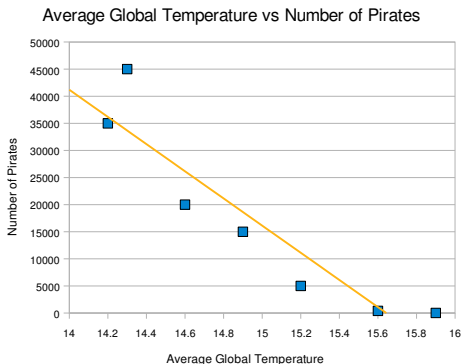
and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Linear Regression Example

For the Global Temperature/ Pirate data, we have that $\beta_1 = -25062.23$, and $\beta_0 = 392061.8$, so the Least Squares line is given by

$$\hat{y} = 392061.8 - 25062.23x.$$



For Next Time

- Quiz 01! Covers Sections 2.1, 2.2, 2.3, 2.4, and 2.5. Problems are taken directly from the book (with small changes!).
- Read Section 3.4, 3.5, and 3.6 from Johnson and Bhattacharyya
- Online homework 3.6: 3.35, 3.37, 3.39, 3.41, 3.43
- Group Problems:

Group	1	2	3	4	5	6	7	8	9	10
Problem	3.38	3.40	3.42	3.46	3.49	3.51	3.52	3.53	3.54	3.55