SOC 3811/5811:
BASIC SOCIAL STATISTICS

Associations Between Continuous Variables

Associations Between Variables

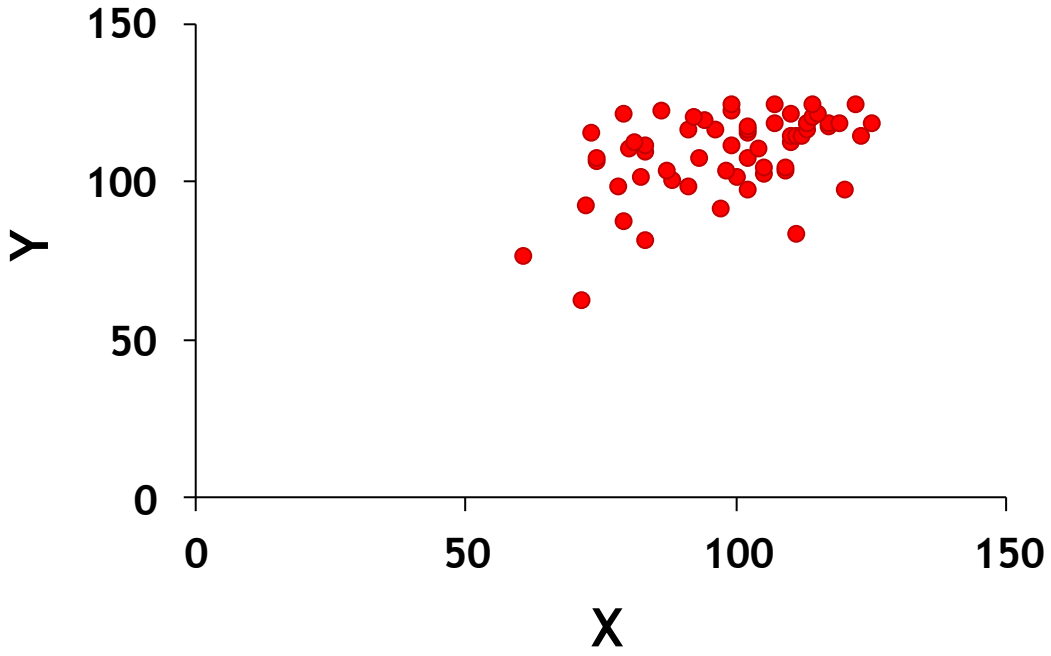
Between now and the 3rd exam we will focus on measuring the association between two variables, X & Y

1. When X is discrete and Y is continuous, we will use "analysis of variance" techniques (Last Tuesday)
2. When X and Y are both discrete, we will use cross-tabular and χ^2 analyses (Last Thursday)
3. When X and Y are both continuous, we will use correlation & regression analyses (This Week)

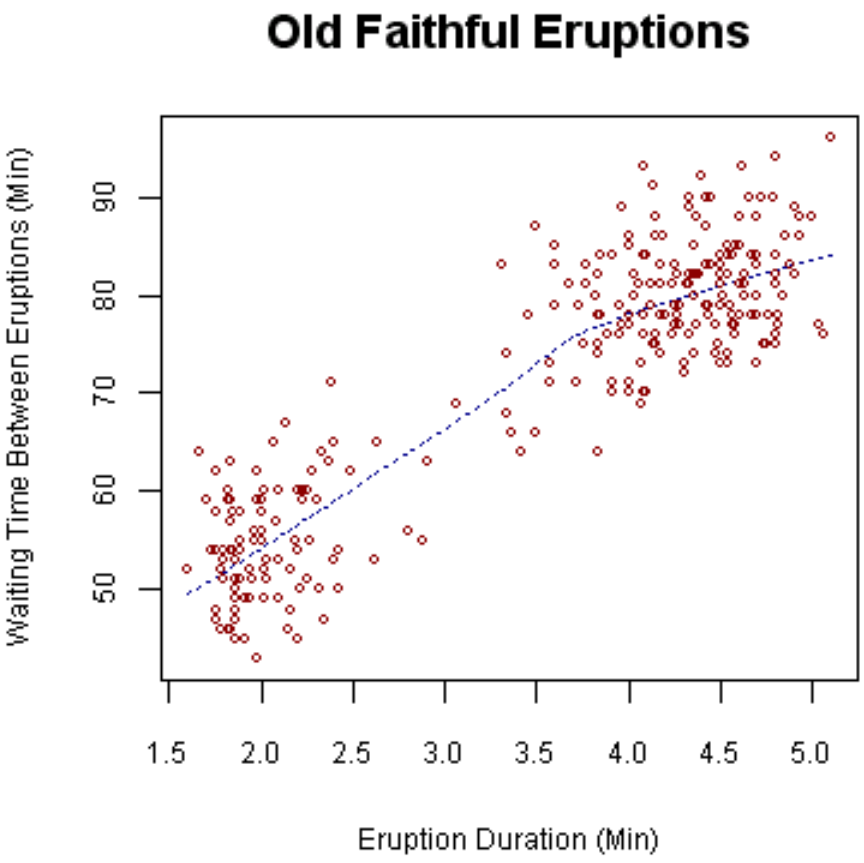
Scatterplots

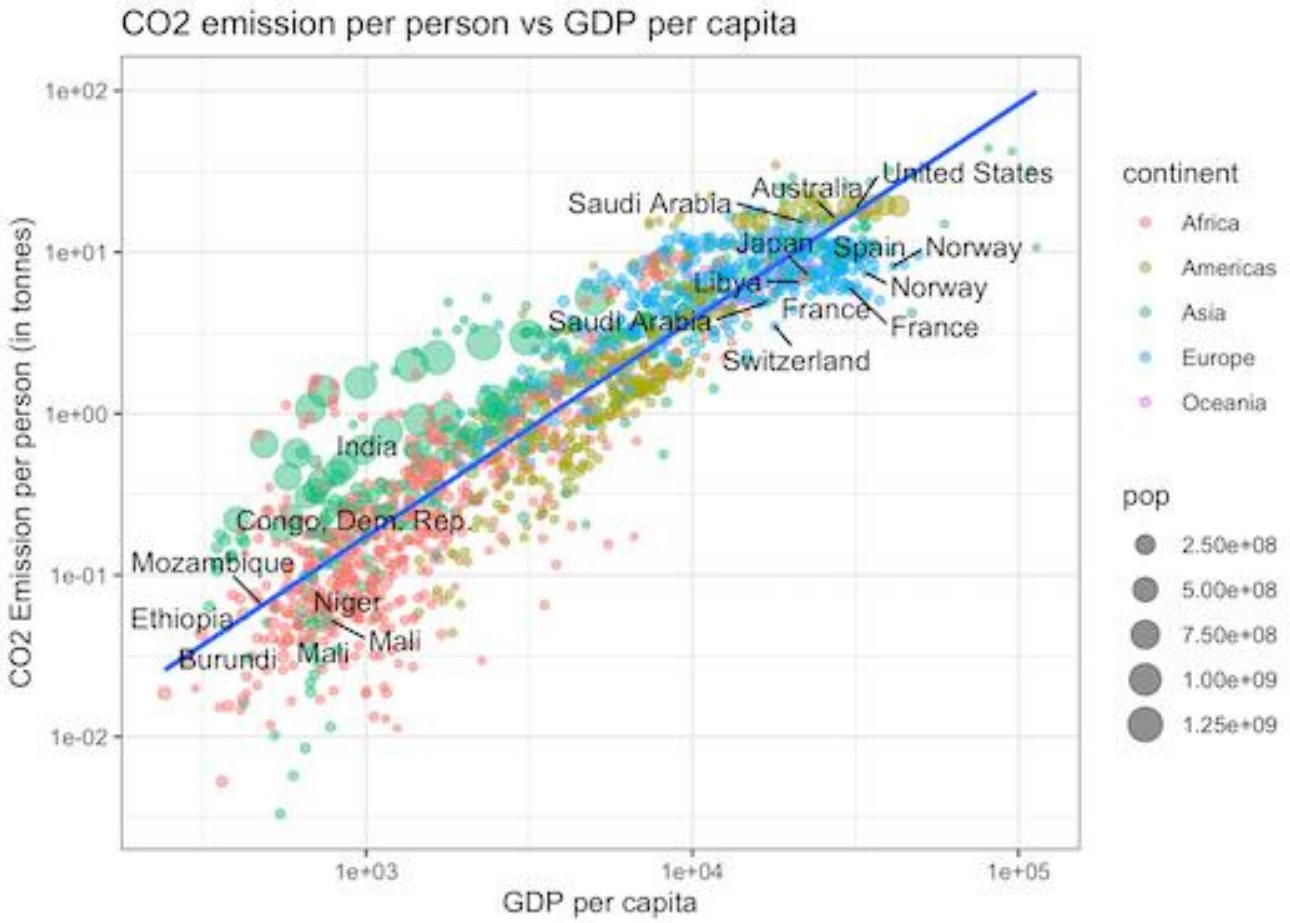
Scatterplot

A diagram that displays the covariation of two continuous variables as a set of points on a Cartesian coordinate system

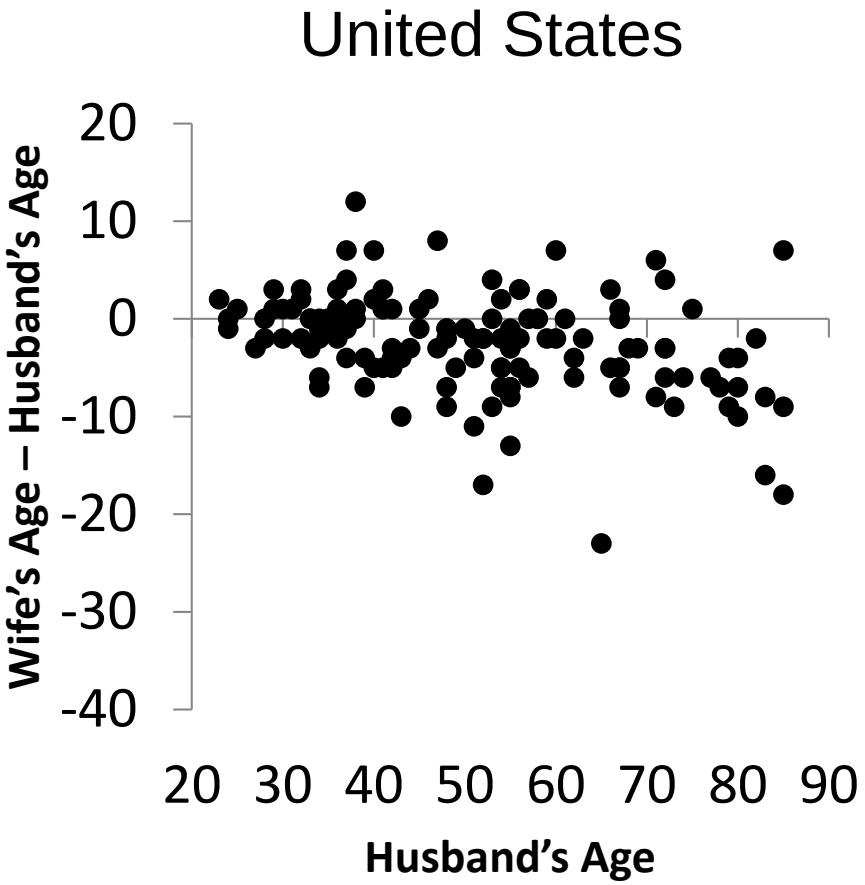


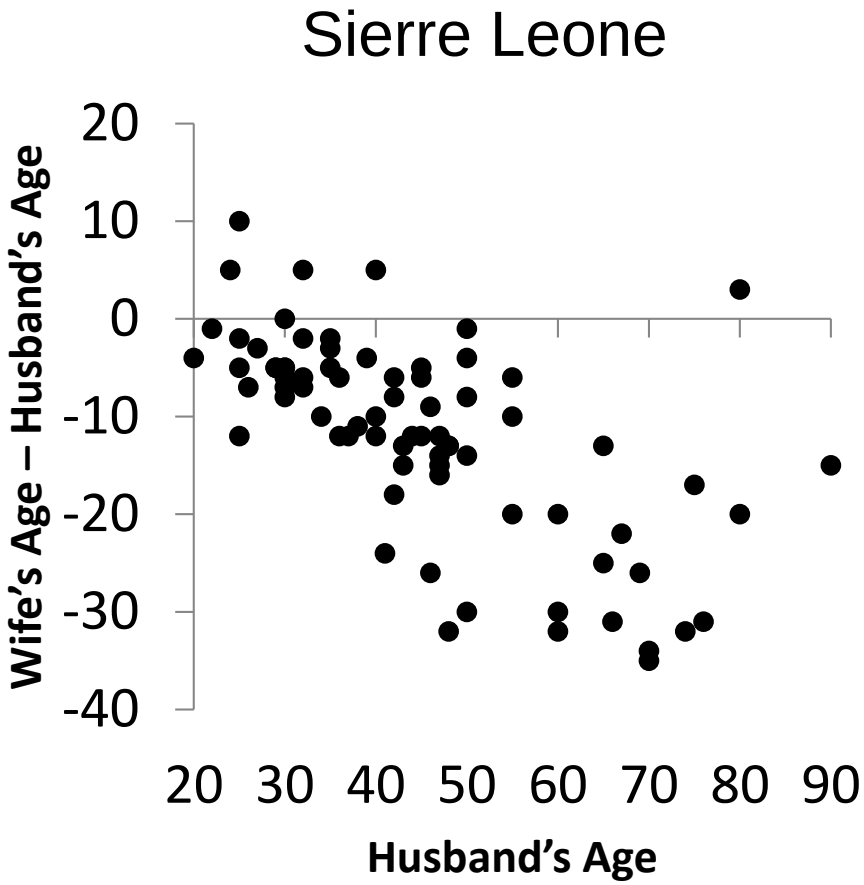
Scatterplots




CO2 emission per person vs GDP per capita
CO2 Emission per person (in tonnes)
1e+02
1e+01
1e+00
1e-01
1e-02
1e+03
1e+04
1e+05
GDP per capita
United States
Australia
Saudi Arabia
Japan
Spain
Norway
Libya
Norway
France
Saudi Arabia
France
Switzerland
India
Congo, Dem. Rep.
Mozambique
Niger
Ethiopia
Mali
Burundi
Mali
continent
Africa
Americas
Asia
Europe
Oceania
pop
2.50e+08
5.00e+08
7.50e+08
1.00e+09
1.25e+09

Scatterplots





Scatterplots

What is the basic shape of the relationship between the two variables? A straight line? A curve? A blob?

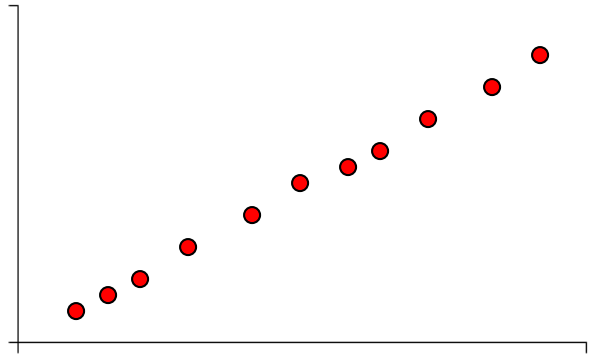
What is the direction of the relationship? Positive, negative, or uncertain?

How much variability is there? How many points deviate from the basic pattern?

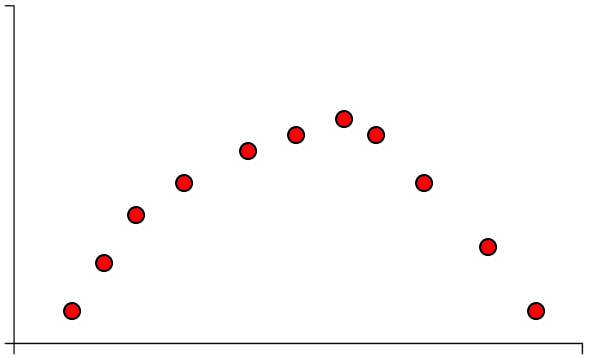
Are there outliers? Unusual observations?

Scatterplots: Shape

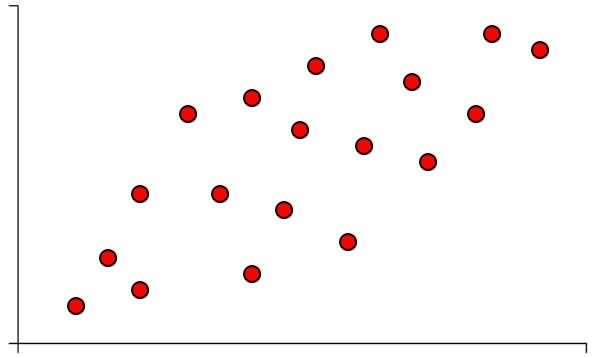
Straight Line



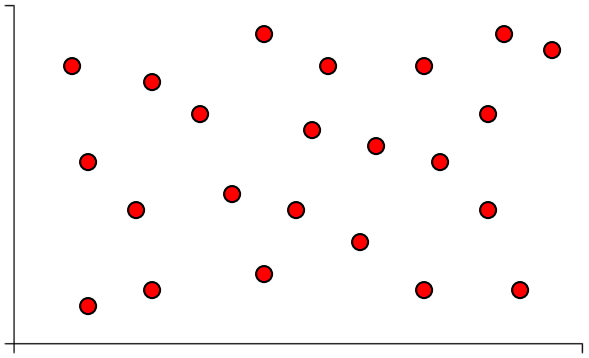
Curved Line



“Football” Shape

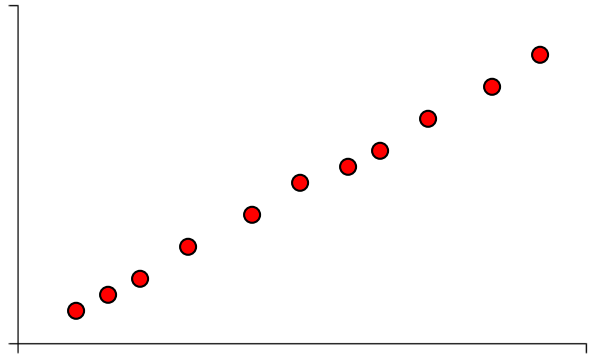


Shapeless Blob

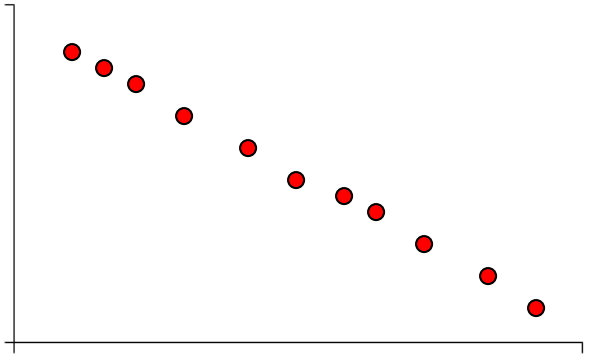


Scatterplots: Direction

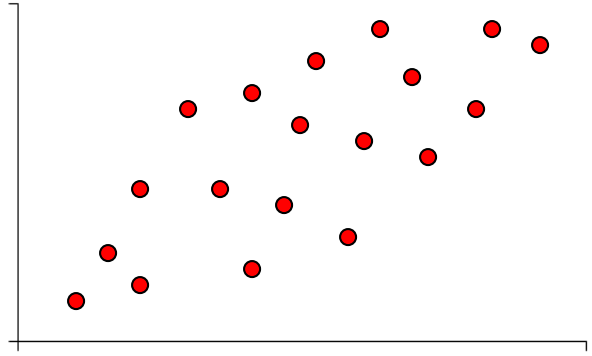
Positive



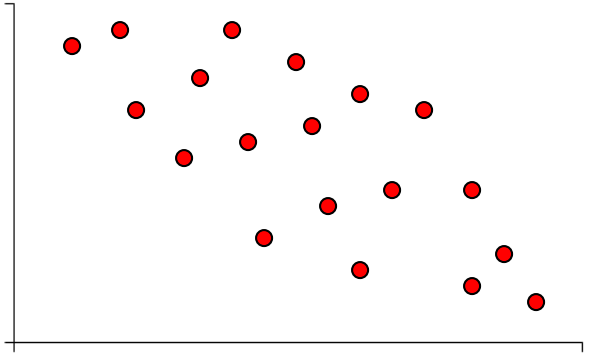
Negative



Positive

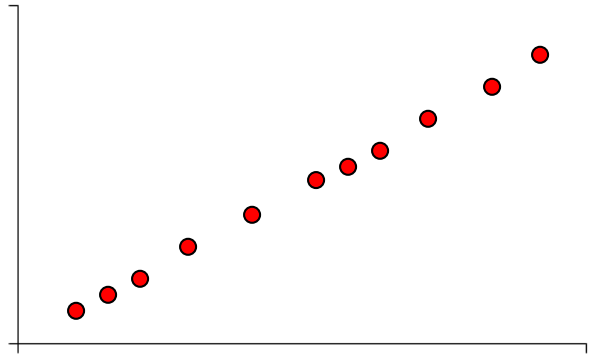


Negative

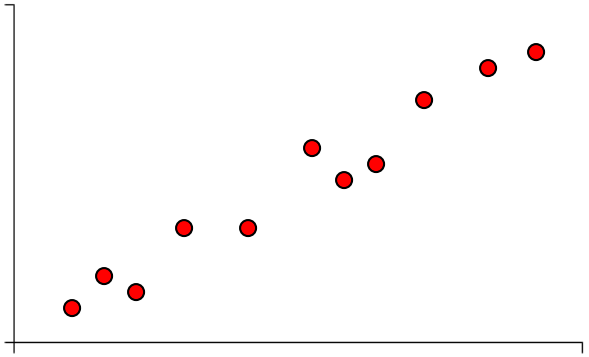


Scatterplots: Variability

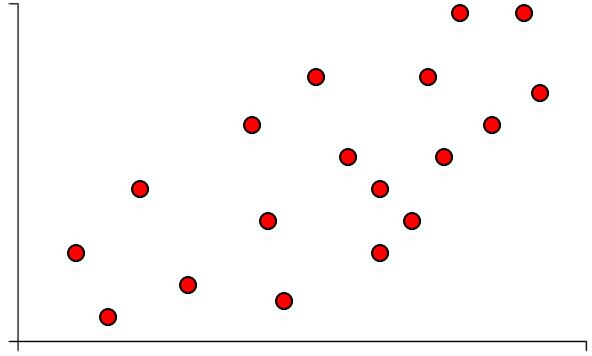
No Variability



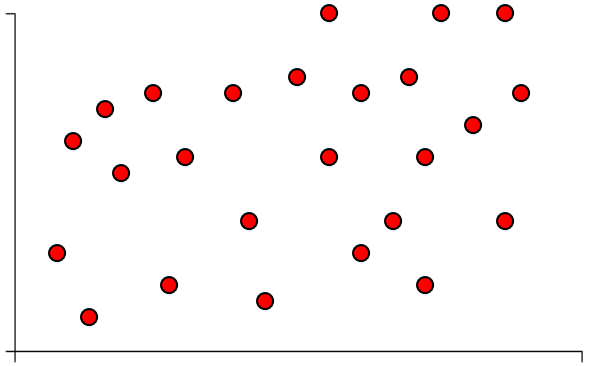
Some Variability



More Variability

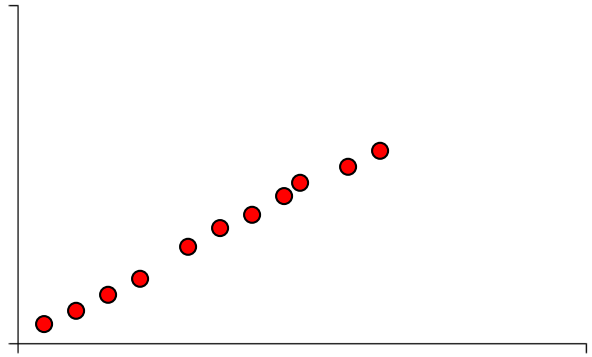


Almost a Shapeless Blob

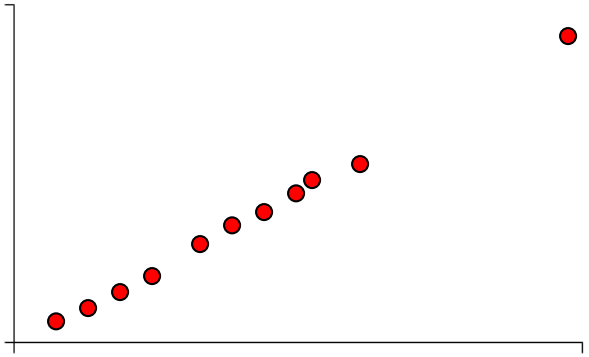


Scatterplots: Outliers

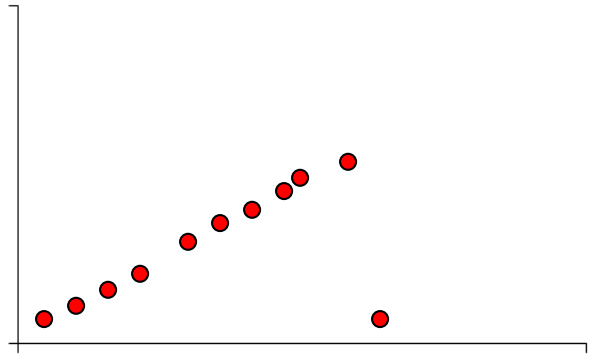
No Outliers



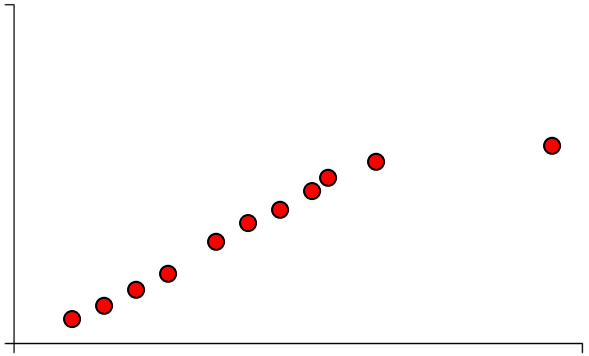
Maybe an Outlier



One Clear Outlier

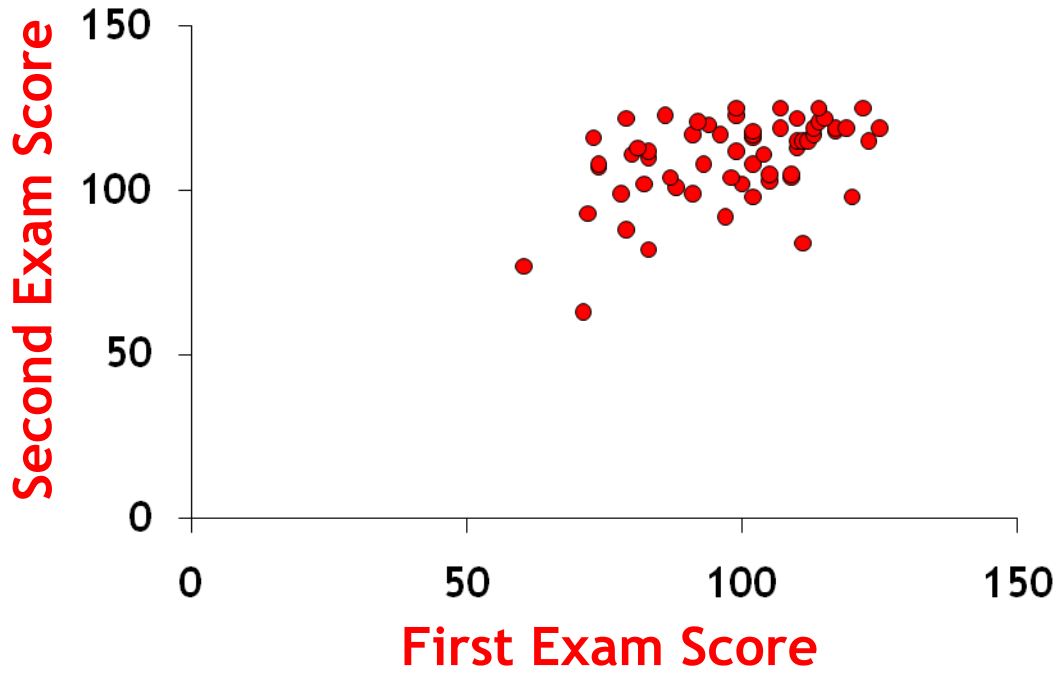


One Likely Outlier



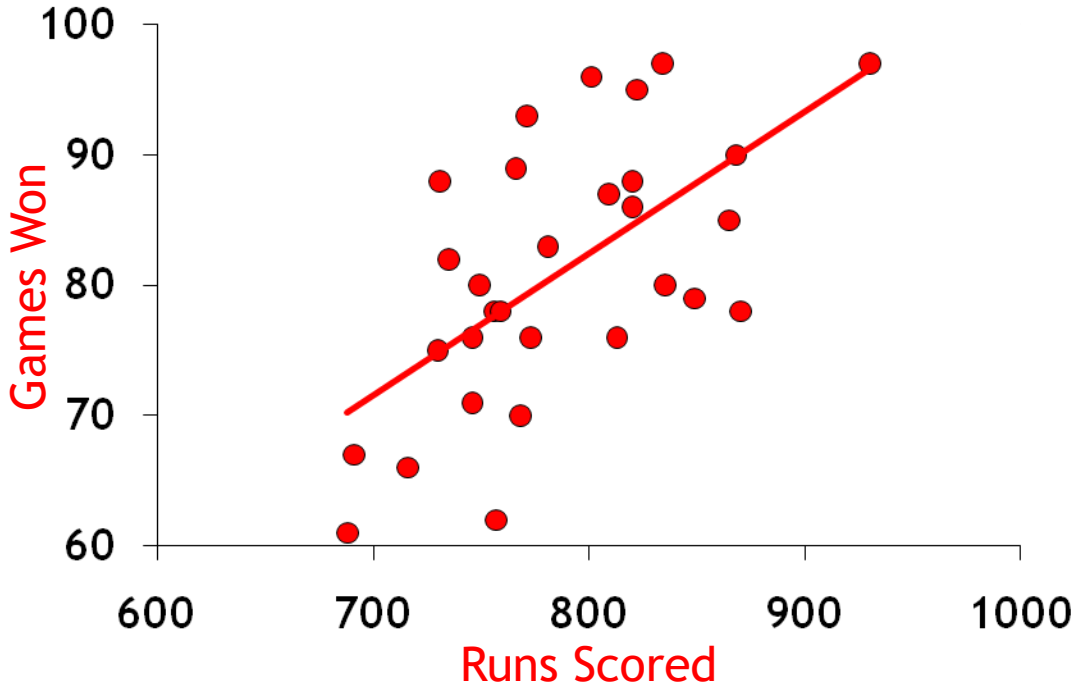
Worksheet

Describe the relationship depicted in this scatterplot



Bivariate Regression

Example: The scatterplot below relates the number of runs scored during the 2019 baseball season to the number of games won by MLB teams in 2019



Bivariate Regression

Conceptually speaking, bivariate regression involves drawing a line through the points on the scatterplot that comes closest to the points on the Y dimension

Regression analysis involves estimating an equation that…

- …describes how, on average, the response variable (Y) is related to the predictor variable (X)
- …allows us to make predictions about the value of the response variable (Y) given a specified value of the predictor variable (X)

When we “regress Y on X” we produce a model that predicts Y on the basis of X

Bivariate Regression

The algebraic equation for a line:

$$Y = a + bX$$

The prediction equation, which expresses the i^{th} individual's value of dependent variable Y as a function of predictor variable X, is:

$$\hat{Y}_i = a + b_{YX} X_i$$

The linear regression model recognizes deviations (or errors, e_i) from the prediction equation:

$$Y_i = a + b_{YX} X_i + e_i$$

Bivariate Regression

Given the prediction equation:

$$\hat{Y}_i = a + b_{YX} X_i$$

and the linear regression model:

$$Y_i = a + b_{YX} X_i + e_i$$

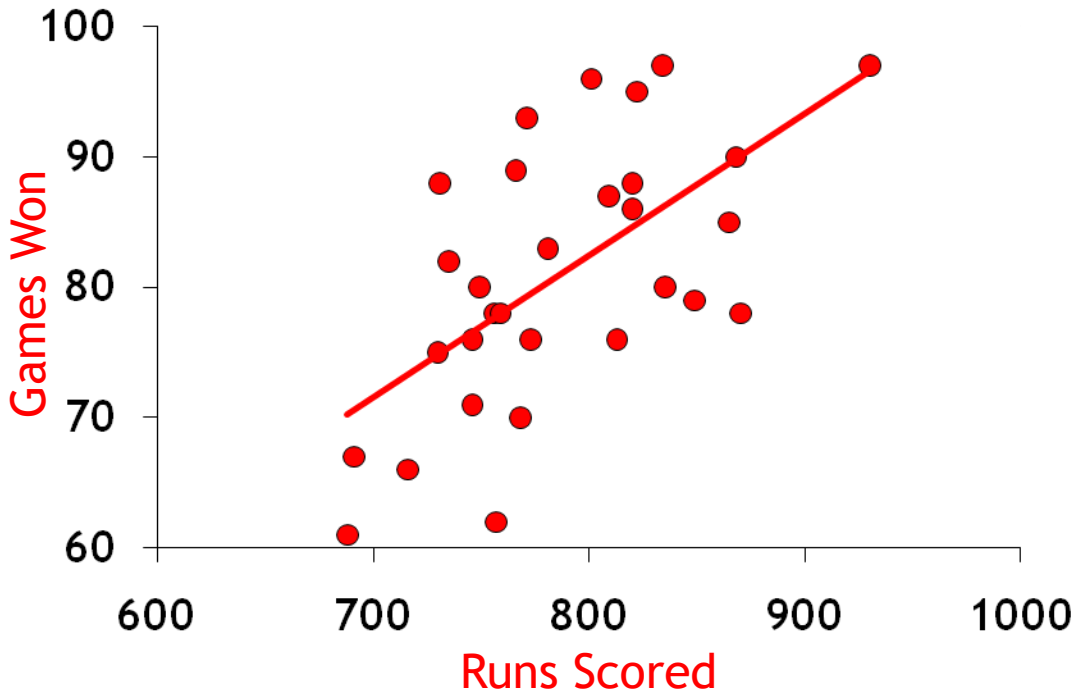
we see that the error term e_i (also know as the residual) can be expressed as the difference between the observed value of Y (from the linear regression model) and the predicted value of Y (from the prediction equation):

$$Y_i - \hat{Y}_i = [a + b_{YX} X_i + e_i] - [a + b_{YX} X_i] = e_i$$

Bivariate Regression

How do we know how to draw the regression line?

There are an infinite number of lines that one could draw through these points ... Why is the middle (red) line best?



Estimating a Regression Equation

The line that we draw … the values of intercept a and slope b_{YX} that we choose … maximizes our ability to predict the value of Y (and thus minimizes the prediction errors)

Mathematically, we choose the line for which

$$\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{N} e_i^2$$

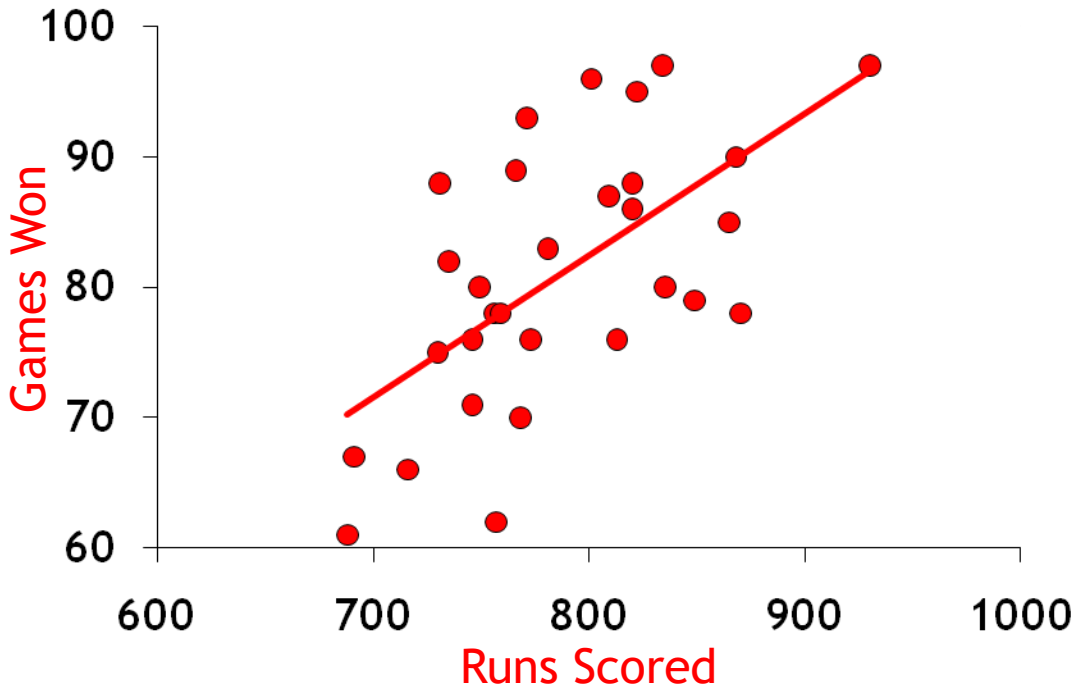
is smallest

This is the "least squares error sum" criterion and produces ordinary least squares (OLS) estimates of intercept a and slope b_{YX}

Estimating a Regression Equation

For the baseball example, the prediction equation is:

$$\hat{Y}_i = -4.76 + 0.11X_i$$



Interpreting the Regression Equation

$$\hat{Y}_i = -4.76 + 0.11X_i$$

How do we interpret this regression equation?

It says that for every one unit increase in X (runs) we should observe a 0.11 unit increase in Y (wins)

It also literally says that if a team were to score zero runs in a season … such that X=0 … we should observe that the team would win -4.76 game (more on this later)

Using the regression equation we can…

…estimate the average value of Y for a given value of X

…predict an individual's value of Y for a given value of X

Interpreting the Regression Equation

The equation $\hat{Y}_i = -4.76 + 0.11X_i$ means (in English) that

$$\text{Expected Number of Wins} = -4.76 + 0.11\text{Runs}$$

How many wins would we predict a team to win if they scored 801 runs?

$$\text{Expected Number of Wins} = -4.76 + 0.11(801) = 83.35$$

What is the average number of wins among teams that score 750 runs?

$$\text{Expected Number of Wins} = -4.76 + 0.11(750) = 77.74$$

How Well Does X Predict Y?

How well does a particular regression equation do in predicting values of the response variable Y?

- How strong the association is between X and Y
- If the association is extremely strong, then knowing X allows you to almost perfectly predict Y
- If the association is weak, then knowing X does nothing to allow you to predict the value of Y

As with ANOVA, we can ask how much of the variation in Y can be attributed to X and how much is random error

- That is, we can partition the variance in Y into the part attributable to X and the part attributable to error

How Well Does X Predict Y?

Start with the deviation:

$$Y_i - \bar{Y} = Y_i - \bar{Y}$$

Then add *and* subtract the predicted value from the right-hand side and rearrange the equation:

$$Y_i - \bar{Y} = Y_i - \bar{Y} + (\hat{Y}_i - \hat{Y}_i)$$

$$Y_i - \bar{Y} = \left(\hat{Y}_i - \bar{Y}\right) + \left(Y_i - \hat{Y}_i\right)$$

Deviations from the mean can be expressed as the sum of (1) deviations of the predicted value from the mean and (2) individual deviations from the predicted value

How Well Does X Predict Y?

$$Y_i - \bar{Y} = \left(\hat{Y}_i - \bar{Y}\right) + \left(Y_i - \hat{Y}_i\right)$$

Portion of the deviation from the mean that is attributable to X

Portion of the deviation from the mean that is attributable to random error

Deviations from the mean can be expressed as the sum of (1) deviations of the predicted value from the mean and (2) individual deviations from the predicted value

How Well Does X Predict Y?

If we square each side and then sum across cases we get:

$$\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2=\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2+\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$$

Total Sum of Squares

Regression Sum of Squares

Error Sum of Squares

$$SS_{TOTAL} = SS_{REGRESSION} + SS_{ERROR}$$

If there is no association between Y and X, then knowing X does not help predict Y

In this case, our best guess about Y-hat is Y-bar; thus $SS_{REGRESSION}$ equals zero and SS_{TOTAL}=SS_{ERROR}

How Well Does X Predict Y?

The coefficient of determination (R^2_{YX}) indicates the proportion of the total variation in Y that is determined by its linear relationship with X

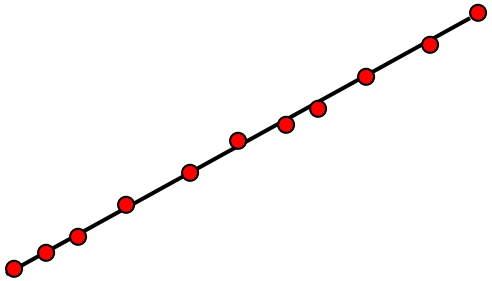
$$R^2_{YX} = \frac{SS_{TOTAL} - SS_{ERROR}}{SS_{TOTAL}} = \frac{SS_{REGRESSION}}{SS_{TOTAL}}$$

If $R^2_{YX}=0$, then $SS_{REGRESSION}=0$, which suggests that there is no association between Y and X

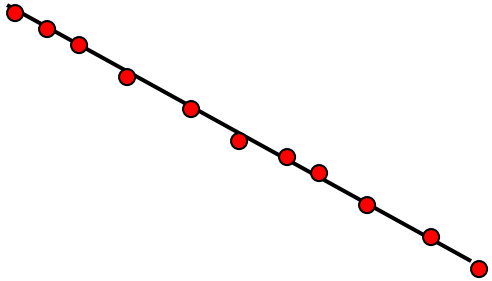
If $R^2_{YX}=1$, then $SS_{REGRESSION}=SS_{TOTAL}$, which suggests that there is no error variation and that we can perfectly predict Y based on X

How Well Does X Predict Y?

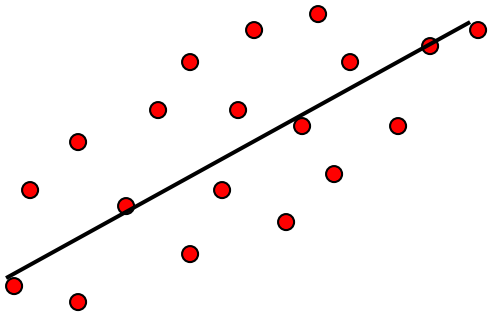
$R^2_{YX} = 1.0$



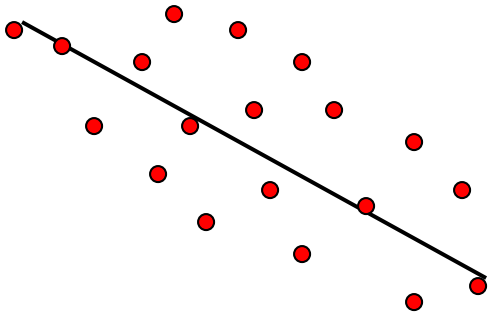
$R^2_{YX} = 1.0$



$R^2_{YX} = 0.25$

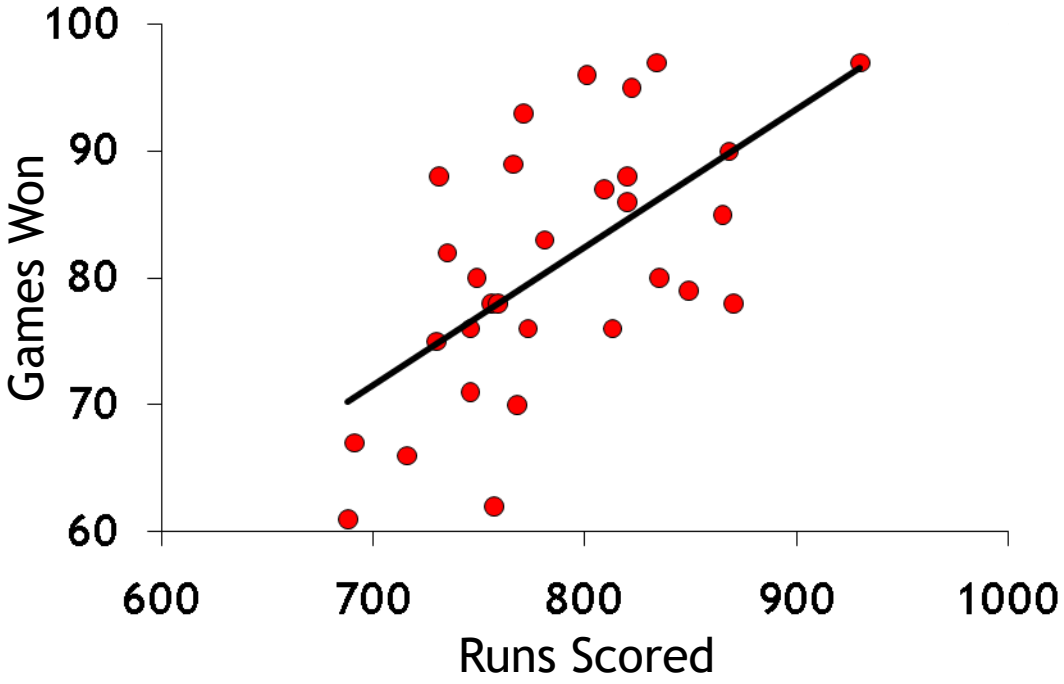


$R^2_{YX} = 0.25$



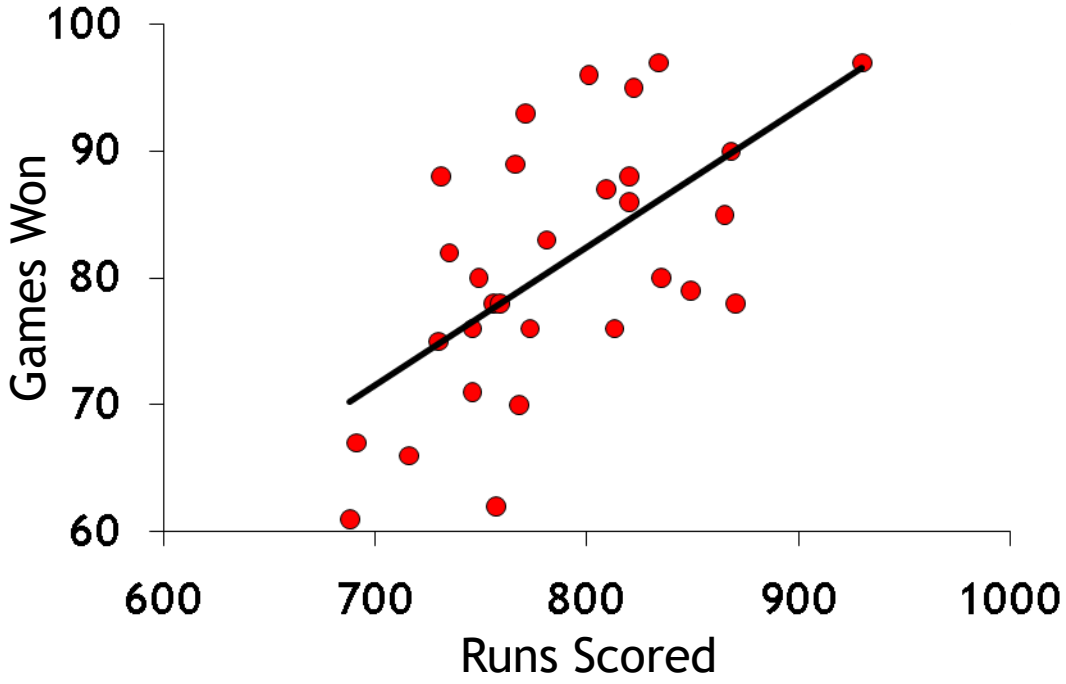
How Well Does X Predict Y?

For the baseball example: $R^2_{YX} = \frac{SS_{REGRESSION}}{SS_{TOTAL}} = \frac{1128.163}{2948.967} = 0.383$

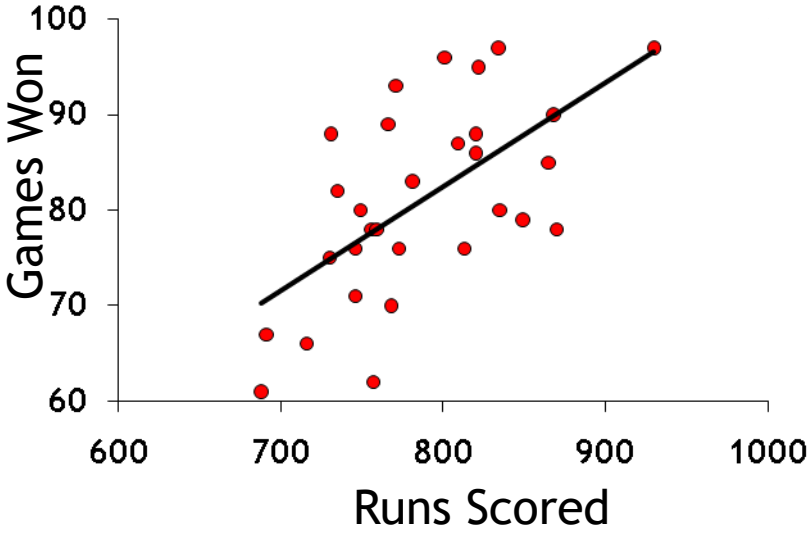


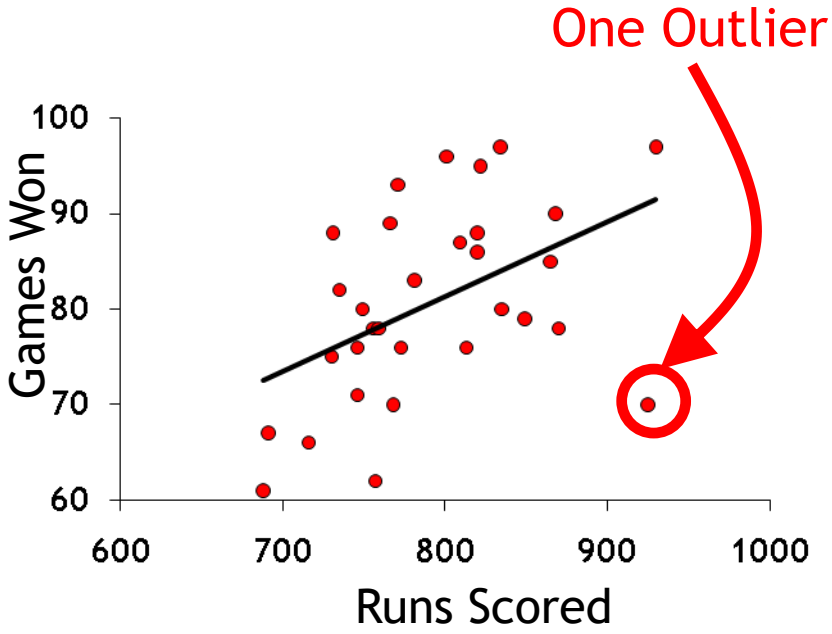
How Well Does X Predict Y?

An R^2_{YX} of 0.383 means that 38.3% of the variation in Y (Wins) is "explained" by X (Runs Scored)



Caution I: Outliers?





$R^2_{YX} = 0.383$

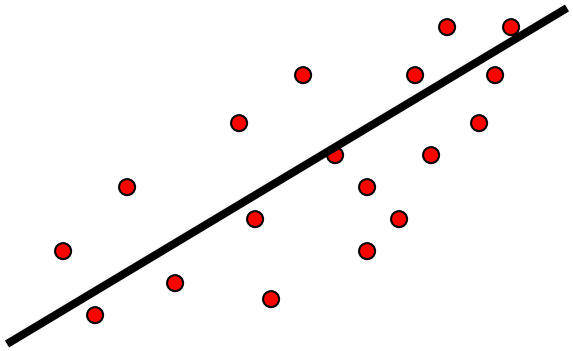
$\hat{Y}_i = -4.764 + 0.109X_i$

$R^2_{YX} = 0.227$

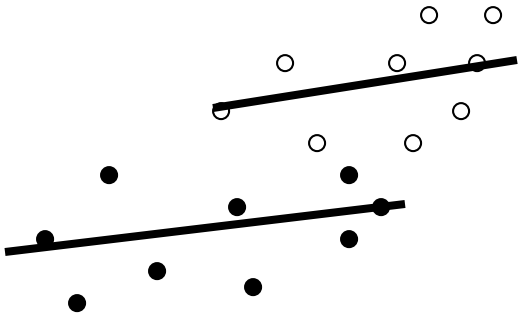
$\hat{Y}_i = 18.703 + 0.078X_i$

Caution II: One Population?

All Cases Included in One Group

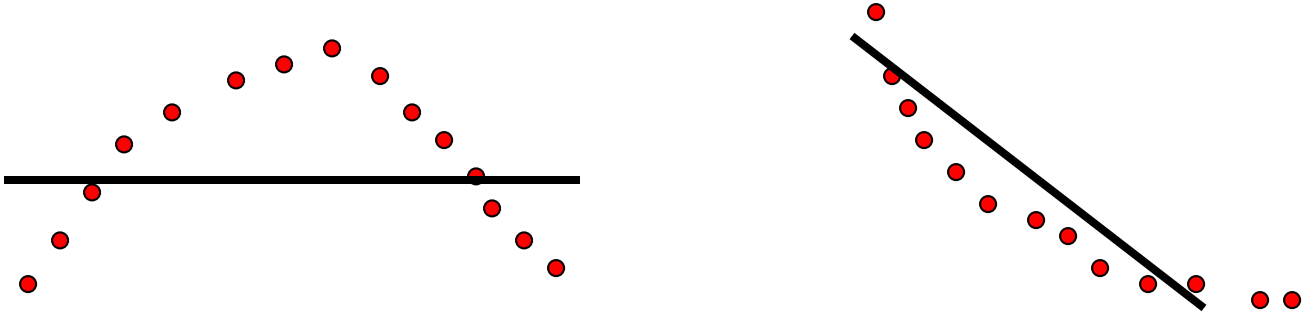


Cases Analyzed as Two Distinct Groups



Caution III: Linear Association?

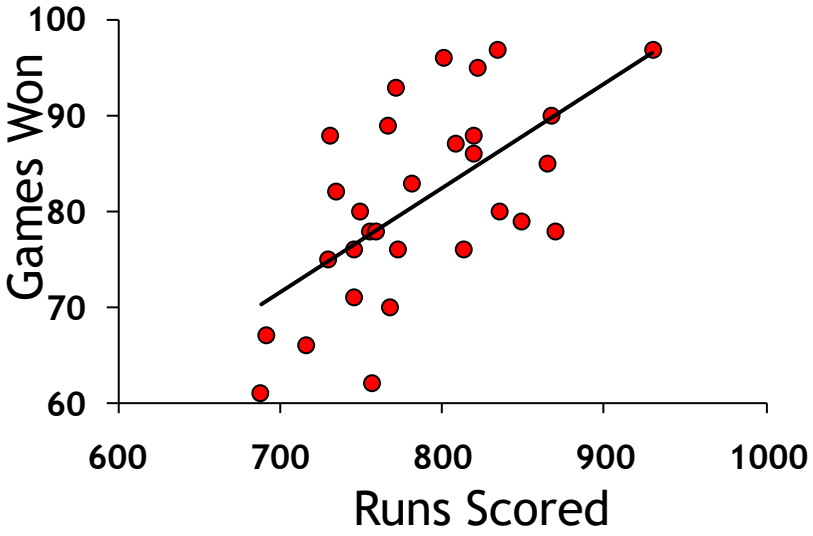
Linear Regression Equations
Fit to Curvilinear Data



Caution IV: Predicting Beyond X

Literally, the equation to the right says that a team would win -4.764 games if they scored zero runs

<u>Lesson</u>: Don't make predictions beyond the observed range of X (~700 to ~900 runs in this case)



$$R^2_{YX} = 0.383$$

$$\hat{Y}_i = -4.764 + 0.109X_i$$

Caution V: Association is Not Causation

Today we are discussing methods for describing the association between continuous variables

In order to make causal statements—for example, that X affects Y … several other things have to be true—more on all of this later

Just because X and Y are highly associated does not necessarily mean that X causes Y …

- It may be that Y causes X instead
- Some third variable may completely account for the correlation
- Some third variable may partially account for the correlation

Correlation

The correlation coefficient (r_{YX}) summarize the strength and direction of the association between two continuous variables

Correlation equals the square root of R^2_{YX}, but it can also be computed as

$$r_{YX} = \left(\frac{1}{n-1}\right)\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_x}\right)\left(\frac{Y_i - \bar{Y}}{s_y}\right)$$

Correlation

Correlation always ranges from -1 to +1

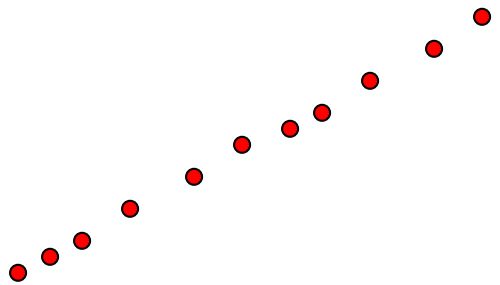
Correlations between 0 and +1 indicate a positive relationship; if r_{YX}=+1, then there is a perfect positive association

Correlations between -1 and 0 indicate a negative association; if r_{YX}=-1, then there is a perfect negative association

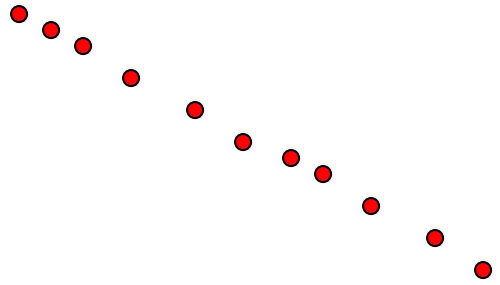
If r_{YX}=0, there is absolutely no association

Correlation

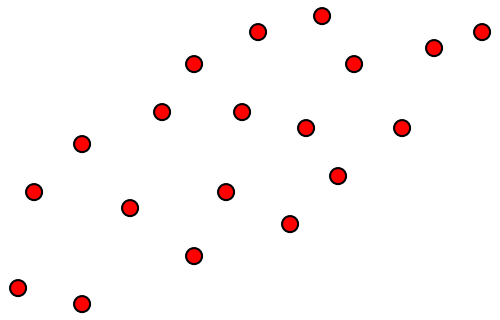
r_{YX} = +1.0



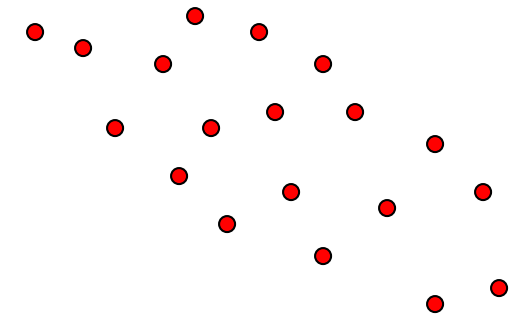
r_{YX} = -1.0



r_{YX} = +0.5



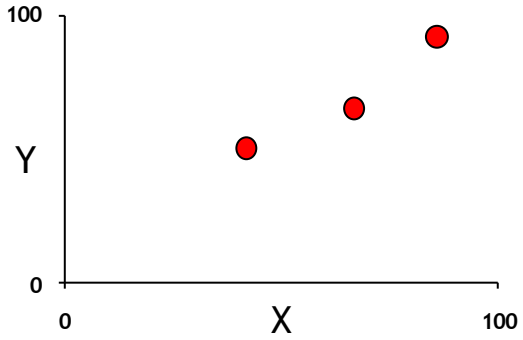
r_{YX} = -0.5



Correlation: Example

Consider the following data on three students' scores on a mid-term exam and on a final exam:

Student:	1	2	3
X: Mid-term:	86	67	42
Y: Final Exam:	92	65	50



Here are the summary statistics for each variable:

	X	Y
Mean	65	69
Standard Deviation	22.1	21.3
N = 3		

Correlation: Example

The formula for correlation looks complicated, but it only involves the means and standard deviations of the two quantitative variables

$$r_{YX} = \left(\frac{1}{n-1}\right)\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_x}\right)\left(\frac{Y_i - \bar{Y}}{s_y}\right)$$

In our example:

$$r_{YX} = \left(\frac{1}{3-1}\right)\left[\left(\frac{86-65}{22.1}\right)\left(\frac{92-69}{21.3}\right) + \left(\frac{67-65}{22.1}\right)\left(\frac{65-69}{21.3}\right) + \left(\frac{42-65}{22.1}\right)\left(\frac{50-69}{21.3}\right)\right]$$

$$r_{YX} = \left(\frac{1}{2}\right)\left[(.95)(1.08) + (0.09)(-0.19) + (-1.04)(-.89)\right] = \left(\frac{1}{2}\right)(1.935)$$

$$r_{YX} = 0.97$$

Correlation and Slope

r_{YX} is not the same as the slope b_{YX}, but there is a close relationship between the two:

$$b_{YX} = r_{YX} \frac{s_Y}{s_X}$$

Recommended Formulas

The easiest way (and order in which) to compute the statistics we covered today:

Compute the mean and standard deviation of each variable

Compute r_{YX} as $r_{YX} = \left(\frac{1}{N-1}\right)\sum_{i=1}^{N}\left(\frac{X_i - \bar{X}}{s_x}\right)\left(\frac{Y_i - \bar{Y}}{s_y}\right)$

Compute the slope b_{YX} as $b_{YX} = r_{YX}\frac{s_Y}{s_X}$

Compute the intercept a as $a = \bar{Y} - b\bar{X}$

Compute R^2_{YX} as $R^2_{YX} = r^2_{YX}$

Worksheet

Here are values of 2 variables, X and Y, for n=4 people:

	X	Y	
Person #1	3	10	$\bar{X} = 4$
Person #2	5	8	$s_X = 1.826$
Person #3	2	11	$\bar{Y} = 10$
Person #4	6	11	$s_Y = 1.414$

Using these data and summary statistics:

1. Draw a scatterplot
2. Compute and interpret r_{YX}
3. Compute and interpret the least-squares regression line; draw it on your scatterplot
4. Compute and interpret R^2_{YX}

Want More?

David Lane's Books

http://onlinestatbook.com/2/regression/regression.html

http://davidmlane.com/hyperstat/prediction.html

Stat Trek

http://stattrek.com/regression/linear-regression.aspx

Lowry's Book (Chapter 3)

http://vassarstats.net/textbook/

Dallal's Book (see "Simple Linear Regression" section)

http://www.jerrydallal.com/LHSP/LHSP.htm