

Correlation and Introduction to Regression

Tom Ilvento
FREC 408

What's Next?

- Correlation and Regression
- These techniques are an extension of Analysis of Variance (ANOVA)
- Correlation is a measure of association between two variables
- Regression allows us to make estimates of the relationship between two or more variables
 - The estimates, called coefficients, are estimates of population parameters and can therefore be tested for statistical significance

What's Next

- Regression allows us to make estimates of relationships
 - Between factors (independent variable) and a response variable (dependent variable)
 - Independent variables which can be continuous or dichotomous
- **While controlling for the simultaneous effects of other independent variables**

Correlation and Regression

- A focus on the variance
- A focus on the co-variance
- A focus on the equation of a line

Measure of Association

- Measures designed to show the relationship between two variables
- We already looked at R^2 and odds ratios as measures of association
- Important criteria
 - What is the range (low to high)
 - Is it bounded?
 - Is it symmetrical?
 - How to interpret?

Correlation and Covariance

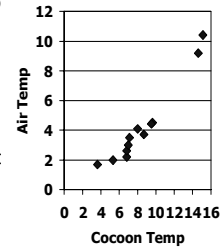
- Suppose we had two variables
- And we want to understand the relationship between them
 - Mean daily air temperature (X_1) and Mean daily cocoon temperature (X_2) of Woolly-bear caterpillars in the High Arctic
 - Data is collected for each variable for 12 days

Data for Woolly-bear caterpillars in the High Arctic

Day	Air Temp	Cocoon Temp
1	10.4	15.1
2	9.2	14.6
3	2.2	6.8
4	2.6	6.8
5	4.1	8.0
6	3.7	8.7
7	1.7	3.6
8	2.0	5.3
9	3.0	7.0
10	3.5	7.1
11	4.5	9.6
12	4.4	9.5

Look at the data as a X-Y Scatter Plot

- This shows how Air Temp and Cocoon Temp vary together
- As one increases the other also increases
- As one decreases the other decreases
- But the relationship is not exact
- This concept is called **Covariance**



Covariance

- We have been interested in how variables vary with respect to their means
- $\sum(x-\bar{x})^2$ = TSS - Total Sum of Squared Deviations
- $\sum(x-\bar{x})^2/(n-1)$ = MS - Mean Squared Deviation

Covariance

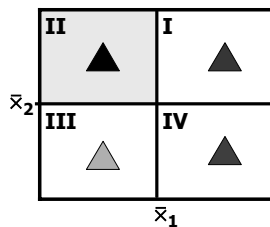
- Covariance** looks at how two variables vary about their means together, divided by n

$$Cov_{x_1x_2} = \frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{n}$$

$$Cov_{x_1x_2} = \frac{SS_{x_1x_2}}{n}$$

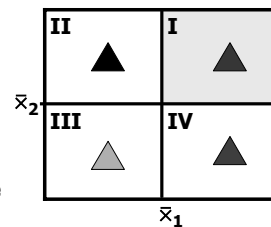
Covariance

- In quadrant II
- We have positive deviations around \bar{x}_2 and negative deviations about \bar{x}_1



Covariance

- In quadrant I we have positive deviations around \bar{x}_2 and positive deviations about \bar{x}_1



Covariance

- In quadrant III we have negative deviations around \bar{x}_2 and negative deviations about \bar{x}_1

Covariance

- In quadrant IV we have negative deviations around \bar{x}_2 and positive deviations about \bar{x}_1

Negative Covariance

Positive Covariance

Covariance

- The covariance between two variables is a useful concept
- But as a measure of association it has limits
 - It is unbounded – no high or low
 - Expressed in cross product units
- One way to normalize the covariance is to divide through by the product of the standard deviations
- This is called a **Pearson Correlation, r**

Correlation

$$r = \frac{Cov_{x_1x_2}}{\sigma_{x_1} \sigma_{x_2}}$$

$$r = \frac{\sum (x_1 - \bar{x}_1)(x_2 - \bar{x}_2)}{\sqrt{\sum (x_1 - \bar{x}_1)^2 \sum (x_2 - \bar{x}_2)^2}}$$

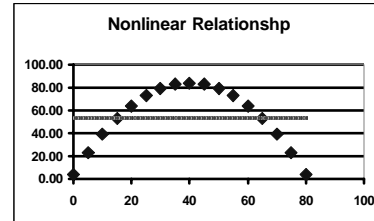
$$r = \frac{SS_{x_1x_2}}{\sqrt{SS_{x_1} SS_{x_2}}}$$

Note: the sample size for each variable must be equal for correlation and regression.

Pearson Correlation

- Also known as the Pearson Product Moment Correlation
- r has a range from -1 to 1
 - 1 means perfect negative correlation
 - 1 means perfect positive correlation
 - 0 means no linear association

Pearson r is a Linear Measure of Association



In this case, the Pearson r equals zero

Pearson Correlation Coefficient

- Properties of r**
 - Bounded -1 to 1
 - Symmetrical $r_{x_1x_2} = r_{x_2x_1}$
 - Invariant to scale –
 - if you add/subtract a constant or multiply/divide by a constant to every value in the distribution, it does not change the relationship
 - r will remain the same

Computation formula for r

$$r = \frac{\sum x_1x_2 - \frac{\sum x_1 \sum x_2}{n}}{\sqrt{\left(\sum x_1^2 - \frac{(\sum x_1)^2}{n}\right) \left(\sum x_2^2 - \frac{(\sum x_2)^2}{n}\right)}}$$

Note: correlation requires the same number of observations for both variables – in essence matched data pairs

Computational formula for r

$$r = \frac{\sum x_1x_2 - \frac{\sum x_1 \sum x_2}{n}}{\sqrt{\left(\sum x_1^2 - \frac{(\sum x_1)^2}{n}\right) \left(\sum x_2^2 - \frac{(\sum x_2)^2}{n}\right)}}$$

- sum cross products $\sum x_1x_2$
- sum(x1) * sum(x2)/n $\sum x_1 \sum x_2/n$
- Difference #1- #2 $\sum x_1x_2 - \sum x_1 \sum x_2/n$
- Total SS x_1 $\sum x_1^2 - (\sum x_1)^2/n$
- Total SS x_2 $\sum x_2^2 - (\sum x_2)^2/n$
- SQRT(SS x_1 *SS x_2)
- #3/#6

Cocoon Example

- $\sum x_1 = 51.30$
- $\sum x_2 = 102.10$
- $\sum x_1^2 = 302.65$
- $\sum x_2^2 = 996.21$
- $\sum x_1 x_2 = 536.56$

$$r = \frac{536.56 - \frac{(51.30)(102.10)}{12}}{\sqrt{\left(302.65 - \frac{51.30^2}{12}\right) \left(996.21 - \frac{102.10^2}{12}\right)}} = .970855$$

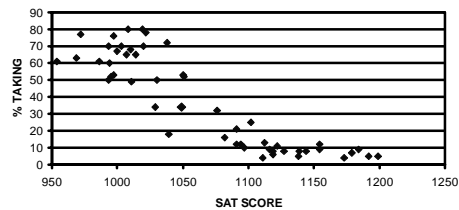
Covariance and Correlation with Excel

- Tools, Data Analysis, Correlation
- Data must be in columns next to each other
- You can (and should) include labels

Correlation			
	Air	Air	Cocoon
	Cocoon	0.970855	1
Covariance			
	Air	Air	Cocoon
	Air	6.945208	
	Cocoon	8.340208	10.62576

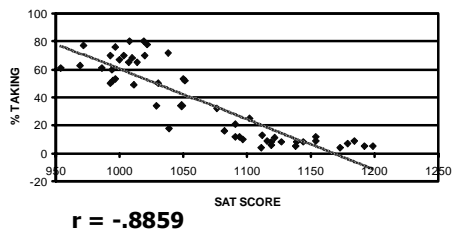
Correlation example: SAT data – shows a negative relationship

Scatter Plot of State SAT SCORES vs % Taking



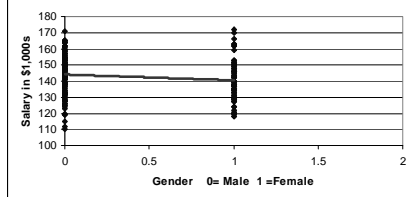
Correlation Example: SAT Data

Scatter Plot of State SAT SCORES vs % Taking



Correlation example – using a Dichotomy for Gender

Income by Gender



$r = -.138$

Student Health Data Example

- On the web site is an Excel file with data on a study of student health
- The variables for 312 students are:
 - Blood pressure
 - Cholesterol Level
 - Weight
 - Sex (1 = Female, 0 = male)

Student Health Data Example

	Weight	Cholesterol	Blood Pressure	Gender
Mean	152.567	198.093	117.449	0.474
Standard Error	1.789	0.668	0.694	0.028
Median	150.000	197.500	117.000	0.000
Mode	126.000	196.000	113.000	0.000
Standard Deviation	31.606	11.791	12.267	0.500
Sample Variance	998.947	139.036	150.480	0.250
Kurtosis	-0.700	-0.176	-0.079	-2.002
Skewness	0.212	-0.060	0.228	0.103
Range	163	61	68	1
Minimum	86	166	88	0
Maximum	249	227	156	1
Sum	47601	61805	36644	148
Count	312	312	312	312

Correlation Matrix of Health Data

	BLOOD			
	WEIGHT	CHOLEST	PRESSURE	GENDER
WEIGHT	1			
CHOLESTEROL	0.15847302	1		
BLOOD PRESSURE	0.48222098	0.188642212	1	
GENDER	-0.8067176	0.179513648	-0.32357819	1

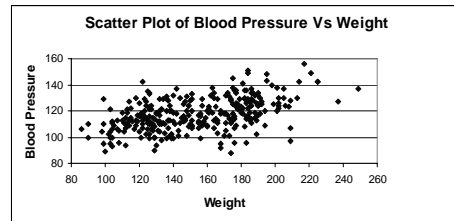
Remember: Gender is a dummy variable where

1 = female

0 = male

N = 312

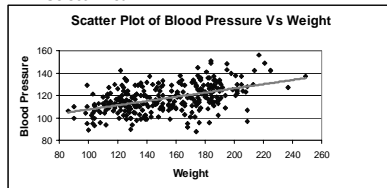
Plot of Blood Pressure v.s. Weight



$r = .482$ As weight increases, so does Blood Pressure

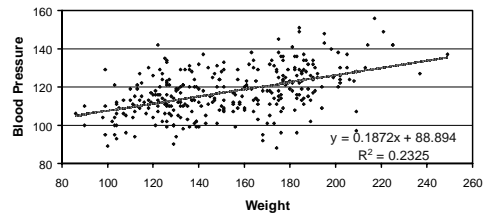
Excel will add a Trend Line based on a regression

- Click on the chart
- Under "Chart" at the top menu bar
 - Trendline
 - Select Linear



There are options to specify the regression equation and R-square

Weight vs Blood Pressure



Thoughts about correlation

- It is a **linear** measure of association
- It is invariant to scale
- It works well with a scatter plot of the data
- Correlation IS NOT Causality
 - The fact that two things are correlated does not mean one causes the other
- Example:
 - In summer, there is a correlation between ice cream sales and the number of people who drown
 - This does not mean that eating ice cream causes people to drown