Correlation and Predicting Quality of wine using correlations

Presented by:

Avish Jadwani

CORRELATION

- Denotes the relation between variables
- Ranges from -1.00 to 1.00
- The closeness to 1.00 or -1.00 determines the closeness among the variables
- The formula that we have used in our study of correlation is Pearson's correlation
- $\rho = \dfrac{\mathrm{cov}(X,Y)}{S_X S_Y}$, S_X & S_Y are standard deviations
- Usually inferences made by using scatter plots
- More the formation is linear on a scatter plot the close is the relation

TYPES OF CORRELATION

- **Pearson's** – evaluates linear relationship between two continuous variables
- **Spearman rank order correlation** : evaluates relationship in form of rank amongst the monotonic variables.
- Used for variables with non linear relationship
- Monotonic variables change but not necessarily at a constant rate i.e they could either increase or decrease
- $r_S = \frac{6\sum d_i^2}{n(n^2-1)}$, $d_i = difference\ of\ ranked\ models$, n = number of samples

- **Kendall's Tau :** used to measure the degree of correspondence between sets of rankings where measures are not equidistant
- $\boldsymbol{\tau = \frac{C-D}{\binom{n}{2}}}$ **,** C = Concordant pairs, D = discordant pairs
- Concordant pair: rank of 2nd variable > rank of former variable
- Discordant pair: rank≤ rank of first variable
- Please use computer for Kendall's tau method

PEARSON SCATTER PLOTS

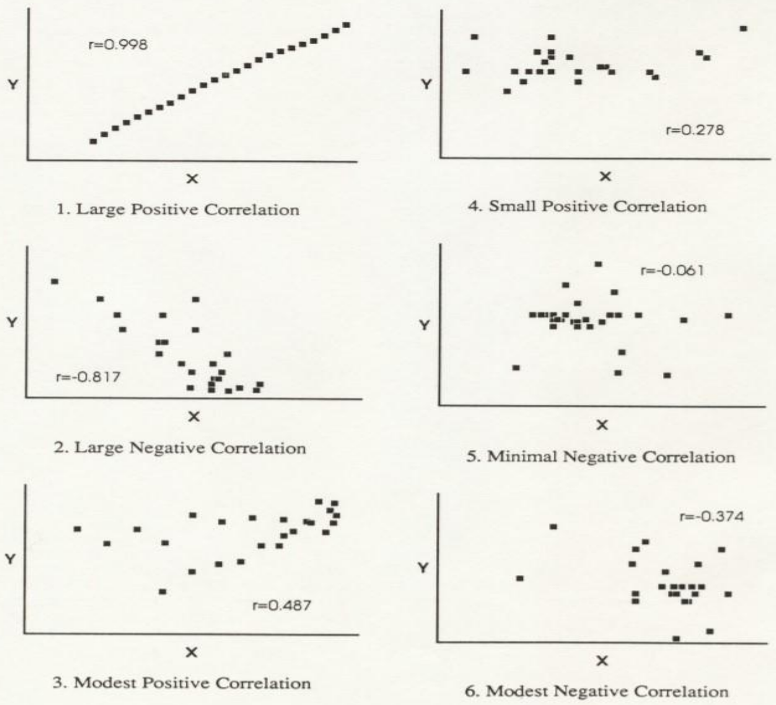
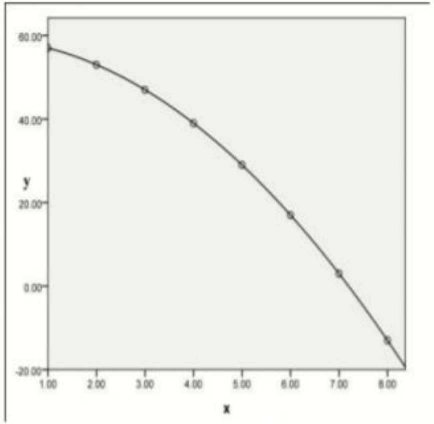


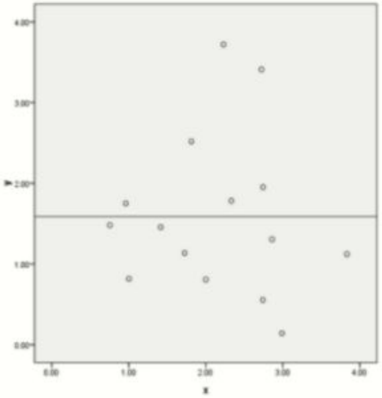
Figure 11.3: Scatter Diagram and Correlation Coefficients for Six Different Relationships

SPEARMAN RANK ORDER CORRELATION PLOT

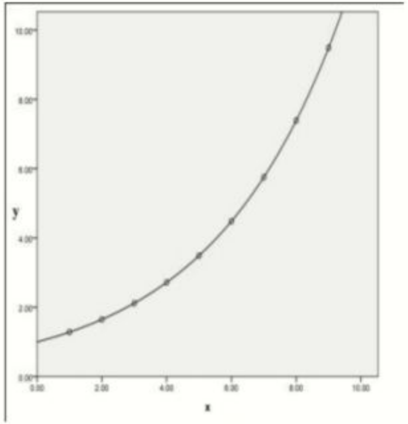
In the figures below various samples and their corresponding sample correlation coefficient values are presented. The first three represent the “extreme” monotonic correlation values of -1, 0 and 1:



$r_s = -1$
perfect –ve
monotonic correlation



$r_s = 0$
no correlation



$r_s = 1$
perfect +ve
monotonic correlation

Spearman Rank Order Correlation

Here is a quick example for spearman correlation the data is ordinal i.e. ordered.

Method - calculating the coefficient

- Create a table from your data.
- Rank the two data sets. Ranking is achieved by giving the ranking '1' to the biggest number in a column, '2' to the second biggest value and so on. The smallest value in the column will get the lowest ranking. This should be done for both sets of measurements.
- Tied scores are given the mean (average) rank. For example, the three tied scores of 1 euro in the example below are ranked fifth in order of price, but occupy three positions (fifth, sixth and seventh) in a ranking hierarchy of ten. The mean rank in this case is calculated as $(5+6+7) \div 3 = 6$.
- Find the difference in the ranks (d): This is the difference between the ranks of the two values on each row of the table. The rank of the second value (price) is subtracted from the rank of the first (distance from the museum).
- Square the differences (d^2) To remove negative values and then sum them ($\sum d^2$).

Convenience Store	Distance from CAM (m)	Rank distance	Price of 50cl bottle (€)	Rank price	Difference between ranks (d)	d^2
1	50	10	1.80	2	8	64
2	175	9	1.20	3.5	5.5	30.25
3	270	8	2.00	1	7	49
4	375	7	1.00	6	1	1
5	425	6	1.00	6	0	0
6	580	5	1.20	3.5	1.5	2.25
7	710	4	0.80	9	-5	25
8	790	3	0.60	10	-7	49
9	890	2	1.00	6	-4	16
10	980	1	0.85	8	-7	49
						$\sum d^2 = 285.5$

Data Table: Spearman's Rank Correlation

CORRELATION AND INDEPENDENCE

If the variables are independent, Pearson's correlation coefficient is 0, but the converse is not true because the correlation coefficient detects only linear dependencies between two variables.

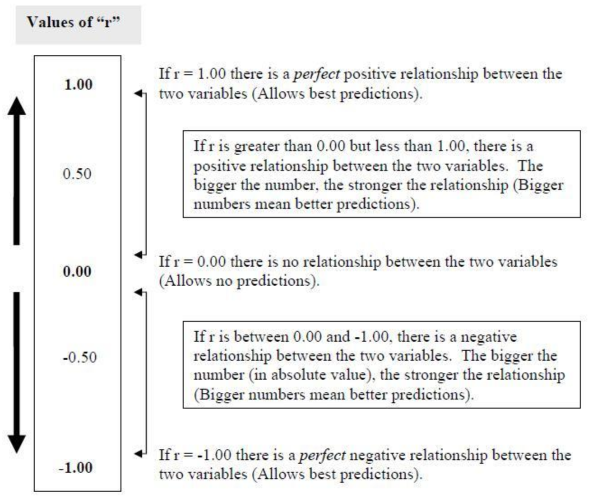
$$
\begin{aligned}
X, Y \text{ independent} \quad &\Rightarrow \quad \rho_{X,Y} = 0 \quad (X, Y \text{ uncorrelated}) \\
\rho_{X,Y} = 0 \quad (X, Y \text{ uncorrelated}) \quad &\not\Rightarrow \quad X, Y \text{ independent}
\end{aligned}
$$

For example, suppose the random variable X is symmetrically distributed about zero, and $Y=X^2$. Then Y is completely determined by X, so that X and Y are perfectly dependent, but their correlation is zero; they are uncorrelated. However, in the special case when X and Y are jointly normal, uncorrelatedness is equivalent to independence.

WHY ARE CORRELATIONS USEFUL?

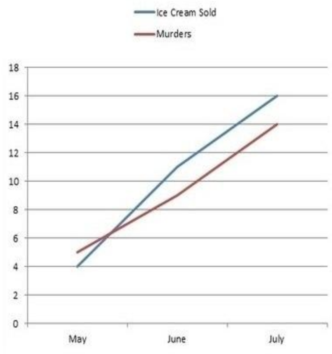
Important Things Correlation Coefficients Tell You

- The Direction Of A Relationship
- Correlation Coefficients Always Fall Between -1.00 and +1.00
- Larger Correlation Coefficients Mean Stronger Relationships
- Making Statistical Inferences



Correlation VS Causation

Correlation is not causation!



Eating ice cream causes shark attack deaths?

Significance Test of Correlation

Null Hypothesis:
There is not a significant correlation between the two variables.

Alternative Hypothesis:
There is a significant correlation between the two variables.

Significance Test of Correlation

- The Correlation Coefficient that you calculated
- Something called the **"degrees of freedom"** which is simply the number of pairs of data in your sample minus 2.
- A table of "Critical Values" of the correlation coefficient.

Values of *r* for the .05 and .01 Levels of Significance

df(*N* − 2)	.05	.01	*df*(*N* − 2)	.05	.01
1	.997	1.000	31	.344	.442
2	.950	.990	32	.339	.436
3	.878	.959	33	.334	.430
4	.812	.917	34	.329	.424
5	.755	.875	35	.325	.418
6	.707	.834	36	.320	.413
7	.666	.798	37	.316	.408
8	.632	.765	38	.312	.403
9	.602	.735	39	.308	.398
10	.576	.708	40	.304	.393
11	.553	.684	41	.301	.389
12	.533	.661	42	.297	.384
13	.514	.641	43	.294	.380
14	.497	.623	44	.291	.376
15	.482	.606	45	.288	.372
16	.468	.590	46	.285	.368
17	.456	.575	47	.282	.365
18	.444	.562	48	.279	.361
19	.433	.549	49	.276	.358
20	.423	.537	50	.273	.354
21	.413	.526	60	.250	.325
22	.404	.515	70	.232	.302
23	.396	.505	80	.217	.283
24	.388	.496	90	.205	.267
25	.381	.487	100	.195	.254
26	.374	.479	200	.138	.181
27	.367	.471	300	.113	.148
28	.361	.463	400	.098	.128
29	.355	.456	500	.088	.115
30	.349	.449	1000	.062	.081

CORRELATION ANALYSIS IN BIOLOGICAL DATA

To know the relation between systolic blood pressure (SBP)(continuous) and risk factors such as age (continuous) and weight (continuous), **Pearson's** correlation analysis would be used.

To understand the relation between maternal age (continuous) and parity (ordinal) or number of hospitalization (ordinal) and history of stroke (ordinal), **Spearman's** correlation analysis would be used.

	SBP	WC
SBP		
Pearson's correlation	1	0.395**
Significance (two tailed)		0.000
n	967	967
WC		
Pearson's correlation	0.395**	1
Significance (two tailed)	0.000	
n	967	967

**Correlation is significant at the 0.01 level (two tailed). SBP: Systolic blood pressure, WC: Waist circumference

Spearman's rho	BMI status	WC
BMI status		
Correlation coefficient	1.000	0.398**
Significance (two tailed)		0.000
n	936	936
WC		
Correlation coefficient	0.398**	1.000
Significance (two tailed)	0.000	
n	936	936

**Correlation is significant at the 0.01 level (two tailed). BMI: Body mass index, WC: Waist circumference

Predicting the quality of wine

- **Bordeaux** is a region in France popular for producing wine
- While wine has been produced in much the same way for hundreds of years, there are differences in price and quality from year to year that are sometimes very significant.
- Wines are aged to its hard to predict the quality of wine when its young
- Wine tasters and experts are helpful and predict which ones better

Can analytics be used to predict the quality of wine ?

What Do I Hear for This 1988 Burgundy? Fine Wine on the Block.

Forget the Corner Liquor Store.
Stop Off at the Local Auction House.

Sold! To the Oenophile in the Blue Dress.

- On March 4, 1990, the New York Times announced that Princeton economics professor **Orley Ashenfelter** can predict the quality of Bordeaux wine without tasting a single drop
- Ashenfelter used a method called linear regression which measures a dependent variable using a set of independent variable

Passell, Peter. “Wine Equation Puts Some Noses Out of Joint.” *The New York Times*, The New York Times, 4 Mar. 1990, www.nytimes.com/1990/03/04/us/wine-equation-puts-some-noses-out-of-joint.html.

Experts Reaction

Robert Parker : The worlds most influential wine expert

“Ashenfelter is an absolute total sham. Rather like movie critic who never goes to see the movie but tells you how good it is based on the actor and director ”

Our Data

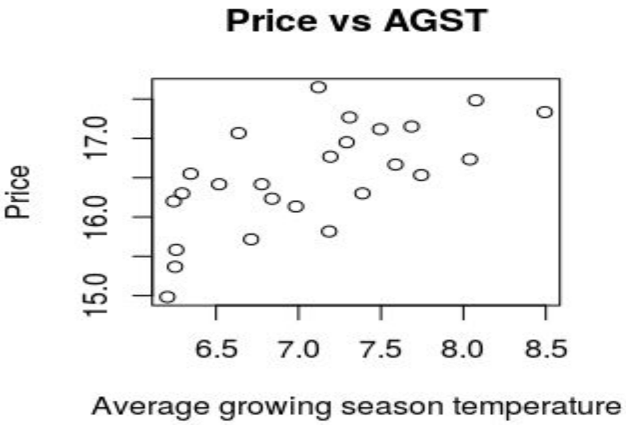
Dependent Variables

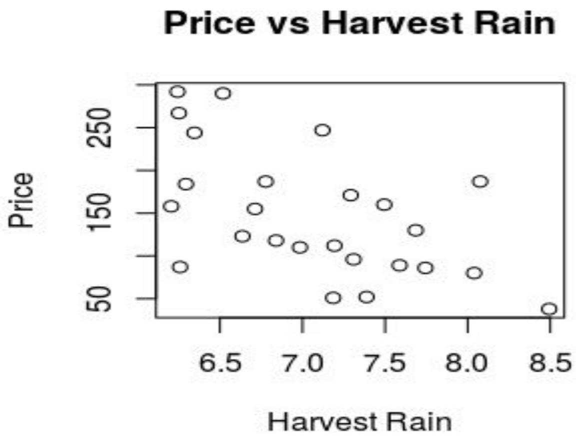
Price: The price at which the wine is sold at an auction

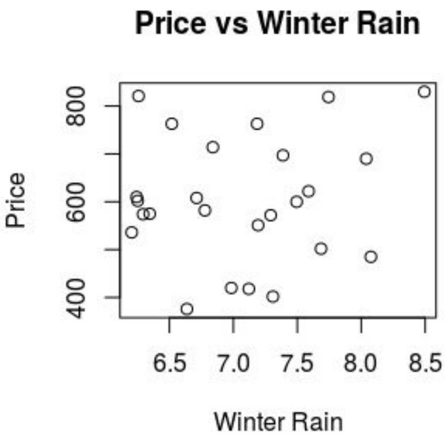
Independent Variables

Age : Older wines are more expensive
Average Growing Season Temperature: measured in celcius
Harvest Rain : Rain measred in mm
Winter Rain: Rain measred in mm
France population : population of France in that particular year

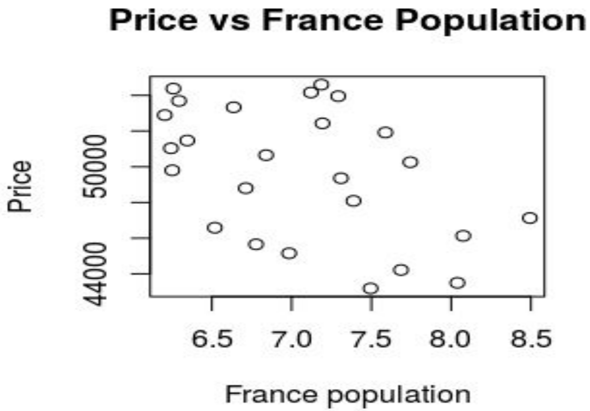
EdX| MIT-x,
courses.edx.org/courses/course-v1:MITx+15.071x+2T2017/courseware/f8d71d64418146f18a066d7f0379678c/35b789067e9b469caed457cfff1645b7/3?activate_block_id=block-v1%3AMITx%2B15.071x%2B2T2017%2Btype%40vertical%2Bblock%405623146ed3b44e21b1765ea19e630d64


Price vs AGST
Price
15.0
16.0
17.0
6.5
7.0
7.5
8.0
8.5
Average growing season temperature


Price vs Harvest Rain
Price
50
150
250
6.5
7.0
7.5
8.0
8.5
Harvest Rain


Price vs Winter Rain
Price
400
600
800
6.5
7.0
7.5
8.0
8.5
Winter Rain


Price vs Age
Price
5
10
20
30
6.5
7.0
7.5
8.0
8.5
Age


Price vs France Population
Price
44000
50000
6.5
7.0
7.5
8.0
8.5
France population

Correlation Matrix

```
> cor(winedata)
                   Year      Price   WinterRain        AGST
Year         1.00000000 -0.4477679  0.016970024 -0.24691585
Price       -0.44776786  1.0000000  0.136650547  0.65956286
WinterRain   0.01697002  0.1366505  1.000000000 -0.32109061
AGST        -0.24691585  0.6595629 -0.321090611  1.00000000
HarvestRain  0.02800907 -0.5633219 -0.275440854 -0.06449593
Age         -1.00000000  0.4477679 -0.016970024  0.24691585
FrancePop    0.99448510 -0.4668616 -0.001621627 -0.25916227
            HarvestRain         Age     FrancePop
Year         0.02800907 -1.00000000  0.994485097
Price       -0.56332190  0.44776786 -0.466861641
WinterRain  -0.27544085 -0.01697002 -0.001621627
AGST        -0.06449593  0.24691585 -0.259162274
HarvestRain  1.00000000 -0.02800907  0.041264394
Age         -0.02800907  1.00000000 -0.994485097
FrancePop    0.04126439 -0.99448510  1.000000000
```

Our Model

```
> #Building a linear Regression model.
> model2<-lm(Price~AGST+HarvestRain+Age+FrancePop , data = winedata)
> #summary(model2)
> #Getting test data
> testdata<- read.csv("./wine_test.csv")
> #summary of test data
> #Predicting the values using the predict function
> prediction<- predict(model2, newdata = testdata)
> #Getting the predicted values
```

The Results

Our Model vs Actual price

Year	Predicted Price($)	Actual price ($)
1979	6.8039	6.6979
1980	6.9291	6.9541

Robert- The wine expert

1986: “Very good to sometime exceptional”

Ashenfelter - The Princeton Professor

1986: “Wine will be mediocore”

1989: “It will be the wine of the century and even better in 1990”

Actual Auction

1989: wine sold for twice the price of 1986

1990 : wine sold for even higher prices

Thank You

Any Questions

Github: https://github.com/avishjadwani/Correlation-Presentation

How did you like the presentation ? EDM

Respond at **PollEv.com/dehghanimoha266**

Text **DEHGHANIMOHA266** to **37607** once to join, then **A, B, C, D, or E**

Not at all interesting	**A**
Not very interesting	**B**
Neutral	**C**
Somewhat Interesting	**D**
Very Interesting	**E**