Lecture Notes in Quantitative Biology

Correlation

Chapter 20.1

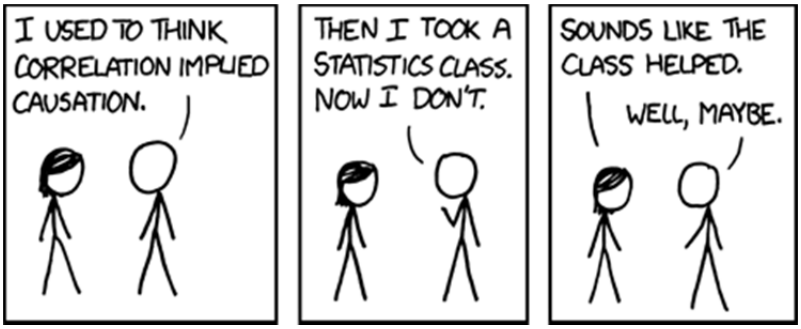
Part VI Extensions
Ch19 Model Selection
Ch19.1 EDA (Tukey)
Ch19.2 Penalized Likelihood
Ch19.3 Forward Selection
Ch19.4 Backward Selection
Ch20 Correlation
Ch20.1 Correlation by lurking variable
Ch20.2 Correlation by outliers
Ch20.3 Autocorrelation
Ch20.4 Multivariate Analysis
Ch20.5 MANOVA

Cavy data 8 Nov 2016
Revised 4 Nov 2017
Revised 3 Nov 2018

Today: Correlation
Example of lurking variable
Contrast with regression
Graphical model
The correlation coefficient ρ
Formal model $\rho = \cos(\theta)$
Estimate of ρ
Likelihood ratio from r^2
Likelihood ratio test

Wrap-up:

Correlation measures association between two randomly distributed variables.
The correlation coefficient is ρ (greek rho), which is estimated by the statistic r.
r^2 measures goodness of fit and explained variance.
Likelihood ratios are calculated from r^2
Correlation arises from lurking variable and from outliers.



Introduction

Correlation is used to measure the strength of association between two variables. Correlation was developed by Francis Galton (1888 Co-relations and their measurement. *Proceedings of the Royal Society London Series* 45:135-145). Statistical theory was developed by Karl Pearson, who introduced the measure of correlation we use today (1896 Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia, *Philosophical Transactions of the Royal Society* A 187: 253-318).

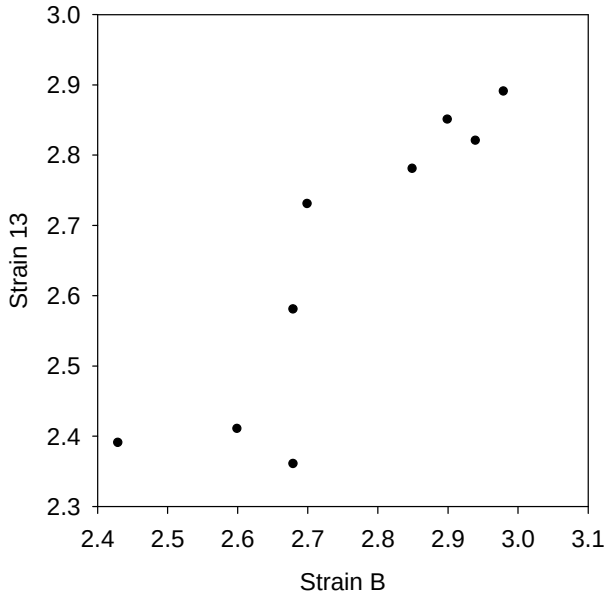
Example

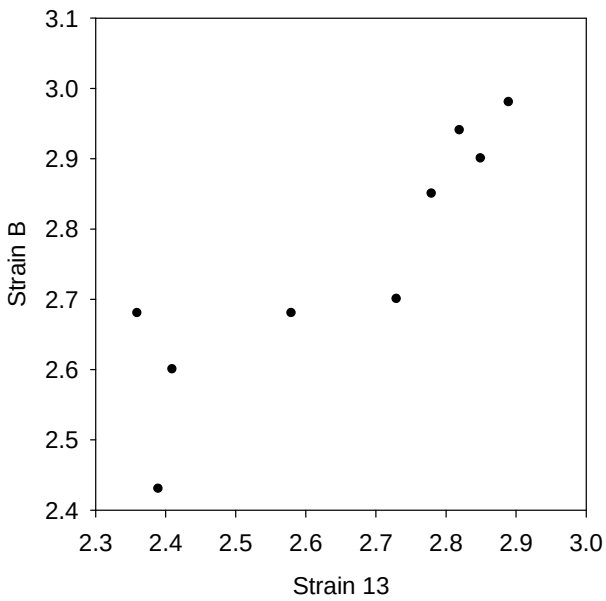
Sokal and Rohlf (2012 Box 13.12) reported data on litter size in two strains of Guinea pigs from the genetics lab of Sokal's thesis supervisor, Sewall Wright. The data can be viewed as two time series. Do the two time series covary?

StrainB	Strain13	Year
2.68	2.36	1916
2.60	2.41	1917
2.43	2.39	1918
2.90	2.85	1919
2.94	2.82	1920
2.70	2.73	1921
2.68	2.58	1922
2.98	2.89	1923
2.85	2.78	1924

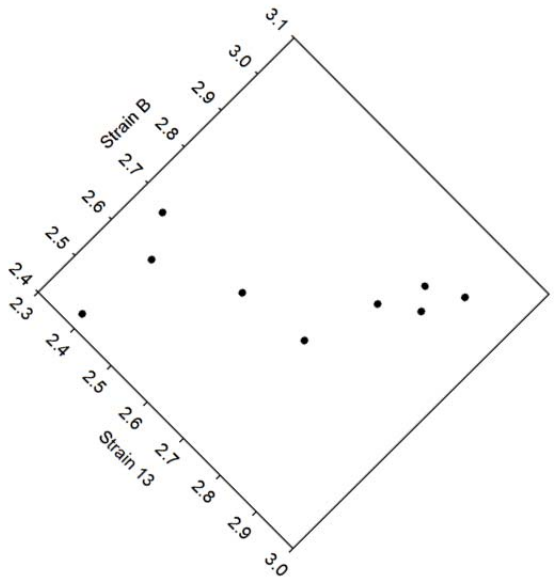
Contrast with Regression

There is no causal ordering of the two time series, so we can graph Strain B versus 13 or we can graph Strain 13 versus B.

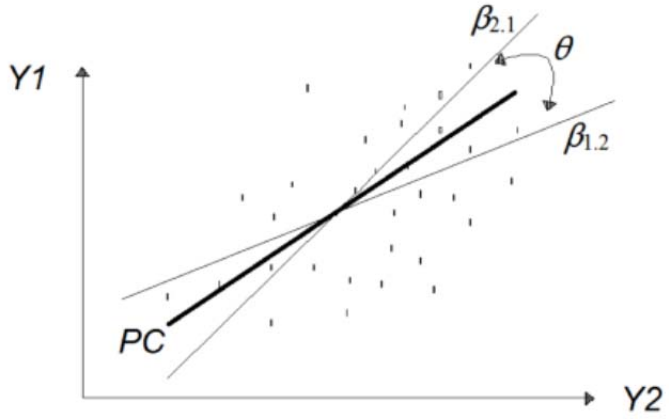




Contrast with Regression



To remove the sense that one quantity (on the vertical axis) is a function of the other (on the horizontal axis) we rotate the graph by 45^o.



Graphical Model

If we regress *Y1* (Strain B) on *Y2* (Strain B),
we obtain a regression equation that differs from the regression of *Y2* (Strain 13) on *Y1* (Strain 13). To confirm this for yourself write the regression equations for:
Regression of *Y1* (Strain B) on *Y2* (Strain 13)

Strain B = _____ (Eq 1a)

Regression of *Y2* (Strain 13) on *Y1* (Strain B)

Strain 13 = _____ (Eq 1b)

Now solve Eq 1b for Strain B.

Strain B = _____ (Eq1c)

Compare Eq 1a to Eq 1c.

Instead of the regression of *Y1* on *Y2* (with slope $\beta_{Y1.Y2}$) or the regression of *Y2* on *Y1* (with slope $\beta_{Y2.Y1}$) our model of association is a line (*PC1*) that splits the difference between the two slopes.

The correlation coefficient.

We use both regression slopes to obtain a measure of association, the correlation coefficient ρ (rho).

Population: $\rho = \text{sqrt}(\beta_{1.2}\ \beta_{2.1})$ $\hat{\rho} = r$

Sample $r = sqrt(\hat{\beta}_{Y1.Y2}\hat{\beta}_{Y2.Y1})$ $r = sqrt(0.7447 * 1.047) = 0.883$

ρ depends on the angle (greek letter θ) between the two regression lines.

$\rho = \cos(\theta \cdot \pi /180)$ $\text{acos}(\rho \cdot 180 / \pi)$ $\text{acos}(0.883 \cdot 180 / \pi) = 27^o$

If $\theta = 0^o$ perfect positive association $\cos(\theta \cdot \pi /180) = 1$

If $\theta = 90^o$ no association $\cos(\theta \cdot \pi /180) = 0$

If $\theta = 180^o$ perfect negative association $\cos(\theta \cdot \pi /180) = -1$

The correlation coefficient ranges from –1 to +1

Equivalanetly $-1 \leq \rho \leq +1$

Here is a derivation of the correlation coefficient ρ

$\text{Var}(Y_1 + Y_2)$	=	$\text{Var}(Y_1)$	+	$\text{Var}(Y_2)$	+	$2\ \text{Cov}(Y_1\ Y_2)$
$\text{Var}(Y_1 + Y_2)$	=	$\text{Var}(Y_1)$	+	$\text{Var}(Y_2)$	+	$2\ \rho\ \sqrt{\text{Var}(Y_1)}\ \sqrt{\text{Var}(Y_2)}$
$\sigma^2_{(Y1 + Y2)}$	=	σ^2_{Y1}	+	σ^2_{Y2}	+	$2\ \rho_{12}\ \sigma_{Y1}\ \sigma_{Y2}$

Solving for ρ yields the formula for ρ.

$$\rho = \frac{\sigma^2_{(Y1 + Y2)} - \sigma^2_{Y1} - \sigma^2_{Y2}}{2\ \sigma_{Y1}\ \sigma_{Y2}}$$

The correlation coefficient ρ is the variance of the pairs, adjusted (by subtraction) for the variance of each quantity, then scaled to (divided by) the standard deviations (σ_{Y1} and σ_{Y2}) of both quantities.

Formal model $\rho = \cos(\theta)$

In regression we have a randomly distributed response variable as a function of a fixed explanatory variable. In correlation we have two random response variables and no observed explanatory variable. Instead of a measured explanatory variable, we obtain a single unobserved explanatory variable T_{PC1} consisting of scores on a single principal component *PC1*.

$$[\ Y_1\ Y_2\ ] = T_{PC1} + \epsilon$$

The scores T_{PC1} fall on a straight line through the cloud of points.

Estimate of the correlation coefficient

We use the sample to estimate the value of ρ. The estimate of ρ is called r, in keeping with the convention of using greek letters for the true (population) values of a parameter, and using roman letters for sample estimates of these parameters. Instead of r, we could also use the symbol $\hat{\rho}$ (rho-hat) for the estimate of ρ.

To obtain the formula to estimate ρ, we substitute estimates into the formula for ρ. Here is a listing of standard notation for parameters and estimates.

parameter name	symbol	estimate	
mean of Y_1	μ_1	$\hat{\mu}_1$	$\bar{Y}_1$
mean of Y_2	μ_2	$\hat{\mu}_2$	$\bar{Y}_2$
standard deviation of Y_1	σ_1		s_1
standard deviation of Y_2	σ_2		s_2
correlation of Y_1 Y_2	ρ	$\hat{\rho}$	r

Sums of square deviations of Y1 and Y2 are used to estimate the covariance.

$$Cov(Y1,Y2) = \frac{1}{n-1} \sum (Y_1 - \bar{Y}_1)(Y_2 - \bar{Y}_2)$$

The degrees of freedom (n–1) are used instead of the sample size n. This corrects for the fact that the population will have a wider range of values than the sample. The estimate from the sample will underestimate the true value ρ if we use n. The correction is the same as we have seen for s^2, which is the sample estimate of the variance of the population. The covariance has units of Y1 * Y2.

The estimate of the correlation is the covariance estimate standardized by s_1 and s_2

$$r = \frac{1}{n-1} \frac{\sum (Y_1 - \bar{Y}_1)(Y_2 - \bar{Y}_2)}{s_1 \; s_2}$$

The result is a dimensionless ratio.

The Likelihood Ratio *LR*.
How good is the evidence for correlation? Equivalently what is the likelihood ratio?
$L(0.883 \mid Y1, Y2)$ The likelihood that $\rho = 0.883$, given the data, normal error.
$L(0.00 \mid Y1, Y2)$ The likelihood that $\rho = 0$, given the data Y, normal error.
$LR = L(0.883 \mid Y) \; / L(0.00 \mid Y)$
The likelihood ratio is calculated from the correlation $LR = (1 - r^2)^{-n/2}$
where n = number of pairs (9)
r^2 = explained variance
1- r^2 = unexplained variance
$LR = (1 - 0.883^2)^{-9/2} = 905 : 1$
Given the data, $\rho = 0.883$ is 905 times more likely than $\rho = 0$.
The strength of the evidence warrants further use of the model.

The likelihood ratio depends on the degree of association (measured by r) and the sample size n. For example, had we obtained the same correlation from a larger sample size (n = 10 years) the LR increases to
$LR = (1 - 0.883^2)^{-10/2} = 1927 : 1$

Likelihood Ratio Test. Litter size data
The customary inferential approach is a likelihood ratio test (LRT). We have already seen LRTs – *t*-tests and *F*-tests. LRTs yield the probability of a likelihood ratio, given the sample size. We apply the generic recipe for hypothesis testing with any statistic.

1. State the population
The data are from the guinea pig population in Sewall Wright's genetic lab at Harvard. We will not infer to other Guinea pig colonies. Nor will we infer to other years. We will restrict inference to the population of measurements of litter size, as generated by the protocols in this lab for maintaining a colony of guinea pigs.

2. State the model or measure of pattern (statistic).
The measure of pattern will be ρ the correlation of the two variables.
The sample size is small so we compute a *t*-statistic to compare r to $\rho = 0$.

$$t = \frac{r - \rho}{s_r} \qquad s_r = \sqrt{\frac{1 - r^2}{df}} = \sqrt{\frac{1 - r^2}{n - 2}}$$

3. State H_A about statistic H_A: $\rho \neq 0$

4. State H_o about statistic H_o: $\rho = 0$

5. **State tolerance for Type I error**. In a lab setting where we have no reason to limit ourselves to a fixed Type I error rate of 5% we will use Fisher's four levels of "definite significance" (Fisher 1954 p 154):
P = 0.10, 0.05, 0.02, and 0.01.

6. **State frequency distribution.** We will use the *t*-distribution.

7. **Calculate statistic** $r = 0.883$ $r^2 = 0.780$

The standard error for *r* is: $s_r = \sqrt[2]{\frac{1-r^2}{n-2}}$

$$s_r = \sqrt[2]{\frac{1-0.883^2}{9-2}} = \sqrt[2]{\frac{0.220}{7}} = 0.314$$

The *t*-statistic is the correlation over its standard error $t = r / s_r$
$t = 0.883 / 0.314 = 4.98$

The probability of obtaining this *t*-statistic by chance alone, on 7 df, is
p = 0.0016

8. Recompute p-value if assumptions are not met.
As with regression, we assume a straight line (*PC1*) accurately describes the relation of *Y1* to *Y2*. The assumptions for the estimate of the correlation and for the t-test are that the residuals are homogeneous, normal, and independent.

In this example the sample size is small, which warrants caution when assumptions are not met. However the *p*-value is far from α and so a better *p*-value by randomization is unlikely to change the decision.

9. **Report Fisher significance**
The correlation is $r = 0.883$, n = 9, with a low Type I error, $p = 0.0016$

10. **Report and interpret parameters of biological interest.**
Litter sizes of two strains of guinea pigs in Sewall Wright's genetics lab were correlated over a 9 year period. There is no reason to expect a direct causal relation of one time series on the other. In this case an external variable can be identified: reduced supply of fresh vegetables in the US (R.R. Sokal, pers. comm) during World War I, ending November 11, 1918. During the war Herbert Hoover organized a voluntary rationing program toward humanitarian food relief for Belgian civilians (Whyte, Kenneth. 2017. Hoover: An Extraordinary Life. NY, Knopf) . Rationing reduced the supply of fresh vegetables, the source of Vitamin C, an essential nutrient for guinea pigs.