Regression and Correlation

Joseph J. Luczkovich, PhD

January 23, 2015

Introduction

We often ask the following kinds of questions: "Is variable X related to variable Y?" or "Is profit of an ecotourism company related to park visitation rates?" or "Is temperature related to production of food in an ecosystem?" or "Is salinity related to the abundance of shrimp?" or "How much variation in Y is explained by variation in variable X?" In such cases, we may want to use either correlation or regression methods, or both. Correlation and regression are widely used techniques in the sciences. We are often interested the association between variables. When two variables are associated or related in a linear manner, they are said to be **correlated**. Correlations can be positive (directly related) or negative (inversely related). **Regression** is a technique to mathematically model the linear association between two or more variables, the predictor (or independent) variable and the response (or dependent variable). Linear regression results in a least-squares fit of the variation in the response variable to the predictor variable.

History - Karl Pearson

- Statistician, Lawyer, Socialist, Eugenicist
- Rival of R.A. Fisher
- Editor of Biometry
- Invented Pearson Correlation Coefficient (r)
- Chi-Square distribution (like the normal curve, but for discreet data)
- P-values and hypothesis tests
- Principal Components Analysis (PCA)

Correlation

When there are two variables (x and y) measured on the same subjects, you can see if they are correlated. Correlation is the linear association between two variables. When the correlation coefficient r is close to 1.0, there is a near

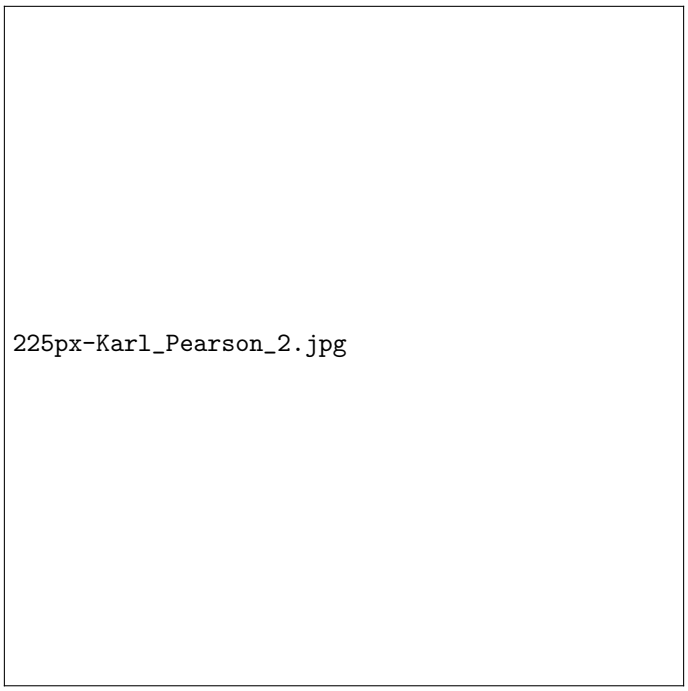


Figure 1: Karl Pearson, B: 27 March 1857 D: 27 April 1936

perfect association between the variables. When x goes up by 1 unit, y also goes up by 1 unit (or some multiple of units). A negative correlation of -1.0 would indicate just the opposite: when x goes up by 1 unit, y goes down by 1 unit (or multiples of 1 unit). The formula for r (or Pearson Product-Moment)is:

$$r = \frac{\sum_i^n (x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum_i^n (x_i - \bar{x})^2 * \sum_i^n (y_i - \bar{y})^2}}$$

It varies between -1.0 and +1.0.
If you look at this equation, it says that when the differences from the mean of x are multiplied by the differences from the mean of y and summed then are divided by the square root of the sum of the squared differences from the mean of x multiplied by the sum of the squared differences from the mean of y, and you get a value of unity (1.0), this can only mean that x and y form a straight line when plotted against one another. The more that x and y are off the line, the more that the denominator will be large relative to the numerator. When x and y are unrelated to one another, and not on a line at all, but more randomly distributed, the numerator is small and the denominator is large in this equation.

Example

Pearson studied the heights of brothers and sisters from different families (in inches):

```
> brother<-c(71,68,66,67,70,71,70,73,72,65,66)
> sister<-c(69,64,65,63,65,62,65,64,66,59,62)
> length(brother)# the vector of brothers measured is 11

[1] 11

> length(sister)# the vector of sisters measured is 11, same as brothers

[1] 11

> #Are height of sisters and brothers in the same family correlated?
> cor(brother,sister)

[1] 0.5580547
```

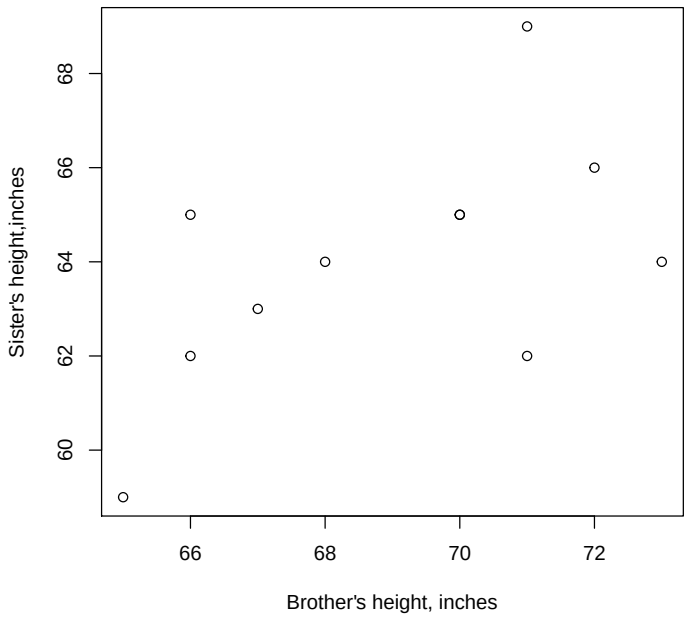
Note that the two vectors must be the same length to have a correlation. The measurements must be paired (by family relationship in this case). It does not make sense to have unpaired, unequal number of measured variables in correlation.

We can apply the Pearson Product-Moment equation manually in R:

```
> x<-brother-mean(brother)# this is x - mean x
> x

 [1]  2 -1 -3 -2  1  2  1  4  3 -4 -3

> y<-sister-mean(sister)# this is y - mean y
> y

 [1]  5  0  1 -1  1 -2  1  0  2 -5 -2

> num<-sum(x*y)# this is the numerator
> num

[1] 39

> dem<-sqrt(sum(x^2)*sum(y^2))# this is the denominator
> dem

[1] 69.88562

> r=num/dem #divide the numerator by the denominator
> r

[1] 0.5580547
```

We can see that the differences from the mean (the residuals) are all that matter here, and we get rid of the mean of x and mean of y in the numerator. We are only concerned with the variation around the means. If the product of the residuals is small, then the denominator will also be small.

```
> #the plot of brother heighs versus sister heights
> plot(brother,sister,xlab="Brother's height, inches",
+      ylab="Sister's height,inches")
```



We can see the linear relationship is there, underlying the two measures, but it is not 1:1, not a perfect linear association, so r < 1.0 (0.55 in this data set). It turns out that the brothers and sisters from the same family have a tendency to have the same height due to the fact they have a genetic relationship, the same mother (almost always) and the same father (mostly this is true), but there are adoptions (unrelated parents), infidelities (Mom!), gene expression differences, and environmental factors (food) that may cause some additional lack of correlation. There is also measurement error, but that is likely to be small in this case. Measurement error is not always so small, and can sometimes cause an additional source of variation. If we could capture all the multiple factors that cause the points to fall off the line, we could explain completely a sister's height by measuring her brother's height, or vice versa. This is a objective of multivariate statistics, and correlation forms the basis of many multivariate statistics (like PCA, we will see more of this later). Note that correlation does

not mean there is a causal relationship between x and y. If a brother grows tall, it does not cause his sister to grow tall. He might have been taking steroids. They share a common genetic and environmental history, that is all. Remember: **Correlation does not equal causation**. Let's test the significance of the relationship, attempting to reject the null hypothesis that there is no linear relationship, positive or negative, between the brother's and sister's heights:

```
> #the correlation test is specified as method = pearson
> cor.test(brother,sister,method="pearson")

        Pearson's product-moment correlation

data:  brother and sister
t = 2.0175, df = 9, p-value = 0.07442
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.06286527  0.86751705
sample estimates:
      cor
0.5580547
```

The null hypothesis is not rejected, and the p = 0.07. There may not be a linear relationship here. But, it was close, and what if there was a bigger sample size?

Non-parametric Correlation

Pearson was a normal guy, with parametric equations used (he used squared deviations from the mean, etc.) depending on assumption of normal distribution of errors. There are non-parametric correlations that do not assume normal distribution of errors. One of them is called the Spearman Rank Correlation test (also called the Spearman rho ρ test). The numbers are converted to ranks (x and y are ranked from low to high, separately), then the correlation is computed using the Person Product moment approach. If ties exist in the ranked data, one should avoid this test and use the Pearson correlation coefficient. Rho varies between -1.0 and +1.0.

Example

Here, we use an example with wormy apples. Each orchard's apple yield is measured in bushels and the number of worms in a sample of apples from each orchard is counted:

```
> apple<-c(8,6,11,22,14,17,18,24,19,23,26,40)
> worms<-c(59,58,56,53,50,45,43,42,39,38,30,27)
> cor.test(apple,worms, method="pearson")#Pearson correlation

        Pearson's product-moment correlation
```

```
data:  apple and worms
t = -5.8842, df = 10, p-value = 0.0001543
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9662791 -0.6207642
sample estimates:
       cor
-0.8808547
```

```
> cor.test(apple,worms, method="spearman")#Spearman correlation
```

```
        Spearman's rank correlation rho

data:  apple and worms
S = 542, p-value = 5.942e-06
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.8951049
```

```
> cor.test(apple,worms, method="kendall")#Kendall correlation
```

```
        Kendall's rank correlation tau

data:  apple and worms
T = 7, p-value = 0.0001074
alternative hypothesis: true tau is not equal to 0
sample estimates:
       tau
-0.7878788
```

```
>
```

Apple yield is negatively correlated with worminess and it is highly significant. All three methods of correlation agree. More worms, fewer apples.
The Kendall's Tau was also computed above and likewise indicted a significant dependence between worms and apples. Tau is a sightly different non-parametric correlation. It defines discordant and concordant x's and y's after ranking them jointly, then re-pairing them. A concordant association is when x1 $>$ x2 and y1 $>$ y2. But if x1 $>$ x2 and y1 $<$ y2, they are discordant. A formula relates the concordance of x and y, that is it tests the null hypothesis that the x's and y's are dependent upon one another, when one increases, the other increases, when one decreases, the other decreases, i.e., they all exhibit concordance. It varies between -1.0 and +1.0. See: [**?**]

Linear Regression

When there is an association between two variables, one can proceed next to do a linear model or regression analysis. The first question one should ask

before starting this analysis is: "Which variable is dependent upon the other? Which variable is the response variable and which is the predictor (independent) variable?" Correlation does not assume either variable is a predictor of the other, but regression does. So you must have some idea of what is known before your start this analysis, preferably through previous experimentation or common sense. For example, in the Wysocking Bay dredge spoil study, we know that dredge spoil depth had a negative effect on *Juncus* biomass - the variables were negatively correlated. But we also know that dredge spoil addition *caused* the decline in *Juncus* biomass. Dredge spoil depth is a good predictor of *Juncus* biomass. We can proceed to regression analysis with the dredge spoil as a predictor of *Juncus* biomass, and fit a linear model to the data.

History

Regression was a term coined by Francis Galton (first cousin of Charles Darwin, founder of the field of eugenics) in the nineteenth century to describe a biological phenomenon - the fact that people were getting less variable in height, converging on the mean, due to random matings. His motto: "whenever you

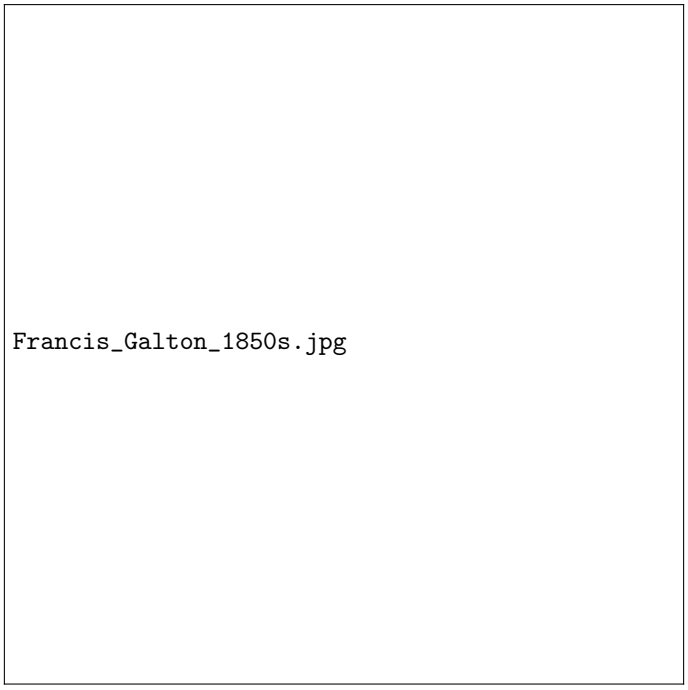


Figure 2: Sir Francis Galton, 1822-1911

can, measure and count". He was interested in all kinds of measurements: head circumference, thumb size, arm length, fingerprints (his analysis was used by Scotland Yard to solve crimes), even rating pretty girls in London, England or

Aberdeen, Scotland (London women were prettier, according to Galton). He published "Regression towards mediocrity in hereditary stature" in the journal " The Journal of the Anthropological Institute of Great Britain and Ireland", Vol. 15 (1886), pp. 246-263. Regression towards the mean was bad thing in his view, i.e, "mediocrity" or the mean was ordinary, and selective breeding of humans could be used to make men taller (I guess this was a good thing). He wanted to improve the human condition by using breeding to select for traits like intelligence, strength, height. See: [**?**] When you plot x and y of father's heights and

Figure 3: Galton's figure of "regression toward the mean" in his 1886 paper

son's heights, you get a positive relationship. But over time, this relationship decreases, so that the mean height is more common in sons. Descendants of tall fathers tend toward the mean for all persons, given many reproductive events, and random mating. The sons get taller if the fathers are taller, but only if there is selective mating with tall women. Galton's definition was very limited and the basis of all this eugenics stuff, but regression has become applicable to all sorts of problems in statistics.

A Formal Model

Regression is similar to correlation (two variables are associated in a linear model), but:

- x is independent, assumed to be under investigator's control

- y is dependent, responds to variations in x due to a causal effect of x on y

Model:

$$y = \alpha + \beta x + \epsilon$$

Where y is the response variable, x is the predictor variable, and α is the y-axis intercept of the linear regression line, β is a coefficient or slope of the linear regression line, and ϵ is the error or residual of the model. We need to find the values of α and β that minimize the sum of the squared residuals:

$$SS_{resid} = \sum_{i}^{n} (y_i - (\alpha + \beta))^2$$

This is known as the least squares method of fitting a regression line. We can compute the slope β and intercept α using the least squares method to minimize the SS_{resid}:

$$\hat{\beta} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n ((x_i - \bar{x}))}$$

and

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

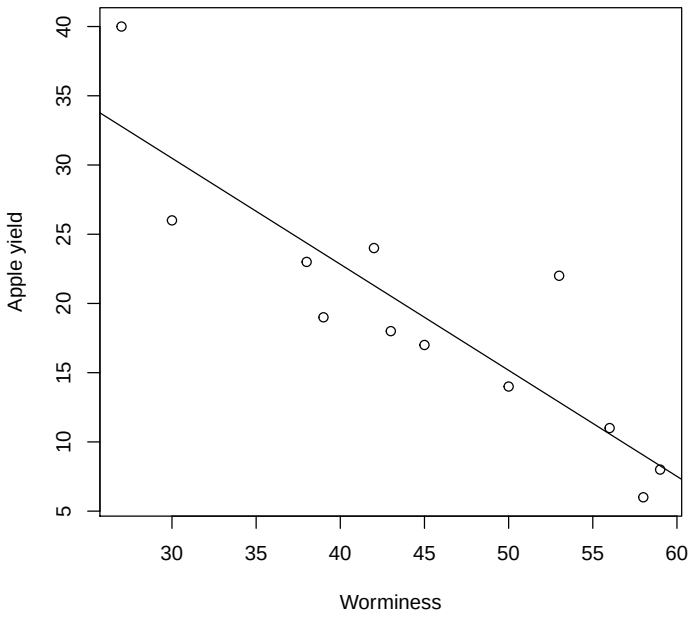
Example

Use the wormy apples example, we want to predict the amount of apples (yield) we can expect from different levels of worminess. We call a linear regression model in R, using the **lm()** command:

```
> fit1<-lm(apple~worms)
> summary(fit1)

Call:
lm(formula = apple ~ worms)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5957 -2.6596 -1.2660  0.9947  9.1277

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  53.4681     6.0032   8.907 4.54e-06 ***
worms        -0.7660     0.1302  -5.884 0.000154 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.55 on 10 degrees of freedom
Multiple R-squared:  0.7759,        Adjusted R-squared:  0.7535
F-statistic: 34.62 on 1 and 10 DF,  p-value: 0.0001543
```

We plot the result:

```
> plot(worms, apple, xlab = "Worminess", ylab="Apple yield")
> abline(fit1)
```



There are lower apple yields when we have more worms. The equation defines the relationship:

$$y = -0.766 * x + 53.468$$

The relationship is significant (slope not equal to 0, p = 0.001) and the R2 = 0.75, or 75 percent of the apple yield is explained by worminess.

References