Correlation

ANTH 3720-001 Archaeological and Forensic Science Lab Methods: Data Analysis with R

Elic Weitzel

Jan. 29-31, 2021

If you would like the original R Markdown file, find it on my GitHub page at https://github.com/weitzele/Basic-R-Tutorial

1 Correlation

Having now learned the basics of linear modeling, we can introduce a related concept: correlation. This is an approach we can use when we're looking to assess the relationship between two continuous variables.

As you now know, a linear model is a mathematical equation expressing the effect of the predictor variable on the response variable: of x on y. We can use this equation to determine the direction of the effect (whether the beta coefficient is positive or negative), the strength of the effect (our r^2 value), the significance of the relationship (our p-value), and several other things.

However, there is a separate (but related) statistical approach called correlation analysis that one commonly encounters in archaeology and anthropology, as well as in many other fields. While linear modeling, or regression, aims to predict whether a change in x leads to a change in y. In contrast, correlation aims to express whether a change in one variable is associated with a change in the other variable. It might seem like I just said the same thing two different ways, but there are differences.

First, there is an implication of causality in linear regression. While practically speaking, we know that there might not actually be a causal relationship between two variables, linear regression models the relationship as if there were one for mathematical purposes. Thus, when using linear models, we talk about predictor and response variables and how a change in x leads to a change in y. One cannot flip the predictor and response variables and obtain the same result. In correlation analysis however, there is no assumption of causality: the statistic expresses the same thing whether a change in x occurs alongside a change in y, or whether changing y changes x. There is technically no predictor or response as they are mathematically interchangeable in a correlation analysis.

Second, the word "predict" also differentiates the definitions of linear modeling and correlation. Linear models are predictive: they are mathematical equations expressing a relationship, and you can use those equations to predict values of y for which you might not have observed values of x. We did this in the previous tutorial when we predicted the value of the Fish-Mammal Index for the date of 2200 BP: a date that was not in our dataset. While linear models allow you predict unobserved values, including those in the past or future, correlation analysis does not. It is restricted to the data you observed and does not produce a mathematical equation expressing the relationship between your variables. It only tells you whether there is an association between them, which direction that association is, and whether it's a strong association or not. Linear modeling does all of those things too, and then some. Linear models include information about

stochasticity in the errors/residuals, as one important example, which is something we can use to investigate confidence intervals. Correlation analyses don't do this.

But sometimes a simpler correlation analysis is all that is needed, and for that reason, it's worth knowning about the most common variety of correlation analysis, Pearson's Correlation.

2 Pearson's Correlation

2.1 Prepare the Data

As with the previous tutorial on linear modeling, let's load the zooarchaeological data from Jack Broughton's 1994 paper (Broughton, J.M. (1994) Journal of Arch. Sci. 21(4):501-514) on the abundance of various animal species through time at 9 sites in the Sacramento Valley of California.

Load this .csv file, which is provided to you along with this tutorial, and assign it to the new object `sac.data`.

```
sac.data <- read.csv("Broughton1994JAS_SacramentoValley.csv")
```

Inspect the data to refresh your memory.

```
head(sac.data) #check the first six rows for each column
```

```
##   Site Date Sylvilagus Lepus    LS_Index Mammals Fish     FM_Index
## 1   68 3665         10    31  0.51219512     213    7 -0.936363636
## 2  105 2875          7     8  0.06666667     336   19 -0.892957746
## 3  101 2650         27    13 -0.35000000     441   53 -0.785425101
## 4  288 1650          9     5 -0.28571429     385   51 -0.766055046
## 5   99 1470        671   263 -0.43683083    3749 3699 -0.006713212
## 6   12  400         10     5 -0.33333333     102 1849  0.895438237
```

```
colnames(sac.data) #check the column names
```

```
## [1] "Site"       "Date"       "Sylvilagus" "Lepus"      "LS_Index"
## [6] "Mammals"    "Fish"       "FM_Index"
```

```
nrow(sac.data) #see how many rows this data frame contains
```

```
## [1] 9
```

This is a data frame with 8 columns and 9 rows. Each row represents one of nine sites while the second column contains the mean occupation date for each site.

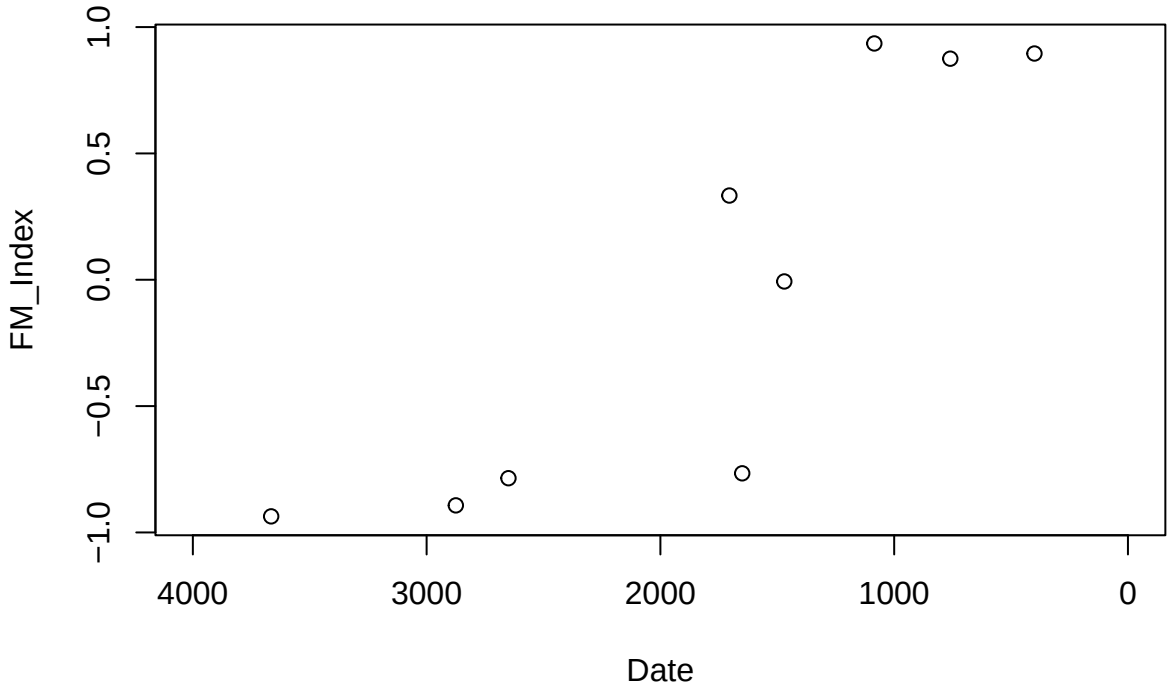
For the sake of consistency, let's once again model the "FM_Index" (*Fish-Mammal Index*) through time. I calculated it based on the specimen counts for fish and mammals reported in Broughton (1994). It represents the relative abundance of fish versus mammals in each site. When the index value is positive, there are more fish bones than mammals. When the index is negative, there are more mammals than fish.

Inspect this column again, and plot the FM Index through time.

```
sac.data$FM_Index
```

```
## [1] -0.936363636 -0.892957746 -0.785425101 -0.766055046 -0.006713212
## [6]  0.895438237  0.935188715  0.875000000  0.333333333
```

```
plot(FM_Index ~ Date, data = sac.data, xlim = c(4000, 0))
```



We once again see the positive relationship between the FM Index and Date, as one moves towards the present (0 BP). But of course we want to know if this trend is actually meaningful or not, so we apply a statistical test.

2.2 Run the Pearson's Correlation

The FM Index is comprised of continuous values (they have decimals) ranging both above and below zero (possibly unbounded). These data can therefore likely be modeled using a normal/Gaussian distribution. It is for this reason that we used the `lm()` function in the previous tutorial, but this also matters here because Pearson's Correlation makes the same general assumptions as Gaussian linear regression. The variables being analyzed must both be continuous, errors must be normally distributed, and the use of this test must make real-world sense. In the context of Pearson's Correlation, that means that the relationship between the two variables must be linear. If your data violate these assumptions, perhaps Pearson's Correlation isn't the best choice of statistical test. But since the FM Index and time both likely conform to these assumptions, we can proceed with a Pearson's Correlation.

Pearson's Correlation is very simple to run in R. We can simply plug our two variables into the `cor.test()` function as follows. Note that the `cor.test()` function has slightly different syntax than `lm()`. If you try

to use `~` or `data = sac.data` here, it will throw an error and not compute the correlation for you. You'll have to specify `sac.data$FM_Index` in order for this function to work.

```
cor.test(sac.data$FM_Index, sac.data$Date)
```

```
##
##  Pearson's product-moment correlation
##
## data:  sac.data$FM_Index and sac.data$Date
## t = -4.7548, df = 7, p-value = 0.002072
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9731846 -0.4997027
## sample estimates:
##        cor
## -0.8738316
```

In the output of this function, you can see that we have run a "Pearson's product-moment correlation", and the data we used is listed as well. Then you see a t score, degrees of freedom, and a p-value. This is the hypothesis test for the Pearson's Correlation, showing our significant p-value of 0.002. This means that if there were no correlation between FM Index and Date, there would only be a 0.2% chance of obtaining our data.

At the bottom of the output is the value of -0.8738, labeled as "cor" under "sample estimates:". This is our correlation coefficient, denoted as r. The correlation coefficient is a value ranging from -1 to 1. If a correlation does not exist, r = 0. If there is a perfect negative correlation, in which as Date goes up, the FM Index goes down, r = -1. If there is a perfect positive correlation, in which as Date goes up, FM Index also goes up, r = 1. Our r value of -0.874 is indicative of a strong negative correlation (remember that Date goes 'up' from 0 to 4000, so an increasing Date value is moving into the past).

To report the results of this test, you would state something akin to the following:

There is a significant negative correlation between FM Index and Date (r = -0.87; t = -4.75, df = 7, $p < 0.01$).

3 Pearson's Correlation as a Linear Model

As discussed in the lecture on linear models, pretty much every common statistical test you'll encounter is simply a type of linear model. Pearson's Correlation is no exception, and is perhaps even more closely related to the basic linear model than many other tests.

To demonstrate, let's model the FM Index through time as we did in the previous tutorial, using the `lm()` function to create a linear model.

```
sac.mod <- lm(FM_Index ~ Date, data = sac.data)
```

So now, let's inspect the model summary.

```
summary(sac.mod)
```

```
##
## Call:
## lm(formula = FM_Index ~ Date, data = sac.data)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.83325 -0.17650 -0.01691  0.30332  0.48598
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.1827966  0.2938171   4.026  0.00502 **
## Date        -0.0006761  0.0001422  -4.755  0.00207 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4278 on 7 degrees of freedom
## Multiple R-squared:  0.7636, Adjusted R-squared:  0.7298
## F-statistic: 22.61 on 1 and 7 DF,  p-value: 0.002072
```

As you can see from this output, our regular linear model reports a significant p-value for the effect of Date on FM Index (t = -4.755, df = 7, p = 0.00207). These values are identical to those reported by the `cor.test()` function. This is because correlation is just a simplified version of a linear model: it's all just linear regression.

This is further indicated by the r^2 value reported in the summary output for the linear model. Our r^2 is 0.7636 for this model. In contrast, our r value for our Pearson's Correlation is -0.8738. Both of these measures - r^2 and r - are showing similar things: the goodness of fit of our model. An r^2 value of 0.76 means that 76% of the variance in FM Index can be explained by variation in Date. An r value of -0.87 means that there is a strong negative correlation between FM Index and Date.

But to fully understand the relationship between these two values, let's take our r value from the Pearson's correlation test and square it.

```
cor.test(sac.data$FM_Index, sac.data$Date)$estimate ^2
```

```
##       cor
## 0.7635817
```

When we square our r value of -0.8738, we of course get our r^2 value of 0.7636 - just as the name "r squared" implies. The correlation coefficient r and the coefficient of determination r^2 are therefore closely related to each other: the latter is simply the square of the former.

Pearson's Correlation is therefore very simply a stripped down version of a linear model. It does away with some of the mathematical density of a linear model and simply reports the bare minimum that most folks are interested in. This certainly has its advantages, but also its costs. You should consider using a correlation test instead of a full linear model when you've got a very simple bivariate relationship to analyze and you don't need to predict anything or care much about error ranges. It's a quick and easy statistic to employ when your data meet the necessary assumptions and is very commonly used in archaeology, anthropology, and many other disciplines.