INTRODUCTION TO PLS 206

Emilio A. Laca

1 October 2014

Contents

What are multivariate statistics?

Operationally, we can define multivariate systems as systems whose state and behavior needs more than one variable to be fully described. Clearly, most systems are multivariate. But in research and modeling we sometimes reduce systems to a single variable for practical purposes. We could discuss the concept that the Dow Jones index is a univariate system because all you can and need to know about it is its current state, a single number.

Although the word "*multivariate*" refers to the fact that multiple variables are involved,, there are two types of definitions of multivariate statistics.

- First, in the most general sense, multivariate or multivariable analysis is the branch of statistics that deals simultaneously with more than two variables.
- In a more strict sense, multivariate analysis deals with multiple *response* (or dependent) variables simultaneously.

Therefore, multiple linear regression is a multivariate method according to the first definition, but it is not a "true" multivariate method according to the strict definition. There is a statistical method or approach called *multivariate multiple linear regression* that is a true multivariate method. MANOVA is a true multivariate method because it handles multiple response variables.

Statistic

A statistic is the result of applying a function or algorithm to a set of data. For example, the average of a sample is a statistic, where the algorithm consists of adding all values in teh sample and dividing the result by the sample size. Other typical statistics are the median of a sample, the sample variance, the maximum and minimum, etc.

> Statistics are random variables.
>
> Statistics are different from parameters.
>
> Parameters are fixed quantities that are typically unobservable.

Dependent and independent variables.

In statistics in general we consider sets of variables called independent, explanatory, or X variables (IV's), and sets of variables labeled as dependent or Y variables (DV's). Typically, variables are not intrinsically dependent or independent. The classification depends on the details of the study, including the purpose of the analysis and the way in which the data were obtained.

Consider for example the study of the relationship between animals size and food consumption. If the goal of the study is to determine whether larger animals of a population tend to eat more, we take a stratified random sample of the population in case (say the deer of Stanislaus National Forest, SNF), where each stratum has animals of closely similar weight, and expose all of them to equal food availability. Then, food consumption is measured and the relationship between food consumption (Y) and animal size (X) is studied. If the goal of the study is to determine whether animals that consume more food tend to be larger we could measure the amount of food consumed in "natural" conditions by a random sample of deer, stratify by level of food consumption, and measure deer size. Then, size (Y) can be related to food consumption (X). Notice that in neither case was "effect of X on Y" mentioned. We just said deer with greater X tend to have greater (smaller) Y. This relates to the concept of cause and effect considered below.

Before going on to the next section, you may want to think about the example above where I said that deer of different sizes are exposed to equal food availability. Is that a necessary condition or does it also depend on the goals of the study? Think about what you would be able to say about the expected food consumption of deer of different sizes at SNF.

Models

Most if not all statistical methods and techniques use models. In a deep sense, data can make sense in relation to some model of reality. This concept is summarized in the following:

Data = Model + Error = Signal + Noise = f(predictors, parameters) + random variable

Further, other models that we will consider in this course have different types of errors or random components. For example, models where residuals are correlated because of proximity in space or time can be written as:

Data = Deterministic model + Random structure model + unstructured error

Notice that each data value is partitioned into the components listed. The same data can be partitioned in different ways depending on the goals of the analysis. This is seen, for example, when the same data set is analyzed as a mixed model, where part of the model includes a random component with structure (the random effects), or as a generalized model, where all random components are incorporated into a "noise" that itself has structure. By "structure", we mean that the random components have certain covariance patterns, as opposed to IID errors.

IID means independent identically distributed

True multivariate statistics

Some authors define true multivariate statistics are those methods that deal with multiple Y or response variables, regardless of the number of X's. Others consider that just having multiple X's qualifies as multivariate, but typically exclude any kind of ANOVA that has only one Y or response variable. For example, an experiment with two factors (say N and P fertilization) where yield is the only response variable, is considered univariate.

Most people agree that MANOVA, discriminant analysis (DA), principal components analysis (PCA), canonical correlation (CCA), etc. are multivariate. It should be mentioned that pure PCA only contains X's; no Y's are involved. Multiple linear regression and path analysis are sort of in the gray area. There is a true multivariate version of MLR, where many Y's are regressed on many X's, and everyone agrees this is a multivariate case.

Usually, scientists treat truly multivariate situations in a univariate fashion. For example, when an experiment is performed to study the effect of NxP fertilization on a crop, usually one measures not only yield but a number of other Y's. Yet, we typically analyze each Y separately. The implications of this may range from giving incorrect probability levels to having little impact. Univariate analysis may lead to true error rates being greater than the nominal α level, a situation called "inflated error rates." A multivariate approach is more conservative and may have more true power (1-β). Moreover, some results are only detectable, understandable, or "visible" with a multivariate approach. Let's take a look at an example.

Univariate and multivariate differences

The step from univariate to true multivariate statistics involves one crucial increase in complexity, the use of multivariate distributions. In addition, it requires that we reconsider the way in which we compare populations or objects. In univariate statistics this is not an issue, because objects, samples and populations are compared on a single characteristics at a time. For example, two soils can be compared based on their organic matter content (OM); two plant communities can be compared based on the abundance of a single species. Soils with similar organic matter content are considered more similar to each other than soils with widely different organic matter contents. When we consider more than one variable, say OM and pH things change. Consider the following two pairs of soils. Which pair is more different than the other? When two or more variables are used to compare objects, the concept of degree of difference becomes more complex. It is necessary to introduce the concept of distance, and statistical distance.

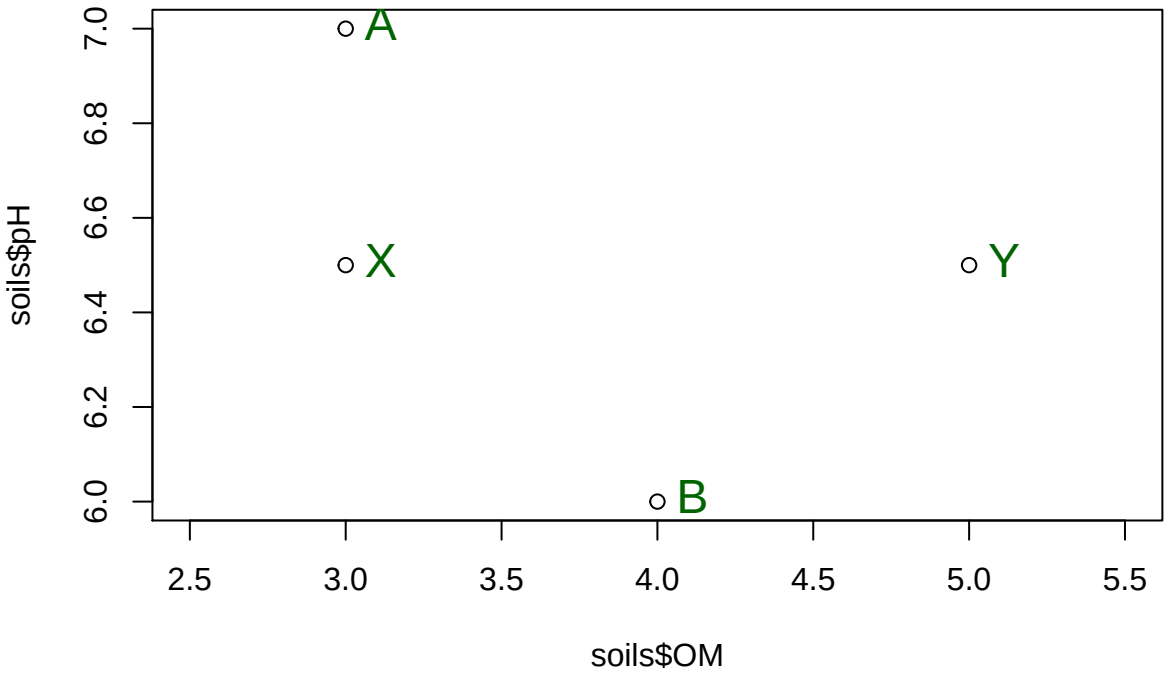
Variable	1st pair		2nd Pair	
_____	Soil A	Soil B	Soil X	Soil Y
OM%	3.0	4.0	3.0	5.0
pH	7.0	6.0	6.5	6.5

One possible solution is to measure or calculate the Euclidean distance between the objects, but this approach ignores the inherent errors and variances of the population. Multivariate analysis uses the statistical distance as a measure of differences in multiple dimensions. One of the consequences of the concept of multivariate differences is that populations can simultaneously be not significantly different in any single characteristic, and be significantly different in a multivariate sense. The concepts of distance, statistical distance, and multivariate differences will be expanded in future chapters.

```
soils <- data.frame(soil = c("A", "B", "X", "Y"), OM = c(3.0, 4.0, 3.0, 5.0), pH = c(7.0, 6.0, 6.5, 6.5
soils
```

```
##   soil OM  pH
## 1    A  3 7.0
## 2    B  4 6.0
## 3    X  3 6.5
## 4    Y  5 6.5
```

```
plot(soils$OM, soils$pH, xlim = c(2.5, 5.5))
text(soils$OM, soils$pH, soils$soil, cex=1.5, pos=4, col="darkgreen")
```




```
dist(soils[,2:3], method = "euclidean")
```

```
##       1     2     3
## 2 1.414
## 3 0.500 1.118
## 4 2.062 1.118 2.000
```

Statistical modeling, causal relationships

Cause-and-effect relationships

Most students have been warned repeatedly, and correctly, that statistical association does not imply that X causes Y. The establishment of cause-and-effect relationships is a philosophical issue deeper than statistics, and worth considering. What does it mean to say that A causes B? How does one put the trite warning to work in the real world? It is generally agreed that A can be considered to cause B if the following conditions are met:

1. Covariance
2. Time ordering
3. Elimination of other possible causes

Covariance means that A and B are statistically associated. When A happens, B tends to happen. Time ordering means that A happens prior to B. In a very deep sense of the way humans understand the world, a cause cannot happen before the effect. Finally, the most restrictive condition is that all other possible causes must be eliminated. From a statistical point of view, this can be approached by randomizing, controlling and manipulative experimentation.

Randomization is of course the basis for the application of any statistics and for the utterance of any statements that refer to probabilities. By taking a random sample of the population in question and randomly assigning individuals or objects ("plots") to each of a series of imposed treatments, the potential impact of

any other variable not associated with the treatments is expected to be eliminated. This ensures (we hope) that no other variable (say Z) that could be the true cause of event Y will covary with the treatments (X's), thus preventing a mistake in which one would say X causes Y while in fact it is Z, through its association with X (who has nothing to do with Y), that causes Y. Note, however, that even in the case where X is a true cause of Y and we randomize, a large random variation in a Z that also affects Y can mask the effects of X and prevent us from identifying X as a cause. This is where the idea of control comes in. In manipulative experiments it is desirable to control for the levels of variables that may affect Y but are not of interest for the researcher.

Relationship between body size and food intake in deer. Without information about the specific manner in which the population was sampled and measured, one cannot determine which one is the dependent and which the independent variable.

What does this mean in the real world? Take the deer example above; suppose that you measure size and food consumption for a random sample of individuals in the wild. In this case, assume that you measured food consumption in the natural conditions. This is a typical observational study. You cannot say anything about what causes deer to consume more or less food. Yet, you have valuable data that indicate that for the sample population, larger deer will eat more food. Moreover, you will be able to take any deer at random from the population, and after you measure its size you will know a lot more about its expected food consumption in the wild than you knew before you measured its size. Now, suppose you take a deer from a different population and measure its size. The relationship in the figure will not help you at all to guess its food consumption; at least not from a strict statistical point of view. In the real world, however, you will have a much better idea about its expected food consumption than before you studied the SNF population.

Model identification

In the vast majority of statistical analyses, a particular model is imposed on the interpretation of the data. Typically, it is assumed that variables are linearly related. In many cases, the linearity can be tested and other models can be tried. For example a scatter plot of absolute population growth rate vs. population size will show that they are not linearly related. In this case, there is an abundance of theories, explanations, and models to support alternative analyses. Those theories and models may not be available in many other more complex situations, thus leaving one with the linearity assumption and its potential rejection.

Data by themselves, without a model or interpretation, have little meaning. The interpretation of a data set is completely dependent on the model imposed or tested.

Proper identification of the model is important, because the interpretation of results is strictly dependent on the model. Thus, whenever possible, alternative models should be selected prior to the analysis of data, on the basis of some theory or hypothesis. Of course, it is also common, valid and correct to try empirical models in a descriptive fashion, or in early stages of the research, in order to generate hypotheses for further experimentation.

Example

```
# Simulate data
n <- 50
x <- 1:n # values for time
y.s <- 150 + 80 * sin(x/3.5) + x # structural part of relationship, in this case it is known.
r.error <- rnorm(length(x), mean = 0, sd = sqrt(2500)) # random component, truly iid normal.
y <- y.s + r.error # simulated observed values.

# Plot the observations and then the true relationship.
plot(y ~ x, pch=19, cex=1.5)
lines(y.s ~ x) # add structural or signal part of data
```

```r
# Use a linear regression as a first guess

my.lr <- lm(y ~ x)
summary(my.lr)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -163.75  -53.02   -6.28   43.49  207.61
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  173.125     21.335    8.11  1.5e-10 ***
## x              1.017      0.728    1.40     0.17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74.3 on 48 degrees of freedom
## Multiple R-squared:  0.039,  Adjusted R-squared:  0.019
## F-statistic: 1.95 on 1 and 48 DF,  p-value: 0.169
```

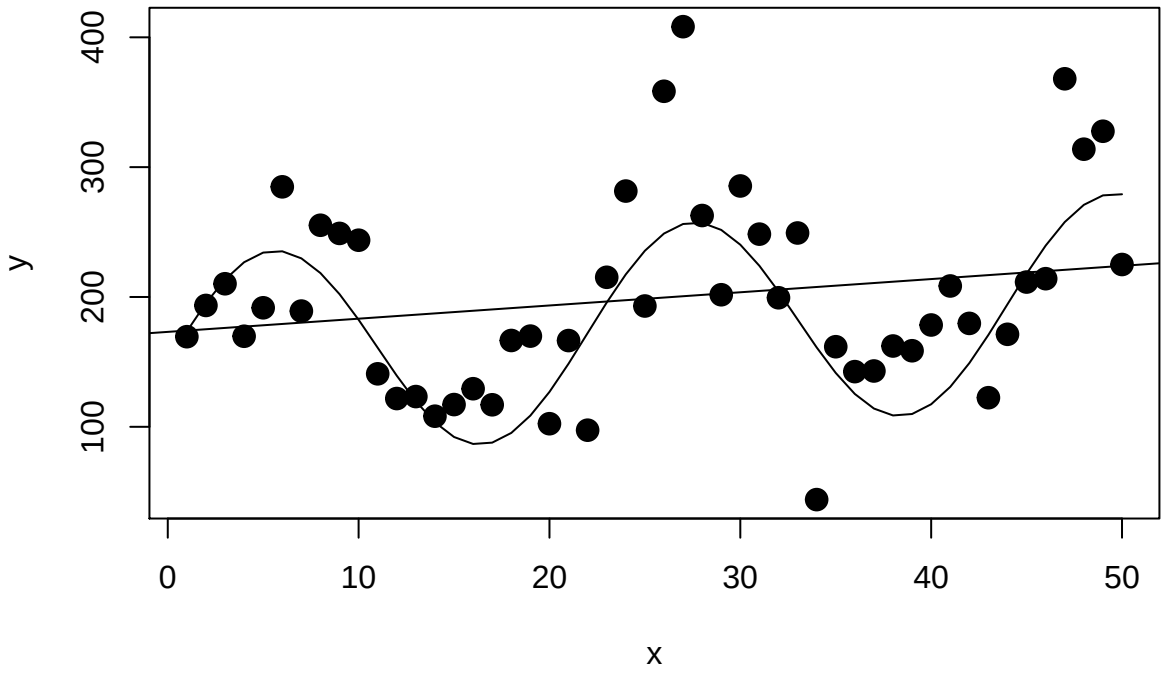
```r
coef(my.lr)
```

```
## (Intercept)           x
##     173.125       1.017
```

```r
str(coef(my.lr))
```

```
##  Named num [1:2] 173.13 1.02
##  - attr(*, "names")= chr [1:2] "(Intercept)" "x"
```

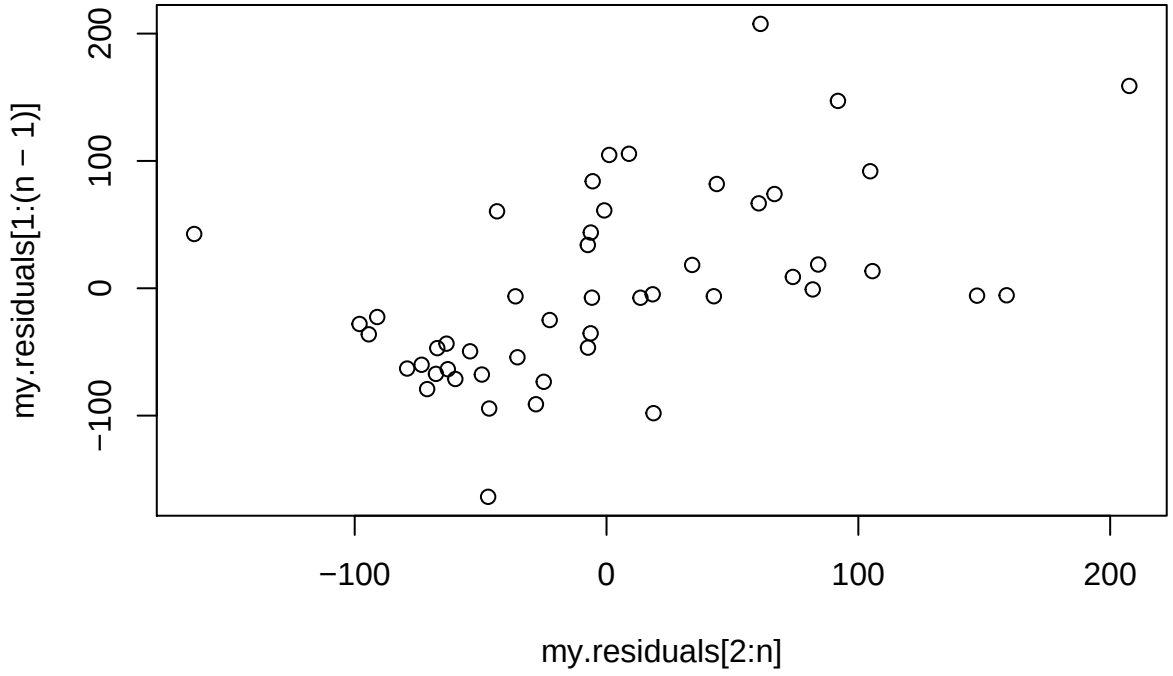
```r
abline(coef=c(coef(my.lr)["(Intercept)"],coef(my.lr)["x"])) # Add the prediction line to the plot
```




```
# How do we know if the model is correct? We don't. But we can find out if it is incorrect.

my.residuals <- residuals(my.lr)

plot(my.residuals[2:n],my.residuals[1:(n-1)])
```




```
cor(my.residuals[2:n],my.residuals[1:(n-1)])
```

```
## [1] 0.5277
```

Populations, samples, manipulative and observational studies

Read Faraway-PRA.pdf section 3.3 and 3.9.

> Samples are measured in order to make generalizations about populations

Statistical inference has two potential bases:

1. Design-based (based random sample from populatino of interest)
2. Model-based (based on an estimated model for the data-generating process)

Manipulative study with proper randomization

Let me present a fictitious example to illustrate the role of randomization and random application of manipulative treatments. This example is based on discussions I have had with one of your fellow students. In a perfect world a typical study proceeds as follows: 1. Identification and definition of the population for which inferences will be made. For example: all the streams in the Sierra Nevada.

2. Statement of goal and hypothesis to be tested. For example: aquatic insect biodiversity of these streams is reduced by proximity to developed areas, but it can be increased by restoration.
3. Experimental design, replications, etc. are carefully selected and planned. Suppose that the best design is a completely randomized design with three treatments: pristine stream (control), stream near development, and restored stream near development and 5 replications in each treatment.
4. A random sample of 15 streams from the population is selected (assuming that there are plenty of pristine streams!) and treatments are randomly assigned to each one of them. Because the streams are a random sample from the populations and treatments were randomly assigned, the analysis has the potential to detect a causal relationship. This will also allow statements like "on the average, streams of the Sierra Nevada will suffer a reduction (increase, or no change) of insect biodiversity if they are impacted by development."
5. Treatments are applied for a specified length of time and then results are measured. By the way, measuring biodiversity in each stream prior to the application of treatments would be a great idea to potentially increase the power of the experiment, a fact that will be discussed when we study ANCOVA.
6. Predetermined statistical analyses and tests of hypotheses are performed and a conclusion is reached. The probabilistic inferences are valid for the whole population (e.g. 95% confidence that the impacted stream will end up with a lower biodiversity than the restored one).

Observational study with random sampling

In the real world, however, things typically do not proceed as above. Let me consider two levels of departure from the example above that determine the validity of the results and how useful the data can be.

It is not realistic to impose development and further restoration treatments manipulatively, but it is possible to randomly sample from the three sub-populations of streams: pristine, impacted and not restored, and impacted and restored. Five streams are randomly selected from all possible candidates within each class, resulting in an observational study. This constraint removes the potential to say anything about the cause-and-effect relationship between stream situation and biodiversity.

But not all is lost! Not even close. Purely on the basis of random sample selection, one can make valid statistical inferences for the population. For example, if significant effects are found, one can make statements such as "The biodiversity is lower (higher, not different) in streams impacted by development that in those not

impacted by development." It would be honest and helpful to also say that one is not statistically implying that development affects biodiversity, but that on average streams in the different sub-populations tend to have different biodiversity, regardless of the true cause. In addition, the study yields a wealth of data that can be used for the assessment and management of 15 specific streams.

Observational study without random sampling

In addition to being impossible to manipulate streams, it may not be realistic to select streams randomly at all, although it would be very strongly recommended. If streams from the sub-populations are selected on some ad hoc basis, the statistical scope of the study suffers significantly. By the way, this happens all the time, and valid, useful information can be obtained in this manner. Because of the ad hoc selection of streams, the target population is no longer clearly defined. Strictly speaking, statistical results become purely descriptive of the 15 streams measured. It would still be valid to make statements such as "The average for the 5 impacted streams was significantly different from the average for the 5 pristine streams," but not much could be said (on a purely statistical basis) about other streams unless one has a means to determine that the other stream is from a population with the same characteristics as the undefined population represented by the 15 streams sampled. Even in this case not all is lost. The study yields valuable data for the comparison and management of the 15 streams measured.

The data might also be sufficient to parameterize a model-based inference, for which random sampling of a population is not strictly necessary. This is a deeply difficult concept. In model-based inference it does not matter whether sampling units were selected randomly. Presumably, the potential indequacy of models estimated from poorly representative samples will be exposed when the variances of predictions are calculated.

All population is sampled

Read Section 2.12 page 23 of Faraway-PRA.pdf.

The example consists of data on the number of species of tortoises in 30 (most or all) of the Galápagos Islands. In this case, all or almost all of the existing population has been measured. The islands measured were not a random sample of a larger target population. If the whole population has been measured, all parameters can be calculated with certainty if we ignore the measurement error. The concept of "significance" can be introduced in this case by considering that the present distribution of species is the result of a **random process** in which there was or was not a certain relationship between geographical characteristics and number of species. Under the null hypothesis that the process of species distribution to achieve the current state was completely independent of geographical variables, the distribution observed should not differ much from random allocation of the observed number of species over islands. If it does, we reject the independence and find "significant" relationships with geographical variables. This can be achieved by a permutation test, and is estimated by the regular F-test, under the assumption of normality and independence of errors. The p-values in this last case apply in regards to the potential ensemble of results or population of islands that "could have been."

```
library(faraway)
```

```
##
## Attaching package: 'faraway'
##
## The following object is masked _by_ '.GlobalEnv':
##
##     wheat
```

```r
data(gala)
str(gala)
```

```
## 'data.frame':    30 obs. of  7 variables:
##  $ Species  : num  58 31 3 25 2 18 24 10 8 2 ...
##  $ Endemics : num  23 21 3 9 1 11 0 7 4 2 ...
##  $ Area     : num  25.09 1.24 0.21 0.1 0.05 ...
##  $ Elevation: num  346 109 114 46 77 119 93 168 71 112 ...
##  $ Nearest  : num  0.6 0.6 2.8 1.9 1.9 8 6 34.1 0.4 2.6 ...
##  $ Scruz    : num  0.6 26.3 58.7 47.4 1.9 ...
##  $ Adjacent : num  1.84 572.33 0.78 0.18 903.82 ...
```

```r
m1 <- lm(Species ~ Area + Nearest, data = gala)
summary(m1)
```

```
##
## Call:
## lm(formula = Species ~ Area + Nearest, data = gala)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -98.93 -52.92 -33.35   5.99 309.81
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  59.1078    21.9598    2.69  0.01206 *
## Area          0.0828     0.0202    4.11  0.00033 ***
## Nearest       0.4435     1.2198    0.36  0.71897
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 93.2 on 27 degrees of freedom
## Multiple R-squared:  0.385,  Adjusted R-squared:  0.339
## F-statistic: 8.44 on 2 and 27 DF,  p-value: 0.00142
```

```r
str(summary(m1))
```

```
## List of 11
##  $ call         : language lm(formula = Species ~ Area + Nearest, data = gala)
##  $ terms        :Classes 'terms', 'formula' length 3 Species ~ Area + Nearest
##   .. ..- attr(*, "variables")= language list(Species, Area, Nearest)
##   .. ..- attr(*, "factors")= int [1:3, 1:2] 0 1 0 0 0 1
##   .. .. ..- attr(*, "dimnames")=List of 2
##   .. .. .. ..$ : chr [1:3] "Species" "Area" "Nearest"
##   .. .. .. ..$ : chr [1:2] "Area" "Nearest"
##   .. ..- attr(*, "term.labels")= chr [1:2] "Area" "Nearest"
##   .. ..- attr(*, "order")= int [1:2] 1 1
##   .. ..- attr(*, "intercept")= int 1
##   .. ..- attr(*, "response")= int 1
##   .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
##   .. ..- attr(*, "predvars")= language list(Species, Area, Nearest)
##   .. ..- attr(*, "dataClasses")= Named chr [1:3] "numeric" "numeric" "numeric"
```

```
##   .. .. ..- attr(*, "names")= chr [1:3] "Species" "Area" "Nearest"
##  $ residuals    : Named num [1:30] -3.45 -28.48 -57.37 -34.96 -57.95 ...
##   ..- attr(*, "names")= chr [1:30] "Baltra" "Bartolome" "Caldwell" "Champion" ...
##  $ coefficients : num [1:3, 1:4] 59.1078 0.0828 0.4435 21.9598 0.0202 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:3] "(Intercept)" "Area" "Nearest"
##   .. ..$ : chr [1:4] "Estimate" "Std. Error" "t value" "Pr(>|t|)"
##  $ aliased      : Named logi [1:3] FALSE FALSE FALSE
##   ..- attr(*, "names")= chr [1:3] "(Intercept)" "Area" "Nearest"
##  $ sigma        : num 93.2
##  $ df           : int [1:3] 3 27 3
##  $ r.squared    : num 0.385
##  $ adj.r.squared: num 0.339
##  $ fstatistic   : Named num [1:3] 8.44 2 27
##   ..- attr(*, "names")= chr [1:3] "value" "numdf" "dendf"
##  $ cov.unscaled : num [1:3, 1:3] 5.55e-02 -1.54e-05 -1.81e-03 -1.54e-05 4.68e-08 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:3] "(Intercept)" "Area" "Nearest"
##   .. ..$ : chr [1:3] "(Intercept)" "Area" "Nearest"
##  - attr(*, "class")= chr "summary.lm"
```

```
obs.F <- summary(m1)$fstatistic["value"]
fstats <- numeric(5000)
for(i in 1:5000){
  fstats[i] <- summary(lm(sample(Species) ~ Area + Nearest, data = gala))$fstatistic["value"]
}

(critical.F <- qf(p = 0.95, df1 = summary(m1)$fstatistic["numdf"], df2 = summary(m1)$fstatistic["dendf"]
```

```
## [1] 3.354
```

```
length(fstats[fstats > obs.F])/5000
```

```
## [1] 0.0082
```

```
sum(fstats > obs.F)/5000
```

```
## [1] 0.0082
```