Chapter 21

Hypothesis Testing

21.8 Analysis of Variance (ANOVA) in R

In this section, we shall learn about the `anova()` function in R using the `gala` data set in the `faraway` package.

First, we have to install and load the `faraway` package in R. To do so, we use the following commands.

```
> install.packages("faraway")
> library(faraway)
```

We can now access the `gala` data set. Let us see what's in it.

```
> gala
             Species Endemics    Area Elevation Nearest Scruz
    Adjacent
Baltra            58       23   25.09       346     0.6   0.6
    1.84
Bartolome         31       21    1.24       109     0.6  26.3
    572.33
Caldwell           3        3    0.21       114     2.8  58.7
    0.78
Champion          25        9    0.10        46     1.9  47.4
    0.18
Coamano            2        1    0.05        77     1.9   1.9
    903.82
Daphne.Major      18       11    0.34       119     8.0   8.0
    1.84
Daphne.Minor      24        0    0.08        93     6.0  12.0
    0.34
Darwin            10        7    2.33       168    34.1 290.2
    2.85
Eden               8        4    0.03        71     0.4   0.4
    17.95
Enderby            2        2    0.18       112     2.6  50.2
    0.10
Espanola          97       26   58.27       198     1.1  88.3
    0.57
```

```
Fernandina          93        35  634.49       1494      4.3  95.3
    4669.32
Gardner1            58        17    0.57         49      1.1  93.1
    58.27
Gardner2             5         4    0.78        227      4.6  62.2
    0.21
Genovesa            40        19   17.35         76     47.4  92.2
    129.49
Isabela            347        89 4669.32       1707      0.7  28.1
    634.49
Marchena            51        23  129.49        343     29.1  85.9
    59.56
Onslow               2         2    0.01         25      3.3  45.9
    0.10
Pinta              104        37   59.56        777     29.1 119.6
    129.49
Pinzon             108        33   17.95        458     10.7  10.7
    0.03
Las.Plazas          12         9    0.23         94      0.5   0.6
    25.09
Rabida              70        30    4.89        367      4.4  24.4
    572.33
SanCristobal       280        65  551.62        716     45.2  66.6
    0.57
SanSalvador        237        81  572.33        906      0.2  19.8
    4.89
SantaCruz          444        95  903.82        864      0.6   0.0
    0.52
SantaFe             62        28   24.08        259     16.5  16.5
    0.52
SantaMaria         285        73  170.92        640      2.6  49.2
    0.10
Seymour             44        16    1.84        147      0.6   9.6
    25.09
Tortuga             16         8    1.24        186      6.8  50.9
    17.95
Wolf                21        12    2.85        253     34.1 254.7
    2.33
```

But what do these numbers mean? A brief description of the `gala` data set is given below.

```
> ?gala

There are 30 Galapagos islands and 7 variables in the dataset. The
    relationship between the number of plant species and several
    geographic variables is of interest. The original dataset
    contained several missing values which have been filled for
    convenience. See the galamiss dataset for the original version.
```

One may use the command `?gala` to obtain the above description and an in depth explanation of all the variables in the data set. We skip this step here.

Since the data set is quite large and occupies most of our screen, we shall only work with the variable names. The `names()` function in R allows us to see only the names of the variables in a data set.

```
> names(gala)
[1] "Species"   "Endemics"  "Area"      "Elevation" "Nearest"   "
    Scruz"     "Adjacent"
```

Naturally, we are interested in learning about the relationship between these variables. Say we are interested in testing the effect of the variable `Area` on the value of `Species` when the variable `Endemic` is not present in the model i.e, we want to test if the regressor `Area` is of any importance or not.

So, we can formulate our hypothesis test as follows:

$$H_0 : \beta_{area} = 0, \beta_{endemic} = 0$$

against the alternate hypothesis

$$H_1 : \beta_{area} \neq 0, \beta_{endemic} = 0$$

Let us fit a generalised linear model using the `lm()` function in R with the `Species` variable as response and all the other variables except `Endemics` as input. We shall call this model `fit`. This will be our unrestricted model, which we assume is a good fit.

Our restricted model under the null hypothesis is `fit0` where the input `Area` has been removed.

```
# UNRESTRICTED MODEL
> fit = lm(Species ~ . - Endemics, gala)

# RESTRICTED MODEL
> fit0 = lm(Species ~ . - Endemics - Area, gala)
```

Here, instead of writing the names of all the inputs, we simply replace them by `.` and use `-` in front of the input we want to remove from the model. Of course, we can type `fit` and hit Enter (similarly for `fit0`) to see the estimates of the intercepts and the coefficients but we are not interested in them in this case.

Normally, we would have to go through the cumbersome process of computing expressions involving the residual sum of squares (RSS) manually for an F-test. Thankfully, our task is made easier by a built in feature of R, the function `anova()`, which takes the restricted model as the first argument and the unrestricted model as the second argument. Let us go ahead and call this function.

```
> anova(fit0, fit)
Analysis of Variance Table

Model 1: Species ~ (Endemics + Area + Elevation + Nearest + Scruz +
     Adjacent) -
    Endemics - Area
Model 2: Species ~ (Endemics + Area + Elevation + Nearest + Scruz +
     Adjacent) -
    Endemics
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     25 93469
2     24 89231  1    4237.7 1.1398 0.2963
```

As we can see, the output of `anova()` gives us the two models, the residual degrees of freedom (`Res.Df`), the residual sum of squares (`RSS`), the difference in the degrees of freedom (`Df`), the difference in the residual sum of squares $RSS_0 - RSS$ given by (`Sum of Sq`), the F-statistic (`F`) and the p-values (`Pr(>F)`).

The p-value, given by `Pr(>F)` is the probability that we shall find a statistic more extreme than the observed value of the F-statistic if the null hypothesis is true. Here, the p-value is `0.2963` which is much larger than 0.05. Thus, we fail to reject H_0.

Therefore, we may say that the variable `Area` does not have a significant effect on the values of the response variable `Species`, i.e. we can drop this variable and the resulting fit is not much further than the generalised fit.

NOTE: One must be careful to not change the order of the models in the call to the `anova()` function. Let us see what happens if we write the unrestricted model first and the restricted model later.

```
> anova(fit, fit0)
Analysis of Variance Table

Model 1: Species ~ (Endemics + Area + Elevation + Nearest + Scruz +
     Adjacent) -
    Endemics
Model 2: Species ~ (Endemics + Area + Elevation + Nearest + Scruz +
     Adjacent) -
    Endemics - Area
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     24 89231
2     25 93469 -1   -4237.7 1.1398 0.2963
```

We see that while the p-value and the observed value of the F-statistic remain the same, this command gives us negative degrees of freedom and sum of squares. This of course makes no sense in the real world. Hence, one must take care while working with the `anova` function in R.