La Vuelta

Laura Cebollero Ruiz, Alexandre Rodríguez Garau

26 de diciembre de 2018

In this project we intent to predict the duration of the different stages of the cycling race *La Vuelta* using the information provided in the file `tour.xlsx`, `tour.xsls`.

```
library("readxl")
library("knitr")
```

```
data <- read_excel("Tour.xlsx")
```

```
## readxl works best with a newer version of the tibble package.
## You currently have tibble v1.4.2.
## Falling back to column name repair from tibble <= v1.4.2.
## Message displays once per session.
```

```
data[,11:15] <- NULL  #Delete empty columns
```

To predict the length of the stages we will use a set that contains 10 explanatory variables and a response variable called **`ForecastedTime`**. Let's take a look at the summary of the variables:

```
kable(summary(data[,1:6]))
```

ports1	ports2	ports3	year	week	bef_mount
Min. :0.000	Min. :0.0000	Min. :0.0000	Min. :1.000	Min. :1.000	Min. :0.0000
1st Qu.:0.000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:2.000	1st Qu.:1.000	1st Qu.:0.0000
Median :0.000	Median :0.0000	Median :1.0000	Median :4.000	Median :2.000	Median :0.0000
Mean :0.581	Mean :0.4857	Mean :0.8476	Mean :3.619	Mean :2.048	Mean :0.2952
3rd Qu.:1.000	3rd Qu.:1.0000	3rd Qu.:2.0000	3rd Qu.:5.000	3rd Qu.:3.000	3rd Qu.:1.0000
Max. :3.000	Max. :3.0000	Max. :4.0000	Max. :6.000	Max. :3.000	Max. :1.0000

```
kable(summary(data[,6:11]))
```

bef_mount	aft_mount	bef_tt	aft_tt	last	ForecastedTime
Min. :0.0000	Min. :0.0000	Min. :0.0000	Min. :0.0000	Min. :0.00000	Min. :180.0
1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.00000	1st Qu.:263.4
Median :0.0000	Median :0.0000	Median :0.0000	Median :0.0000	Median :0.00000	Median :307.8
Mean :0.2952	Mean :0.2952	Mean :0.1333	Mean :0.1905	Mean :0.04762	Mean :306.7
3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:0.0000	3rd Qu.:0.0000	3rd Qu.:0.00000	3rd Qu.:348.2
Max. :1.0000	Max. :1.0000	Max. :1.0000	Max. :1.0000	Max. :1.00000	Max. :437.5

In the previous tables we can see a short summary of the variables. **`ports1`**, **`ports2`** and **`ports3`** indicate the number of mountain sections in a given stage and their category (1,2 or 3). **`Year`** corresponds to the year in which this data was recorded (1 to 6). **`Week`** is the week of the race in wich the stage takes place (1 to 3). The variables **`bef_mount`** and **`aft_mount`** tell us whether a stage took place before or after a mountain stage, respectively. Similarly, the variables **`bef_tt`** and **`aft_tt`** indicate if a stage took place before or after a time trial stage. Finally, the variable **`last`** tells us if that stage was the last of the whole race.

By the looks of the summary we can't seem to find any outliers or abnormal values. Most of the explanatory variables range from 0 to very low values and are natural numbers. The only variable that has values that

vary in a wider range is the response variable **ForecastedTime**. This makes sense because this variable indicates time and time is a continuous variable. We can also see that the values range from 180 to 437.5.

By taking a closer look at the data we detect some abnormal values in the **last** variable: Since this variable indicates if a stage was the last of the whole race, then there should as many stages with this value equal to 1 as years of data have been recorded. Since the recorded stages are from 6 different years then there should be 6 rows, but istead there are only 5, meaning that there is at least 1 missing stage.

```
data[data$last == 1,]
```

```
## # A tibble: 5 x 11
##   ports1 ports2 ports3  year  week bef_mount aft_mount bef_tt aft_tt  last
##    <dbl>  <dbl>  <dbl> <dbl> <dbl>     <dbl>     <dbl>  <dbl>  <dbl> <dbl>
## 1      0      0      0     6     3         0         0      0      1     1
## 2      0      0      0     5     3         0         0      0      1     1
## 3      0      0      1     4     3         0         0      0      1     1
## 4      0      1      0     2     3         0         1      0      0     1
## 5      0      0      1     1     3         0         1      0      0     1
## # ... with 1 more variable: ForecastedTime <dbl>
```

I we look at the data we can easily see that the missing value belongs to the year 3. However, this will probably not greatly affect our predicting power. There is something to be said about the variables **Year**, **week**, **bef_mount**, **aft_mount**, **bef_tt**, **aft_tt** and **last**: all of these variables are categorical and indicate the group or category a row belongs to. For this, we should transform these variables into factors.

```
data$year <- as.factor(data$year)
data$week <- as.factor(data$week)
data$bef_mount <- as.factor(data$bef_mount)
data$aft_mount <- as.factor(data$aft_mount)
data$aft_tt <- as.factor(data$aft_tt)
data$bef_tt <- as.factor(data$bef_tt)
data$last <- as.factor(data$last)
```

Now that we have transformed the categorical columns to factors we should take another look at the summary of these variables:

```
kable(summary(data[,4:10]))
```

year	week	bef_mount	aft_mount	bef_tt	aft_tt	last
1:15	1:33	0:74	0:74	0:91	0:85	0:100
2:18	2:34	1:31	1:31	1:14	1:20	1: 5
3:17	3:38	NA	NA	NA	NA	NA
4:17	NA	NA	NA	NA	NA	NA
5:18	NA	NA	NA	NA	NA	NA
6:20	NA	NA	NA	NA	NA	NA