Second Assignment

Miguel Alcón Doganoc & Roger Pujol Torramorell

1 Executive summary

In this assignment we want to simulate the behaviour of a runner at a marathon, specifically the Boston marathon.

In order to do the task, we analyzed the Boston marathon information and, with the results of the analysis, we have chosen and extracted what we expect to be the most relevant factors. The main goal that we want to achieve is to generate data that provides accurate levels of randomization, replication and statistical homogeneity

Our solution starts with the simulation of the marathon, and afterwards it continues with the validation of the model and the data. Finally, it finishes with the hypothesis checking and the extraction of some valid conclusions.

2 System description

- **Entities**:
  In our system the only entity is the *runner*.
- **Attributes**:
  - *Age*: The age of the runner.
  - *Gender*: The gender of the runner.
  - *Fitness*: The level of fitness of the runner based on his Bib number (since lower Bib means better performance in other marathons).
- **Activities**:
  We have mainly one activity, which is the runner running, but since this can have different behaviours for different parts of the marathon, we subdivided it in 8 parts. Then we could say that we have 9 activities:
  - Run stage 0 (from 0km to 5km)
  - Run stage 1 (from 5km to 10km)
  - Run stage 2 (from 10km to 15km)
  - Run stage 3 (from 15km to 20km)

  - Run stage 4 (from 20km to 25km)
  - Run stage 5 (from 25km to 30km)
  - Run stage 6 (from 30km to 35km)
  - Run stage 7 (from 35km to 40km)
  - Run stage 8 (from 40km to 42km)

- **States of the system**:
  Since we only have one entity the states will be the runner running each stage. So we will have from state Stage 0 to Stage 8.

- **System progress**:
  The progress of the system is lineal between the nine stages mentioned before.

3 Problem description

The problem we want to solve with the simulation is how some factors of a runner affect his performance in a marathon.

3.1 Simplification Hypotheses

- **SH_01**: We will not use the elevation changes as a factor since we are targeting only a specific marathon and this factor won't change.

- **SH_02**: We will not use any of the weather related attributes (weather, temperature, humidity) since the values for all the data is almost the same and doesn't seem to have any major influence to the results.

- **SH_03**: We will not use the distance as an explicit factor since we are using it implicitly for each stage.

- **SH_04**: We will not use the nationality of the runner as a factor because it is not significant in most cases and when it could have some importance, it is taken into account with the fitness level.

3.2 Structural Hypotheses

- **EH_01**: The age will be used as a factor expressed in years.

- **EH_02**: The gender will be used as a Boolean factor where 0 means women and 1 means men.

- **EH_03**: The fitness level will be used as a factor where 0 means best elite runner and 1000 means amateur runner with no marathon experience. This value will be relative to the Bib number.

- **EH_04**: At each stage we add a normal distribution with average 0 and with a standard deviation of 120 (seconds) for the 5km long stages and 60 (seconds) for the 2km long stage.

4 Model specification

Since we are only going to simulate the runners individually and we use the same inputs for each state, the model will look like the following.

Figure 1: Specification of our model

5 Codification

The model has been programmed in Python3 because it is a very versatile language with many useful packages for statistics. We have done 3 main functionalities: pre-process, the model and experiments.
The pre-process gets the data from the different dataset and store it in a single dataset with only the data we want to use. The data that we have after the pre-processing is: the age of the runner, the gender, the fitness level (based on the Bib) and the time for every stage in seconds.
The model code uses all the data to create linear models for each stage and adds a random number (with normal distribution) to the result of each stage.
The experiments code uses the models created and checks some statistics on them (more detailed at the next section).

6 Definition of the experimental framework

Our design of the experiment (DOE) is based in a 2^k factorial design plus the Yates algorithm. With this DOE we want to observe the effect of each factor within the model and between them.

6.1 2^k factorial design

We have selected the 2^k factorial design because of its easy analysis and interpretation, to acquire a global view of factors in our model. These factors are the following (see table 1).

	Age	**Gender**	**Fitness**
-	18	0	0.043478
+	84	1	1000

Table 1: Extreme values of model's factors

Once we define the highest and lowest values of each factor, we will need 2^k experiments to obtain realistic values. Since we have 3 factors, we must perform 8 experiments (see table 2), each of them repeated 10 times.

Age	**Gender**	**Fitness**	**Values**			**Mean**
-	-	-	18	0	0.043478	9988.895918
+	-	-	84	0	0.043478	9779.528944
-	+	-	18	1	0.043478	10553.542673
+	+	-	84	1	0.043478	10170.772298
-	-	+	18	0	1000	15956.170771
+	-	+	84	0	1000	15802.245597
-	+	+	18	1	1000	16418.229829
+	+	+	84	1	1000	15962.181358

Table 2: 2^k configuration and the resulting mean of each experiment

6.2 Yates algorithm

After the 2^k factorial design, we can continue with the Yates algorithm. The results obtained with it are the following (see table 3)

(1)	**(2)**	**(3)**	**Effect**	**Id**
19768.424863	40492.739834	104631.567388	13078.945924	Mean
20724.314971	64138.827554	-1202.110994	-300.527748	Age
31758.416368	-592.137349	1577.884928	394.471232	Gender
32380.411187	-609.973645	-475.526698	-118.881675	Age · Gender
-209.366974	955.890109	23646.087720	5911.521930	Fitness
-382.770375	621.994819	-17.836297	-4.459074	Age · Fitness
-153.925174	-173.403400	-333.895290	-83.473822	Gender · Fitness
-456.048472	-302.123298	-128.719897	-32.179974	Age · Gender · Fitness

Table 3: Results of the Yates algorithm

6.3 Confidence interval

Before extracting interpretations about the results of Yates algorithm, we must be sure that these results have significance. In order to do that, we have performed a Student's t-test with 9 degrees of freedom. With this, we have obtained the self-confidence intervals of each experiment. We want that these intervals have a confidence of 95%, i.e. $\alpha = 0.05$ and $n = 10$. As you can see in table 4, the desired h is greater than h, i.e. the confidence interval is higher than 5%. Hence, 10 replications are enough to ensure that our results are significant.

Mean	S	h	Desired	Low	High
9988.895918	324.044946	231.791685	499.444796	9757.104234	10220.687603
9779.528944	276.740888	197.954752	488.976447	9581.574193	9977.483696
10553.542673	251.675928	180.025605	527.677134	10373.517068	10733.568278
10170.772298	287.451227	205.615934	508.538615	9965.156364	10376.388233
15956.170771	237.287979	169.733801	797.808539	15786.436969	16125.904572
15802.245597	324.958071	232.444850	790.112280	15569.800747	16034.690447
16418.229829	190.298675	136.122013	820.911491	16282.107816	16554.351843
15962.181358	369.465636	264.281432	798.109068	15697.899926	16226.462789

Table 4: Results of the model for a runner with the minimum value of each factor as data

The formula used to compute h is written below:

$$h = t_{1-\frac{\alpha}{2},n-1} \cdot \frac{S}{\sqrt{n}} = t_{0.975,9} \cdot \frac{S}{\sqrt{n}} = 2.262 \cdot \frac{S}{\sqrt{10}}$$

6.4 Conclusions

Now, we are able to extract interpretations of the results of Yates algorithm (see table 3).

- The *Age* affects positively to the performance of a runner, i.e. the older the runner the better the performance. It is a little bit weird, but we suppose that it is caused because we are using extreme values. The best marathon results are from people between 30 and 40 years, so ages that are around the middle are most likely to obtain good times.
- The *Gender* affects negatively to the performance of a runner, i.e. women obtain better results than men. This can be caused because we have more samples of men than of women, and because a woman tends to participate in a marathon when she is prepared for it.
- The *Fitness* affects negatively to the performance of a runner, i.e. the greater the fitness, the worst the result. It is logic, and it matches with our expectations with this factor. It is the factor that explains more variability in the results, by far.
- All the combinations of the factors affects positively to the performance, but with not so much impact.

7 Model validation

To validate the model we have partitioned the data in two parts (train and test), with one part of 99.5% of the runners and a 0.5% with the other part. With the test part we tried to predict and compare it with the real data. Using this method we got the

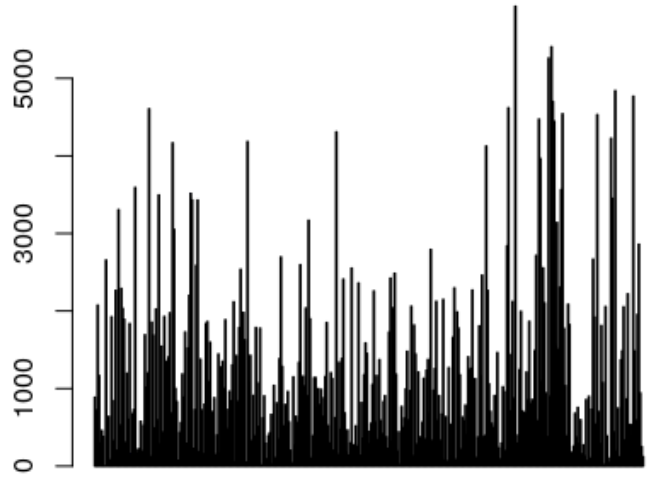

Figure 2: Barplot with the deviation between the predicted and the real values

differences between the predicted and the real times of the test dataset.

In the figure 2 we can see that in most of the cases the error is under 1000 seconds (around 15 minutes) which is quite good. What is not that good is that we have some of the results even over 3000 seconds (around 50 minutes) which is quite a huge number compared even with the total final value.

Since we only have used three factors to predict the final time, is normal that the results are not very accurate. With that in mind we could say that we have quite accurate results since in many cases the error is not very high. If we wanted better results the only way would be to increase the number of factors.

8 Results/Conclusions

On the one hand, we have realised that our model is not as good as we would like. The factors *Age* and *Gender* have weird behaviours, as we described in section 6. During the validation process, we have observed that in some cases the model can obtain results that differs 50min with the reality, which is really bad. We think that it can be possible because the unbalance of our data set (we have more data about men than women, about middle-aged than young and old people, etc). Furthermore, running a marathon depends on lots of things, as we described in sections 2 and 3. Using only 3 factors must have affected in the behaviour of the model.

On the other hand, not all things are bad. We create a very good factor, which it is known before starting the marathon. It is the factor that explains more variability in our results. Furthermore, our model obtains results with an error around 15min in most of the cases, so it is not as bad as we thought before validating the model.

At the start of this assignment we gather factors like the country of the runner, terrain elevation changes, temperature, humidity and whether (you will see them within the code). They were prepared to be used, but finally we decide to not use them to simplify the experimentation. Now, we could use them in order to improve the resulting model.

To finish, we are proud to say that we were able to create the model, develop an experimental factorial design to extract conclusions and, finally, validate it.