BootstrapExampleReport

November 9, 2017

0.1 Assignment 1

I provided the following experiments with DS-1 dataset:

1. Estimated mean, median of y_i and found error for my estimation using 100 bootstrap samples. The results are:

```
Mean of the original sample: [ 4.03  5.51  7.31  1.53  5.63]
Mean confidence interval:
From:[3.19, 4.72, 6.37, 1.3, 4.84]
To:[4.78, 6.67, 8.21, 1.71, 6.46]
Median of the original sample: [ 4.    6.25  8.21  1.86  6.22]
Median confidence interval:
From:[2.74, 5.24, 7.99, 1.64, 5.34]
To:[5.64, 7.5, 8.93, 2.0, 7.28]
```

All is ok, every parameter fits to its confidence interval, as expected in 95% cases.

2. Estimated β parameter of linear regression model for predicting each of the y_i variables and found their confidence intervals using 100 bootstrap samples. Concretely linear regression model is the following:

$$y = \beta^T x + \epsilon, \text{ where } \beta \in \mathbb{R}^{n+1}, \beta_{n+1} \text{ is intercept, so } x_{n+1} = 1$$

The results for y_0 for β_i, where $i \in [1, 7]$, are:

```
beta for y0 for b from 1 to 7 of the original sample:
[ 0.35 -0.09 -0.47  1.12 -0.33  0.92 -1.07]
beta for y0 for b from 1 to 7 confidence interval:
From:[-0.19, -0.55, -1.1, -0.03, -0.95, -0.46, -1.28]
To:[1.17, 0.51, 0.51, 1.25, 0.3, 1.3, 0.27]
```

Here also everything is fine. Each parameter fits to its 95% bootstrap confidence interval.

One note on how I do bootstrapping: I select rows from the initial dataset based on randomly uniformly distributed indexes.

0.2 Assignment 2

For estimating parameters on dataset DS-2 I took all three S-shaped models mentioned in http://www.hpl.hp.com/techreports/tandem/TR-96.1.pdf

1. G-O S-shaped model:

$$bugs = a(1-(1+bt)e^{-bt}), \text{ where } a \geq 0, b > 0$$

2. Gompertz S-shaped model:

$$bugs = a(b^{c^t}), \text{ where } a \geq 0, 0 \leq b \leq 1, 0 < c < 1$$

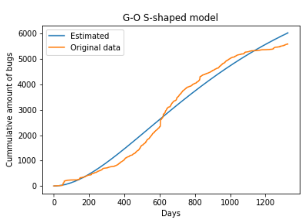
3. Yamada Raleigh S-shaped model:

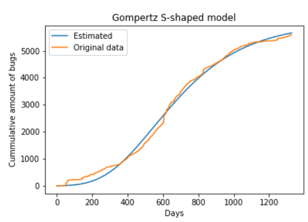
$$bugs = a(1-e^{-r\alpha(1-e^{-\beta t^2/2})}), \text{ where } a \geq 0, r\alpha > 0, \beta > 0$$

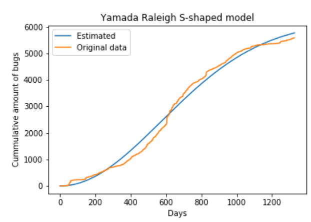
For estimating parameters of the models I took maximum likelihood estimator. If we assume that our noise is Gaussian, we can define MLE task as sum of squares. Here is an example for the G-O model:

$$a, b = argmax_{a,b} L(a, b, \sigma) = argmax_{a,b} log(L(a, b)) = -\sum_{i=1}^{n}(y_i - a(1-(1+bt_i)e^{-bt_i}))^2$$

Let's try to fit all the models and compare results visually:







In order to estimate parameters of the models I calculated them based on the original sample of bugs in a software system and calculated 95% confidence interval based on 100 bootstrap samples. The results are the following:

1. For G-O model:

```
a, b from initial sample: [8541.1262, 0.0019]
Bootstrap confidence interval for a, b:
From: [5729.6659, 0.0018]
To: [8230.305, 0.0027]
```

   Here is a strange thing that needs to be discussed on the seminars - all the parameters from initial sample slightly don't fit to the 95% bootstrap confidence intervals.

2. For Gompertz model:

```
a, b, c from initial sample: [6031.5207, 0.0008, 0.9964]
Bootstrap confidence interval for a, b, c:
From: [6232.21, 0.0295, 0.9978]
To: [9433.5938, 0.0616, 0.9987]
```

Here is the same thing.

3. For Yamada Raleigh model:

```
a, r, alpha, beta from initial sample: [281352680.9316, 0.0659, 0.0003, 0.0]
Bootstrap confidence interval for a, r, alpha, beta:
From: [29214103.7872, 0.0058, 0.0, 0.0]
To: [1104614364.766, 0.4071, 0.0011, 0.0]
```

Here everything is ok, all parameters fit to their bootstrap confidence intervals.