Assignment 1 Numerical Methods for EOR (EBB115A05)

due: 6 Oct 2018, 1800hrs

- This assignment is to be completed in the groups that you are assigned to. No changes are permitted at this time.
- Please provide the following information on the first page of your solutions: group number as stated on Student Portal and the names and student numbers of members of the group.
- Submit your solutions through Student Portal. Please call the file that you submit `group1.pdf`, `group2.pdf` etc, so that we are not stuck with 25 files each called `assignment.pdf`
- You need to type your answers and include R code in your answers.
- You need to show your work. Simply providing R code without any explanation will not result in full marks even if the code and answer are correct. Explain your reasoning clearly.
- Read all the questions carefully and email us if something is not clear.
- Answer all 3 questions.
- Late submissions will not be graded.

GOOD LUCK!

1. Consider the dataset with tennis data that you created in the first problem set. **To make sure though that every group uses the same data, use the dataset as provided with this assignment.** Players are ranked on a world ranking, that information is in the variables `ARank` and `BRank`. To assess how higher ranked players have a larger probability to win a match against a lower ranked opponent, you could estimate a logit model

$$\Pr(A \text{ wins against } B) = \frac{1}{1 + \exp(-\beta_0 - \beta_1 \log R_A/R_B)}$$

where R_A is the rank of player A and R_B is the rank of player B.

   (a) Suppose both players are equally good (ie, have the same rank), what is the prediction of the model? What estimate do you expect for β_0? Do you expect a positive or negative sign for β_1?

   (b) What is the contribution of one match to the loglikelihood function?

   (c) What is the loglikelihood function? Program the loglikelihood function. The first line should be

```
loglik <- function(p,z){
```

   (d) Estimate β_0 and β_1 using your data. Give 95%-confidence intervals for both parameters.

2. Use the dataset of the previous problem. Winning probabilities are inversely related to payout: a high payout means that the winning probability is small. In fact, if the bets were fair (ie, if they would have expected value 0 to both the punter and the betting agency), we would have

$$\Pr(\text{A wins}) \approx \frac{1}{\text{B365A}},$$

and

$$\Pr(\text{B wins}) \approx \frac{1}{\text{B365B}}.$$

Since one of the players has to win the match, these probabilities should add up to 1. They don't, that is due to the take of the bookmaker. Proper implied probabilities are

$$\Pr(\text{A wins}) \approx \frac{\frac{1}{\text{B365A}}}{\frac{1}{\text{B365A}} + \frac{1}{\text{B365B}}},$$

and

$$\Pr(\text{B wins}) \approx \frac{\frac{1}{\text{B365B}}}{\frac{1}{\text{B365A}} + \frac{1}{\text{B365B}}}.$$

(see for example Koning (2011)). These two estimates add up to 1 by construction. Use `group_by`, `summarise`, and `cut_width` to create ten groups, with implied winning probabilities in $[0, 0.10]$, $(0.10, 0.20]$, etc. (all three functions are part of the `tidyverse` library). Within each group, calculate the average implied win probability of player A, and the observed win frequency of player A. Make a graph of these ten datapoints, with (horizontal) the average implied win probabilities, and (vertical) the actual win frequency. Also, draw the 45-degree line in this graph. How do you interpret your results?

3. **(Optimization)** In practice, sometimes data are presented in the form of a table with (relative) frequencies. For example, in Klugman et al. (1998), a table with individual payments on a general liability insurance is presented. The data are from a specific region. This table is reproduced as table 1.

Table 1: General liability payments, table 2.10 from Klugman, Panjer, and Willmot (1998).

table2.10	lower	upper	frequency	average
1	0	2.5	41	1.389
2	2.5	7.5	48	4.661
3	7.5	12.5	24	9.991
4	12.5	17.5	18	15.482
5	17.5	22.5	15	20.232
6	22.5	32.5	14	26.616
7	32.5	47.5	16	40.278
8	47.5	67.5	12	56.414
9	67.5	87.5	6	74.985
10	87.5	125	11	106.851
11	125	225	5	184.735
12	225	300	4	264.025
13	$\geq$ 300		3	

Table 2: Additional general liability payments, table 2.12 from Klugman, Panjer, and Willmot (1998).

table2.12	lower	upper	frequency
1	0	2.5	101
2	2.5	7.5	132
3	7.5	12.5	61
4	12.5	17.5	50
5	17.5	22.5	29
6	22.5	32.5	50
7	32.5	47.5	28
8	47.5	67.5	40
9	67.5	87.5	22
10	87.5	125	27
11	125	175	19
12	175	225	6
13	225	325	10
14	325	500	4
15	$\geq$ 500		5

(a) Fit a Gamma distribution to the data in table 1 by optimizing a log-likelihoodfunction.

(b) Suppose now a second sample is available, say, from another region. The classes of this second sample are different. The additional data are listed in table 2. Test the hypothesis that the parameters of the Gamma distribution are equal between both regions. (hint: section 4.3 in Azzalini (1996)).

(c) Provide a point estimate for the expected payment, and a 95% confidence interval for that parameter.

References

Azzalini, A. (1996). *Statistical Inference*. London: Chapman & Hall.

Klugman, S.A., H.H. Panjer, and G.E. Willmot (1998). *Loss Models*. New York: John Wiley & Sons.

Koning, R.H. (2011). Home advantage in professional tennis. *Journal of Sports Sciences 29*(1), 19–27.