

Assignment 4

Instructions

- This assignment is due on Saturday 28th November 2020 at 11:59pm.
- You should submit it to the ‘Assignment 4’ assignment object in [redacted]
- You should submit two files only:
 1. Rmd file detailing the commented code you used to obtain your answers.
 2. final document in either pdf or Word which should contain answers to the questions below.
 - If you created an HTML file, please convert it to pdf. You can use Google Chrome: **File > Print > Destination [Change...] > select Save as PDF.**
- You may submit it multiple times before the deadline, but only the last version will be marked.
- There is a maximum of 10 marks for this assignment. This assignment is worth 10% of your final grade.
- Late submissions will score 0, unless a “Late Submission of Coursework” form is submitted.
- Assignment 4 is broken up into two tasks: exploring and modelling.
- You may have to discover and learn some new functions. Use `help()` and `help.search()` to find what you need.
- Complete your assignment using R Markdown, check that all the output and code are correctly shown in your final document. Knit your document frequently to fix errors. Once completed, submit the Rmd file and the resulting pdf or word document which shows all your code.

Data

The spreadsheet `CLGoals.xlsx` contains the number of goals scored in each UEFA Champions League game to-date this season (three match weeks of sixteen games). The data are count data that take the values 0, 1, 2,

Modeling

Poisson Model

The Poisson distribution is probably the most standard model for count data.

The Poisson model, with parameter λ , assumes that

$$P\{X = x\} = \frac{\lambda^x e^{-\lambda}}{x!}.$$

Thus, $P\{X = 0\} = \exp(-\lambda)$, $P\{X = 1\} = \lambda \exp(-\lambda)$, $P\{X = 2\} = \lambda^2 \exp(-\lambda)/2$, ...

The expected value (mean) of the Poisson distribution is λ and the variance is also λ (thus, the standard deviation is $\sqrt{\lambda}$).

Hurdle Model

The Hurdle model, with parameters θ and λ , assumes that

$$P\{X = x\} = \begin{cases} \theta & \text{if } x = 0 \\ (1 - \theta) \frac{\lambda^x e^{-\lambda}}{x!(1 - e^{-\lambda})} & \text{if } x = 1, 2, \dots \end{cases}$$

Thus, $P\{X = 0\} = \theta$, $P\{X = 1\} = (1 - \theta)\lambda \exp(-\lambda)/(1 - \exp(-\lambda))$, ...

If $\theta = e^{-\lambda}$ then the Hurdle model is the same as the Poisson model. If $\theta < e^{-\lambda}$, then zeros are less likely than under a Poisson model. If $\theta > e^{-\lambda}$, then zeros are more likely than under a Poisson model.

Likelihood

The likelihood function is defined to be the probability of the observed data for a given parameter value. If we have independent observations x_1, x_2, \dots, x_n , then the likelihood is

$$L = \prod_{i=1}^n P\{X = x_i\}.$$

The log-likelihood is (natural) logarithm of the likelihood, thus it takes the form

$$\ell = \log L = \sum_{i=1}^n \log P\{X = x_i\}.$$

Task 1: Exploring

1. Read the data into R.

Hint: The `read.xlsx()` function in the `openxlsx` R package is useful for doing this.

2. Produce a table that tabulates frequency of each number of goals.
3. Produce a plot of the frequency of each number of goals.
4. Calculate the mean and the standard deviation of the number of goals.

Task 2a: Poisson Modelling

1. Write a function that calculates the log-likelihood function (for a specified value of λ) for the Poisson model for the UEFA Champions League data.
2. Plot the log-likelihood function for a range of values of λ .

Hint: Make sure that $\lambda = \bar{x}$ is in the range.

3. Add a vertical line to the plot at the value \bar{x} and visually verify that this maximizes the log-likelihood function.
4. Simulate 48 values from a Poisson model with $\lambda = \bar{x}$ and summarize the resulting values (contrasting them with the summaries produced in Task 1).
5. Simulate 48 values from a Poisson model for other values of λ and summarize the resulting values (contrasting them with the summaries produced in Task 1).

Task 2b: Hurdle Modelling

1. Create a `dHurdle()` function that has arguments `x`, `param` that computes $P\{X = x\}$ for the Hurdle model, where the first element of the vector `param` is θ and the second element of the vector `param` is λ . Ensure that the function can handle `x` being a vector of values.
2. Write a function that calculates the log-likelihood function (for a specified value of `param`) for the Hurdle model for the UEFA Champions League goal data.
3. Use the `optim` function to find the value of θ and λ that maximizes the log-likelihood.

Hint: `optim` minimizes functions, by default, so you may want to write a function that computes minus the log-likelihood and minimize that.

Alternatively, you can set `control=list(fnscale=-1)` as an argument in `optim` to make it maximize.

4. Comment on the value of θ found and compare the log-likelihood values found for the Poisson and Hurdle models.