ST5226 Assignment

AY22/23 Semester 1

Instructions:

- The submission dateline is 10 Oct at 2359pm. If you are unable to make the dateline please let me know early and give me a valid reason.
- Upload your solutions to Assignment → Assignment Submission in Canvas. Upload a pdf or word file solution set with a txt file containing the codes that you used. I will be reading your pdf file and checking your txt file only if I have doubts on the codes that you use, so do make sure that the full solutions are in your pdf file. If unsure you can always reproduce your codes in both the pdf and txt file. Under Canvas you can submit multiple files within the same submission. The solutions to Question 3 can be handwritten however make sure that I am able to read your handwriting.
- Make sure that you have contributed substantially to the solutions that you have submitted. Penalties will be imposed if the solutions of two students are unreasonably similar.
- This assignment is worth 30 marks. There are 4 questions with a total of 29 parts. Each part is worth 1 mark. The last mark is for following all the instructions.

1. The R object in `abc.rds` contains the number of births (BIR74) and birth defects (SID74) in 100 counties of a state of US for the period July 1, 1974 to June 30, 1978.

   (a) What type of spatial object is `abc`?

   (b) Find the CRS of `abc`.

   (c) What commands in R transform the coordinate system of `abc` to WGS84? Specify the exact code.

   (d) How would you compute the rate of birth defects per 10,000 births over the four year period?

   (e) How about the rate of birth defects per 10,000 births in a typical year during the period?

(f) The entry for SID74 is "NA" for one of the counties. Change it to 0 so that it does not cause problems in the computations below. Take note of the name of this county.

(g) Provide a choropleth map of the raw rates per 10,000 births over the four year period. The map should be simple and visually appealing. Mark the county which you made the change in (f) with the letter A on the county in the map.

(h) Are there counties with very few birth defect counts? What is the problem with displaying the raw rates for such counties? What is the commonly referred to name for this problem?

(i) One way of dealing with the problem in (h) is to group counties so that the groups have roughly equal population sizes. Comment on the advantages and disadvantages of doing so.

(j) The following ratio smoother was discussed in class:

$$\widehat{\xi}_i = \frac{\sum_{j=1}^{100} w_{ij} Y_j}{\sum_{j=1}^{100} w_{ij} n_j},$$

where Y_j is the number of defect cases and n_j the number of births in county j. Consider smoothing weights

$$w_{ij} = \exp(-\frac{d_{ij}^2}{1000}),$$

where d_{ij} is the distance in km between centroids i and j. Assuming that $Y_j \sim \text{Poisson}(\xi n_j)$ for all j, show that

$$\text{Var}(\widehat{\xi}_i) \leq \text{Var}(\frac{Y_i}{n_i}).$$

Remark: I understand there is a tutorial problem on this and solutions were provided. However it is for a different w_{ij}. If there are similar steps involved then you need to say exactly why those steps apply to the w_{ij} here.

(k) Provide a choropleth map of the smoothed rates using the ratio smoother and weights in (j).

(l) Provide a choropleth map using the empirical Bayes estimator.

(m) Discuss some of the differences between the ratio smoother and the empirical Bayes estimator.

2. The R object in `def.rds` is a `ppp` object containing the locations of individuals with a certain health condition. The R object in `dist.rds` is a list contains two objects `d1` and `d2`. These objects are images containing the distances to two different sources of pollution: Source 1 and Source 2 respectively. Of interest is whether these distances are predictors of the intensity function.

   (a) Display a plot of the F function of the point pattern against radius r (careful!). Set the envelope so that 199 datasets are simulated and the envelope corresponds to 95% pointwise confidence intervals. Write the code so that the same envelope is generated each time the code is run.

   (b) Write down the null and alternative hypotheses of the test corresponding to $r = 0.06$ and say whether the null hypothesis is rejected.

   (c) Explain in your own words the differences between the F and G functions.

   (d) Why is a Monte Carlo test required in this setting?

   (e) Assume we are interested in finding envelopes corresponding to simultaneous confidence bands instead of pointwise confidence intervals. Typing `global=TRUE` within the `envelope` command allows you to access this option. By making use of online resources (e.g. R documentation), repeat (a) with this option. Explain in your own words how the Monte Carlo tests are executed when you apply this option. The essential idea here is a one-sided Monte Carlo test based on the test statistic

   $$T_{\text{obs}} = \sup_r |\widehat{F}_{\text{obs}}(r) - F(r)|,$$

   where $F(r)$ is the theoretical function under CSR.

   (f) Fit a log-linear model of the intensity using covariates `x`, `y`, `d1` and `d2`. Write down the model that has been fitted.

   (g) Perform an ANOVA test, at level $\alpha = 0.05$, of the model fitted in (f) with a model using only `x` and `y`. Compare the AIC of the two models and provide suitable conclusions.

   (h) Is it true that if models A and B are such that the covariates of B are a subset of A, the AIC of model A is always smaller than that of model B? Why?

   (i) Perform a Monte Carlo quadrat test for the log-linear model with covariates `x` and `y`, with 4 × 4 quadrats. Intepret the output of the test.

   (j) The `spatstat` package allows you to apply the `envelope` function on `Kinhom`. You can use the same code as for the F and G functions with `def` as the `ppp` object and `Kinhom` as the function. Plot the Kinhom function of `def` together with an envelope corresponding to 95% pointwise confidence intervals with 199 simulated datasets. Comment on the plot.

3. Consider the log-linear model of the intensity function

$$\log \lambda(u) = \beta_0 + \sum_{j=1}^{p} \beta_j z_j(u),$$

where $z_1(u), \ldots, z_p(u)$ are the covariates at location u and $\beta_0, \ldots, \beta_p$ are the unknown parameters. In Topic 5 Slide 48 we stated that given a point pattern $u_1, \ldots, u_n$ the maximum likelihood estimates $\widehat{\beta}_j$ maximizes

$$L(\beta) = \sum_{i=1}^{n} \log \lambda(u_i) - \int_A \lambda(u) du.$$

Expand on the key ideas below which justifies $L(\beta)$ using properties of the inhomogeneous Poisson point process.

(a) The number of events N in A follows a distribution with mean $C_\lambda = \int_A \lambda(u)du$. What distribution is this? Write out the expression for $P(N = n)$.

(b) The location of each event has density $f(u)$ proportional to the intensity $\lambda(u)$. Find the exact expression for $f(u)$ in terms of $\lambda(u)$. Remember that a density function must integrate to 1. That is $\int_A f(u)du = 1$.

(c) Show that finding β maximizing $L(\beta)$ is the same as finding β maximizing the likelihood

$$P(N = n) \prod_{i=1}^{n} f(u_i).$$

4. (a) Describe `bei` and `bei.extra` with the help of the `class` and `str` commands.

   (b) The `as` function is a general function that lets you convert R objects from one class to another. For example

   `jkl <- as(ghi,“SpatialPoints”)`

   creates an object `jkl` of class `SpatialPoints` from an object `ghi` of similar class. Similarly

   `jkl <- as(ghi,“SpatialGridDataFrame”)`

   creates an object of class `SpatialGridDataFrame`. Convert `bei` to `SpatialPoints` and the gradient (slope) in `bei.extra` to `SpatialGridDataFrame`.

   (c) Create a `SpatialPointsDataFrame` with the tree locations in `bei` as the point pattern and the slope at the tree locations as the covariates. Display the first few entries of the `SpatialPointsDataFrame`.