Biostatistics 685/Statistics 560 Nonparametric Statistics Final Project

Xinye Jiang

December 18, 2018

(1) There is an association between 'age' and 'hemoglobin'.

Kendall's tau is one nonparametric measure of association that could deal with the problem that whether there is an association between 'age' and 'hemoglobin'. I prefer this one because kendall's tau is an appropriate nonparametric method that could measure the association between paired data well, as 'age' and 'hemoglobin' were clearly paired observations collected from each person. And kendall's tau is easy to interpret, as it takes advantage of the probability of concordance and discordance.

Kendall's tau's corresponding null hypothesis is that 'age' and 'hemoglobin' are independent, i.e., there is no association between 'age' and 'hemoglobin'.

```
##
##  Kendall's rank correlation tau
##
## data:  allergy$age and allergy$hemoglobin
## z = -2.0576, p-value = 0.03963
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##        tau
## -0.1149615
```

By R function 'cor.test()', we can get a p-value smaller than 0.05 and sample kendall's tau estimate -0.1149615. To derive the exact null distribution and corresponding p-value, use permutation method (hold 'age' fixed, shuffle 'hemoglobin') to make inference.

```
## [1] "p-value using permutation method: 0.04"
```

We can see that p-value using permutation method is smaller than 0.05 . Thus we reject the null hypothesis and conclude that hemoglobin is associated with age by the nonparametric measure of association - kendall's tau.

(2) There is no difference in days among the four treatment groups.

Kruskal-Wallis test would be an appropriate test to address investigators' concern that there may be difference in days among the four treatment groups. This is because we are comparing days among four independent treatment groups and there is no block design regarding it.

Kruskal-Wallis test's corresponding null hypothesis is that there is no difference in days among the four treatment groups.

Using the 'kruskal.test()' function in R, we can get:

```
##
##  Kruskal-Wallis rank sum test
##
## data:  allergy$day by allergy$treatnew
## Kruskal-Wallis chi-squared = 3.7073, df = 3, p-value = 0.2949
```

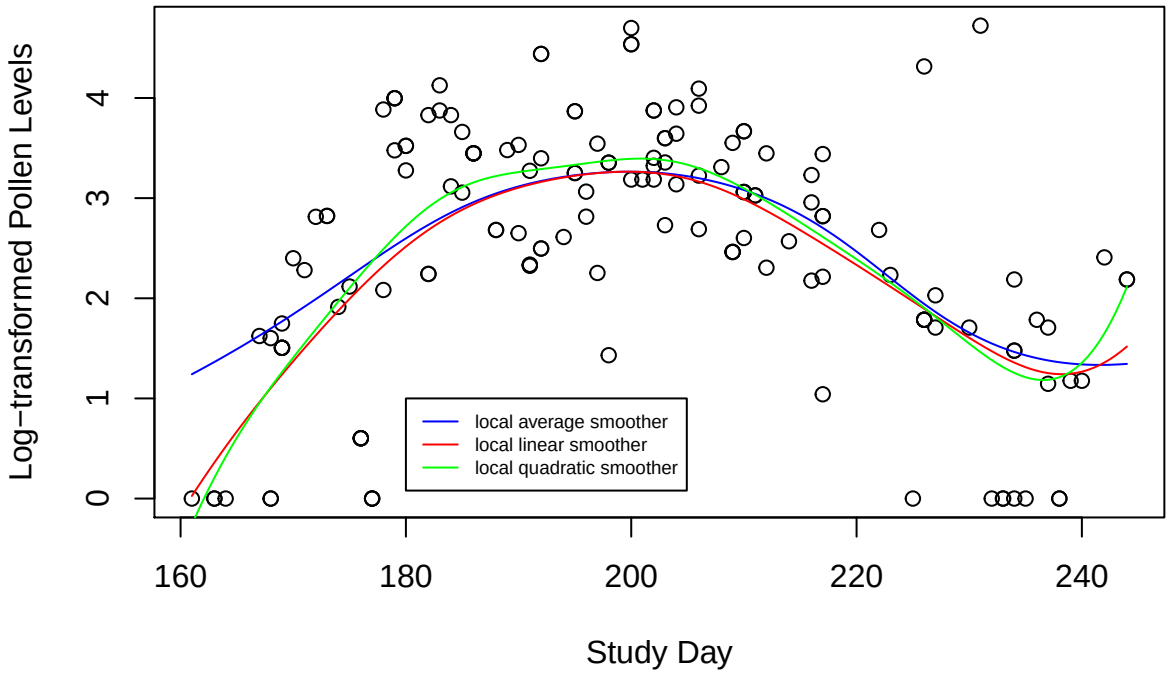
To derive the exact null distribution and corresponding p-value, do Kruskal-Wallis test with permutation distribution (by permuting the group assignment many times) instead of chi-squared approximation to make inference.

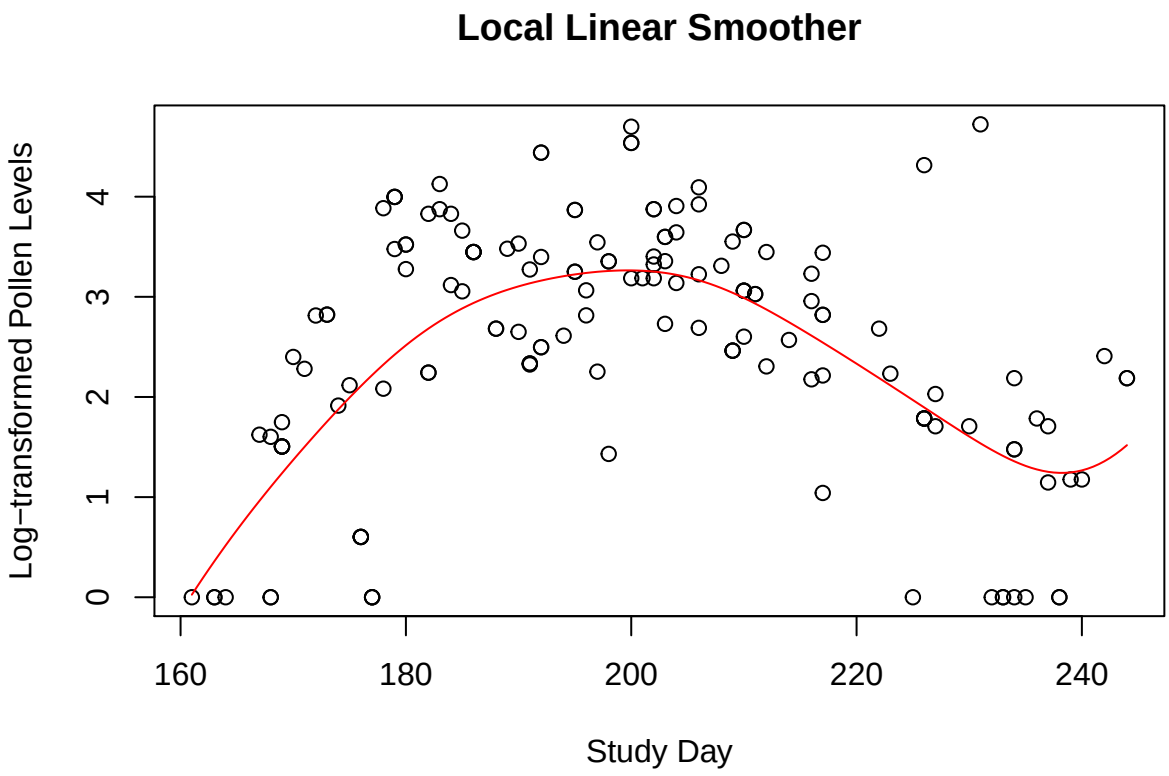
```
## [1] "p-value using permutation method: 0.28"
```

The p-values above from chi-squared approximation and permutation method are both much larger than 0.05. So we fail to reject the null hypothesis and reach the conclusion that there is no difference in days among the four treatment groups, the days are chosen randomly.

(3) The pattern of log-transformed pollen levels over the course of the study:

Make a scatterplot of the study day and log-transformed pollen levels observations, and superimpose local average kernel smoother, local linear kernel smoother and local quadratic kernel smoother onto it. Here use 'lsmooth()' function and set xmin = minimum value of day, xmax = maximum value of day, npoints = 167 and h using the default setting (In order to do question 4, set the x at which to compute fitted values cover all the original x points).



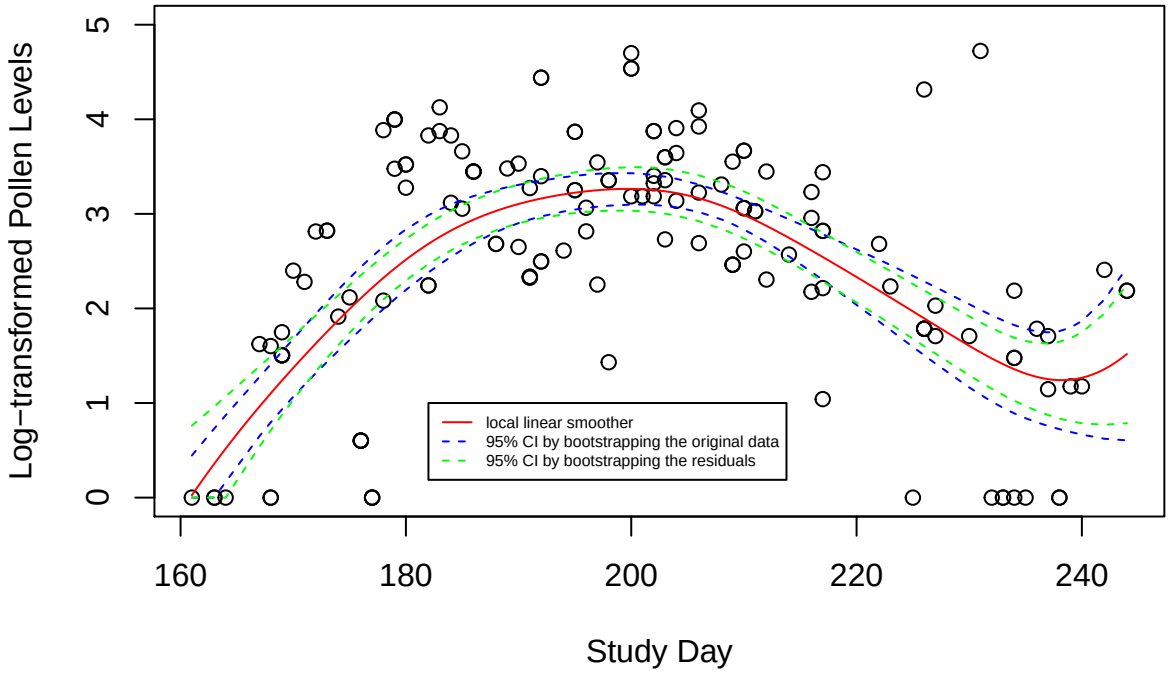


The above plot visually shows the pattern of log-transformed pollen levels over the course of the study and how each local kernel smoother fits the data. The log-transformed pollen levels increase over the first half of the course of the study with the increasing rate slowing down over time, decrease evenly over the last half of the study course and tends to increase at the last phase of the study course.

Local linear kernel smoother describes the pattern well and should be used to produce the final answer. This method fits local linear models within the window which a certain study day is in, and gets the estimate of its corresponding log pollen level by plugging the study day value into the fitted model.

Choosing this method is because we can see from the first plot that local average kernel smoother fails to catch the pattern as it addresses bias at boundaries much more poorly than other two smoothers do. The local quadratic kernel smoother is affected by the data at boundaries too much due to its higher order, and thus seems to have the issue of overfitting. Local kernel smoothers which have order larger than 2, may have the same overfitting problem. The local linear kernel smoother fits the observations and describes the pattern well. So the pattern of log-transformed pollen levels over the course of the study should be local linear kernel smoothing. What's more, I chose this over spline methods because it saves the efforts to choose knots.

(4) Possible 95% confidence bands provided by bootstrapping the original data and bootstrapping the residuals:



The plot above shows a visual summary of the variability of the local linear smoother of log-transformed pollen levels over the course of the study. The blue dashed lines show the 95% confidence band provided by bootstrapping the original data and the green dashed lines show the 95% confidence band provided by bootstrapping the residuals. I prefer the 95% confidence band provided by bootstrapping the original data. This is because bootstrapping the residuals may bring large variance to those areas that originally do not have large variance, which is not a problem by bootstrapping the original data. In this problem, the 95% confidence bands provided by these two bootstrapping methods do not have much difference.

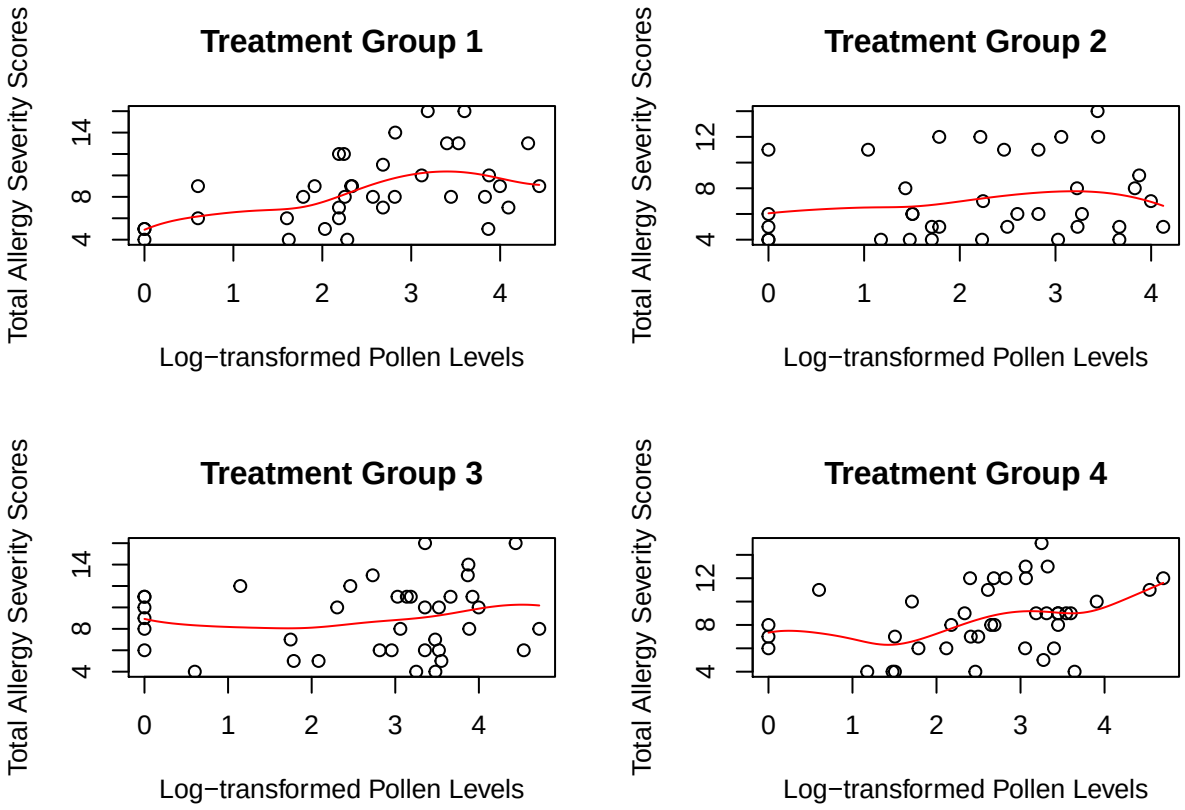
(5) An estimate of the day in the study when peak log-transformed pollen levels were observed with a 95% confidence interval:

Estimated day when peak log-transformed pollen levels were observed is chosen by finding the day corresponding to the maximum fitted log-transformed pollen levels. And use the method of bootstrapping the original data to provide the 95% confidence interval.

```
## An estimate of the day in the study when peak log-transformed pollen levels were observed: 199.5
```

```
## Its 95% confidence interval: ( 194 , 202.5125 )
```

(6) The pattern of total allergy severity scores varied with log-transformed pollen levels for each of the four treatment groups:



The above plot visually shows how the pattern of total allergy severity scores varied with log-transformed pollen levels for each of the four treatment groups. The red imposed lines show fitted local linear kernel smoother.

We can see that there is an obvious linearly increasing trend in total allergy severity scores when log-transformed pollen levels increase in the treatment group 1. And there seems to exist a trend of slow increase in total allergy severity scores when log-transformed pollen levels increase in treatment group 4. No obvious pattern can be seen regarding total allergy severity scores with the varying log-transformed pollen levels in treatment group 2 and 3.

(7) The treatment group 2 and 3 appear to be "effective".

In order to find which, if any, of the four treatment groups appear to be "effective", defined as keeping average total allergy severity scores stable regardless of pollen level, test whether there is an association between pollen levels and total allergy severity scores for each group.

The null hypothesis H0 is that for each group, the expectation of total allergy severity scores given any pollen level is the same for all pollen levels. The test statistic is F = {(RSS0 - RSS1) / (df0 - df1)} / {RSS1 / df1}, where RSS0 and RSS1 are residual sum of squares from the reduced model when H0 is true and the full model fitted by local linear kernel smoothers respectively, and df0 and df1 are corresponding degrees of freedom of the two models. Use 'lsmooth()' function to fit the data by local linear kernel smoother for each group and derive the test statistic. And use permutation methods to get the null distribution for F and corresponding p-value. This method is appropriate because now the model is fitted by local average kernel smoother.

```
## F statistics for the treatment group 1 : 4.907746
## The corresponding p-value: 0.01
##
## F statistics for the treatment group 2 : 0.7356713
## The corresponding p-value: 0.695
##
## F statistics for the treatment group 3 : 0.6540995
## The corresponding p-value: 0.82
##
## F statistics for the treatment group 4 : 2.876617
## The corresponding p-value: 0.035
```

The corresponding p-values for the four treatment groups are shown above. We can see that group 1 has a very small p-value, which indicates the null hypothesis does not hold. For treatment group 1, there is an association between pollen levels and total allergy severity scores, and thus this treatment does not appear to be effective. As group 4 also has a p-value smaller than 0.05, we get the same conclusion as what we get for group 1, the treatment fails to keep average total allergy severity scores stable regardless of pollen levels. For group 2 and 3, we can see p-values much larger than 0.05 . So we fail to reject the null hypothesis and conclude that these two treatment groups appear to be effective and keep average total allergy severity scores stable regardless of pollen levels.