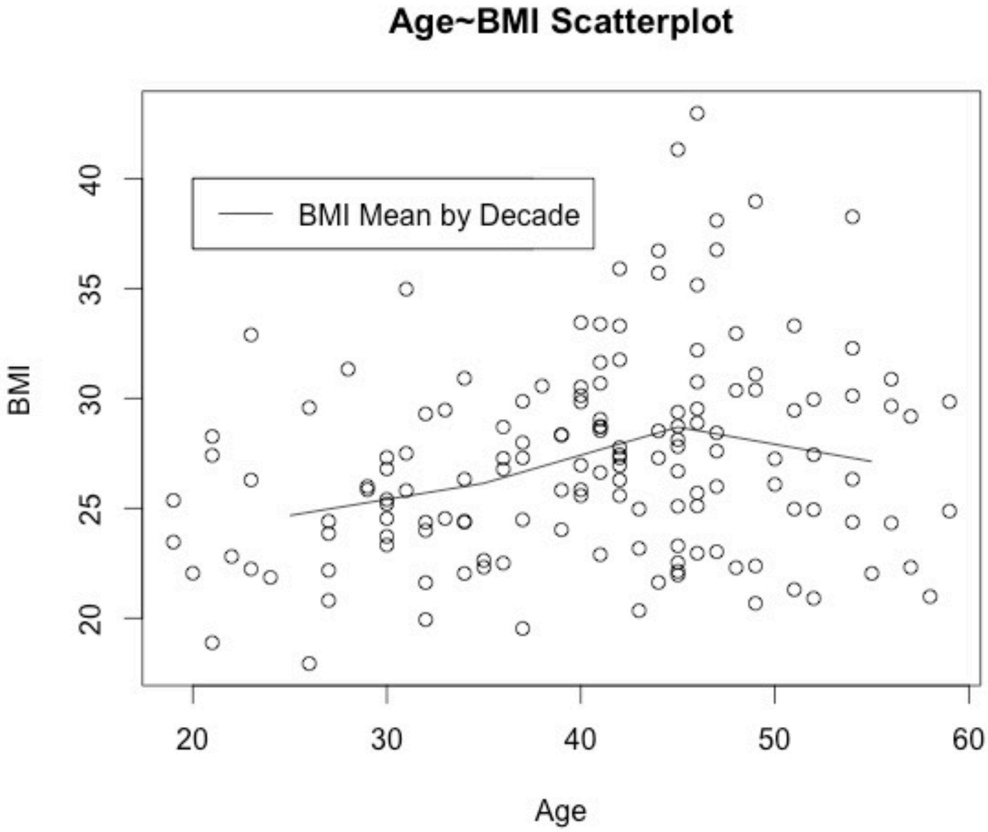
Yu Zhou

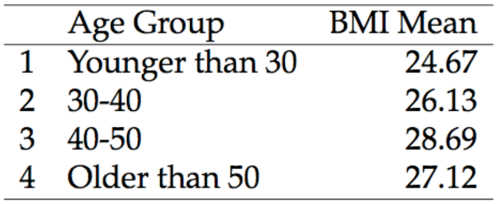
(1) Is there any association between age and bmi? Please report one nonparametric measure of association and explain why you prefer this one. Does the chosen measure give evidence that bmi is associated with age? Please also include any descriptive statistics and plots that you think can help answer this question.

To explore if age and BMI are associated by a nonparametric method, I can consider Spearman’s Rho or Kendall’s Tau, since they both can measure linear associations nonparametrically (scatterplot below hint a linear association). I prefer Kendall’s Tau, because Spearman’s Rho has no direct sample analog of a “population coefficient” when we make inference on Spearman’s Rho. Additionally, compared to Rho, Tau is also more tractable mathematically, particularly when ties are present, and variable age has ties in this dataset. Finally, I find it easier to explain the idea of Tau to a non-statistician versus the idea of Rho.

The sample Kendall’s Tau coefficient is 0.15 and the p value of the null hypothesis that Tau=0 is 0.005; therefore this data suggests BMI is associated with age. Actually, Spearman’s Rho gives me a similar result.



The Age-BMI scatter plot above suggests a moderate positive association between age and BMI. I calculated the mean BMI for grouped subjects as follow:



	Age Group	BMI Mean
1	Younger than 30	24.67
2	30-40	26.13
3	40-50	28.69
4	Older than 50	27.12

The BMI mean in the above table is also superimposed onto the scatterplot, connected by lines. This descriptive statistics suggests an overall increasing linear trend of BMI with age.

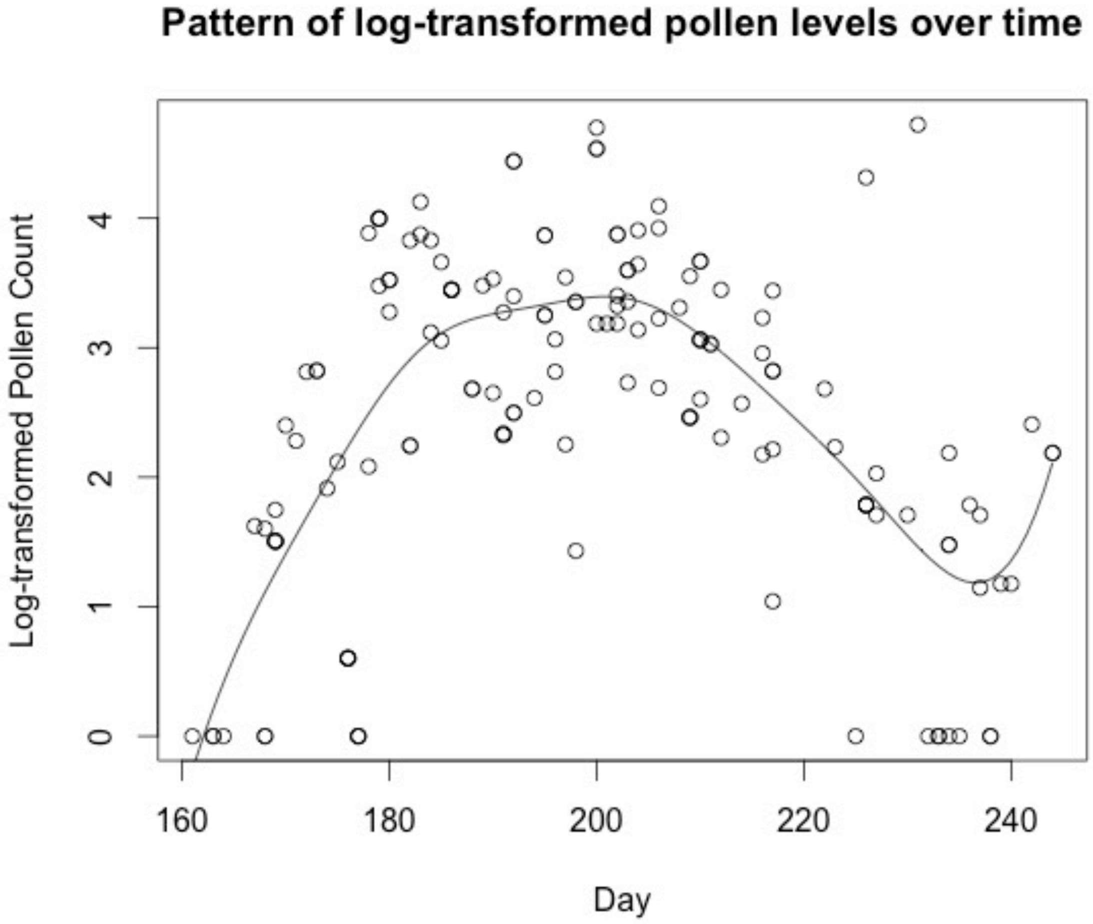
(2) Although days are supposed to be chosen randomly, investigators are concerned that there may be difference in days among the four treatment groups. Please use appropriate test to address investigators concern. Please clearly statement the null and alternative hypotheses.

Intuitively the mean of days among the four treatment groups should be close to each other if days are chosen randomly.
My null hypothesis is that the means of days among the four treatment groups are equal. My alternative hypothesis is that at least one mean of the days among the four treatment groups is different from the others.
Let the mean of days of the four groups be u1,u2,u3 and u4.
H0: u1=u2=u3=u4, Ha: u1, u2, u3 and u4 are not equal.

Kruskal-Wallis test should be used to calculate the p value of H0 nonparmetrically, because Kruskal-Wallis test can conduct multiple group nonparametric comparison. Kruskal-Wallis test gives a p value of 0.29, which is not significant; therefore we fail to reject the null hypothesis that the four means are equal. Investigator should not worry about difference in days among the four treatment groups.

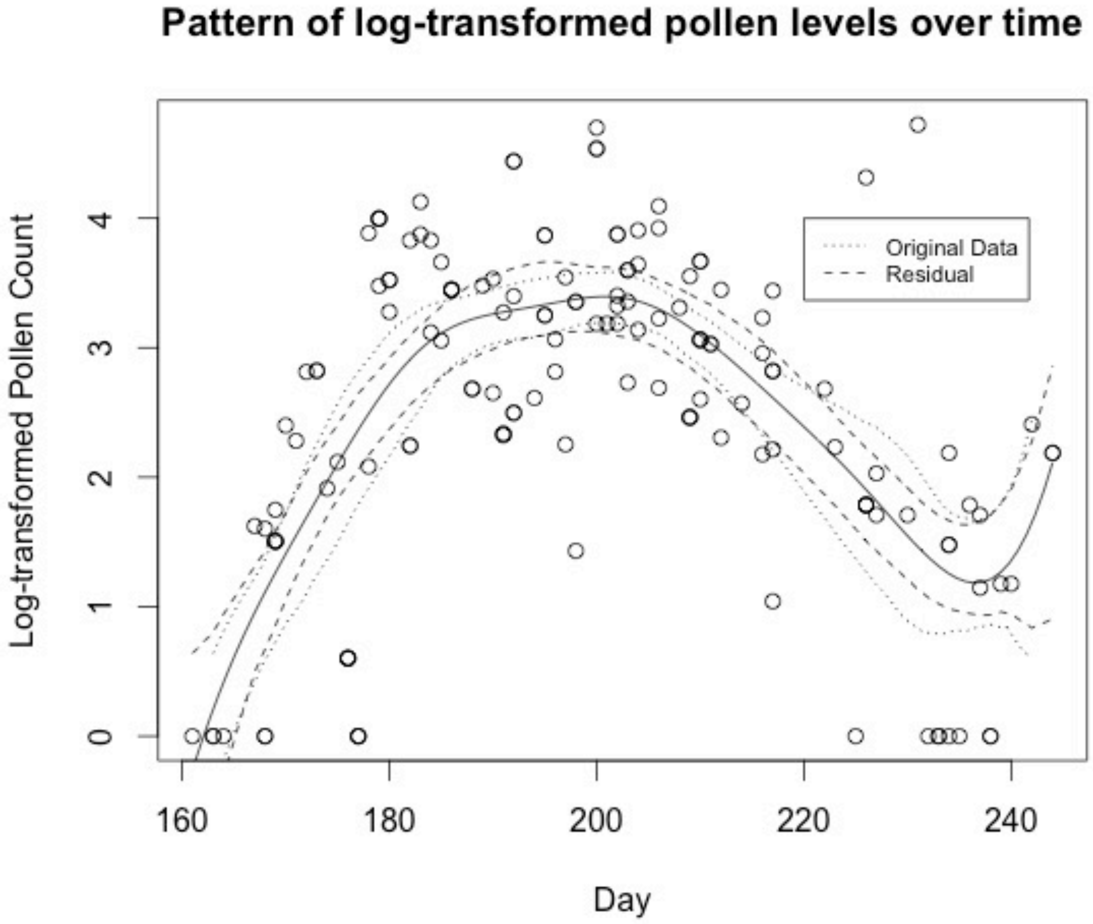
(3) Visually and verbally describe to investigators the pattern of log-transformed pollen levels over the course of the study. Explain what method you used to produce your final answer and why you chose this over all the methods we covered in lecture. You need to superimpose the fitted line to the scatterplot of the data.

The log-transformed pollen level increase until day turns 200, decrease until days turns 240 and increase again a little beyond day 240. The increase at the end of study may or may not reflect truth due to a limited amount of observations. The graph below can visually show the pattern of log-transformed pollen levels over the course of the study.



To describe this pattern nonparametrically, I choose local quadratic normal kernel estimation. I choose local regression because I do not want to assume an overall model, but I want to estimate log-transformed pollen count mostly relying on observations within a narrow local window of days. Normal kernel is chosen for convenience, as choice of kernel does not matter much. I choose to fit quadratic regression locally, because I want to not only capture the linear trend at ends, but also capture curvatures well. Local linear regression and local average should be fine as well. I use lsmooth() function given in class to calculate local quadratic normal kernel estimates and I keep using lsmooth() in this project for local regression.

(4) Investigators would like a visual summary of the variability in your answer to (3) above. Provide two possible 95% confidence bands by: (a) bootstrapping the original data, and (2) bootstrapping the residuals resulting from your answer in (3) above. Do you prefer one confidence band to the other? Defend your answer.



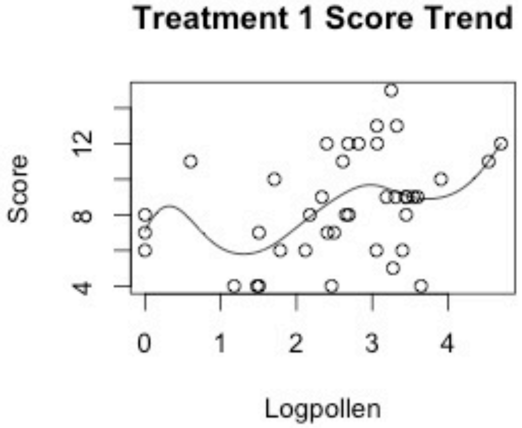
I prefer the band from bootstrapping original data.

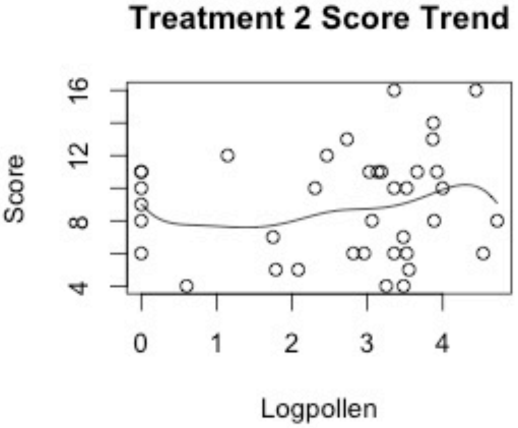
Bootstrapping residual assumes that our model fits the data well, but does not make any assumption about how residuals are distributed. Bootstrapping original data assumes nothing about model or residuals, which is safer.

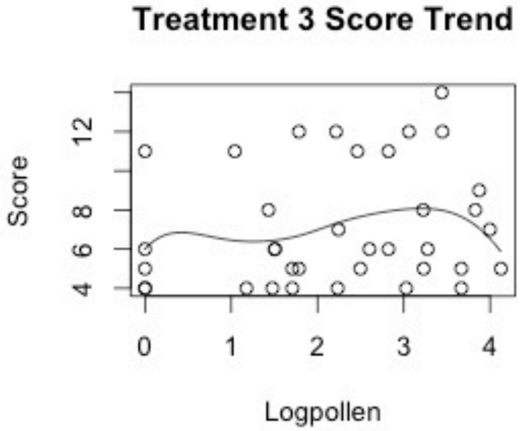
The confidence interval band of bootstrapping from residual is narrower than the band of bootstrapping from original data.

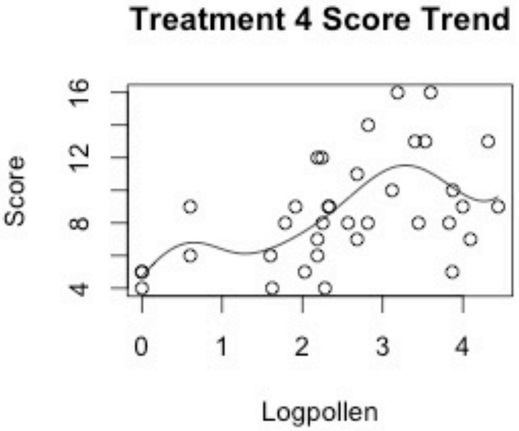
I prefer the band from bootstrapping original data, because I choose to make no assumption about either model or residuals.

(5) A "total allergy severity score" for each subject is computed as the sum of his/her values for the variables itchy, sneezy, runny and stuffy. Visually and verbally describe to investigators the how the pattern of total allergy severity scores varied with log-transformed pollen levels for each of the four treatment groups. Please at least superimpose the fitted line on the plot.







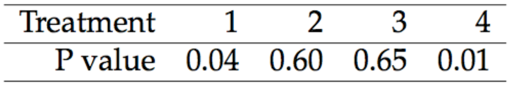


The graphs above are score-logpollen scatterplots for four treatment groups. The fitted lines are local polynomial regression estimates.
For treatment 1 and 4, the allergy severity scores fluctuate but have increasing trends overall. For treatment 2 and 3, the allergy severity scores are flat overall.

(6) 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? Provide statistical evidence for your answer. (note the function lsmooth() will return mse and df as output)

The treatment 2 and treatment 3 appear to be effective.
For each treatment, I want to test the null hypothesis that score and logpollen have no association. My alternative hypothesis is that score and logpollen have some association. H0: E(Y|X)=E(Y)= θ_0 , H1: E(Y|X)= $\theta(x)$
For each treatment, I can get RSS0 and df0 under H0 and I can get RSS1 and df1 under H1, then I calculate F statistics by F=[(RSS0-RSS1)/(df0-df1)]/(RSS1/df1). With calculated F statistics, I can get the p value of null hypothesis. The table below gives the p values of each treatment. Because the p values of treatment 2

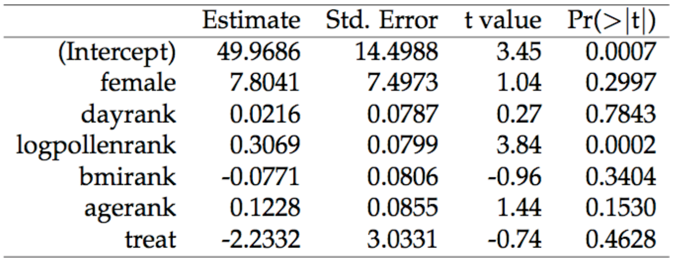
and treatment 3 are not significant, we fail to reject their null hypotheses. We can claim that treatment 2 and treatment 3 seem to be effective.



Treatment	1	2	3	4
P value	0.04	0.60	0.65	0.01

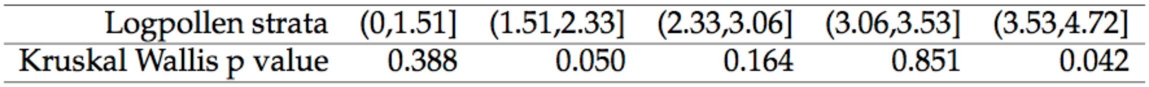
(7) Do female and male differ in their total allergy severity score? Note, several things might be affecting allergy severity score (e.g., treatment, day). Here we are only interested in the difference due to gender, but you may or may not need to account for/consider other factors. Please justify your choice of method.

The covariates available for adjustment include day, logpollen, age, bmi and treatment. If we want to explore the relationship between gender and severity score parametrically adjusting for other covariates, we can consider multiple linear regression, therefore we can rank transform day, logpollen, age, bmi and score, then we can fit the rank transformed data to a multiple linear regression model with ranked score as outcome, and the following as covariates: gender, rank of logpollen, rank of bmi, rank of age, rank of day and treatment. The table below is the modeling result. Because gender(female) is not significant, we can claim that female and male do not differ in their allergy severity score, adjusting for other factors. We notice that rank of logpollen achieves significance.

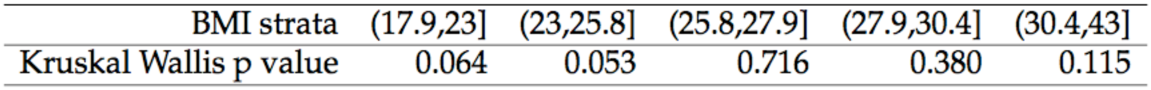


	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	49.9686	14.4988	3.45	0.0007
female	7.8041	7.4973	1.04	0.2997
dayrank	0.0216	0.0787	0.27	0.7843
logpollenrank	0.3069	0.0799	3.84	0.0002
bmirank	-0.0771	0.0806	-0.96	0.3404
agerank	0.1228	0.0855	1.44	0.1530
treat	-2.2332	3.0331	-0.74	0.4628

Alternatively, we can stratify the subjects into different strata and then apply Kruskal Wallis test within each stratum to test if males' scores are different from females' score. By stratification, we can explore the relationship between gender and score, while adjusting for one covariate. The covariates logpollen, bmi and day are broken into 5 equal sized strata respectively (5 is arbitrarily chosen so that the number of strata and the size of each strata is not too small), while covariates age is broken into 3 equal sized strata (3 is chosen instead of 5 because each strata should have males and females). Tables below gives p values of Kruskal Wallis test applied to each stratum to test the relationship between gender and score.



Logpollen strata	(0,1.51]	(1.51,2.33]	(2.33,3.06]	(3.06,3.53]	(3.53,4.72]
Kruskal Wallis p value	0.388	0.050	0.164	0.851	0.042



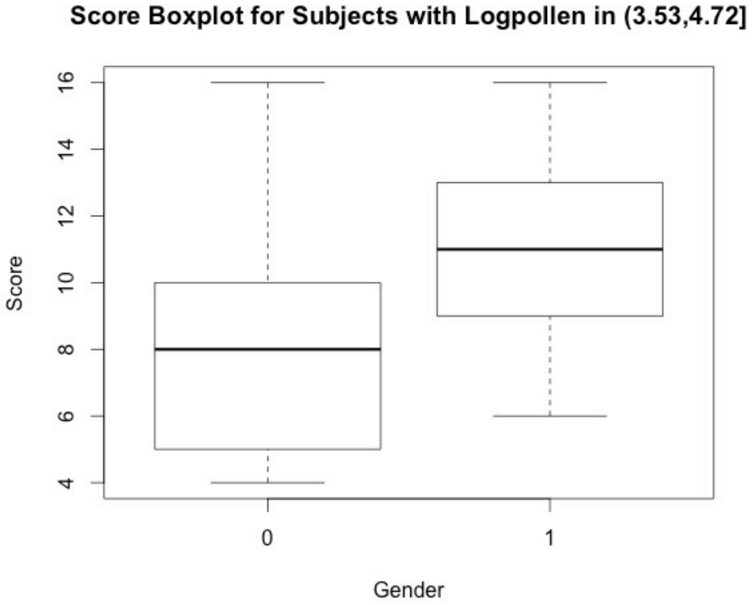
BMI strata	(17.9,23]	(23,25.8]	(25.8,27.9]	(27.9,30.4]	(30.4,43]
Kruskal Wallis p value	0.064	0.053	0.716	0.380	0.115

Age strata	(19,37]	(37,45]	(45,59]
Kruskal Wallis p value	0.251	0.506	0.322

Day strata	(161,179]	(179,192]	(192,204]	(204,218]	(218,244]
Kruskal Wallis p value	0.766	0.527	0.890	0.766	0.070

Treatment	1	2	3	4
Kruskal Wallis p value	0.853	0.427	0.476	0.398

Males and females do not differ in their total allergy severity score mostly, except when they have logpollen between 3.53 and 4.72, which corresponds to the fact in the rank transformation multiple regression result above that gender is not significant and logpollen is significant. This boxplot provides a visual evidence of difference in score between males and females when they have logpollen in (3.53, 4.72].



When I test the relationship between gender and score without adjusting any other covariates by Kruskal Wallis test, I get a p value of 0.14, therefore I want to conclude that female and male do not differ in their allergy severity score, although they seem to have a difference in score when logpollen is very high.