Detecting Routes whose Average Time to Travel Changes Using Capital Bikeshare Data Set

1 Methods: Two-step Procedure

The following is the procedure we decided to implement to answer the question:

1. Cluster the data by route, and run a linear model on confounders only, i.e. Days_since_Jan1_2010 is not part of the model.
2. Within each cluster, use the residuals for permutation test, but group the residuals by day and permute the groups instead of the individual residuals, in order to account for the dependency of data belonging to the same day.

1.1 First Step: Regressing out Global Trends Using Linear Regression

The model we have decided to use for our linear regression is:

```
Duration = (Distance) + Season + Weekday + Member_type +
           Rush_hour + (Distance:Season) + (Distance:Member) + Member:Rush_hour
```

where the variables in parentheses are assimilated into the intercept term since the way in which we cluster our data already controls for the distance as the distance is the same for each route.

1.2 Second Step: Group Permutation Test

We decided to utilize a permutation test that retains the structure of the dependency among data points within the same day, in order to determine whether the observations behave in a way that would suggest that there has been a change in the duration it takes to travel the route with respect to the time passed. The permutation test that we use, instead of permuting all data points randomly, permutes the days at random. Therefore, all the data points that belong to a certain day will be permuted to a different day as a group. In order to run such permutation, first we have to make sure that all the other variables that do not pertain to the day-to-day changes (weather) would have to be accounted for, hence the linear regression that is done initially, which would regress out those confounders before running the permutation test. One example of such variable is the seasonal effects. Because there are differing effects on the other covariates and the duration across different seasons, we would need to account for those differing effects before permuting the days at random, or else a permutation that moves a day in January to a day in June would not be valid. In a given route, if we input the individual data points, then the residuals would contain both the day-to-day variance and the noise. Now, if we order the residuals by the date, then under the null (there isn't any change in duration over time), the residuals will not show a particular trend, as weather is independent of the dates (and the seasonal effects have already been accounted for). However, if the null hypothesis is false, and there in fact is a certain trend in duration over time, then the residuals will show a certain trend (decreasing, increasing, etc.). Now, we permute the days as a group. A simple example of such permutation would be as follows:

Original Data	Permuted Data
(Day1, 0.1)	(Day2, 0.1)
(Day1, 0.3)	(Day2, 0.3)
(Day2, -0.5)	(Day1, -0.5)
(Day3, 0.7)	(Day3, 0.7)

where if the permutation of the days changes (1, 2, 3) to (2, 1, 3), then now the resulting permuted data would be of the form of the right column. Under the null hypothesis, there would not be a clear distinction between the original data and the permuted data in terms of the correlation, and therefore our p-value would be higher. However, if there is a certain trend in the original data, then permuting the days would remove this trend, and therefore the permuted data would not have as strong of a correlation in general than our original data, and therefore our p-value would be lower. Our way of permuting the groups, which maintains the correlation structure within each group, is similar to Scheme B of our first assignment, which also preserves the correlation structure among the genes of the same individual. We ran 100 permutations, which is a choice that took into account the computation time.

2 Data

2.1 Definition of Route

We define a route as a set {Start_station, End_station}, i.e. the order does not matter: routes are considered to be identical if they share the same pair of locations, regardless of which is defined as a starting station or the end station. Such routes serve an integral role in our report because we cluster our entire data set by route. We exclude routes that have the same start and end stations, as such data conflict with our way of estimating the distance, which is a covariate in our null model and will be discussed further in a later section.

Potential Issues and Questions Posed

Assuming that a trip from A-B is the same as a trip from B-A may be too much of an approximation, but we believe that this estimation would not be too problematic given that Washington DC is generally flat. Furthermore, by collapsing the two routes in opposite directions together, we have more observations for each cluster.

2.2 Data Collected

We collected additional data on rush hour, on season, and on distance. The rush hour and season are specific to the period 2010-2011 and to Washington D.C. We will explain each data below.

- Season
  Using an article on the weather of Washington DC that we found online as a guide (which can be found in the Citations section), we grouped the 12 months into 6 levels: 12 1 / 2 3 / 4 5 / 6 7 / 8 9 / 10 11. As an example, the months June and July had significantly higher temperatures and multiple heat waves, and the months August and September had high precipitation and rainstorms. We also assumed that seasonal effects do not vary much between years, so we grouped the months from 2010 with the months from 2011: for example, September 2010 and September 2011 have the same level. We thought that grouping Season this way would capture the seasonal trends adequately without sacrificing too many degrees of freedom.

- Rush_hour
  We used Google to find that rush hours in Washington are Monday to Friday from 6 until 9:30am and from 3:30 until 6:30pm. This categorical variable has two levels.

- Distance
  We compute the distance using R's `distHaversine` function, which takes two pairs of coordinates as inputs. We therefore found the coordinates of each station, and we organized that data in a textfile called `data.txt`. We included a portion of it below as an example.

```
"M St & New Jersey Ave SE" "38 52 35N, 77 00 15W"
"1st & N St  SE" "38 52 29N, 77 00 22W"
"5th & K St NW" "38 54 09N, 77 01 09W"
"19th St & Pennsylvania Ave NW" "38 54 01N, 77 02 37W"
"7th & T St NW" "38 54 56N, 77 01 19W"
"10th & U St NW" "38 55 01N, 77 01 34W"
"Minnesota Ave Metro/DOES" "38 53 57N, 76 56 49W"
"4th & W St NE" "38 55 09N, 77 00 03W"
"1st & M St NE" "38 54 20N, 77 00 22W"
"Park Rd & Holmead Pl NW" "38 55 51N, 77 01 52W"
```

  We learned that this is a "deg, mins, sec" form of representing the coordinates, where 1 degree corresponds to 1 hour. However, since the `distHaversine` function takes the "deg" form of coordinates, we made the necessary conversions.

2.3 Data Table

We organize all of our data in a data table in R, which we thought is better than a data frame structure due to efficiency, considering that our data set is very big. Our data table contains the following information.

- Duration

- Start_date
- End_date
- Start_station_number
- Start_station
- End_station_number
- End_station
- Member type (categorical with 2 levels)
- Days_since_Jan1_2010 (ordinal categorical / continuous)
- Weekday (categorical with 2 levels)
- Rush_hour (categorical with 2 levels)
- Distance (continuous)
- Season (categorical with 6 levels)

The raw data table we have that we use for preprocessing data is called `dtBike`, and the final data table that we use for running diagnostic for our null model is called `dtAggRoutes`. Since our goal is to apply our method to each cluster, whereby the data set is clustered by routes, we create a list of data tables, with each data table corresponding to a specific route. The specific way of generating this list is detailed out in the code. The list of data tables is called `dtListRoutes`.

2.4 Summary Statistics

Some of the data are not used in our method for several reasons, which we discuss here. The table below summarizes the various statistics.

Total Number of Observations	1342364
Number of Routes	6780
Number of Routes Excluding those of the form A_A	6637
Number of Routes Satisfying Conditions	2638

where the form A_A refers to the routes that have the same starting and end points. The conditions in the fourth row refer to the conditions imposed by a linear regression model, which is part of our method. The two conditions are:

- The number of data points should be at least 9, which is the degrees of freedom
- There should be at least one observation for each level of every covariate

Due to the second condition, each data table contains number of observations that are way greater than 9.

3 Hypotheses tested and Multiple Testing Procedure

In this section, we will discuss how we carry out our multiple testing procedure, which, as discussed in class, has to do with combining information of a set of p-values in a valid way, where we assume that each p-value is valid. The p-values we use are the ones we get after implementing our method as described in teh first section.

Since we are testing for each cluster, each being a route, and then selecting which routes to reject based on the p-values, we need to account for the fact that we are doing multiple tests. Therefore, we implement a multiple testing procedure called Benjamini-Hodchberg (BH) procedure. We chose the BH method instead of Holm Bonferroni (HB) method since HB method would be too conservative for data that are correlated. We believe that routes are correlated as some routes might be a subset of a different route, and routes starting from the same station would be dependent. Moreover, even though BH assumes independence of the data, as the professor mentioned in class, dependence does not pose a serious problem in a realistic setting as it has a False Discovery Rate (FDR) that scales as $\alpha \times log(n)$ in the worse case. However, this worst case is not likely to be achieved in a real-world data, and therefore utilizing the BH, even though it assumes independence among data, would be able to keep the FDR low. Because of the dependency among the routes, we believe that the BH method is still too conservative, and in the end we determined a cutoff threshold of alpha = 0.3 to increase the power in exchange for an increased fdr.

Based on the summary statistics of the section above, the number of tests we run as part of our multiple testing procedure is 2638, which is the number we use for when setting the threshold in the BH method. It may be a problem that the number of tests to run is decided in a data dependent way, but it is inevitable in this case as it is a direct result of the constraints imposed by the method we implement.

4 Model

4.1 Null Model

Our assumption is that the full data follows the following linear model under the null hypothesis, which assumes that there is no change in the time taken to travel a route over time:

```
Duration = Distance + Season + Weekday + Member_type +
           Rush_hour + Distance:Season + Distance:Member_type + Member_type:Rush_hour +
           {noise on a given day(day-to-day variance)} + noise
```

Trying our best to be conscious of the problems of data-dependency, we decided on this model prior to looking at the data. We justify our choice of each confounder below.

- Season
  We have included seasonal effects as a covariate, in order to control for seasonal effects that affect the duration of a trip, such as snow in winter causing trips to be longer. Originally, we intended to include all 12 months as levels but decided not to due to the degrees of freedom issues that it might cause, as we have a limited sample size after grouping by routes.

- Weekday
  Similarly to seasonal effects, we also included a binary categorical variable indicating whether the ride took place on a weekday or on the weekend, as we believe that this could have an effect on the duration (i.e. trips on weekends could be longer). Overall, controlling for these 2 variables along with sub-grouping residuals by days will allow us to permute freely across Days_since_Jun1_2010.

- Interaction terms
  For interactions, we included only those we believed to have the strongest association. We included Distance:Season because it is reasonable to expect that the seasonal effects would have an effect on the distance travelled. We included Distance:Member_type because we believe that members are more willing to take trips of shorter distances than non-members, as they have a fixed annual fee, whereas non-members would be more likely to travel longer distances, since they would want to make the most out of the one time fee. Finally, we included member:rush_hour thinking that there would be more members riding during rush hours, as the people who use bikes to commute during rush hours everyday are more likely to be members.

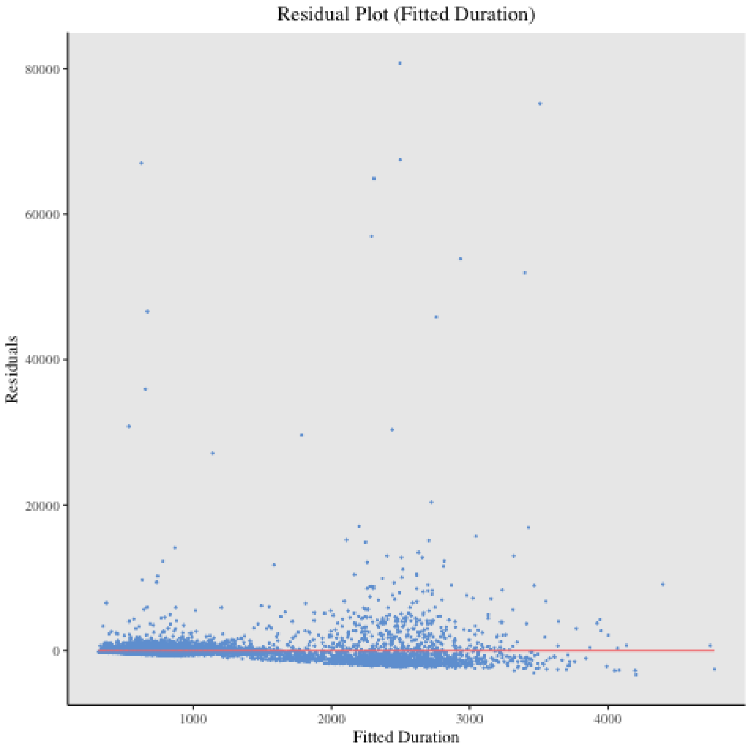
  - Member_type_type:Rush_hour
    We added two other confounders of Member_type and Rush_hour. Member_type is a binary variable on whether the rider is a member or not, and Rush_hour is a binary variable on whether the ride took place during rush hours. We thought that non-members are likely to take longer trips to get value out of what they paid for and that rush hour trips might be longer due to congestion.

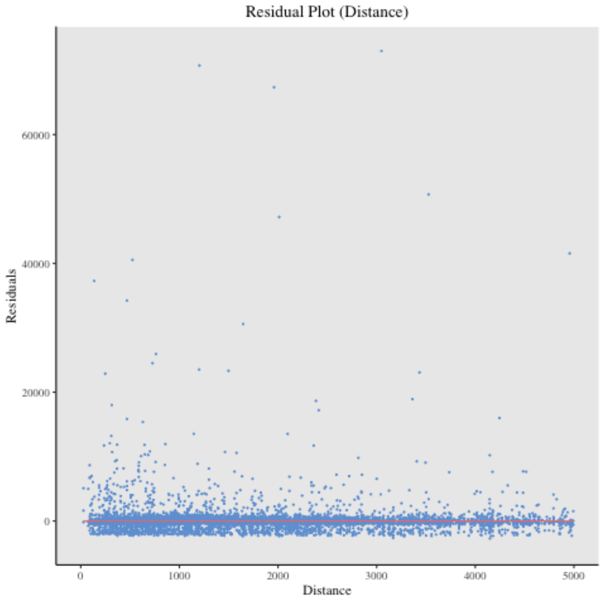
- Terms not included
  We did not include the bike numbers because we felt it has no predictive power for duration and we do not want to fit our model to noise. We also had considered including a city_center categorical variable as traffic in city centers would be more congested and hence trips will take a longer time, but we realized that drawing the boundary would be too arbitrary given our lack of knowledge of Washington DC, and that the boundaries would not carry too much distinction as it is not too problematic to consider Washington DC to have a roughly similar level of traffic congestion.

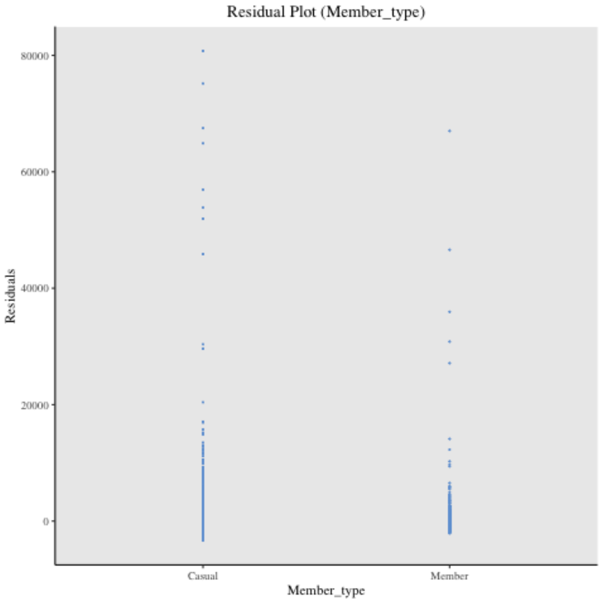
4.2 Diagnostics

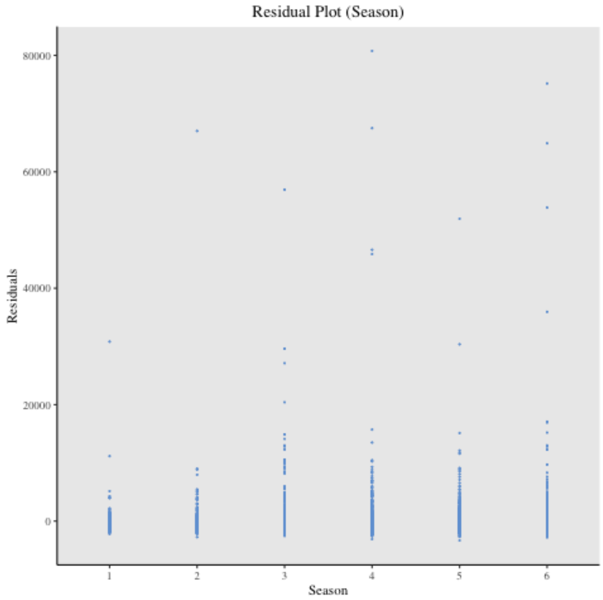
We ran diagnostics of the aforementioned null linear model by sampling 10000 observations from the aggregate data. We chose to sample instead of running on the entire data set for computational reasons, as the data set is very large.

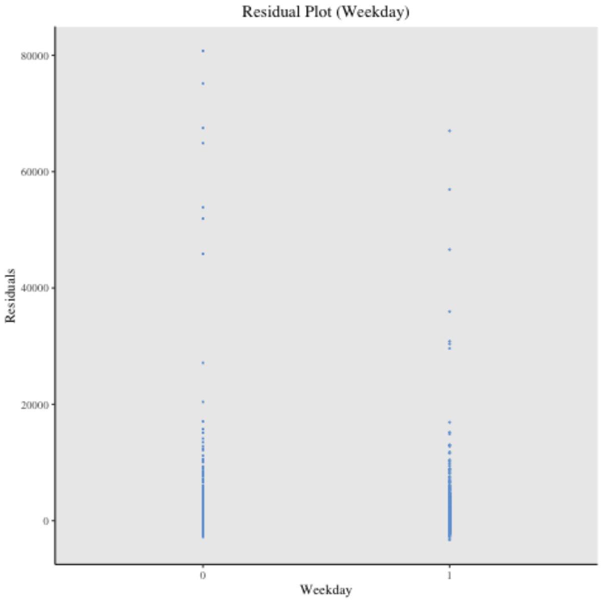
We ran linear regression on our null model specified in the previous section, with one difference being the covariate `{noise on a given day(day-to-day variance)}`, which is not accounted for when we run diagnostics. However, we assume that such variations within a day would be averaged out as random noise in the context of an aggregate data and therefore do not pose a problem. Moreover, by using a sample of the aggregate data, we assume that the aggregate model follows a null distribution, i.e. our data set contains way more routes corresponding to the null hypothesis than those corresponding to the alternate hypothesis. The following are our residuals plots.


Residual Plot (Fitted Duration)
Residuals
80000
60000
40000
20000
0
1000
2000
3000
4000
Fitted Duration


Residual Plot (Distance)
Residuals
60000
40000
20000
0
0
1000
2000
3000
4000
5000
Distance


Residual Plot (Member_type)
Residuals
80000
60000
40000
20000
0
Casual
Member
Member_type





Other than the outliers, there does not seem to be noticeable abnormalities in the residual plots that suggest nonconstant variance. The outliers tend to have a positive residual, which we think can be explained by the fact that we estimated the distance using the geographical length. As a result, we are not able to account for the trips that do not take the shortest path, such as those of people who use the bike to exercise and people who take their time or stop by certain places. Unfortunately, we will not be able to account for the outliers, as we have models for thousands of routes, and it will be computationally infeasible to calculate the outliers and determine whether to remove them or not for every single model given. We wondered whether the residual plots would look better if we removed the outlier points, for example, by implementing Jackknife in each cluster and then combining the clusters to from an aggregate data set, which we would then sample from. However, due to the limitation of time, we did not further explore this idea.

Some other potential problems with this model include dependence between observations that come from the same person representing multiple data points. There is not a way to resolve this issue, given the limited data that we have available. Another problem is that we approximated the weather conditions with seasonal effects. If we could incorporate the weather data, it would have been able to increase the model accuracy.

The following is a summary of the model corresponding to these residual plots.

```
Call:
lm(formula = finModelAgg, data = sampAgg)

Residuals:
   Min     1Q Median     3Q    Max
 -3214   -373   -172     94  64198

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   1.837e+03  1.105e+02  16.620  < 2e-16 ***
Distance                      2.300e-01  4.605e-02   4.993 6.03e-07 ***
Weekday1                     -1.376e+02  4.774e+01  -2.883 0.003952 **
Member_typeMember            -1.455e+03  8.035e+01 -18.112  < 2e-16 ***
Rush_hour1                   -4.666e+02  1.195e+02  -3.906 9.44e-05 ***
Season2                       3.629e+02  1.206e+02   3.009 0.002628 **
Season3                       2.151e+02  1.059e+02   2.030 0.042334 *
Season4                       3.187e+02  1.028e+02   3.101 0.001935 **
Season5                       1.822e+02  1.030e+02   1.769 0.076899 .
Season6                       1.618e+02  9.948e+01   1.627 0.103846
Member_typeMember:Rush_hour1  4.642e+02  1.242e+02   3.738 0.000186 ***
Distance:Season2             -1.484e-01  5.191e-02  -2.859 0.004257 **
Distance:Season3             -4.235e-03  4.718e-02  -0.090 0.928475
Distance:Season4             -1.036e-01  4.587e-02  -2.259 0.023934 *
```

```
Distance:Season5            -6.826e-02  4.656e-02  -1.466 0.142655
Distance:Season6            -7.167e-02  4.439e-02  -1.614 0.106461
Distance:Member_typeMember   1.220e-02  3.136e-02   0.389 0.697327
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1882 on 9983 degrees of freedom
Multiple R-squared:  0.1039,Adjusted R-squared:  0.1025
F-statistic: 72.34 on 16 and 9983 DF,  p-value: < 2.2e-16
```

It seems that the confounders we included do seem to account for much of the global trend.

4.3 Model for Each Subgroup

Having confirmed that our null model seems valid, we make the same assumption on the null model for each cluster, and we use this assumption in the method discussed in the first section: for each route (a cluster), we implement our two-step procedure, where the first step regresses out the global trend, and the second step makes inference while accounting for the local trend, i.e. ‘within-day trend’. It should be noted that when we regress out the global confounders, we do not need Distance and the interactions terms that involve Distance as it is already controlled for by the fact that we are clustering the data by groups.

5 How We Decided on Our Final Method

This section is more of a description of what questions we’ve asked among ourselves and the discussions we’ve had when trying to decide on which method to use.

5.1 Initial Attempt

The question we are trying to answer is an example of a type of problem of determining whether there is an association between X and Y, conditioned on Z. The Y in this case would be the duration of a ride, and the X would be the days that have passed. In order to tackle this problem, our group initially decided to consider a regression approach to determine the coefficient on X. In this particular case, Y would be Duration, X would be Days_since_Jan1_2010, and Z would be confounders such as Season. In this model, under the null hypothesis, Duration and Days_since_Jan1_2010 would have no association, so the coefficient of Days_since_Jan1_2010 would be 0. Our model would be a linear regression model that tests whether the coefficient is 0 or not.

However, we soon realized that there is an important problem that we must address, which is that the individual data points are not independent. Because of the day-to-day variations such as the weather, the data points on the same day would be dependent with each other, and furthermore the duration on different days would vary significantly from one another. Because of these problems, a usual regression and a t-test would not be valid to answer our question, as they assume independence between data points.

Potential Issues and Questions Posed

Our group had a lengthy discussion on this linear model, regarding whether it would be valid to use a linear model when Duration and Days_since_Jan1_2010 would not have a linear relationship if there is a significant change in the route, i.e. if the route is the route that we are trying to reject. For example, if there is a construction on the route on a specific day that shortened the distance between point A and point B, then the function of Duration with respect to Days_since_Jan1_2010 would look more like a step function rather than a linear function. One of us suggested that this would not be a problem because a step function can be approximated by a linear model, even if it is a rough approximation. However, another issue that was raised was that if the relationship cannot be described by a monotonic function, then the linear model would not be able to approximate it, even roughly.

Another confusion we had was about whether we could trust the pvalue we would get after running this model. Even if we assumed that the data are independent (which is a very unlikely assumption), we were concerned that if the relationship is nonlinear, then we would be violating the assumptions of a linear model and therefore the estimate of the coefficient as well as the p-value would be invalid. However, after one of us discussed this particular issue with Professor Barber, we learned that this would not be a problem since the nonlinearity has to do with the alternate hypothesis, not the null hypothesis. This clarification eliminated a lot of the confusions we had. However, the fact that the data points are not iid (due to dependence among data points belonging to the same day) is still a big issue, and using the linear model, which assumes iid data, would have been too strong of an assumption that would have caused the model to be invalid.

5.2 Second Attempt

When we confused ourselves by incorrectly thinking that the nonlinear relationship would be the main issue of not being able to use the linear model, thinking that the approximation of a linear model to a nonlinear relationship between Days_since_Jan1_2010 and Duration would be too rough, we came up with an alternate approach. This approach was to run the linear model as well, but not with the intention of using this linear model as a tool for inference. In fact, the model we conceived of did not even include the covariate Days_since_Jan1_2010, which is the sole difference between the linear model in this attempt and that in the previous attempt, i.e. this model would still have the same set of confounders. Our approach was to run this linear model as a tool to get the residuals, thinking that, under the assumption that the relationship between Duration and other confounders is linear, all the nonlinearity of Days_since_Jan1_2010 would be captured in the residuals. Therefore, in this attempt, under the null hypothesis, the residuals would be iid with the mean of 0. On the other hand if it is the residual plot of a route to be rejected, it would capture the nonlinear trend, which would look like a step function. Then, in order to detect this nonlinearity, we thought about using the permutation test, permuting the residuals with Days_since_Jan1_2010.

Potential Issues and Questions Posed

However, we realized that we would face the same problem as the problem we had in our first attempt, which is that the data are not iid, due to the fact that the routes belonging to the same day would be influenced by a similar weather condition. Even if we regressed all the other confounding effects out by having a valid set of confounders in the linear model, the residuals would still be dependent as those confounders only account for the global trend such as which season it is, whether it is a weekday or not, and whether the rider is a member or not. Since a permutation test also assumes that the data are iid (even though it doesn't make any assumption about which distribution it has to be, i.e. is a valid test for any distribution), we faced the same exact issue that stopped us from using the linear model for inference approach we initially had. We therefore decided that we should continue our search for a better model.

5.3 Final Choice of Method

The method that we finally decided on is the method that is described in the beginning of the report. This method is similar to our second attempt which we just mentioned, but it accounts for the issue of data of the same day being dependent. This method assumes that all the global confounders, by which we mean confounders that account for associations across different days, are valid and have a linear relationship with Duration. As we already ran a diagnostic on our model, we believe that this assumption is valid. Therefore, the residuals capture two things: the nonlinearity and the dependency among the data due to several of them coming from the same day. By grouping the residuals by day, we would account for this dependency, isolating the trend in the residuals to nonlinearity, as well as abiding by the iid assumption of a permutation test. Then, our permutation test would reject hypotheses solely based on the trend in in the residuals, where a lack of trend (iid noise with mean 0 and constant variance) denotes that the average time to travel the route has not changed over time whereas any trend in the residuals would suggest otherwise. We therefore thought that this model addresses the main problems we faced in our first two attempts.

6 Conclusions

We summarize our method again here; the two-step procedure is summarized as below.

1. Runs linear regression on each data table corresponding to a route. After running the model, the residuals are added as a column to each data table

2. Runs grouped permutation test where we permute the residuals with Days_since_Jan1_2010. By grouped, we mean that the residuals corresponding to the same day stay as a single unit throughout the permutation. This is similar to Scheme B of Assignment 1 where we keep X the same and permute Y. This preserves the correlation structure, and we do this in order to account for the confounding effects of the observations that fall under the same day.

Then, we implement the BH multiple testing procedure for our 2648 tests at the rejection level of 0.3.

6.1 Findings

Out of the 2648 p-values we have, our method rejected 90 hypotheses. The routes, the hypotheses corresponding to which were rejected, are listed below. Based on our method, we report that the following routes are the routes whose average time to travel changed over time.

```
"M St & New Jersey Ave SE, Massachusetts Ave & Dupont Circle NW"
"1st & N St  SE, Potomac & Pennsylvania Ave SE"
"5th & K St NW, 20th & E St NW"
"5th & K St NW, Georgetown Harbor / 30th St NW"
"5th & K St NW, 5th & F St NW"
"5th & K St NW, New York Ave & 15th St NW"
"5th & K St NW, 14th St & Spring Rd NW"
"5th & K St NW, 13th St & New York Ave NW"
"7th & T St NW, Massachusetts Ave & Dupont Circle NW"
"10th & U St NW, 4th & W St NE"
"10th & U St NW, Van Ness Metro / UDC"
"4th & W St NE, 13th & H St NE"
"1st & M St NE, Massachusetts Ave & Dupont Circle NW"
"Park Rd & Holmead Pl NW, 14th & R St NW"
"Park Rd & Holmead Pl NW, Massachusetts Ave & Dupont Circle NW"
"Park Rd & Holmead Pl NW, 20th & E St NW"
"Park Rd & Holmead Pl NW, Harvard St & Adams Mill Rd NW"
"Park Rd & Holmead Pl NW, 3rd & H St NE"
"14th & Harvard St NW, Massachusetts Ave & Dupont Circle NW"
"14th & Harvard St NW, 14th & R St NW"
"14th & Harvard St NW, 37th & O St NW / Georgetown University"
"14th & V St NW, 15th & P St NW"
"16th & Harvard St NW, Calvert & Biltmore St NW"
"16th & Harvard St NW, L'Enfant Plaza / 7th & C St SW"
"19th & E Street NW, 14th & Rhode Island Ave NW"
"19th & E Street NW, 17th & K St NW"
"20th & Crystal Dr, 15th & Crystal Dr"
"20th & Crystal Dr, 12th & Army Navy Dr"
"21st & I St NW, Wisconsin Ave & Newark St NW"
"15th & P St NW, Adams Mill & Columbia Rd NW"
"15th & P St NW, Georgia & New Hampshire Ave NW"
"15th & P St NW, Lamont & Mt Pleasant NW"
"15th & P St NW, 14th St Heights / 14th & Crittenden St NW"
"15th & P St NW, Maine Ave & 7th St SW"
"27th & Crystal Dr, Crystal City Metro / 18th & Bell St"
"Calvert & Biltmore St NW, Massachusetts Ave & Dupont Circle NW"
"Calvert & Biltmore St NW, 3rd & H St NW"
"Calvert & Biltmore St NW, Ward Circle / American University"
"Calvert & Biltmore St NW, 4th & M St SW"
"Massachusetts Ave & Dupont Circle NW, Georgia & New Hampshire Ave NW"
"Massachusetts Ave & Dupont Circle NW, 3rd & D St SE"
"Massachusetts Ave & Dupont Circle NW, New York Ave & 15th St NW"
"Massachusetts Ave & Dupont Circle NW, C & O Canal & Wisconsin Ave NW"
"Adams Mill & Columbia Rd NW, L'Enfant Plaza / 7th & C St SW"
"23rd & Crystal Dr, 12th & Army Navy Dr"
"14th & R St NW, 20th St & Florida Ave NW"
"21st & M St NW, Georgia & New Hampshire Ave NW"
"Good Hope Rd & MLK Ave SE, Anacostia Metro"
"Good Hope Rd & MLK Ave SE, 13th & H St NE"
"4th & M St SW, 3rd & H St NE"
"4th & M St SW, North Capitol St & F St NW"
"4th & M St SW, Convention Center / 7th & M St NW"
"20th & E St NW, 18th & M St NW"
"20th & E St NW, S Joyce & Army Navy Dr"
"20th & E St NW, C & O Canal & Wisconsin Ave NW"
"18th & Eads St., Aurora Hills Community Ctr/18th & Hayes St"
"18th & Eads St., USDA / 12th & Independence Ave SW"
"20th St & Florida Ave NW, Van Ness Metro / UDC"
"20th St & Florida Ave NW, Kennedy Center"
"Georgia & New Hampshire Ave NW, 13th St & New York Ave NW"
```

```
"Georgia & New Hampshire Ave NW, North Capitol St & F St NW"
"Georgia & New Hampshire Ave NW, Van Ness Metro / UDC"
"Georgia & New Hampshire Ave NW, 14th St Heights / 14th & Crittenden St NW"
"Georgia & New Hampshire Ave NW, 19th & L St NW"
"Crystal City Metro / 18th & Bell St, Eads & 22nd St S"
"17th & K St NW, 13th & H St NE"
"3rd & H St NW, 3rd & D St SE"
"3rd & H St NW, 19th & East Capitol St SE"
"3rd & H St NW, Florida Ave & R St NW"
"3rd & D St SE, 8th & Eye St SE / Barracks Row"
"3rd & D St SE, Eastern Market Metro / Pennsylvania Ave & 7th St SE"
"3rd & D St SE, 10th & Monroe St NE"
"14th & D St SE, 19th & East Capitol St SE"
"Potomac & Pennsylvania Ave SE, 8th & Eye St SE / Barracks Row"
"Potomac & Pennsylvania Ave SE, Eastern Market Metro / Pennsylvania Ave & 7th St SE"
"19th & East Capitol St SE, 13th & D St NE"
"Florida Ave & R St NW, 14th St & New York Ave NW"
"Florida Ave & R St NW, Eastern Market / 7th & North Carolina Ave SE"
"USDA / 12th & Independence Ave SW, 11th & Kenyon St NW"
"L'Enfant Plaza / 7th & C St SW, 3rd & H St NE"
"12th & Newton St NE, North Capitol St & F St NW"
"8th & H St NW, Columbus Circle / Union Station"
"8th & H St NW, 14th & G St NW"
"US Dept of State / Virginia Ave & 21st St NW, Lincoln Park / 13th & East Capitol St NE "
"3rd & H St NE, Columbus Circle / Union Station"
"3rd & H St NE, Eckington Pl & Q St NE"
"Eckington Pl & Q St NE, 14th & G St NW"
"Eckington Pl & Q St NE, Bladensburg Rd & Benning Rd NE"
"New Hampshire Ave & T St NW, C & O Canal & Wisconsin Ave NW"
"New Hampshire Ave & T St NW, 14th & G St NW"
```

6.2 Analysis of Results

In order to see whether the results are consistent with our expectation, we randomly sampled two routes from the routes which were rejected and two routes from those which were not rejected. The residual plots are shown below.

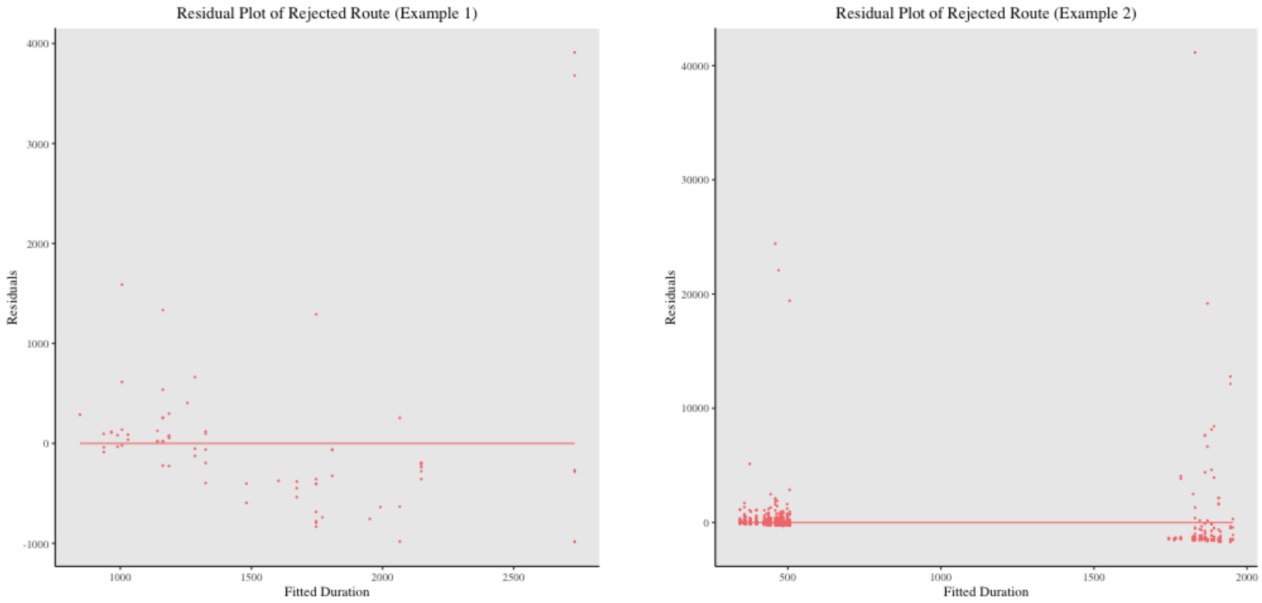


Figure 1: Residual Plots of Routes Corresponding to Rejected Hypotheses

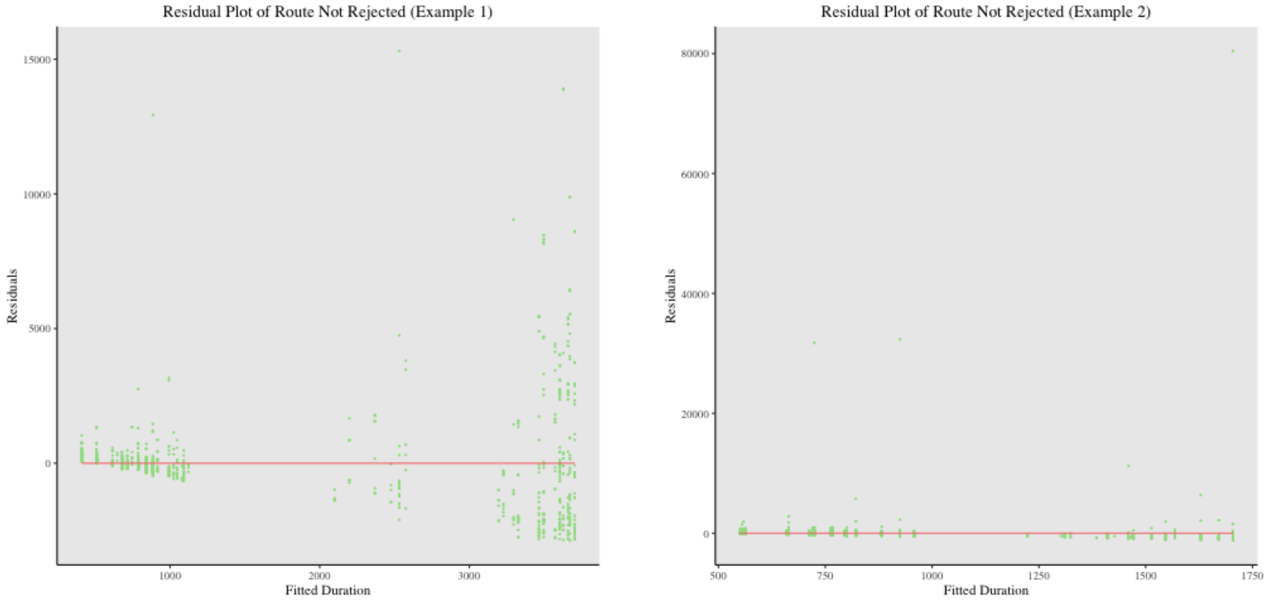


Figure 2: Residual Plots of Routes Corresponding to Hypotheses Not Rejected

It seems that the residual plots are in fact consistent with our expectations. The residual plot of Example 2 of the rejected hypotheses shows a very clear stepwise nonlinear trend in the residuals, as we thought would happen. On the other hand, for the plots corresponding to the hypotheses that were not rejected, it seems to be fair to say the the mean of the residuals is close to 0. The bottom left plot shows signs of nonconstant variance, but the mean still seems to be close to 0, signifying that the road probably did not undergo a structural change. The bottom right residual plot seems consistent with our null model.

6.3 Potential Issues

Throughout our analysis, we have pointed out several potential issues that could not be resolved either due to the nature of the problem or due to the limited time constraints. We would like to mention a few more potential issues here.

- We were not able to incorporate the exact weather data into the analysis, and resolved this issue by instead estimating the seasonal trends and allowing for the day-to-day variation structure to be retained in the permutation tests. However, the model would have been more accurate if we had access to the weather data.

- We assumed that the distance that the riders took were identical given a route. This does not account for those trips that incorporate intentional detours (sightseeing, exercise). Closely connected to this issue, we were also not able to control for outliers, which were mostly those rides that we assume took significantly longer than others due to the riders not taking the shortest route, stopping at some places on the way, or simply taking their time.

- The BH test is conservative given that our routes are dependent to each other, and therefore our power has been diminished. We therefore set the rejection level to be $\alpha = 0.3$, taking this dependency into account, but we have found it difficult to come up with a strong justification for this threshold. It should also be noted that this threshold was chosen in a data dependent way as our method did not reject any hypotheses at $\alpha = 0.1$ or at $\alpha = 0.2$.

6.4 Notes on Codes

The total time taken to run the entire code is 8.61366 mins. The script implements everything described in this report, from data pre-processing and methods to diagnostics and analysis of results. Upon completion, it generates files in the the directory in which the code is run: these files include all the plots that are in this report (they would be different as we randomly sample which routes to plot) and a textfile containing the routes to be rejected. The code has been thoroughly commented, and it is largely divided into the following five different sections.

1. Preprocesses the data and produces the following

    - a list of data tables for each route

    - an aggregate data table containing all observations

2. Runs diagnostics to validate our choice of linear model
3. Runs our model which is composed of a linear regression followed by a permutation test, for each route
4. Implements Benjamini Hochberg(BH) procedure as our multiple testing procedure
5. Analyzes the results

The parts of the code that take a relatively long time is the part where we use R's `distHaversine` function to compute the distance (line 134) and where we run the group permutation test (line 397). Given the large data set, we tried to make the code as efficient as possible by using a data table instead of a data frame, by using list operations, and by vectorizing as many functions as possible.

6.5 Citations

- https://tinyurl.com/vnk9vbh stackechange (weather information)
- https://www.wikihow.com/Avoid-Traffic-Around-Washington,-D.C (for Washinton D.C.'s rush hours)
- stackoverflow.com (for various inquiries)