Diag_reduced

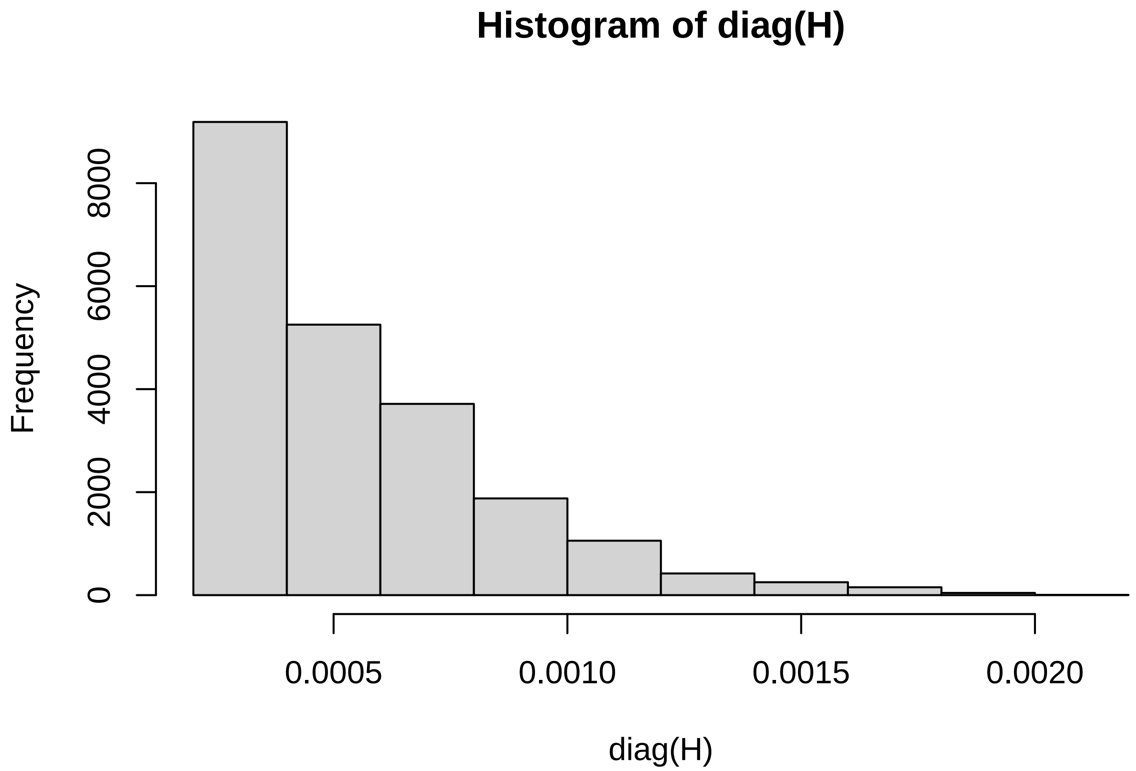
Joanna Ma

2023-02-07

```
## 
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
## 
##     filter, lag

## The following objects are masked from 'package:base':
## 
##     intersect, setdiff, setequal, union

##   Duration          Start.date            End.date Start.station.number
## 1     1012 2010-09-20 11:27:04 2010-09-20 11:43:56                31208
## 2       61 2010-09-20 11:41:22 2010-09-20 11:42:23                31209
## 3     2690 2010-09-20 12:05:37 2010-09-20 12:50:27                31600
## 4     1406 2010-09-20 12:06:05 2010-09-20 12:29:32                31600
## 5     1413 2010-09-20 12:10:43 2010-09-20 12:34:17                31100
## 6      982 2010-09-20 12:14:27 2010-09-20 12:30:50                31109
##                   Start.station End.station.number
## 1      M St & New Jersey Ave SE              31108
## 2                1st & N St  SE              31209
## 3                 5th & K St NW              31100
## 4                 5th & K St NW              31602
## 5 19th St & Pennsylvania Ave NW              31201
## 6                 7th & T St NW              31200
##                            End.station Bike.number Member.type
## 1                        4th & M St SW      W00742      Member
## 2                       1st & N St  SE      W00032      Member
## 3        19th St & Pennsylvania Ave NW      W00993      Member
## 4              Park Rd & Holmead Pl NW      W00344      Member
## 5                       15th & P St NW      W00883      Member
## 6 Massachusetts Ave & Dupont Circle NW      W00850      Member

##   Duration  Start.date Bike.number Member.type      routes weekday
## 1     1012 0.000000000      W00742      Member 31208-31108       1
## 2       61 0.009930556      W00032      Member 31209-31209       1
## 3     2690 0.026770833      W00993      Member 31600-31100       1
## 4     1406 0.027094907      W00344      Member 31600-31602       1
## 5     1413 0.030312500      W00883      Member 31100-31201       1
## 6      982 0.032905093      W00850      Member 31109-31200       1
## # A tibble: 12,179 x 2
##    routes          n
##    <chr>       <int>
##  1 31104-31106  4833
##  2 31106-31104  4831
##  3 31613-31619  4820
##  4 31619-31613  3806
##  5 31200-31201  3674
##  6 31217-31217  3373
##  7 31229-31200  3270
##  8 31201-31200  3229
##  9 31101-31200  2779
## 10 31623-31611  2753
## # ... with 12,169 more rows
```

```r
# Run the outliers test and White test with smaller sample size
count_routes2 <- df %>%
  group_by(routes) %>%
  summarise(count = n()) %>%
  filter(count >= 3500)

df_reduced2 <- df[df$routes %in% count_routes2$routes,]
rm(df)
rm(bikeshare)
rm(bikeshare_2010)
rm(bikeshare_2011)
model3 <- lm(Duration~Start.date*routes+weekday+Member.type,data = df_reduced2)
```

```r
# Outliers
X = model.matrix(model3)
H = X%*%solve(t(X)%*%X,t(X))
hist(diag(H),breaks=10)
```



No outlier visible from the histogram.

```
# White Test
library(skedastic)
skedastic:::white_lm(model3)
```

```
## # A tibble: 1 x 5
##   statistic  p.value parameter method       alternative
##       <dbl>    <dbl>     <dbl> <chr>        <chr>
## 1      124. 3.85e-16        22 White's Test greater
```

Extremely small P-value; strong evidence of heteroscedasticity.