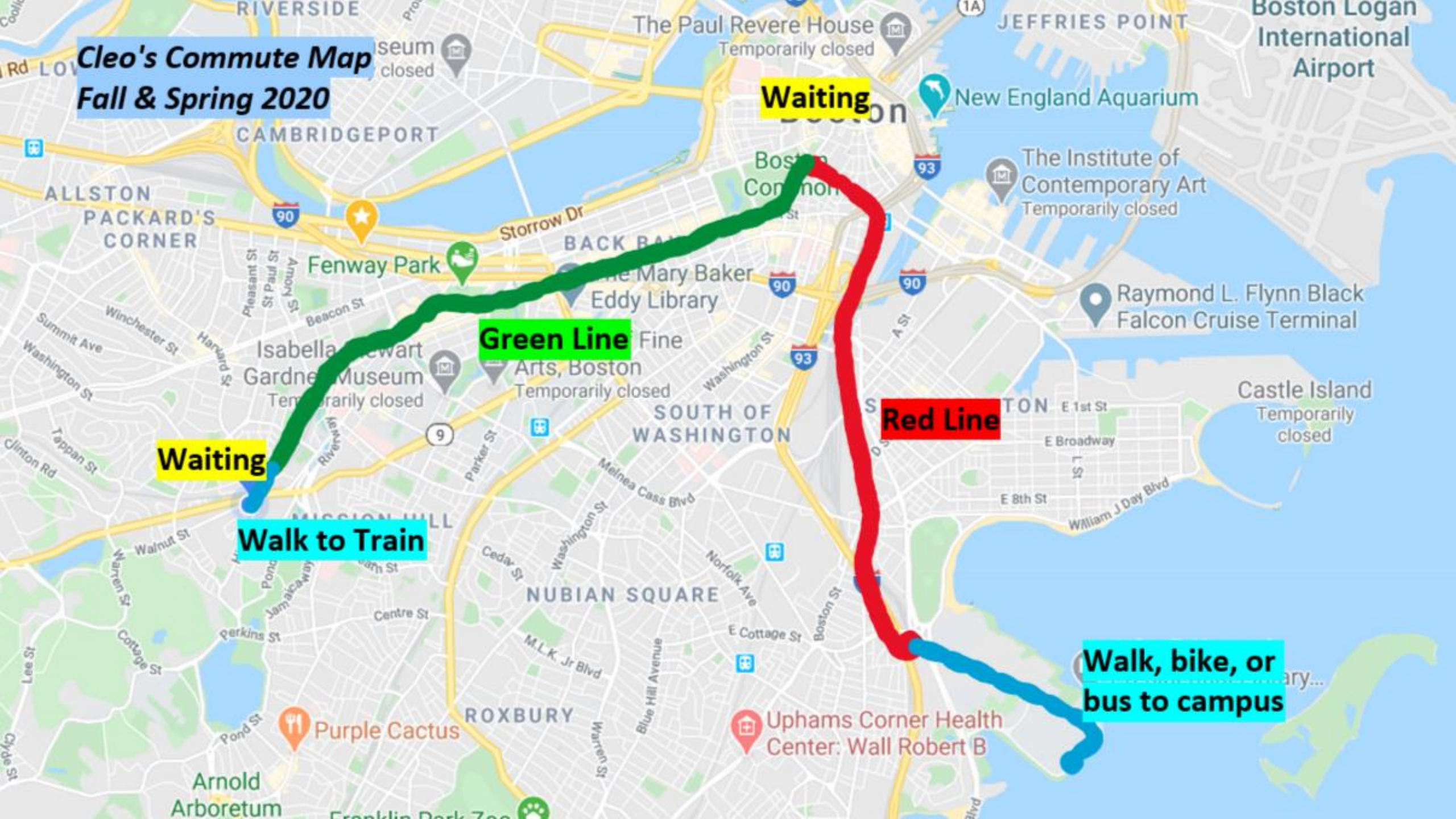
My Commute is a Train Wreck

Cleo Falvey

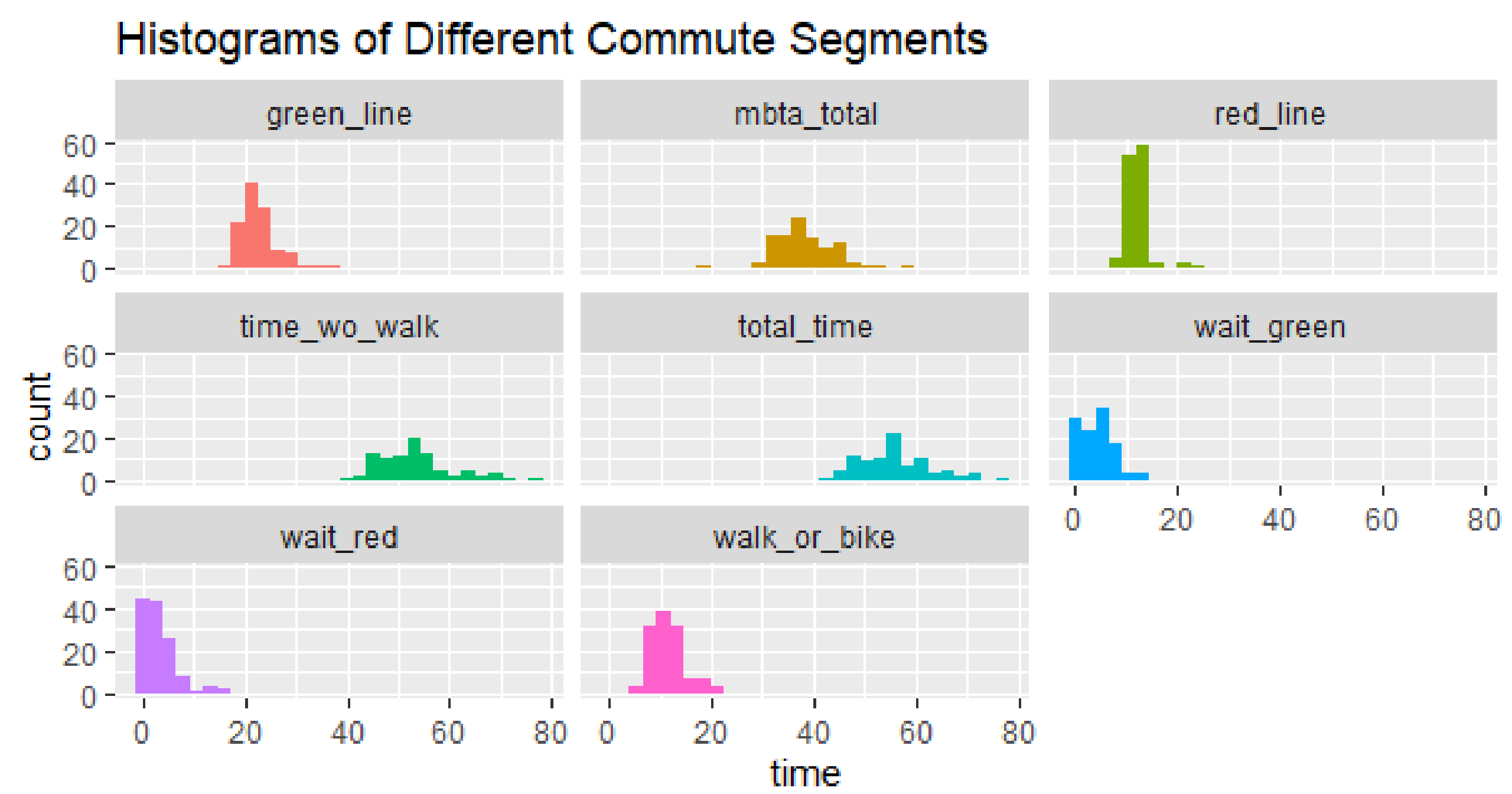
Math 448

Spring 2020


Cleo's Commute Map
Fall & Spring 2020
Waiting
Green Line
Red Line
Waiting
Walk to Train
Walk, bike, or bus to campus
RIVERSIDE
CAMBRIDGEPORT
ALLSTON
PACKARD'S CORNER
Fenway Park
Storrow Dr
BACK BAY
The Paul Revere House
Temporarily closed
JEFFRIES POINT
Boston Logan International Airport
New England Aquarium
The Institute of Contemporary Art
Temporarily closed
Raymond L. Flynn Black Falcon Cruise Terminal
Castle Island
Temporarily closed
SOUTH OF WASHINGTON
NUBIAN SQUARE
ROXBURY
Purple Cactus
Uphams Corner Health Center: Wall Robert B
Arnold Arboretum
Franklin Park Zoo

How long are the different segments of my commute?

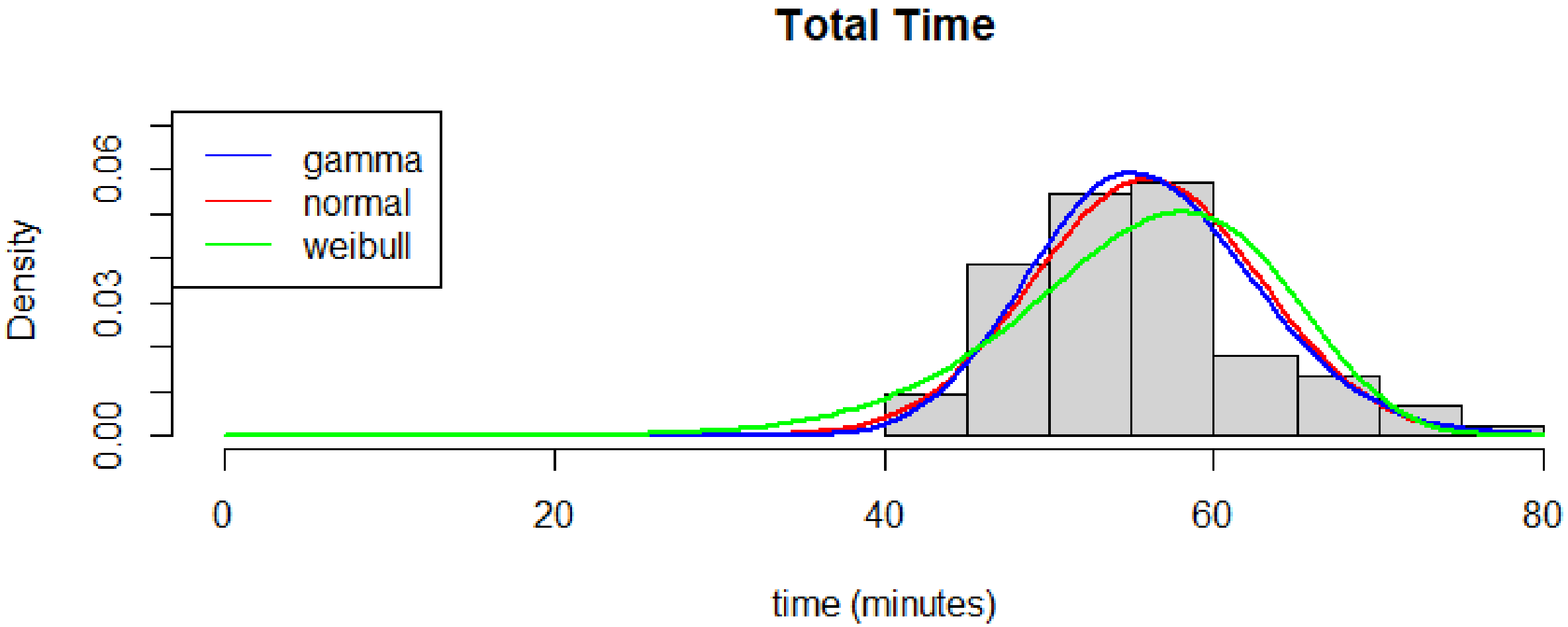
Section I of IV


Histograms of Different Commute Segments
green_line
mbta_total
red_line
time_wo_walk
total_time
wait_green
wait_red
walk_or_bike
count
time
0
20
40
60
80

What probability distribution can we use to model commute segments?

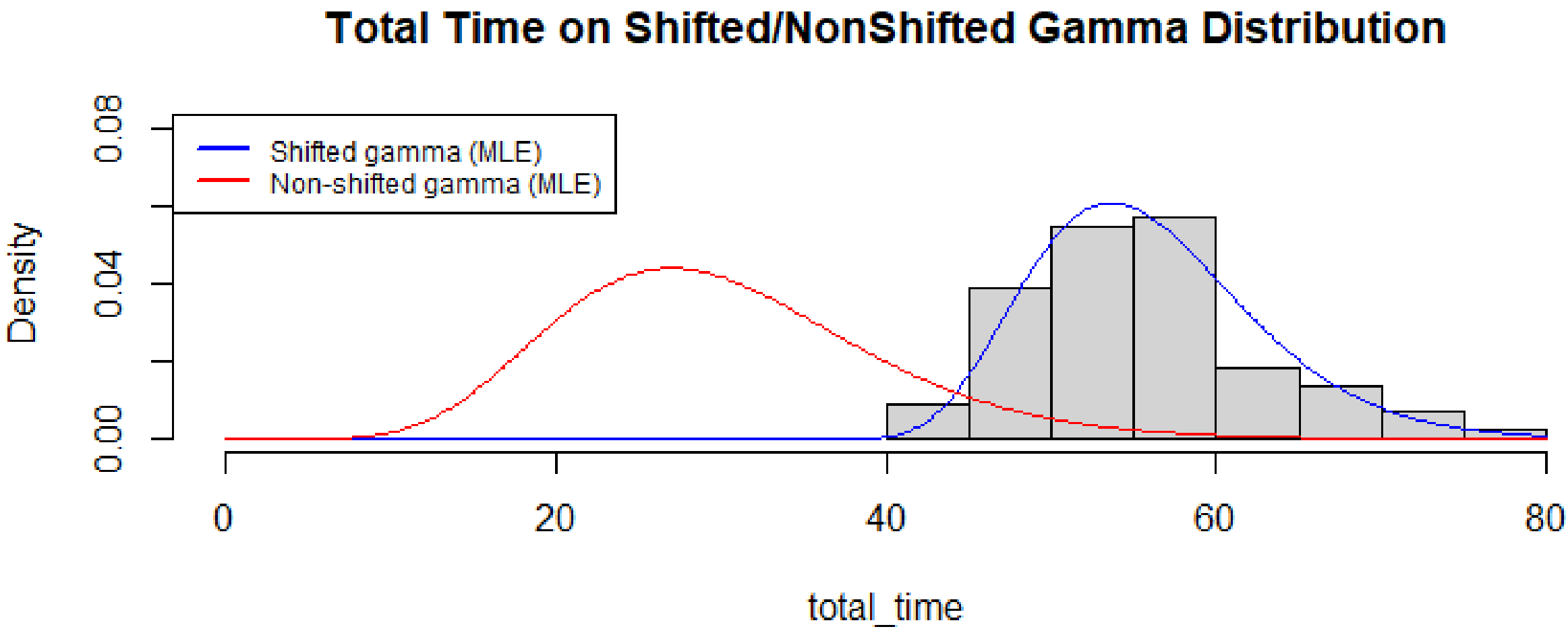
Section II of IV

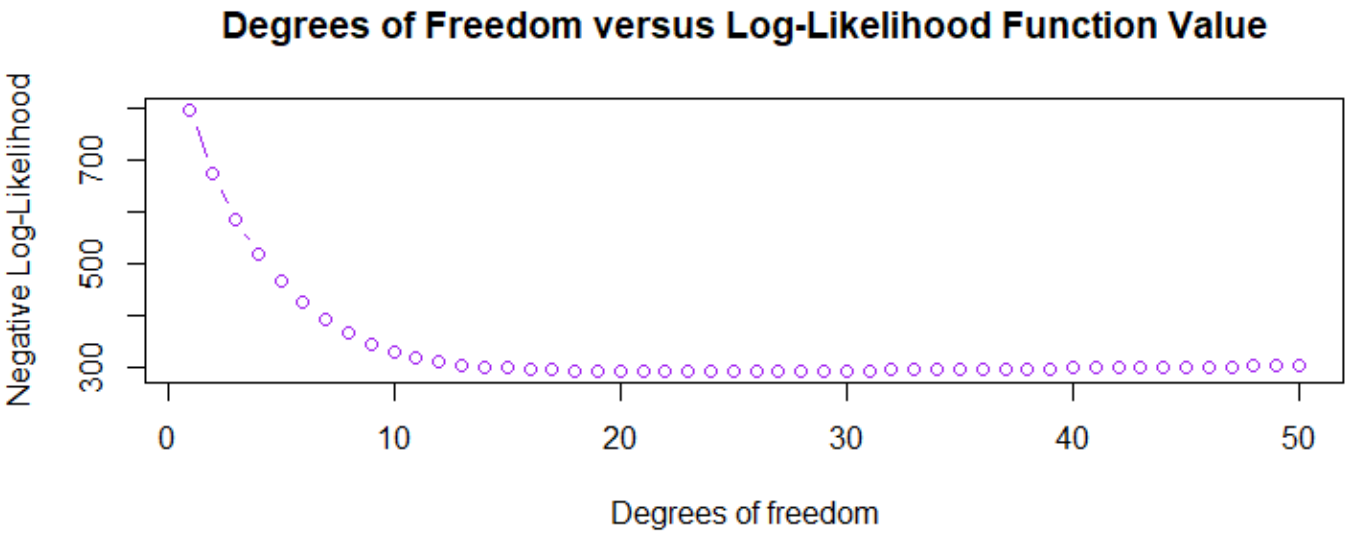
To start, let's try some common distributions using MLE.

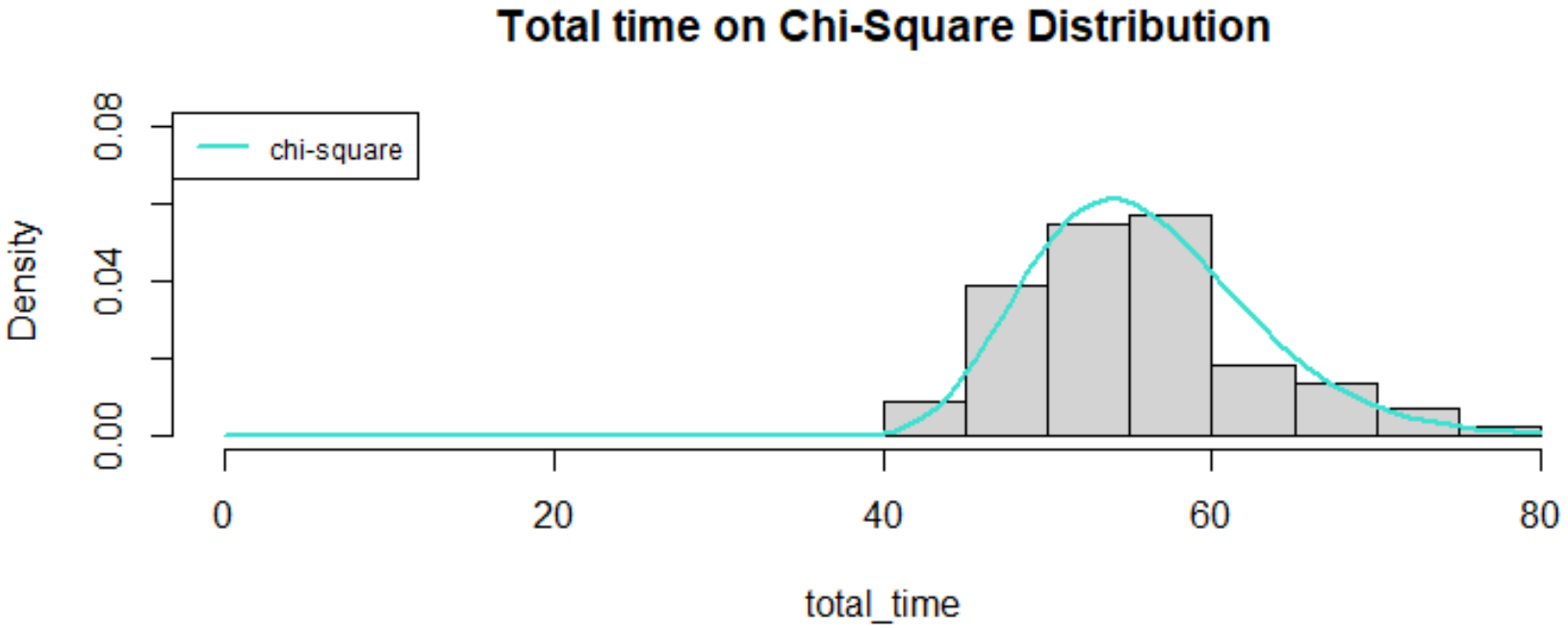


The shifted gamma distribution looked the best.

$\alpha = 9.122924, \beta = 2.281909, c = 35.09$

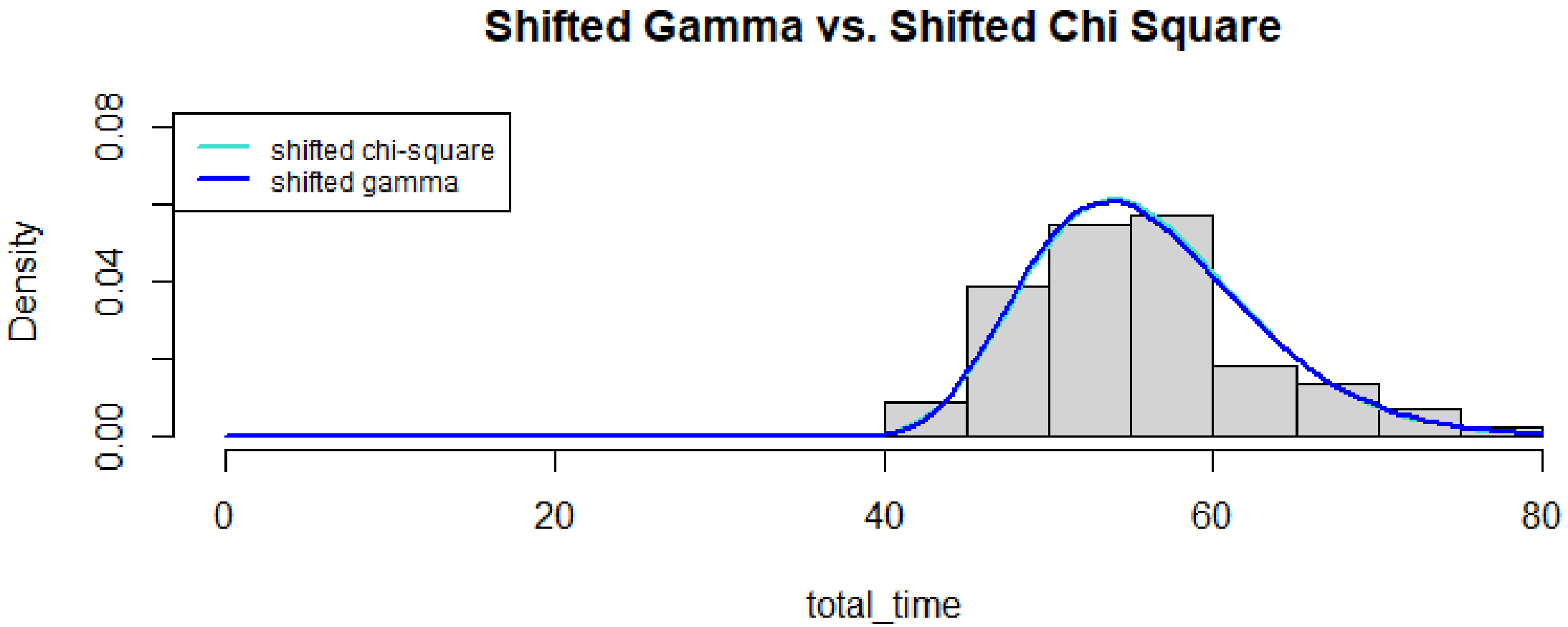






Here's a Chi-square with 23 degrees of freedom (minimized log-likelihood).

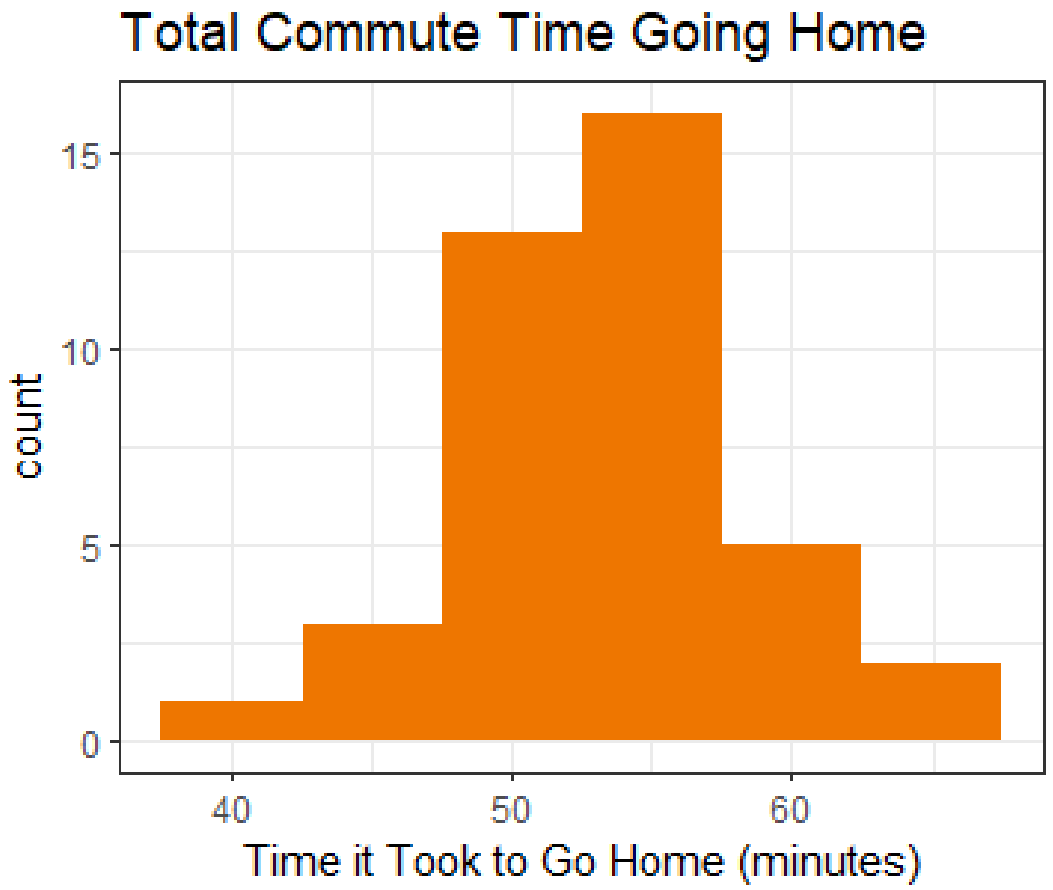
Was the shifted gamma or the shifted chi-square better? *p*=.*75*; the shifted chi-square is better!

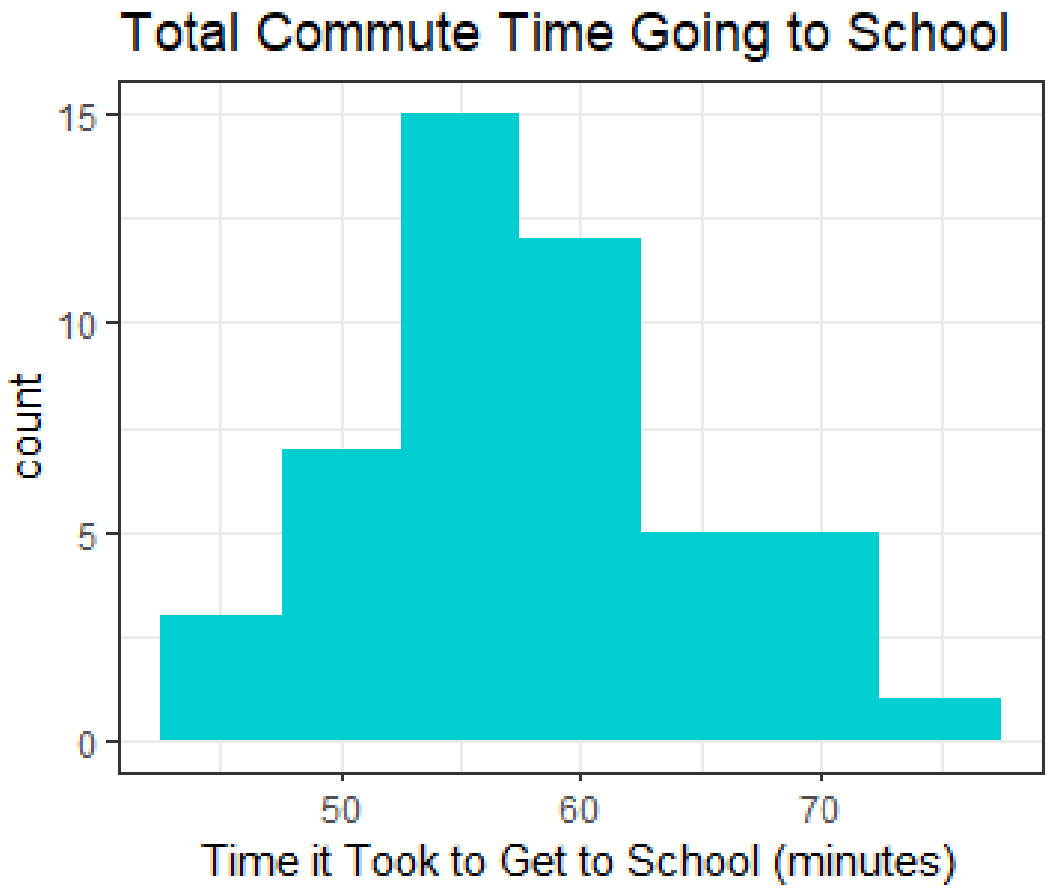


Which direction (going to school vs. coming home) has more variance?

Section III of IV

F=.5366, df=30,47, p=.04811; there is a marginally significant difference (larger variance coming home b/c of rush hour!).



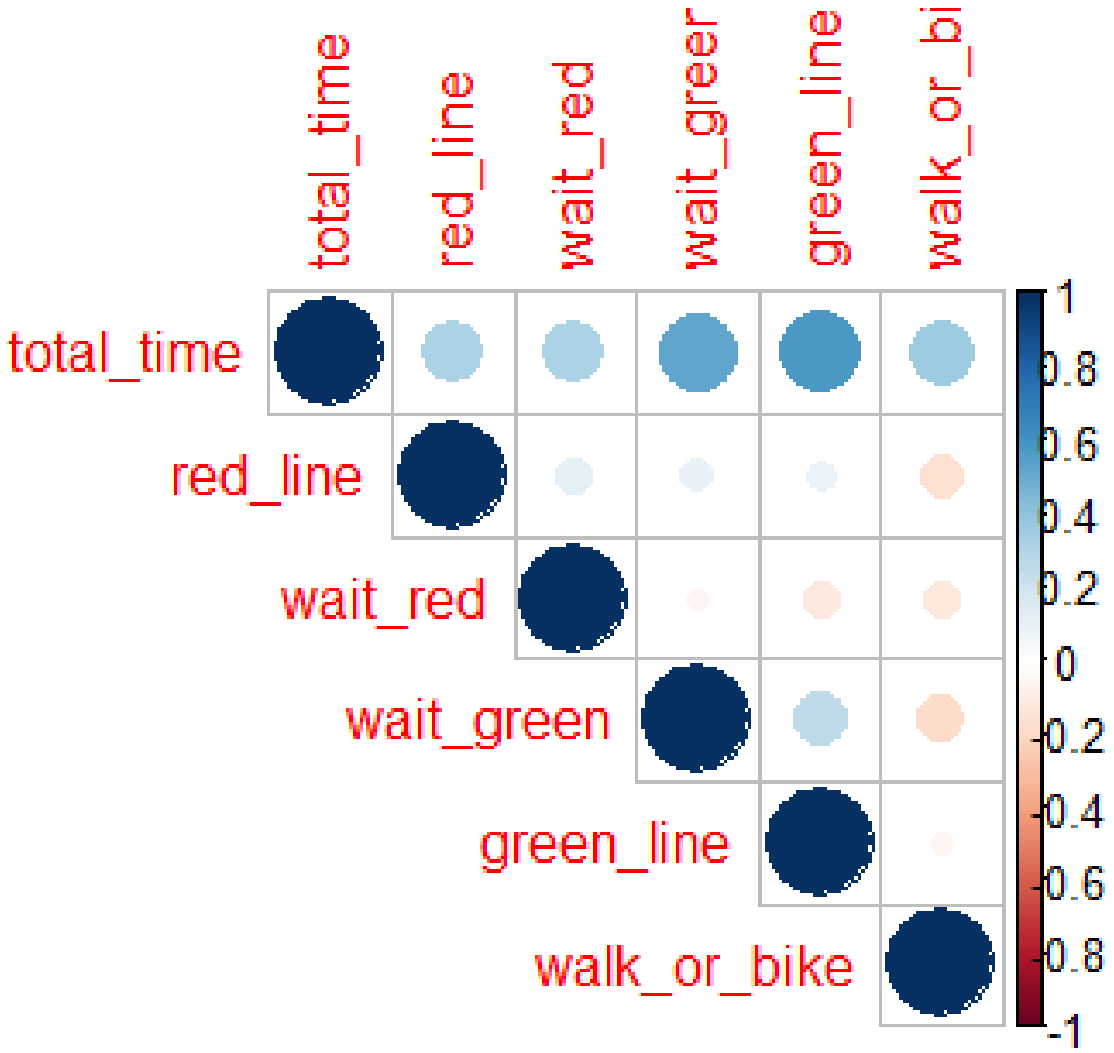


Linear Modelling

Section IV of IV

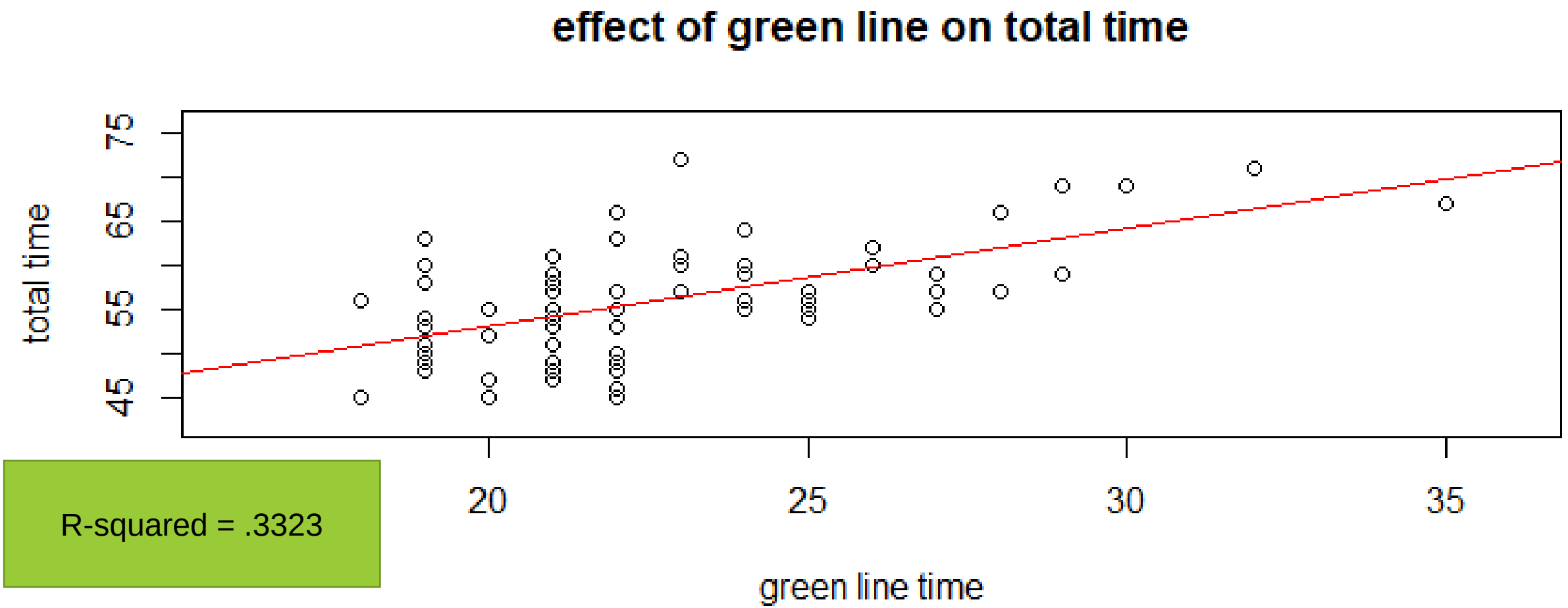
First, let's figure out which segments have the greatest impact on length.

Waiting for the green line and taking the green line had the greatest association with length.



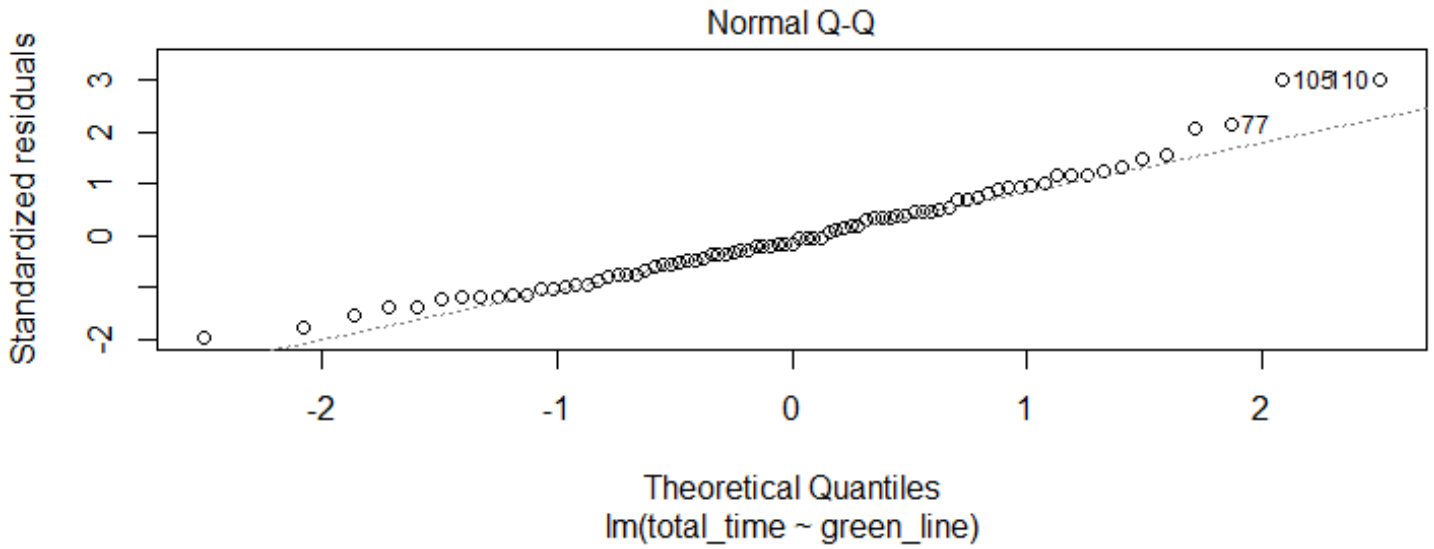
Now, we can plot!

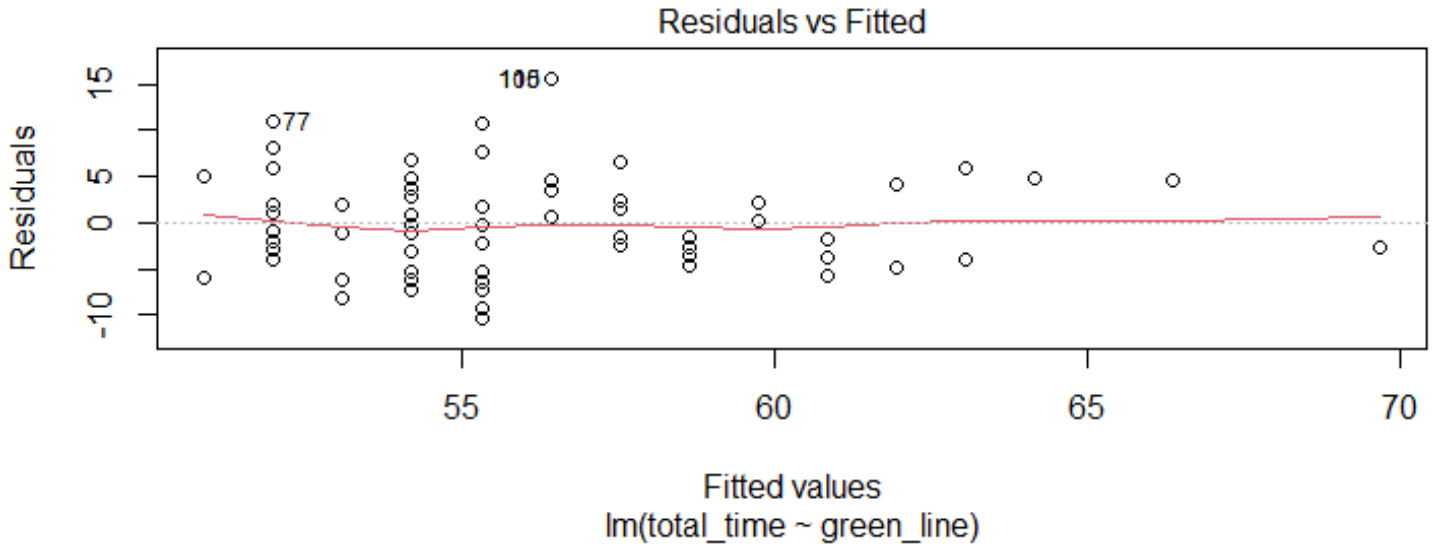
Total time = 30.965+1.1062(green line)



R-squared = .3323

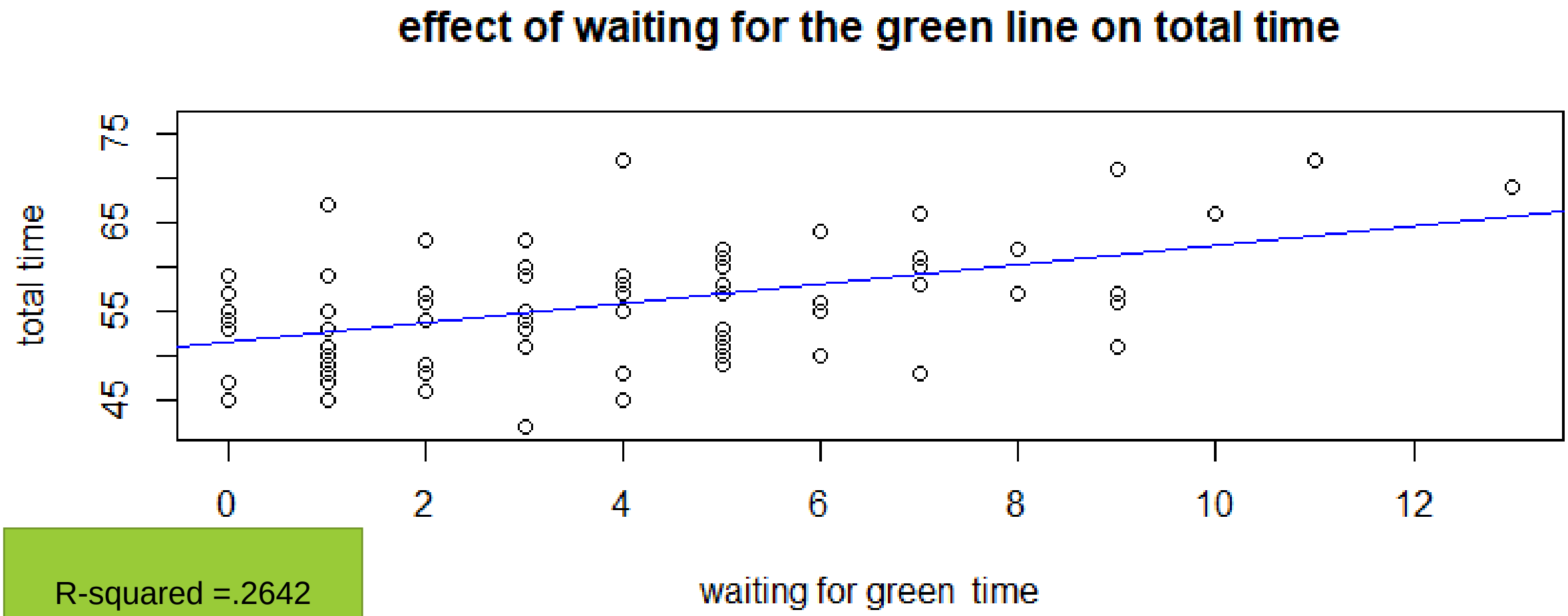
But wait! Are the residuals normal?





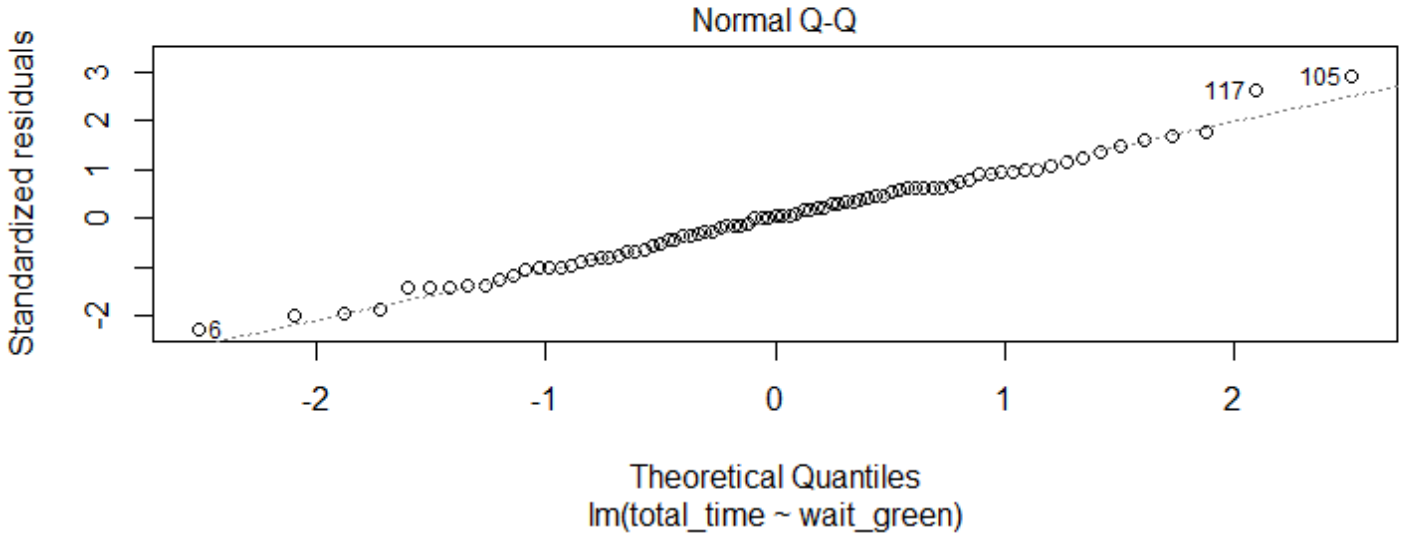
Waiting for the green line was also correlated!

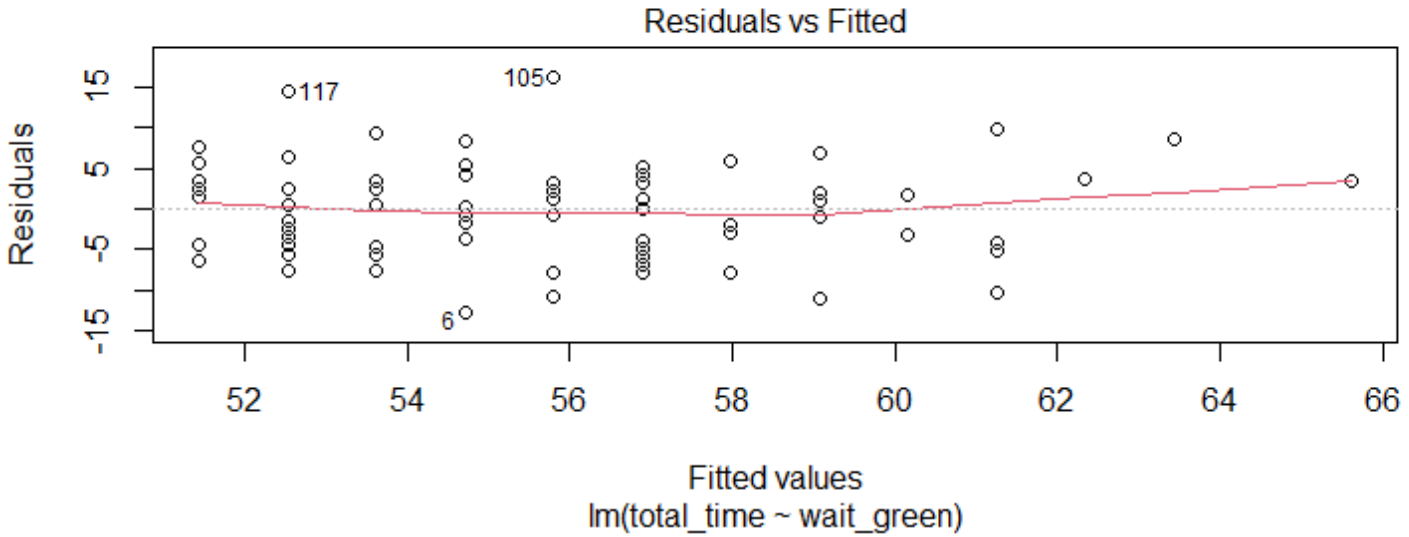
Total time = 51.4505+1.0897(waiting for green line)



R-squared =.2642

Again, we ask if the residuals are normal. And they seem to be!





Takeaways

- My commute is long (between 45 and 70 minutes).
- A good distribution is a shifted chi-square distribution.
- Factors that matter when contributing to the length of it are:
  - The green line
  - Rush hour

Limitations

- Bias / convenience sample
- Inaccuracy in timing
- Covid – 19 limited sample size

Any questions?

Citations

- Marie Laure Delignette-Muller, Christophe Dutang (2015). fitdistrplus: An R Package for Fitting Distributions. Journal of Statistical Software, 64(4), 1-34. URL http://www.jstatsoft.org/v64/i04/.
- Owen, W. J., & Wackerly, D. D. (2008). _Mathematical statistics with applications, seventh edition, Dennis Wackerly, William Mendenhall, Richard L. Scheaffer._ Belmont, CA: Brooks/Cole Cengage Learning.
- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Taiyun Wei and Viliam Simko (2017). R package "corrplot": Visualization of a Correlation Matrix (Version 0.84). Available from https://github.com/taiyun/corrplot
- Thomas Lin Pedersen (2019). patchwork: The Composer of Plots. R package version 1.0.0. https://CRAN.R-project.org/package=patchwork
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- Millard SP (2013). _EnvStats: An R Package for Environmental Statistics_. Springer, NewYork. ISBN 978-1-4614-8455-4, <URL: http://www.springer.com>.
- Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. Rpackage version 1.28.