10: Generalized Linear Models (T-Test)

Environmental Data Analytics | Kateri Salk

Spring 2020

Objectives

1. Describe the components of the generalized linear model (GLM)
2. Apply special cases of the GLM (t-test) to real datasets
3. Interpret and report the results of t-tests in publication-style formats

Set up

```
getwd()
```

```
## [1] "/Users/ks501/Documents/GitHub_Repos/Environmental_Data_Analytics_2020"
```

```
library(tidyverse)

EPAair <- read.csv("./Data/Processed/EPAair_O3_PM25_NC1819_Processed.csv")

# Set date to date format
EPAair$Date <- as.Date(EPAair$Date, format = "%Y-%m-%d")

# Set theme
mytheme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
theme_set(mytheme)
```

Generalized Linear Models (GLMs)

The one-sample test (model of the mean), two-sample t-test, analysis of variance (ANOVA), and linear regression are all special cases of the **generalized linear model** (GLM). The GLM also includes analyses not covered in this class, including logistic regression, multinomial regression, chi square, and log-linear models. The common characteristic of general linear models is the expression of a continuous response variable as a linear combination of the effects of categorical or continuous explanatory variables, plus an error term that expresses the random error associated with the coefficients of all explanatory variables. The explanatory variables comprise the deterministic component of the model, and the error term comprises the stochastic component of the model. Historically, artificial distinctions were made between linear models that contained categorical and continuous explanatory variables, but this distinction is no longer made. The inclusion of these models within the umbrella of the GLM allows models to fit the main effects of both categorical and continuous explanatory variables as well as their interactions.

Choosing a model from your data: A "cheat sheet"

T-test: Continuous response, one categorical explanatory variable with two categories (or comparison to a single value if a one-sample test).

One-way ANOVA (Analysis of Variance): Continuous response, one categorical explanatory variable with more than two categories.

Two-way ANOVA (Analysis of Variance) Continuous response, two categorical explanatory variables.

Single Linear Regression Continuous response, one continuous explanatory variable.

Multiple Linear Regression Continuous response, two or more continuous explanatory variables.

ANCOVA (Analysis of Covariance) Continuous response, categorical explanatory variable(s) and continuous explanatory variable(s).

If multiple explanatory variables are chosen, they may be analyzed with respect to their **main effects** on the model (i.e., their separate impacts on the variance explained) or with respsect to their **interaction effects,** the effect of interacting explanatory variables on the model.

Assumptions of the GLM

The GLM is based on the assumption that the data residuals approximate a normal distribution (or a linearly transformed normal distribution). We will discuss the non-parametric analogues to several of these tests if the assumptions of normality are violated. For tests that analyze categorical explanatory variables, the assumption is that the variance in the response variable is equal among groups. Note: environmental data often violate the assumptions of normality and equal variance, and we will often proceed with a GLM even if these assumptions are violated. In this situation, justifying the decision to proceed with a linear model must be made.

T-Test

One-sample t-test

The object of a one sample test is to test the null hypothesis that the mean of the group is equal to a specific value. For example, we might ask ourselves (from the EPA air quality processed dataset):

Are Ozone levels below the threshold for "good" AQI index (0-50)?

```
summary(EPAair$Ozone)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    5.00   32.00   40.00   40.88   46.00  129.00    2146
```

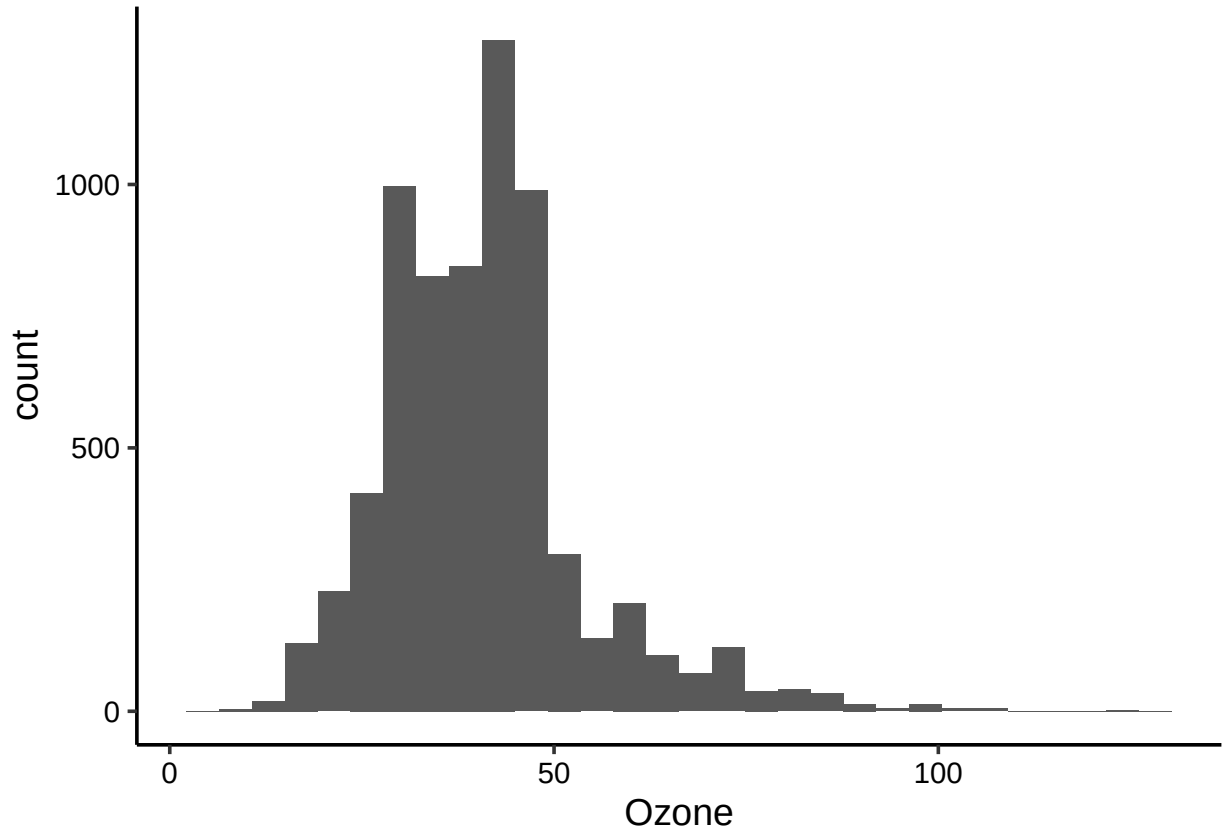
```
EPAair.subsample <- sample_n(EPAair, 5000)

# Evaluate assumption of normal distribution
shapiro.test((EPAair.subsample$Ozone))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  (EPAair.subsample$Ozone)
## W = 0.92145, p-value < 2.2e-16
```

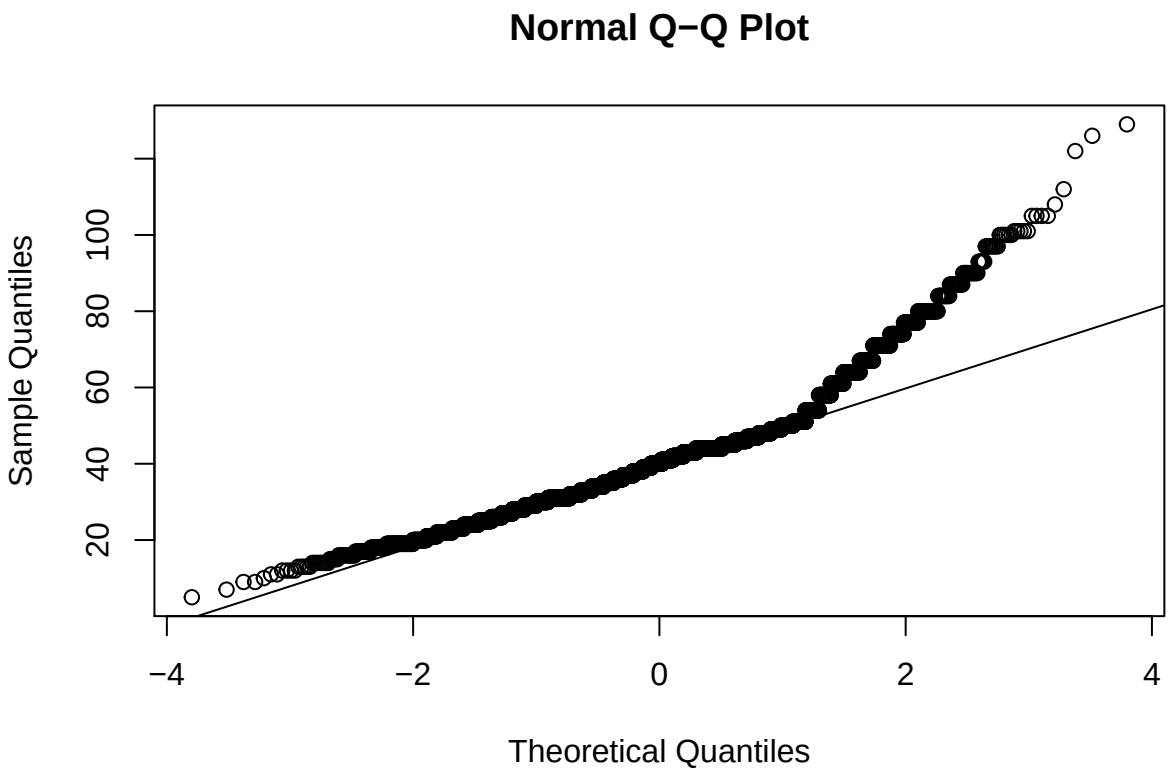
```
ggplot(EPAair, aes(x = Ozone)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2146 rows containing non-finite values (stat_bin).
```




```
qqnorm(EPAair$Ozone); qqline(EPAair$Ozone)
```

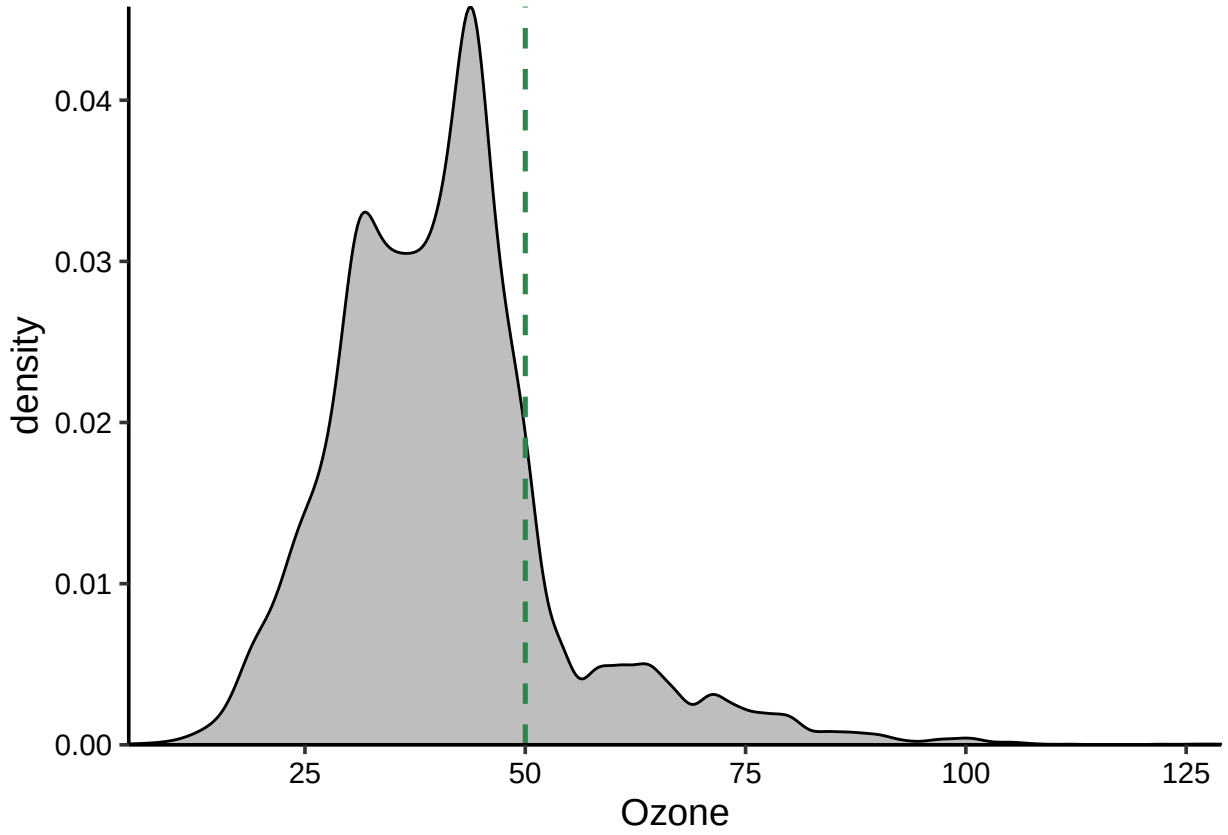



```
O3.onesample <- t.test(EPAair$Ozone, mu = 50, alternative = "less")
O3.onesample
```

```
##
##  One Sample t-test
##
## data:  EPAair$Ozone
## t = -57.98, df = 6829, p-value < 2.2e-16
## alternative hypothesis: true mean is less than 50
## 95 percent confidence interval:
##      -Inf 41.13416
## sample estimates:
## mean of x
##  40.87526
```

```
Ozone.plot <- ggplot(EPAair, aes(x = Ozone)) +
  #geom_density(stat = "count", fill = "gray") +
  geom_density(fill = "gray") +
  geom_vline(xintercept = 50, color = "#238b45", lty = 2, size = 0.9) +
  scale_x_continuous(expand = c(0, 0)) + scale_y_continuous(expand = c(0, 0))
print(Ozone.plot)
```

```
## Warning: Removed 2146 rows containing non-finite values (stat_density).
```



Write a sentence or two about the results of this test. Include both the results of the test and an interpretation that puts the findings in context of the resarch question.

Two-sample t-test

The two-sample t test is used to test the hypothesis that the mean of two samples is equivalent. Unlike the one-sample tests, a two-sample test requires a second assumption that the variance of the two groups is equivalent. Are Ozone levels different between 2018 and 2019?

```
shapiro.test(EPAair$Ozone[EPAair$Year == 2018])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  EPAair$Ozone[EPAair$Year == 2018]
## W = 0.92665, p-value < 2.2e-16
```

```
shapiro.test(EPAair$Ozone[EPAair$Year == 2019])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  EPAair$Ozone[EPAair$Year == 2019]
## W = 0.92132, p-value < 2.2e-16
```

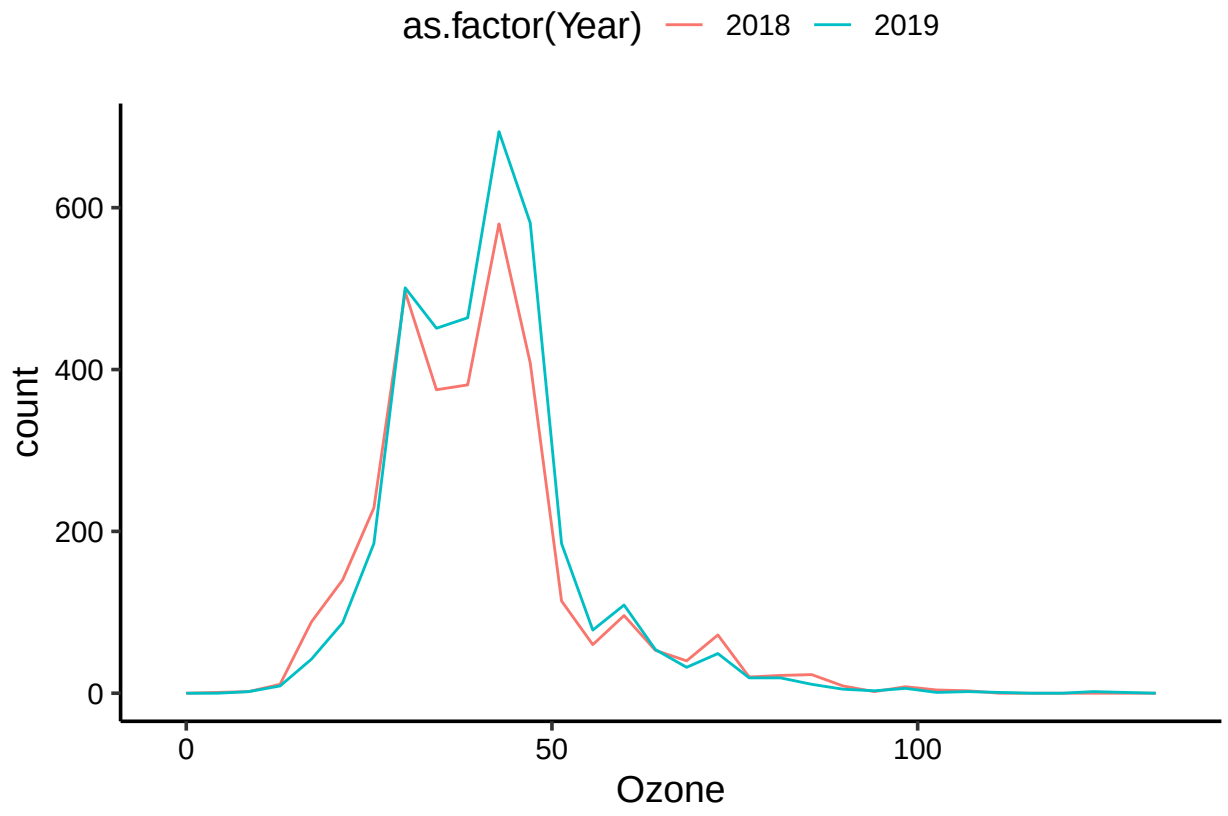
```
var.test(EPAair$Ozone ~ EPAair$Year)
```

```
##
##  F test to compare two variances
##
## data:  EPAair$Ozone by EPAair$Year
## F = 1.3061, num df = 3236, denom df = 3592, p-value = 6.217e-15
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.221265 1.396919
## sample estimates:
## ratio of variances
##           1.306065
```

```
ggplot(EPAair, aes(x = Ozone, color = as.factor(Year))) +
  geom_freqpoly()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2146 rows containing non-finite values (stat_bin).
```




```
# Format as a t-test
O3.twosample <- t.test(EPAair$Ozone ~ EPAair$Year)
O3.twosample
```

```
##
##  Welch Two Sample t-test
##
## data:  EPAair$Ozone by EPAair$Year
## t = -2.6642, df = 6467.7, p-value = 0.007736
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.4670426 -0.2232942
## sample estimates:
## mean in group 2018 mean in group 2019
##           40.43065           41.27581
```

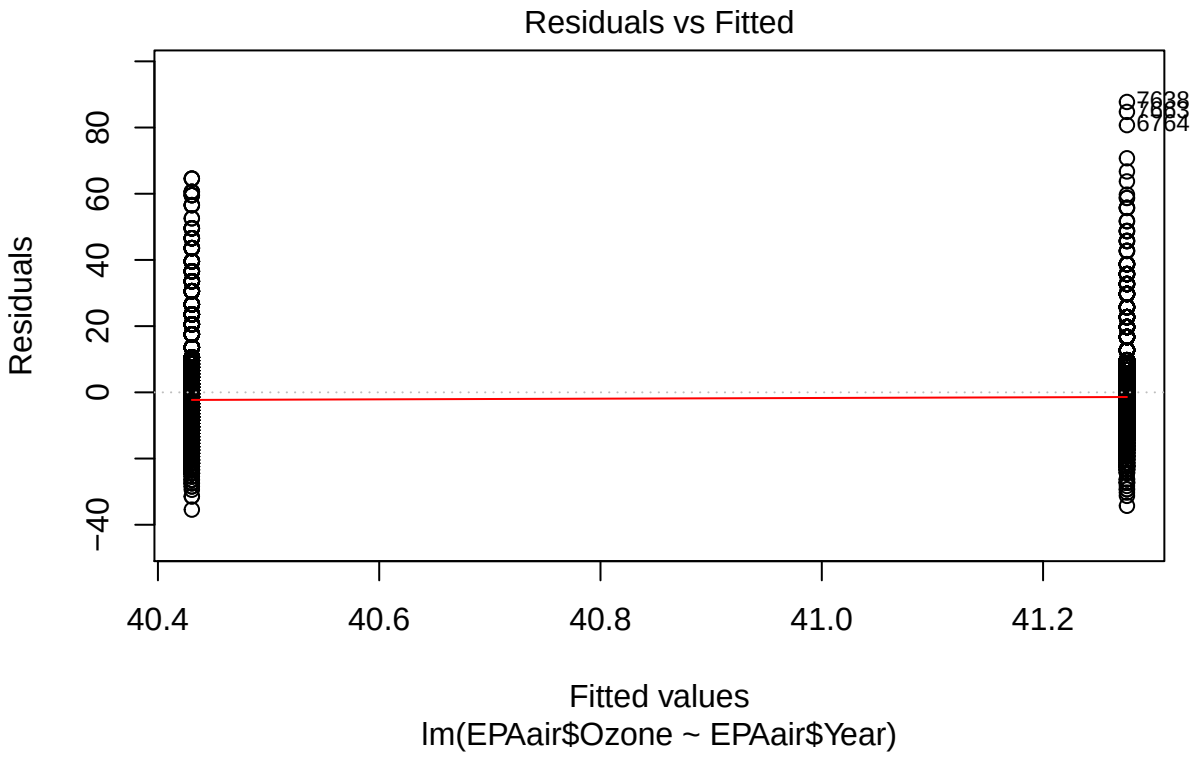
```
O3.twosample$p.value
```

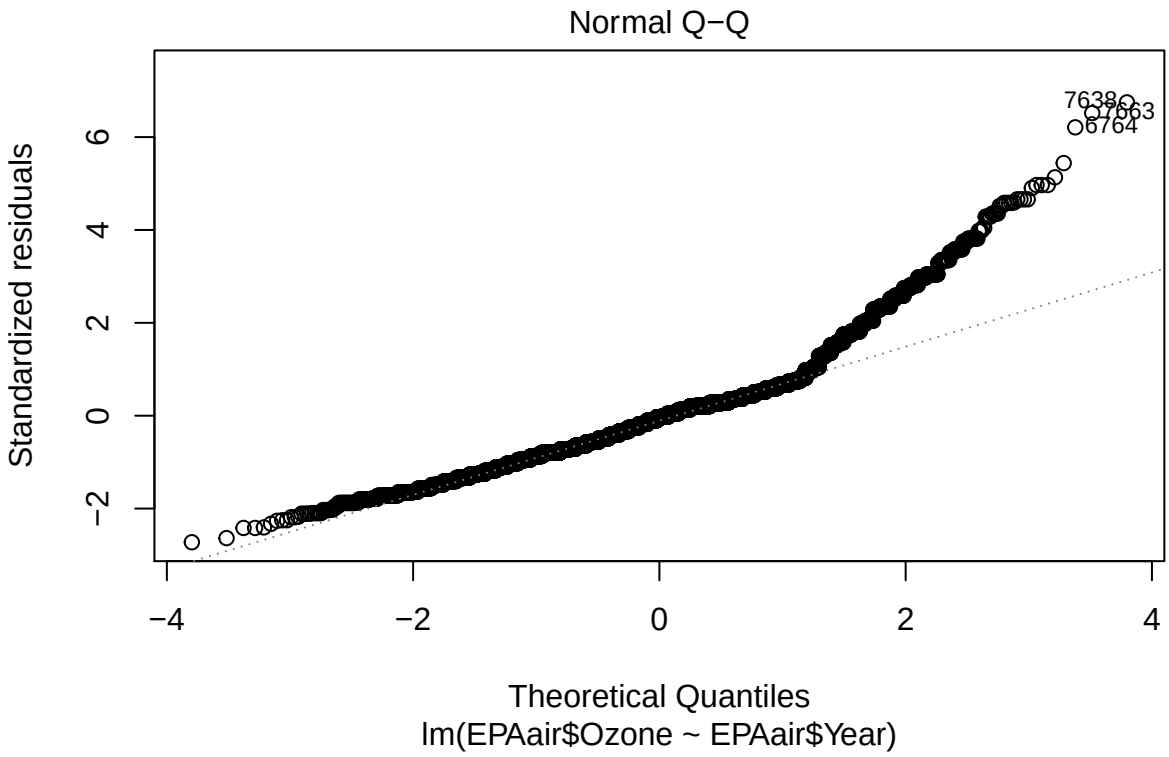
```
## [1] 0.00773585
```

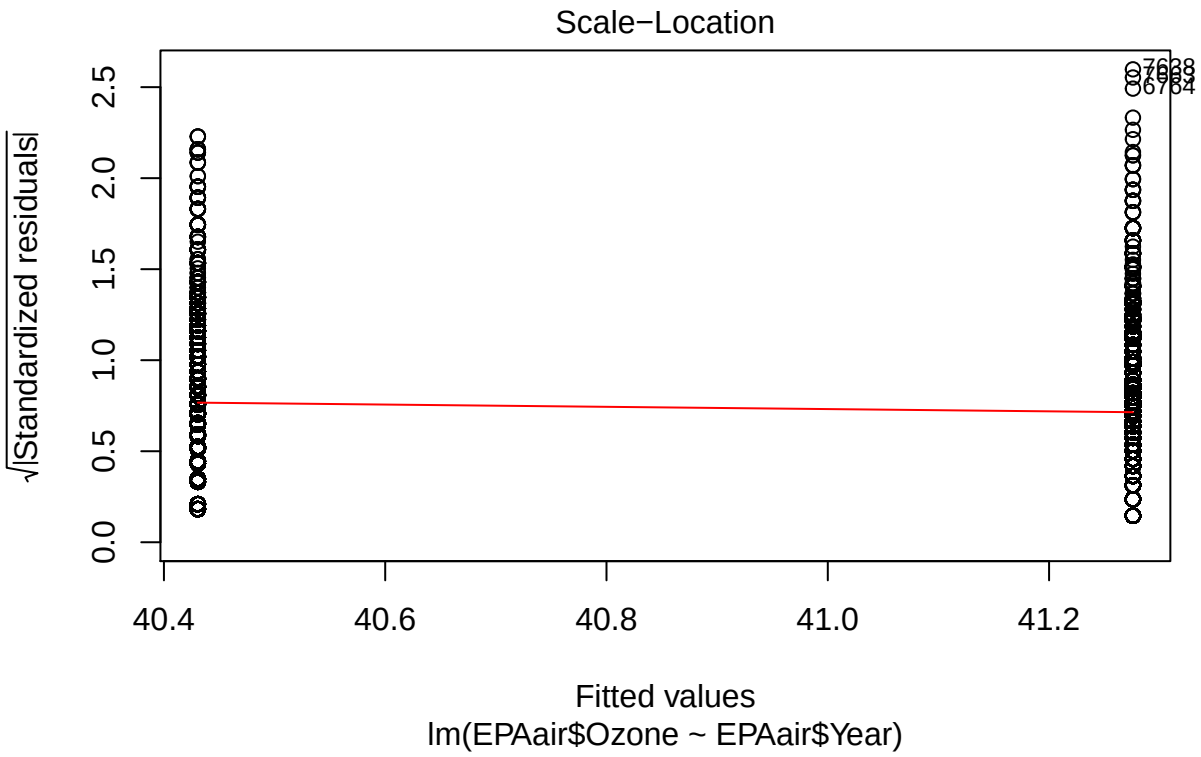
```
# Format as a GLM
O3.twosample2 <- lm(EPAair$Ozone ~ EPAair$Year)
summary(O3.twosample2)
```

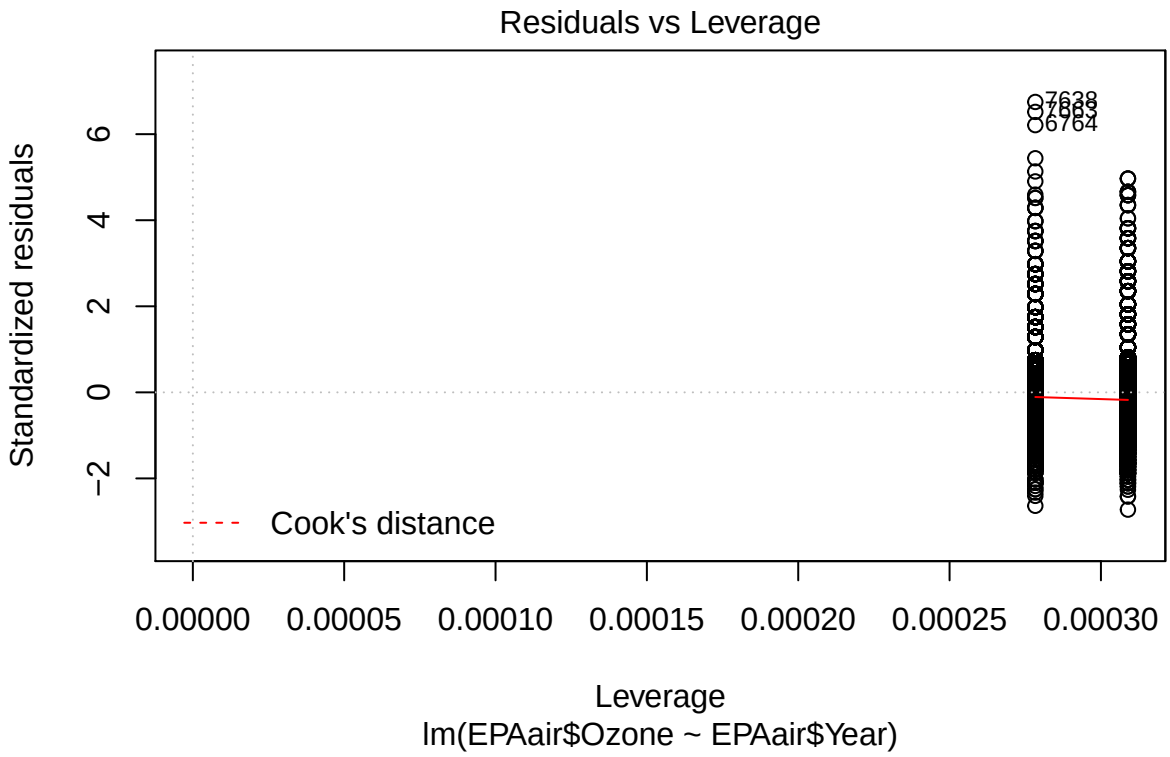
```
##
## Call:
## lm(formula = EPAair$Ozone ~ EPAair$Year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.431  -8.431  -0.431   5.569  87.724
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1665.1192   635.9203  -2.618  0.00885 **
## EPAair$Year     0.8452     0.3150   2.683  0.00732 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13 on 6828 degrees of freedom
##   (2146 observations deleted due to missingness)
## Multiple R-squared:  0.001053,   Adjusted R-squared:  0.0009066
## F-statistic: 7.197 on 1 and 6828 DF,  p-value: 0.00732
```

```
plot(O3.twosample2)
```


Residuals vs Fitted
Residuals
80
60
40
20
0
−40
7638
7663
6764
40.4
40.6
40.8
41.0
41.2
Fitted values
lm(EPAair$Ozone ~ EPAair$Year)


Normal Q−Q
Standardized residuals
−2
0
2
4
6
−4
−2
0
2
4
Theoretical Quantiles
lm(EPAair$Ozone ~ EPAair$Year)
7638
7663
6764


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
40.4
40.6
40.8
41.0
41.2
Fitted values
lm(EPAair$Ozone ~ EPAair$Year)



Non-parametric equivalent of t-test: Wilcoxon test

When we wish to avoid the assumption of normality, we can apply *distribution-free*, or non-parametric, methods in the form of the Wilcoxon rank sum (Mann-Whitney) test. The Wilcoxon test replaces the data by their rank and calculates the sum of the ranks for each group. Notice that the output of the Wilcoxon test is more limited than its parametric equivalent.

```
O3.onesample.wilcox <- wilcox.test(EPAair$Ozone, mu = 50, alternative = "less")
O3.onesample.wilcox
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  EPAair$Ozone
## V = 3116908, p-value < 2.2e-16
## alternative hypothesis: true location is less than 50
```

```
O3.twosample.wilcox <- wilcox.test(EPAair$Ozone ~ EPAair$Year)
O3.twosample.wilcox
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  EPAair$Ozone by EPAair$Year
## W = 5399454, p-value = 3.164e-07
## alternative hypothesis: true location shift is not equal to 0
```

Visualization and interpretation challenge

Create three plots, each with appropriately formatted axes and legends. Choose a non-default color palette.

1. geom_density of ozone divided by year (distinguish between years by adding transparency to the geom_density layer).
2. geom_boxplot of ozone divided by year . Add letters representing a significant difference between 2018 and 2019 (hint: stat_summary).
3. geom_violin of ozone divided by year, with the 0.5 quantile marked as a horizontal line. Add letters representing a significant difference between 2018 and 2019.

Now, write a summary of your findings, incorporating statistical output, reference to the figure(s), and a contextual interpretation.