



ENVIRONMENTAL DATA ANALYTICS: M7 – GENERALIZED LINEAR MODELS

M6.1- Basics of GLMs

- What are GLMs?
- Hypothesis testing
- Simple Linear Regression (“lm”)
 - Principles
 - Running in R
 - Interpreting results: stats and plots

Terminology

Term	Use
Response	Variable we are trying to predict ("dependent variable" or "target")
Independent variable	A variable used to predict the response ("predictor", "feature")
Record	Vector of predictor(s) and outcome value from an observation
Intercept	Predicted value when $X = 0$
Regression Coefficient	Slope of the regression line
Fitted values	Estimates of Y obtained from the regression line (aka "prediction")
Residuals	Difference between observed and fitted values (errors)
Least Squares	Method used to find line that minimizes squared sum of residuals

General workflow

- View data: Scatterplot of Y vs X
 - ▣ Can you see a trend?
 - ▣ Transform an axis?
- Create the linear model
 - ▣ Finds the best fit line (ordinary least squares method)
 - ▣ Assumes residuals are normal; sensitive to outliers
 - ▣ Assumes causation
- Examine the model summary & plots

Interpreting results

```
> summary(irradiance.regression)
```

Call:

```
lm(formula = irradiancewater ~ depth, data = PeterPaul.chem.nutrients)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-456.67	-142.62	-39.85	91.13	1375.43

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	484.5698	3.1509	153.8	<2e-16 ***
depth	-95.6492	0.8947	-106.9	<2e-16 ***

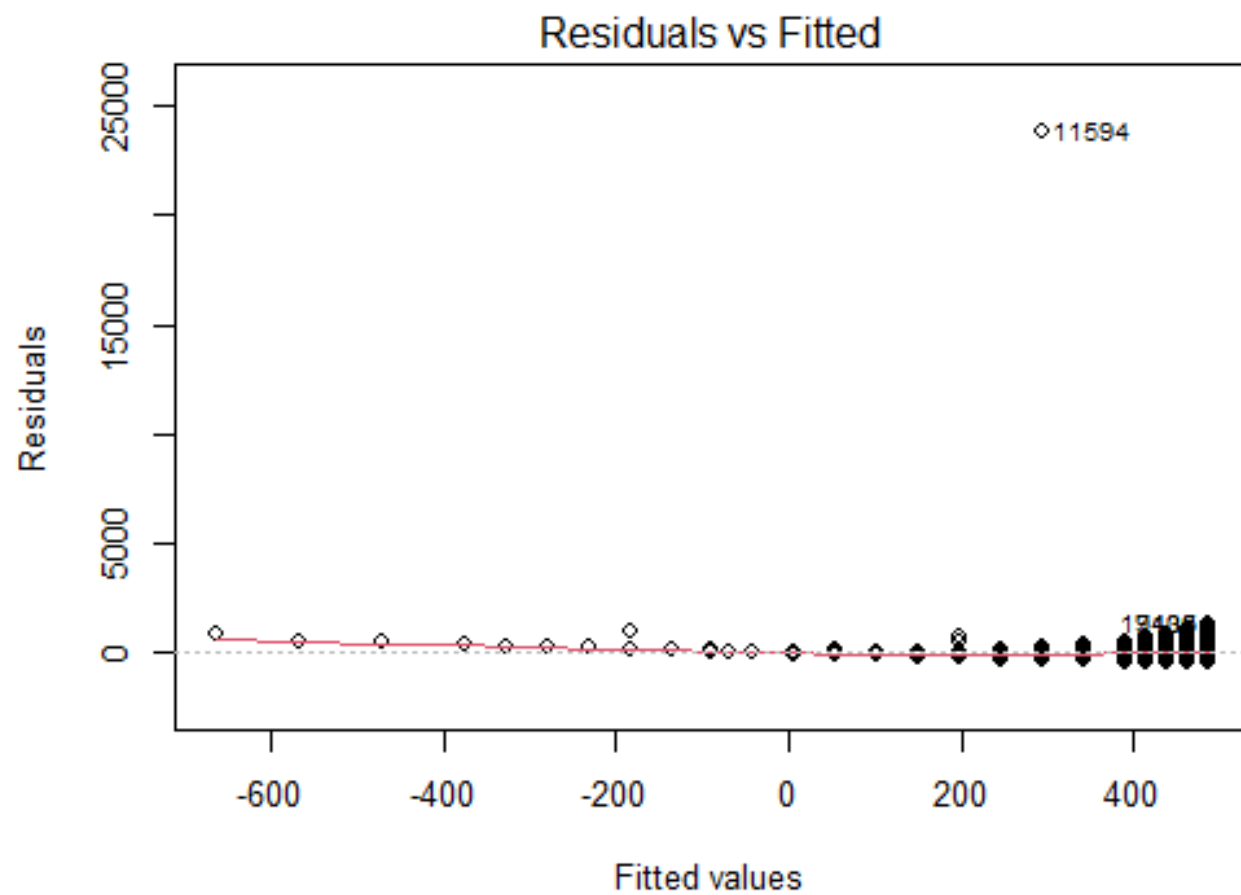
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 235.3 on 15445 degrees of freedom

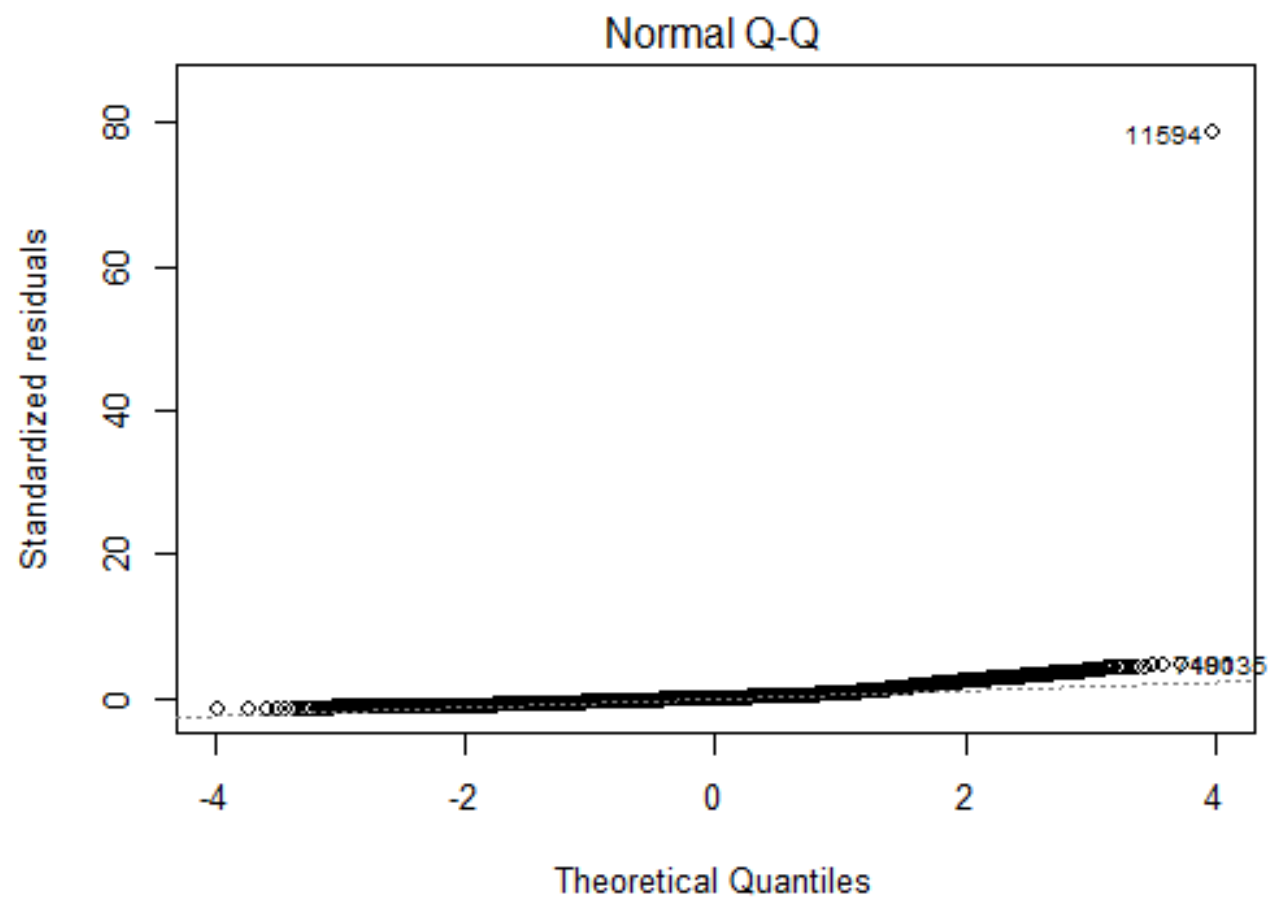
Multiple R-squared: 0.4253, Adjusted R-squared: 0.4252

F-statistic: 1.143e+04 on 1 and 15445 DF, p-value: < 2.2e-16

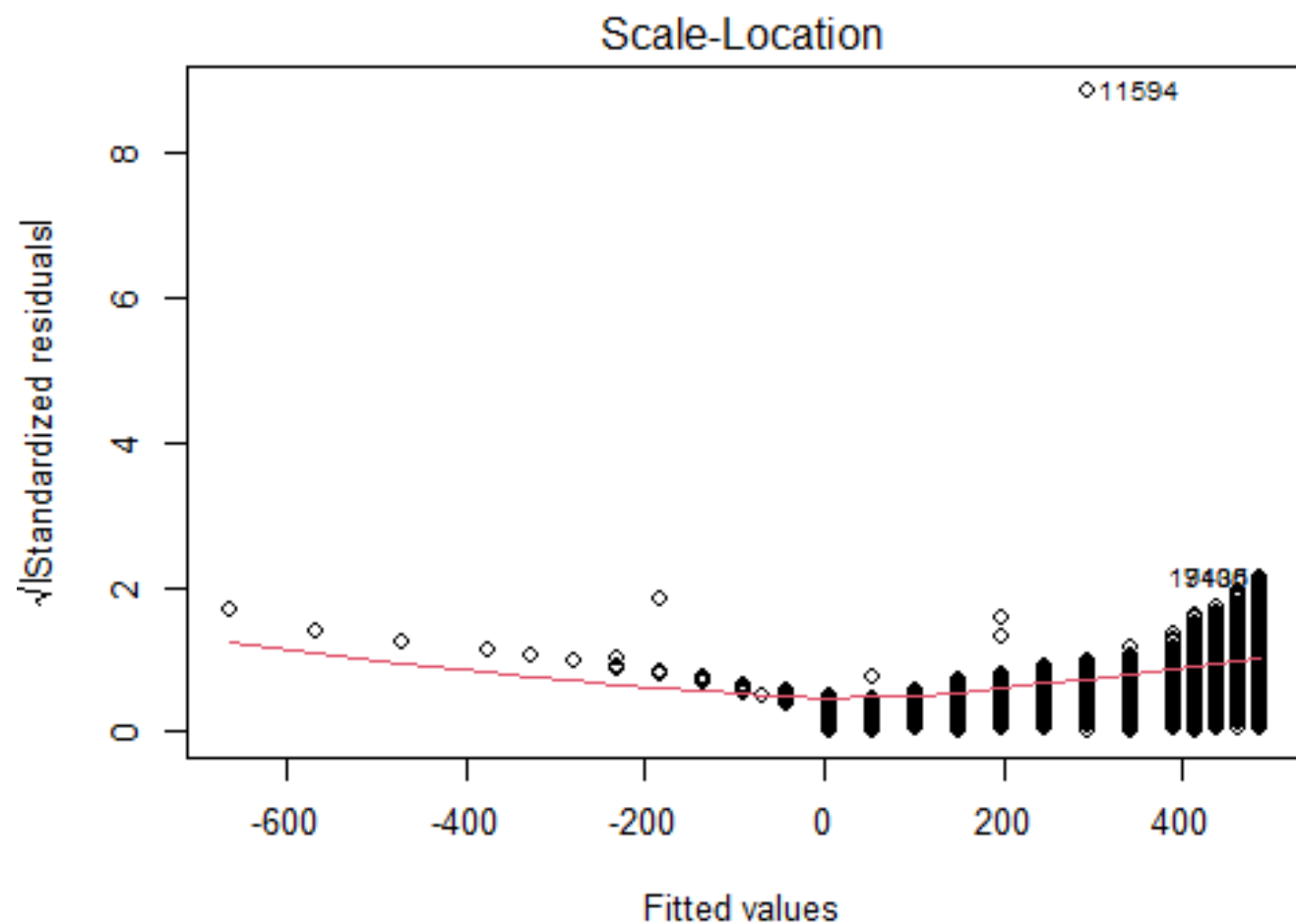
Plots...



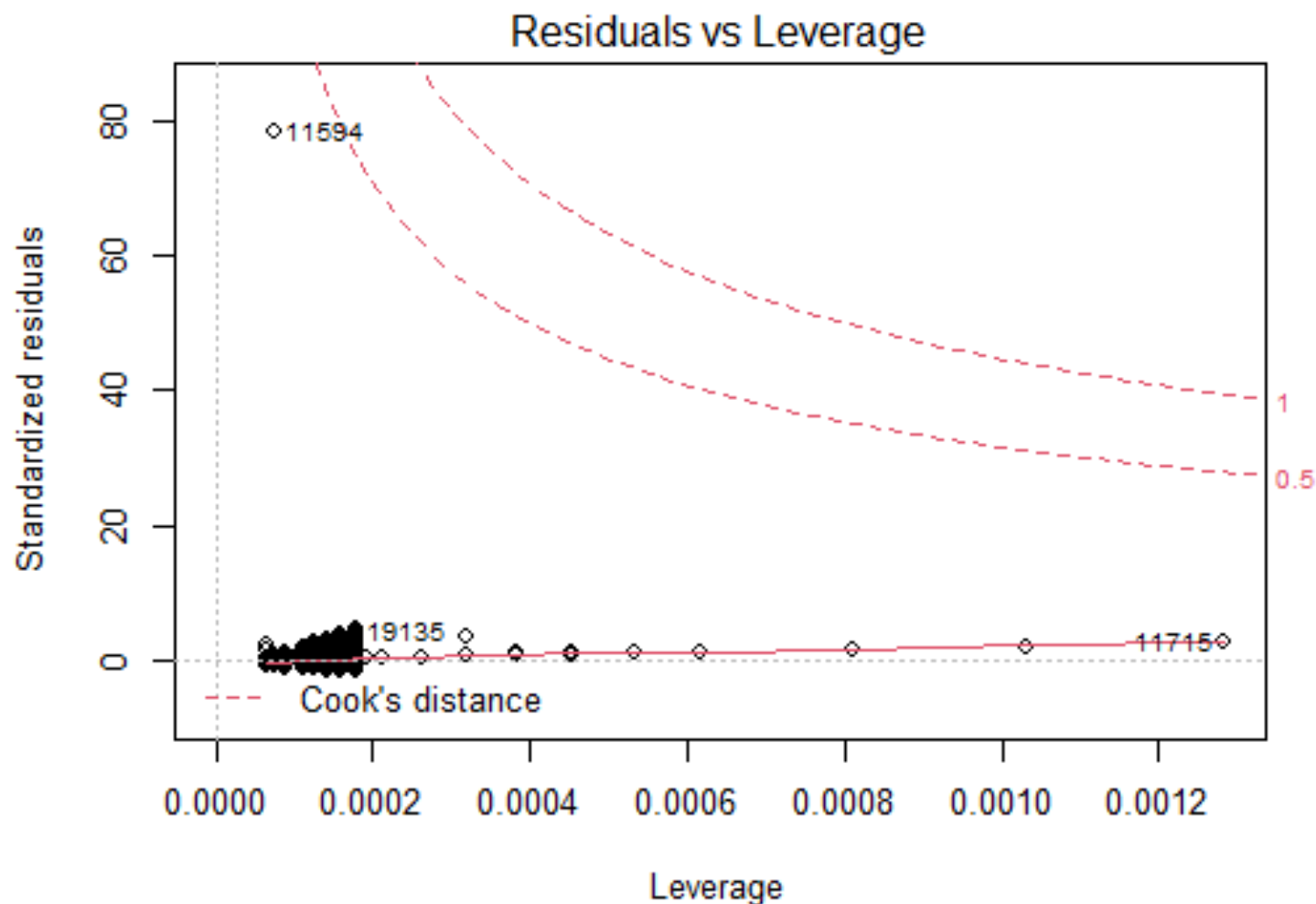
Plots...



Plots...



Plots...



Multiple Linear Regression

- Many independent variables to predict “y”
- Correlation matrices
- Issue of overfitting...
- Akaike's Information Criterion (AIC)

Multiple Linear Regression: Workflow

- Generate linear model (``lm``)
- Apply ``step()`` function to resulting model
 - ▣ Note initial AIC
 - ▣ Note change in AIC with removal (addition) of single terms
 - ▣ If AIC decreases with removal, then remove the term(s) and re-run ``lm``
 - ▣ Repeat: ``step()`` will suggest final linear regression model
- Run suggested model and report findings: Does R^2 increase?

M6.2 – ANOVA

- Predicting Y from categorical variables
- Terminology

Terminology

- **Factor:** A variable used to group data, suspected to explain variability in another [response] variable.
 - ▣ Example: Land cover from which a litter sample was collected

- **Levels:** The different values found in the factor
 - ▣ Example: *Forest, Wetland, Shrub*

- **Balanced Design:**
 - ▣ All *levels* have equal number of observations

ANOVA: Assumptions

- Populations are normally distributed
- Variances are equal
- Observations are independent

ANOVA: Litter biomass across sites

- Group data by factor (plot, date, land cover class)
- Compute sum of dry mass across combos of factors
- Examine summaries
 - ▣ Value ranges and variance, factor levels
- Assess assumptions
 - ▣ Population sizes equal? No...
 - ▣ Normality? Shapiro test → Only two sites..
 - ▣ Normality? QQ Plot → Not normal
 - ▣ Equal variance? Bartlett test → Not normal
- Compute ANOVA: `AOV`

ANOVA: Results

“aov”

```
> Litter.Totals.anova <- aov(data = Litter.Totals, dryMass ~ plotID)
> summary(Litter.Totals.anova)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
plotID	11	7584	689.5	4.813	1.45e-06	***
Residuals	198	28363	143.2			

“lm”

```
Call:
lm(formula = dryMass ~ plotID, data = Litter.Totals)

Residuals:
    Min       1Q   Median       3Q      Max
-18.586  -5.419  -1.529   1.964   59.821

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    15.680     2.746   5.711 4.08e-08 ***
plotIDNIWO_041     1.299     4.061   0.320 0.749396
plotIDNIWO_046    -5.724     3.996  -1.432 0.153580
plotIDNIWO_047   -11.204     4.134  -2.710 0.007315 **
plotIDNIWO_051   -10.011     4.061  -2.465 0.014546 *
plotIDNIWO_057     5.006     3.937   1.272 0.205013
plotIDNIWO_058   -13.282     3.883  -3.420 0.000760 ***
plotIDNIWO_061    -2.494     3.937  -0.633 0.527140
plotIDNIWO_062   -12.632     3.883  -3.253 0.001342 **
plotIDNIWO_063   -13.286     3.937  -3.375 0.000888 ***
plotIDNIWO_064    -7.664     3.883  -1.974 0.049805 *
plotIDNIWO_067    -3.114     4.061  -0.767 0.444110
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.97 on 198 degrees of freedom
Multiple R-squared:  0.211,    Adjusted R-squared:  0.1671
F-statistic: 4.813 on 11 and 198 DF,  p-value: 1.452e-06
```


ANOVA: *Post Hoc* tests

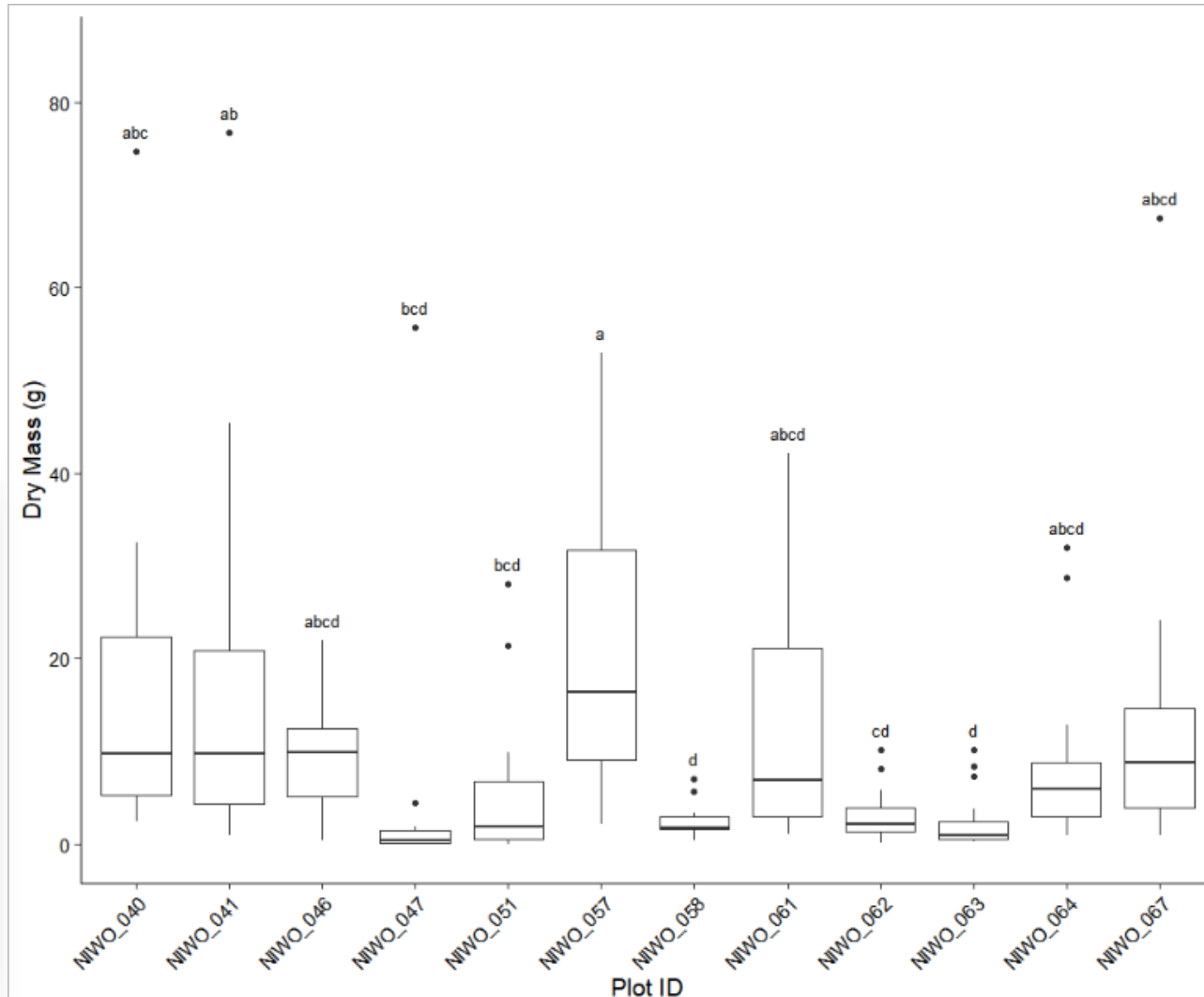
- If means are found not to be the same, which are different?
- Tukey HSD → Compares all pairwise combinations
 - ▣ Computes diff of mean, overall lower and upper values
 - ▣ Finds groups

```
$groups
      dryMass groups
NIWO_057 20.685833      a
NIWO_041 16.979063     ab
NIWO_040 15.680000    abc
NIWO_061 13.186111   abcd
NIWO_067 12.565938   abcd
NIWO_046  9.956176   abcd
NIWO_064  8.015789   abcd
NIWO_051  5.668750    bcd
NIWO_047  4.476333    bcd
NIWO_062  3.047632     cd
NIWO_058  2.398421      d
NIWO_063  2.393889      d
```

ANOVA: *Post Hoc* tests

□ Box plots!

\$groups	dryMass	groups
NIWO_057	20.685833	a
NIWO_041	16.979063	ab
NIWO_040	15.680000	abc
NIWO_061	13.186111	abcd
NIWO_067	12.565938	abcd
NIWO_046	9.956176	abcd
NIWO_064	8.015789	abcd
NIWO_051	5.668750	bcd
NIWO_047	4.476333	bcd
NIWO_062	3.047632	cd
NIWO_058	2.398421	d
NIWO_063	2.393889	d



Two-way ANOVA

- Do samples have different mean dry mass among groupings by **functional group** and **NLCD class**?

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
functionalGroup	7	6193	884.7	71.540	< 2e-16	***
nlcdClass	2	223	111.7	9.033	0.000125	***
Residuals	1682	20800	12.4			

- Interactive effects...

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
functionalGroup	7	6193	884.7	72.445	< 2e-16	***
nlcdClass	2	223	111.7	9.147	0.000112	***
functionalGroup:nlcdClass	14	431	30.8	2.521	0.001444	**
Residuals	1668	20369	12.2			

Two-way ANOVA: Post Hoc

- Tukey's HSD
- Create interaction list (all combinations):
- Run ANOVA on that...
- Run HSD.test on ANOVA result^{†c}
- Find functional groups...

	dryMass	groups
Needles.evergreenForest	7.431888889	a
Needles.grasslandHerbaceous	5.178888889	b
Needles.shrubScrub	4.406288660	bc
Mixed.shrubScrub	2.266184211	cd
Twigs/branches.evergreenForest	2.079294118	d
Mixed.evergreenForest	1.624375000	d
Woody material.evergreenForest	1.203936170	d
Mixed.grasslandHerbaceous	1.129000000	d
Twigs/branches.grasslandHerbaceous	0.949900000	d
Twigs/branches.shrubScrub	0.479583333	d
Woody material.shrubScrub	0.127968750	d
Flowers.evergreenForest	0.119625000	d
Other.grasslandHerbaceous	0.096666667	d
Other.evergreenForest	0.084807692	d
Seeds.evergreenForest	0.073461538	d
Other.shrubScrub	0.066576087	d
Leaves.shrubScrub	0.058936170	d
Woody material.grasslandHerbaceous	0.048877551	d
Leaves.grasslandHerbaceous	0.030471698	d
Seeds.shrubScrub	0.028777778	d
Leaves.evergreenForest	0.016025641	d
Flowers.shrubScrub	0.015505618	d
Flowers.grasslandHerbaceous	0.005425532	d
Seeds.grasslandHerbaceous	0.005416667	d

M6.3 – T-test

- T-tests:
 - 1-sample & 2-sample;
 - 1-sided & 2-sided

Question

- On average, do daily ozone values in our data meet the air quality standards of 50 ppm?



One Sample T-Test

Tests for different response among samples in two groups...

One-sample T-test: Is the mean equal to *50 ppm*

- ▣ H_0 : The difference between sample mean and the value is zero
- ▣ H_a : The difference is NOT zero (two-sided);
The difference is GREATER THAN zero (one-sided);
The difference is LESS THAN zero (one-sided);

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

T-test: Workflow

- **State the hypothesis:**
 - ▣ H_0 : Mean ozone is $\geq 50\text{ppm}$ (*one-sided*)
 - ▣ H_a : Mean ozone is $<$ than 500ppm
- **Examine the data:**
 - ▣ What is the reported mean *of our sample*?
- **Test for normality** (Shapiro-Wilks; histogram; QQplot)
- **T-test** (one-tail?)
- **Summarize results**
 - ▣ Put result into words
 - ▣ Reference the test used, the test-statistic, and the p-value

1-sample, 1-sided T-test: Output

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

Two-Sample T-Tests

- **Do two samples have different means?**
 - ▣ H_0 : Samples have the same mean
 - ▣ H_a : Samples have different means

- **Assumptions:**
 - ▣ Normal distributions
 - ▣ Similar variances

2-sample T-test result

□ As T-test

```
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 between group 2018 and group 2019 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
```

□ As linear model→

```
Call:
lm(formula = EPAair$Ozone ~ EPAair$Year)

Residuals:
    Min       1Q   Median       3Q      Max
-35.431  -8.431  -0.431   5.569  87.724

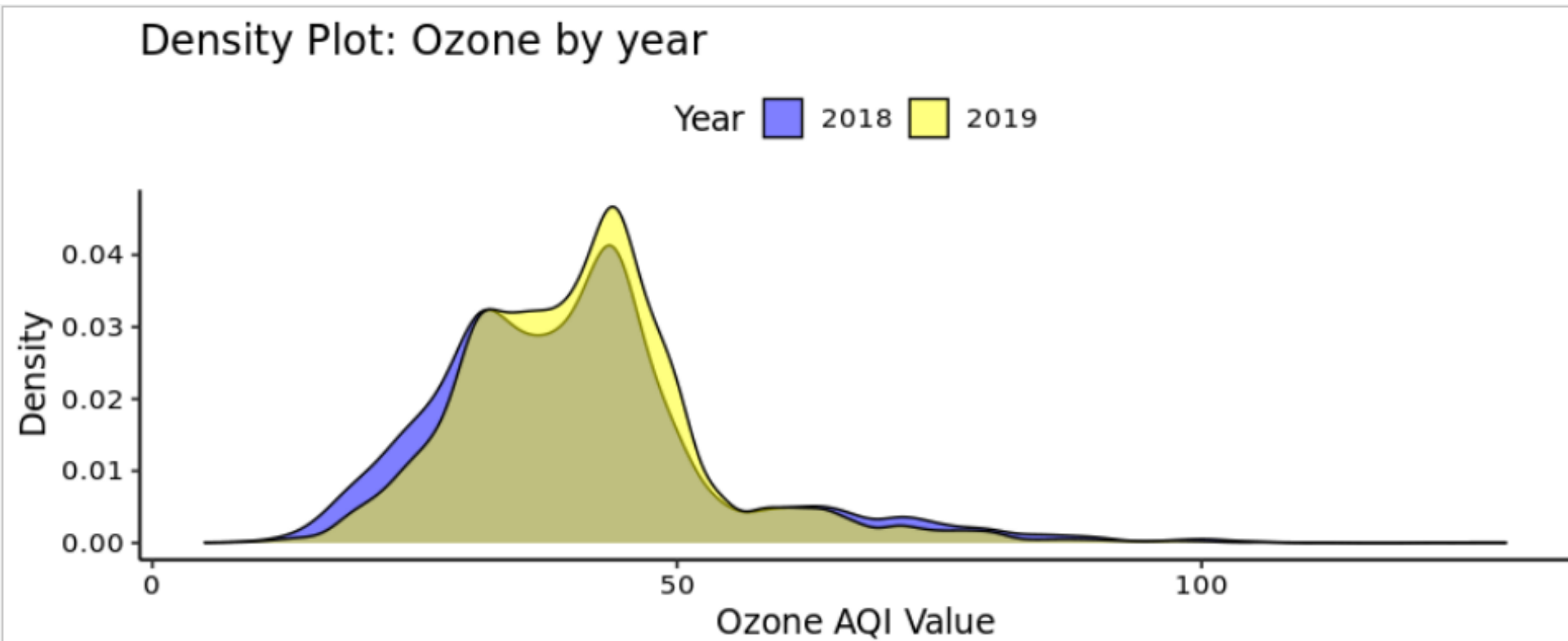
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1665.1192   635.9203  -2.618  0.00885 **
EPAair$Year    0.8452    0.3150   2.683  0.00732 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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
```

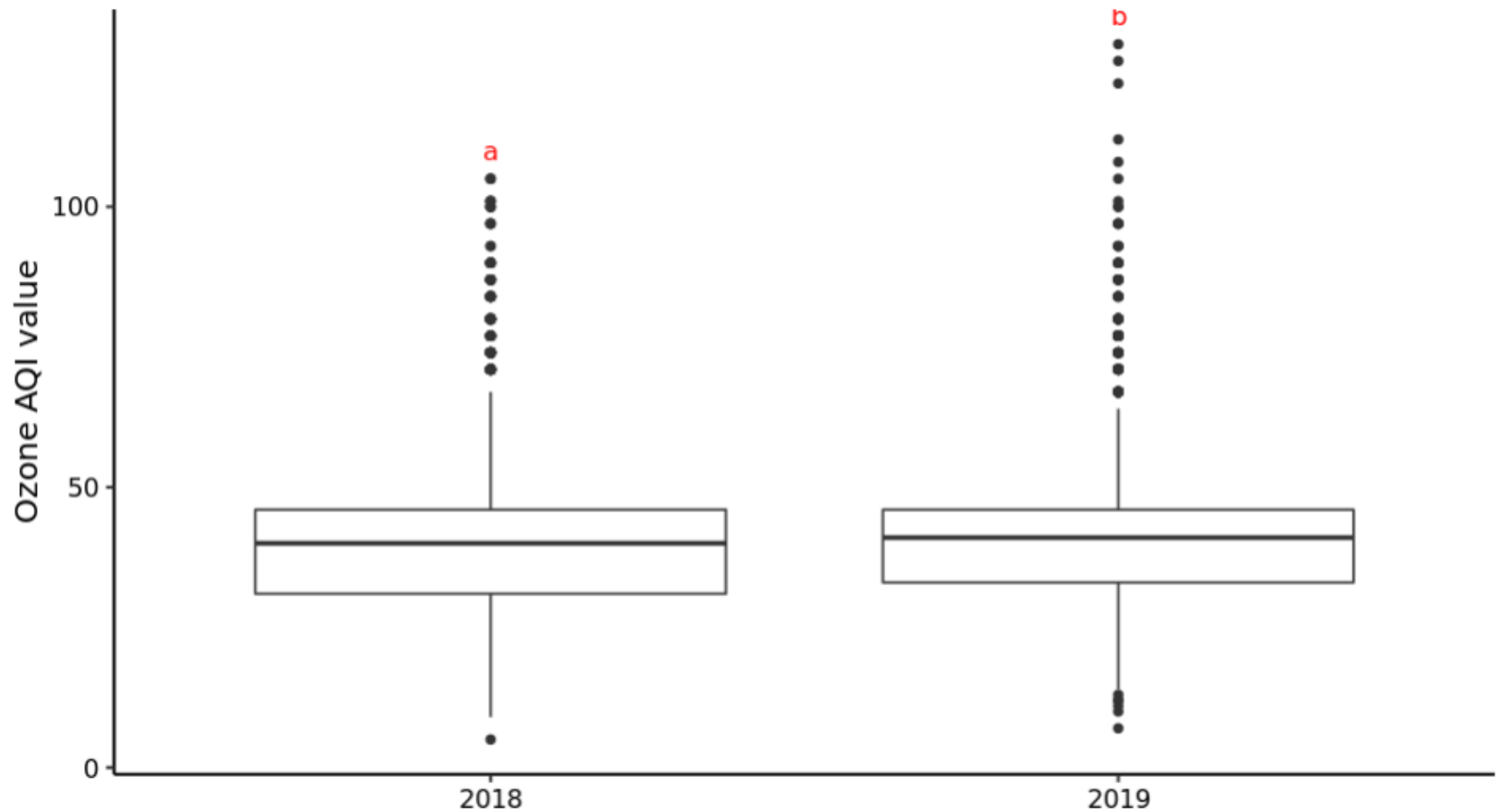
M6.3 - Exercises

- Exercises...
 - Linear regression

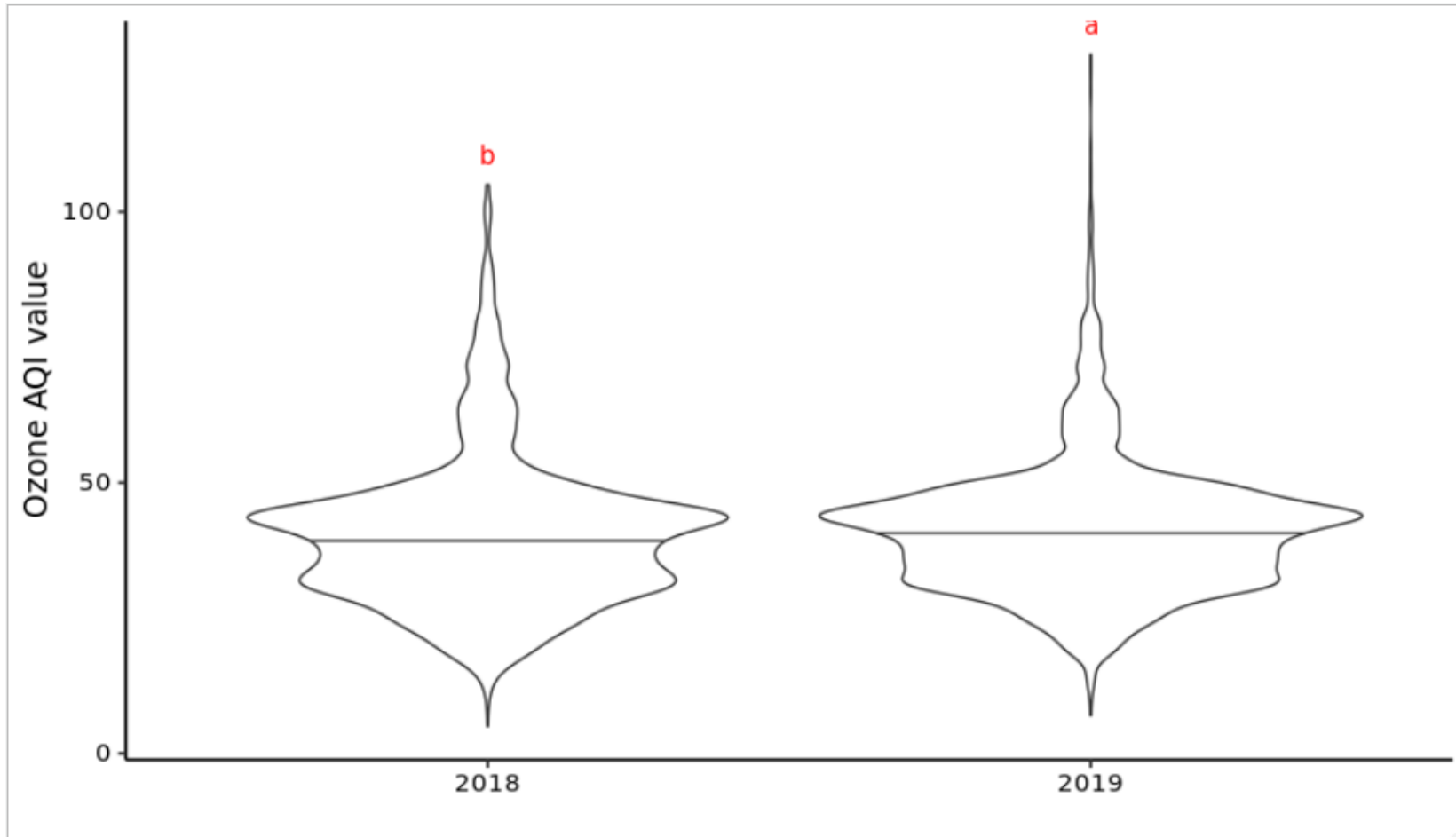
Exercise 2: Density plot



Exercise 2: Box plot

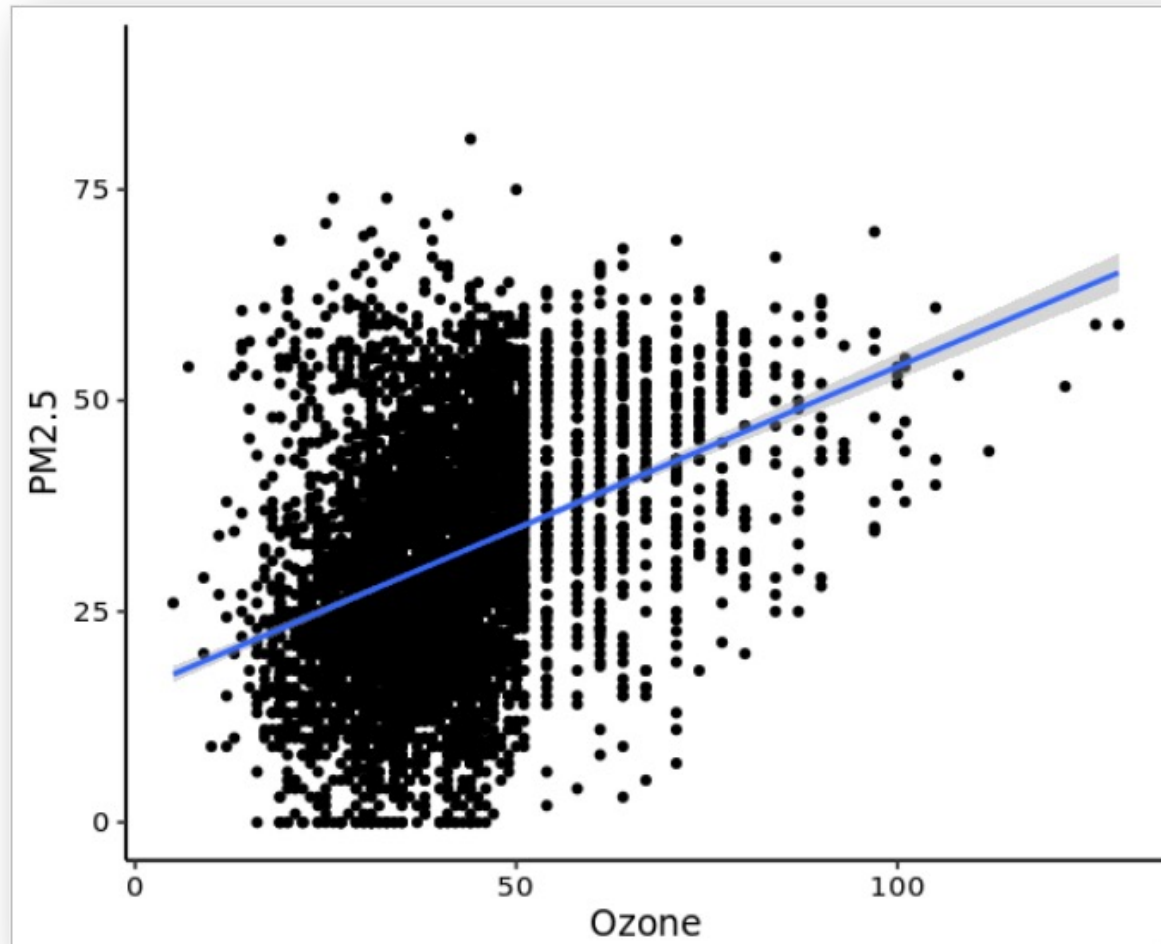


Exercise 2: Violin plot



Exercise 3&4: Linear Regression

- Can we predict PM2.5 from Ozone?



Exercise 3&4

Call:

```
lm(formula = PM2.5 ~ Ozone, data = EPAair)
```

Residuals:

Min	1Q	Median	3Q	Max
-37.204	-8.931	-0.613	8.463	48.473

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	15.63824	0.55556	28.15	<2e-16 ***
Ozone	0.38384	0.01298	29.58	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.06 on 5774 degrees of freedom

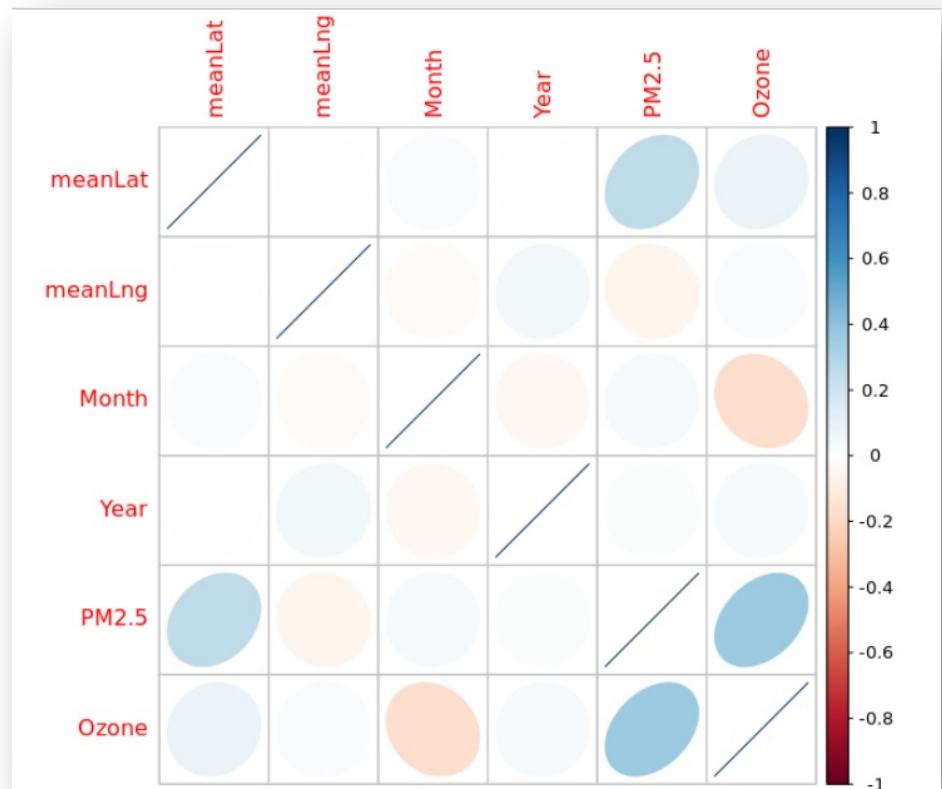
(3200 observations deleted due to missingness)

Multiple R-squared: 0.1316, Adjusted R-squared: 0.1314

F-statistic: 874.9 on 1 and 5774 DF, p-value: < 2.2e-16

Exercise 5: Correlation matrix

- Tip:
 - ▣ Subset dataframe to include numeric columns only
 - ▣ Remove NAs



Exercise 6: Stepwise AIC

PM2.5 ~

Ozone + Year + Month + SITE_LATITUDE + SITE_LONGITUDE

All Terms

```
Residual standard error: 12.6 on 5770 degrees of freedom  
(3200 observations deleted due to missingness)  
Multiple R-squared:  0.1927,    Adjusted R-squared:  0.192  
F-statistic: 275.5 on 5 and 5770 DF,  p-value: < 2.2e-16
```

Trimmed...

```
Residual standard error: 12.6 on 5771 degrees of freedom  
(3200 observations deleted due to missingness)  
Multiple R-squared:  0.1926,    Adjusted R-squared:  0.192  
F-statistic: 344.2 on 4 and 5771 DF,  p-value: < 2.2e-16
```