

# Lab 4: Null Hypothesis Significance Testing and Linear Models

Melissa Pulley;

2024-02-06

## 1. Before you start

Make sure you worked through the week 4 examples! You will need to understand the code and the concepts behind



You will need to submit your code. Please answer the questions by annotating your answers in the code (using the pound # symbol)



These boxes will inform you of things you need to submit or questions you need to answer!

## 2. Confidence Intervals

Before we start, We will simulate a population. This simulated population has 5,438 mussels (Green Floater) in a stream. We did this in class already. Run the following code: Size is represented by  $X \sim N(\mu = 40, \sigma = 6)$

```
set.seed(123)
pop <- rnorm(n=5438, mean=40, sd=6)
```

Now, we already did something similar as well, but we are going out and sampling 15 individuals

```
set.seed(23)
SnorkSample1 <- sample(pop, size=15)
SnorkSample1
```

```
[1] 34.70203 38.78461 45.87414 40.86444 37.31775 38.73252 40.30039 36.54749 [9] 38.15778 49.49907 32.49237
59.74310 35.37099 37.55291 37.10394
```

Now, try to obtain the mean, standard deviation, and n of the sample (check the *classexample2.Rmd* file if you need help):

```
X_bar <- mean(SnorkSample1)
X_bar
```

```
[1] 40.2029
```

```
s <- sd(SnorkSample1)
s
```

```
[1] 6.86906
```

```
n <- length(SnorkSample1)
n
```

```
[1] 15
```

And now, calculate the lower and upper confidence interval using the qt

```
lowerCI <- X_bar - qt(0.975, df=n-1)*s/sqrt(n)
lowerCI
```

```
[1] 36.39894
```

```
upperCI <- X_bar + qt(0.975, df=n-1)*s/sqrt(n)
upperCI
```

[1] 44.00686 OK, now, assume we have a second population of the Green Floater mussel in a different body of water. This population has 1,819 mussels.

You believe that the lower abundance (and in this case, lower density) of mussels results in a larger size. The real population length parameter is represented  $X \sim N(\mu = 41.6, \sigma = 6)$

First, let's simulate this second population.

```
set.seed(111)
pop2 <- rnorm(n=1819, mean=41.6, sd=6)
```

Now, we are going out and sampling 15 individuals. Why 15? In this case let's assume it is because sampling is time consuming and difficult and you only get to sample 15 individuals.

```
set.seed(41)
SnorkSample2 <- sample(pop2, size=15)
SnorkSample2
```

```
[1] 41.93917 47.26936 65.95154 40.94948 53.39416 54.77494 40.33097 30.41462 [9] 43.03565 49.91905 44.70289
42.11576 30.73274 41.96217 30.72916
```



Q1. 3 pts. a) Estimate mean, sd, and n of your sample of population 2. b) estimate the lower and upper CI for this second population. c) assume you don't know the population values, and all you have to go on are your estimates and confidence intervals, would you say these populations are different in size? More importantly, would you say that the mussels from the population 2 are larger?

```
#a
X_bar2<- mean(SnorkSample2)
s2<-sd(SnorkSample2)
n2<-length(SnorkSample2)

#b
lowerCI2 <- X_bar2 - qt(0.975, df=n2-1)*s2/sqrt(n2)
upperCI2 <- X_bar2 + qt(0.975, df=n2-1)*s2/sqrt(n2)

#c
X_bar
```

```
[1] 37.08615
```

```
X_bar2
```

```
[1] 43.17932
```

```
s
```

```
[1] 5.333372
```

```
s2
```

```
[1] 4.683624
```

```
lowerCI
```

```
[1] 34.13262
```

```
lowerCI2
```

```
[1] 40.58562
```

```
upperCI
```

```
[1] 40.03967
```

```
upperCI2
```

```
[1] 45.77302
```



a1 c) I would say that the populations may be close in size, it appears that the second population may be larger.



Try plotting (or drawing in your notebook!) the points and CI's to make a good judgement of whether they are different!

### 3. Two sample t-test

Now, sometimes it's hard to come up with conclusions only using the confidence intervals, so let's do Null Hypothesis Statistical Testing. We are going to repeat the exercise but with an unpaired 2 sample t-test test.



Check last week's presentation and .rmd files for all the code you will need!

Remember the five steps of NHST? They are:

1. State your null and alternate hypothesis
2. Collect data: in this case we will collect data using code. Example: `sample(pop2, size=15)` is a way we collect 15 individuals
3. Perform a statistical test: unpaired 2 sample t-test. Think whether it should be 1 tail or 2 tails!!! Check the code and slides from week 4.
4. Decide whether to reject or fail to reject your null hypothesis
5. Present your findings.



Q2. 5 pts. a) Following the 5 steps of NHST, explore whether the population 2 has larger mussels than population 1. YOUR SAMPLE SIZE FOR BOTH SITES SHOULD BE 15! You should define what you did for each step. b) Are your conclusions different than what you obtained from Q1? c) Check the simulated population values (the real parameter mean and variance). Did your test accurately reflect reality, or did it committed an error? What type of error? Why do you think this happened?



Q2 a) Consider the following hypotheses.

$H_0$ : The average size of mussels in population 2 is less than or equal to the average size of mussels in population 1.

$H_1$ : The average size of mussels in population 2 is greater than the average size of mussels in population 1.

Now, we will collect a new sample from each population.

```
set.seed(321)
#a2 b
newSample1 <- sample(pop, size=15)
newSample2 = sample(pop2,size=15)
```



Q2 a) Since we are examining whether population 2 is larger than population 1, we only need a one-sided tail. We are only looking on "one-side" of the mean

```
#a2 t-test
t.test(x=newSample1, y=newSample2, alternative="less", var.equal=T, paired=F)
```

Two Sample t-test

data: newSample1 and newSample2 t = -0.76689, df = 28, p-value = 0.2248 alternative hypothesis: true difference in means is less than 0 95 percent confidence interval: -Inf 1.882351 sample estimates: mean of x mean of y 38.52561 40.07079

```
t.test(x=SnorkSample1, y=SnorkSample2, alternative='less', var.equal=T, paired=F)
```

Two Sample t-test

data: SnorkSample1 and SnorkSample2 t = -1.2025, df = 28, p-value = 0.1196 alternative hypothesis: true difference in means is less than 0 95 percent confidence interval: -Inf 1.52532 sample estimates: mean of x mean of y 40.20290 43.88144



a2 a). The p-value for this t-test is 0.2248. Since this value is above 0.05, we fail to reject the null hypothesis. This means that we do not have sufficient support to believe that the size of mussels in population 2 is greater than mussels in population 1.



a2 b). The results of the t-test disagree with my initial with my conclusions from question 1.



a2 c). Because we know that average size of mussels in population 2 is larger than the average size of mussels in population one, this is an error. Because we did not reject the null hypothesis, even though it was false, this is a type 2 error.

Now, let's explore what happens when we change the sample size.



Q3. 3 pts. a) Repeat question 2, subsection a, but with a sample size of 125 individuals for site 1, and 250 for site 2. b) Where there any differences with Q2? Why?



a3 a) Consider the same hypotheses:

$H_0$ : The average size of mussels in population 2 is less than or equal to the average size of mussels in population 1.

$H_1$ : The average size of mussels in population 2 is greater than the average size of mussels in population 1.

Now, we will collect a new, larger sample from each population.

```
set.seed(321)
#a3 a
LrgSample1 <- sample(pop, size=125)
LrgSample2 = sample(pop2,size=250)
```



a3 a) We will also use a one-sided test

```
#a3 c
t.test(x=LrgSample1, y=LrgSample2, alternative="less", paired=F)
```

Welch Two Sample t-test

data: LrgSample1 and LrgSample2 t = -2.5015, df = 256.35, p-value = 0.006495 alternative hypothesis: true difference in means is less than 0 95 percent confidence interval: -Inf -0.5260626 sample estimates: mean of x mean of y 39.88882 41.43572



a3 a) The p-value for this t-test is 0.006495. Since this value is below 0.05, we can reject the null hypothesis. This means we have significant support that the mean of mussel size of population 2 is larger than than population 1.



a3 b) This answer disagrees with question 2. Since our sample size for the t-test in question 3 is larger than in question 2, these results are more reliable.

I hope this section helped you think of the effects of sample size on the power of your test!



Before continuing, think... what else might affect the power of the test?

We are now going to run this experiment once again, but follow these directions:

1. Population 1:  $X \sim N(\mu = 37.15, \sigma = 6)$  and your sample size will be 15
2. Population 2:  $X \sim N(\mu = 48.9, \sigma = 6)$  and your sample size will be 15

Simulate your populations, simulate your samples and run a t-test!



Q4. 3 pts. a) Repeat question 2, subsection a, but with the new population values. b) Where there any differences with Q2? Why? Your samples are small!



a4 a) Consider the same hypotheses again:

$H_0$ : The average size of mussels in population 2 is less than or equal to the average size of mussels in population 1.

$H_1$ : The average size of mussels in population 2 is greater than the average size of mussels in population 1.

Now, we will collect a sample from each new population.

```
set.seed(123)
newpop <- rnorm(n=5438, mean=37.15, sd=6)

set.seed(111)
newpop2 <- rnorm(n=1819, mean=48.9, sd=6)

set.seed(321)
#a3 a
NewPopSample1 <- sample(newpop, size=125)
NewPopSample2 = sample(newpop2, size=250)
```



a4 a) We will also use a one-sided test

```
#a3 c
t.test(x=NewPopSample1, y=NewPopSample2, alternative="less", paired=F)
```

Welch Two Sample t-test

data: NewPopSample1 and NewPopSample2 t = -18.915, df = 256.35, p-value < 2.2e-16 alternative hypothesis: true difference in means is less than 0 95 percent confidence interval: -Inf -10.67606 sample estimates: mean of x mean of y 37.03882 48.73572



a4 a) The p-value for this t-test is  $2.2 \times 10^{-16}$ , which is significantly below 0.05. This means we have significant support that the mean mussel size of population 2 is larger than population 1.



a4 b) This answer disagrees with question 2. Despite small samples, because the original means were further apart, the t-test was still able to determine that the mean mussel size of population 2 is smaller than population 1.

I hope this section helped you think of the effects of sample size on the power of your test!

We are now done with t-tests. Hopefully you're happy to be done with this (and not have to run another t-test), but even happier because you now understand some factors that affect the power of your statistical test! You should be thinking of 2 factors that affect it, and you can only control one.

## 4. Linear models

In this lab, we will be building a simple linear regression model, performing diagnostics of assumption violations, and interpreting the regression outputs and visualizing the regression relationship to answer a research question.

## 4.1 Data origin

Today we will use a dataset that has been processed by Dr. Xingli Giam. The dataset is:

LakeMendota.csv, and it contains information on the duration of lake ice cover in a given year at Lake Mendota in Madison, WI, for the years 1884 to 2019. You can download the dataset

Dr. Giam processed the data from two datasets of lake ice cover duration and air temperature in Madison, WI. These two datasets are available in the `lterdatasampler` package in R: `ntl_icecover` and `ntl_airtemp`. He processed and combined the datasets to get data on the annual ice cover duration (in terms of number of days) at Lake Mendota and the average daily air temperature in Madison during the cold season (days in November to April) when the ice is present on the lake, following the approach taken in the `ltersampler` vignette (see here: [Vignette](#)). The only differences are that the vignette averaged the ice cover duration between the two lakes, whereas Dr. Giam used data from only Lake Mendota. Furthermore, the vignette used a tidyverse package framework to process data, instead of basic R functions.

In the reference section of this document, I have posted the code Dr. Giam used to process the data. You will find many useful tools for data processing, and data wrangling!

## 4.2 Data

Download the processed dataset named `LakeMendota.csv` from Canvas.

`LakeMendota.csv` contains information on the duration of lake ice cover in a given year at Lake Mendota in Madison, WI, for the years 1884 to 2019.

For this dataset, there are six columns/variables:

1. `year`: the hydrological year of the lake ice duration and air temperature observation
2. `lakeid`: the name of the lake. There is only one value here: Lake Mendota
3. `ice_on`: ice\_on date for that hydrological year (yyyy-mm-dd format)
4. `ice_off`: ice\_off date for that hydrological year (yyyy-mm-dd format)
5. `ice_duration`: the number of days there is >50% ice cover on Lake Mendota
6. `temp`: average daily air temperature from November to April of that hydrological year

## 4.3 Data Analysis

We are working on three research questions in this lab: (1) Is there climate warming over the past century or so around Lake Mendota? If so, what is the rate of warming?; (2) Is there a decline in ice cover duration over this time? If so, what is the rate of decline in ice cover duration; (3) Is warming linked to a decline in ice cover duration? If so, what is the decline in ice cover duration per 1 deg C increase in air temperature?

Load the data set into `r` in a project named `dat`. We have done this in the past!

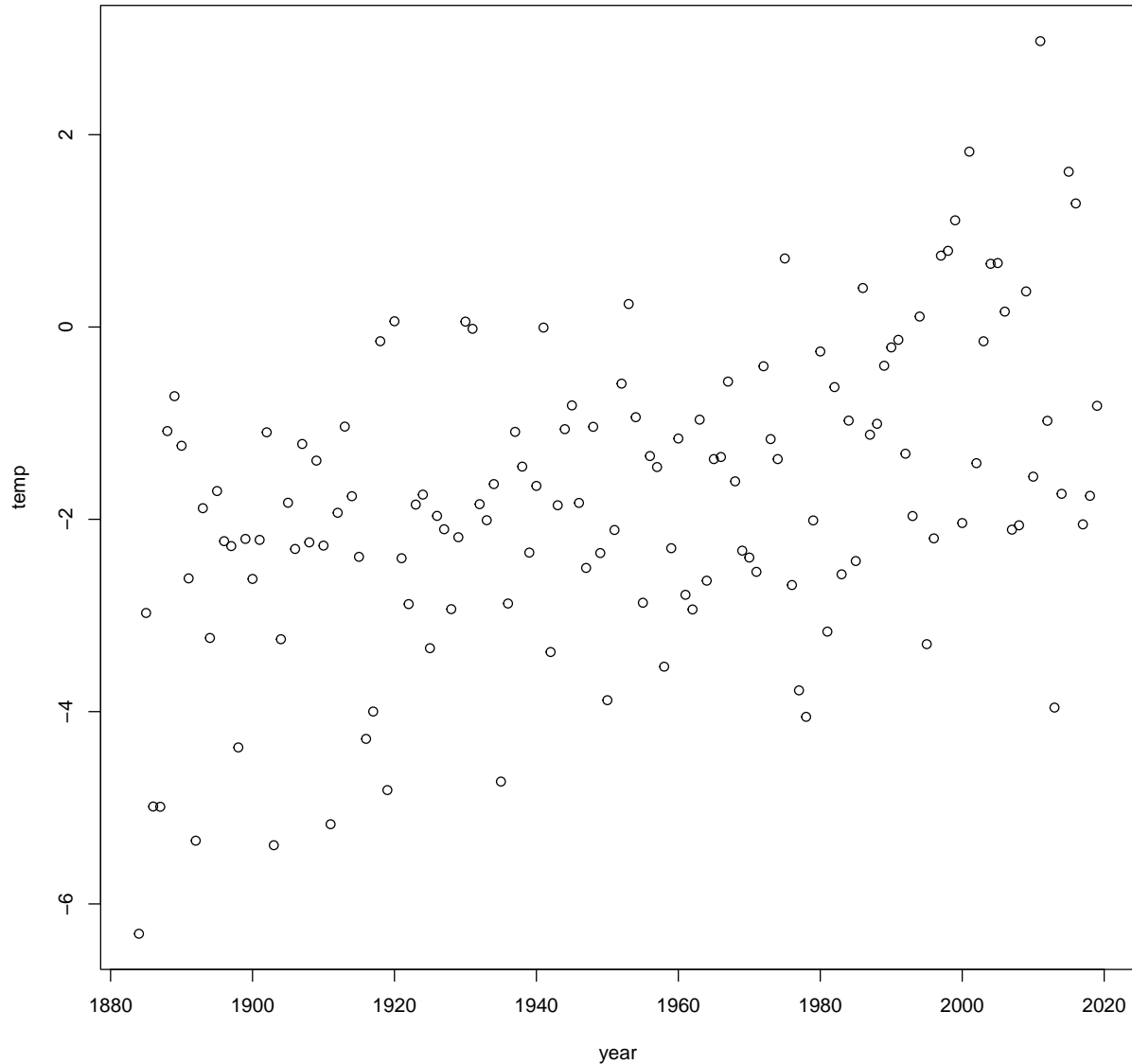


As we move forward, I will be supplying less of the code, however, if you check your previous assignments you can find most of the functions!

If the climate is warming over time, we would see an increase in air temperature over time. Let's plot air temperature on the y-axis and year on the x-axis.

```
dat<-read.csv("LakeMendota.csv")
```

```
plot(temp~year, dat) # this is the simplest plot
```



Let's adjust the axes. You can adjust the axes with `xlim=c(min,max)` and `ylim=c(min,max)`. You might also want to give a more formal axis title using `xlab` and `ylab`

```
plot(temp~year, ylim=c(-8,4), xlab='Hyrdological Observation Year', ylab= 'Average Daily Temperature', c
```



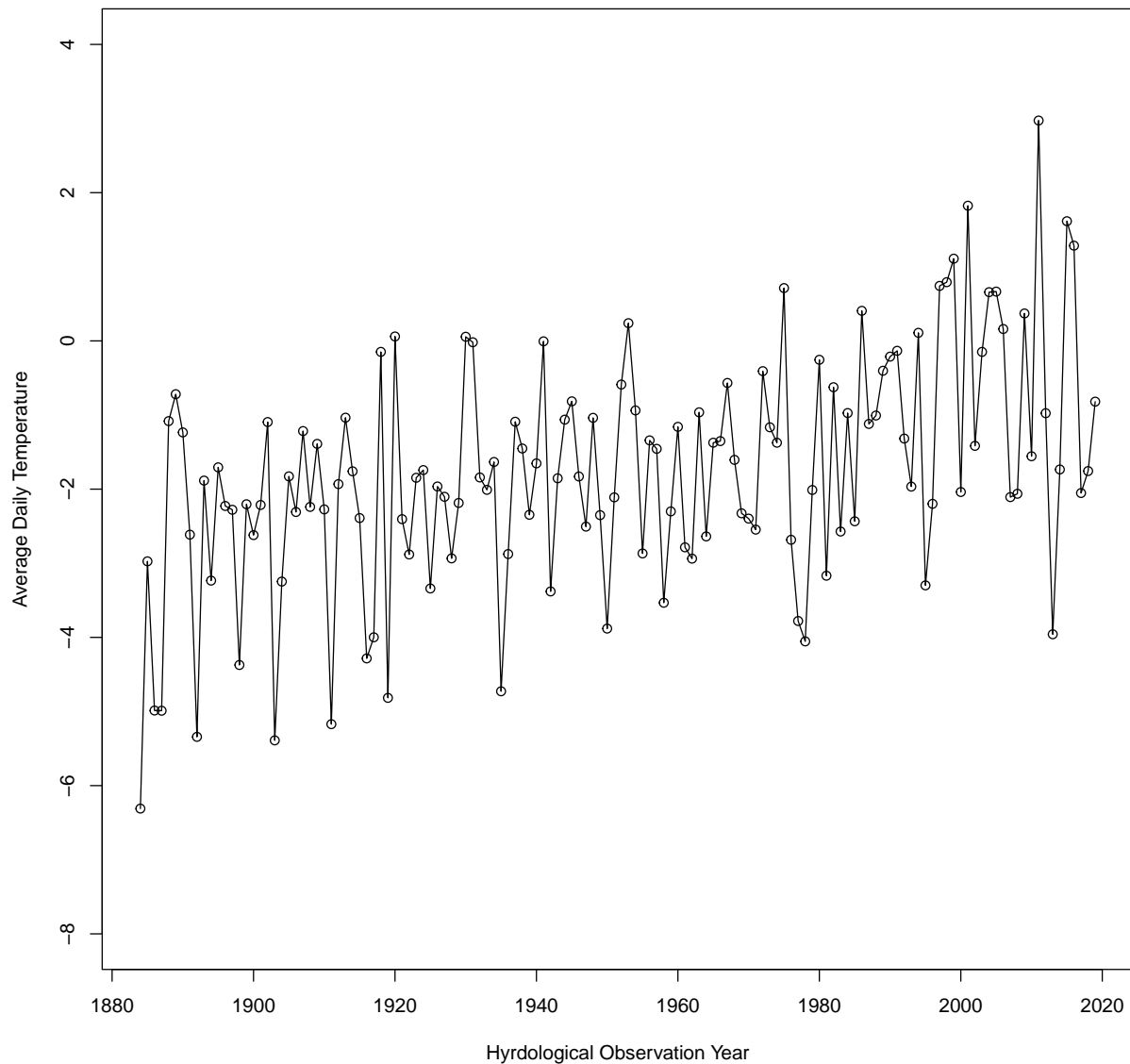
Q5. 1 pt. Upload your plot to Canvas with appropriate x and y labels

From the plot you have just made, you can see that there is an increasing trend through time. This is also called a time-series: repeated measurements of a given variable (air temperature) through regular time



intervals (yearly in this case). Because observations have a temporal (time) order to them, we might want to use lines to join the points through time to make this clear to the readers.

```
plot(temp~year, ylim=c(-8,4), xlab='Hyrdological Observation Year', ylab= 'Average Daily Temperature',  
lines(temp~year, dat))
```



#### 4.3.1 Running a linear model

Running a linear model is very simple. We can use `ml(y~x, data=...)` to fit a SLR model of  $y$  (the response) against  $x$  (the predictor). We can then use `summary` to get more information. In this case:

```
m1<-lm(temp~year,data=dat) #fit SLR model of temp against year, place in object m1  
summary(m1)
```

Call: `lm(formula = temp ~ year, data = dat)`

Residuals: Min 1Q Median 3Q Max -3.3478 -1.0755 0.1454 0.9051 3.6211

Coefficients: Estimate Std. Error t value Pr(>|t|)

(Intercept) -38.66047 5.89528 -6.558 1.09e-09 **year 0.01890 0.00302 6.258 4.88e-09** — Signif. codes: 0 ‘**0.001**’ ‘**0.01**’ ‘0.05’ ‘0.1’ ‘1’

Residual standard error: 1.383 on 134 degrees of freedom Multiple R-squared: 0.2262, Adjusted R-squared: 0.2204 F-statistic: 39.17 on 1 and 134 DF, p-value: 4.883e-09

This summary is super helpful! It gives you 6 main outputs:

1. Model formula (structure) and data.
2. Information about the residuals (observed Y - fitted Y) fitted Y represents the “predicted” or “expected” Y under the model.
3. Coefficients. This represent your  $\beta_0$  and your  $\beta_1$ . The estimate of each value, and a P-value (essentially, this p-val tells you whether it is significantly different from 0)
4. Residual Standard Error: The standard deviation for the normally distributed error term
5.  $R^2$  of the model. Which is the percentage of variance explained by the model
6. Our statistic, and model significance. Essentially, does this model is significantly better at explaining the variance than a null (no-effect) model?

Look at the distribution of the residuals. This tells us something about our assumptions. How do they look?

Look at the coefficients. The intercept is  $\beta_0$  and the slope (effect of year) is  $\beta_1$ . The larger  $\beta_1$ , the stronger of an effect the explanatory variable has on the response variable. Essentially, we can predict mean Air Temp by using the following equation:

$$Temp_i = \beta_0 + \beta_1(year)$$

We remove the error error  $\epsilon$  because the mean error is 0.



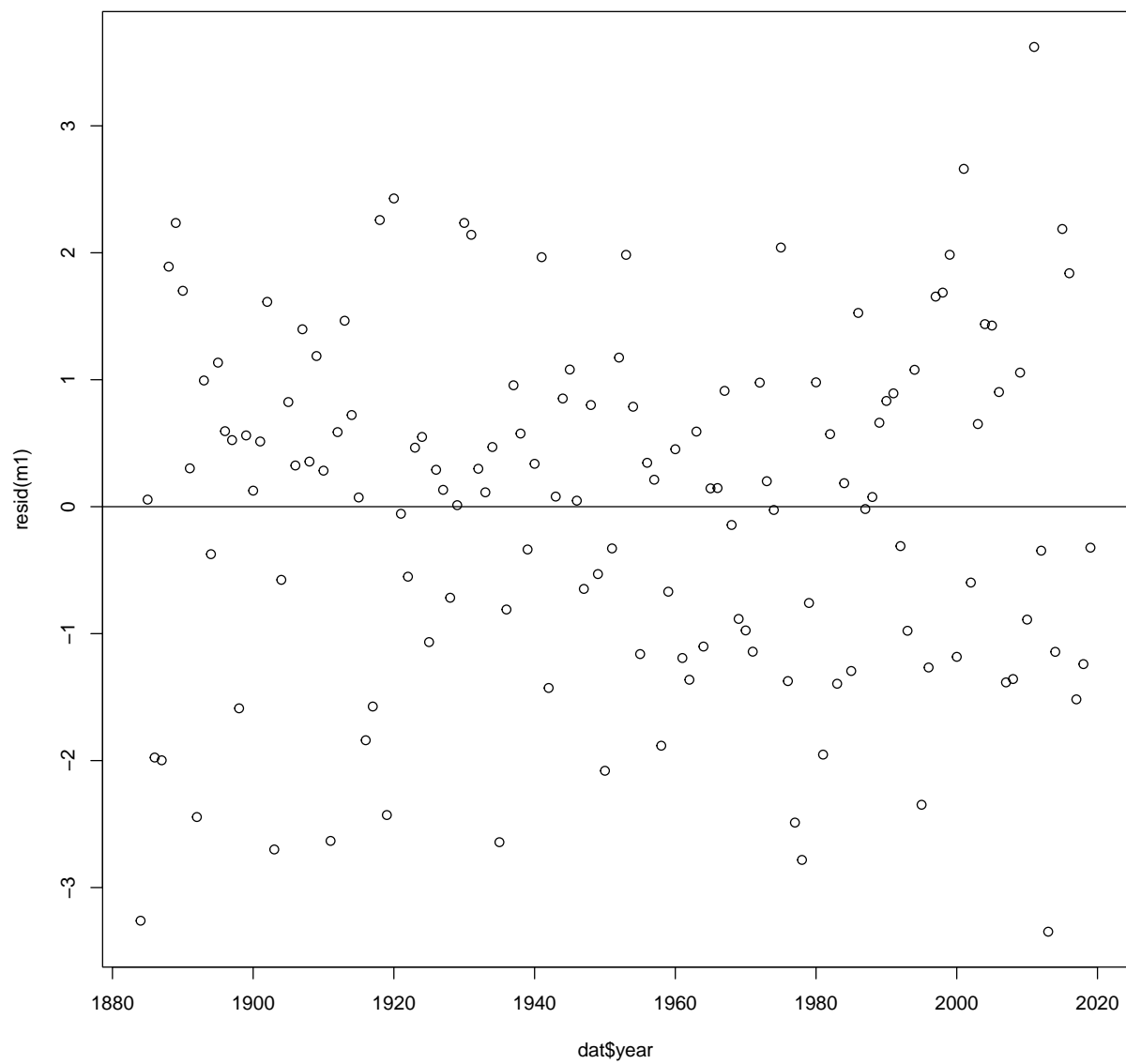
The following section is the most important part of this analysis. Make sure you UNDERSTAND it prior to the exam

Because from the coefficient value of Year (0.0189, which is significantly different from zero because ( $p \leq 0.05$ ), you can conclude that when Year increases by 1 unit (going from one year to the next year), average air temperature increases by 0.0189, and that this effect is statistically significant.



Q6. 3 pts. a) Looking at the residuals, explain whether they look OK, or if there seems to be a problem with their distribution, b) substitute the coefficient values, and using R, obtain and report the predicted temperature for year 2000, and for year 2001. What is the difference in predicted temperature between this years? Does it make sense? c) What is the model fit (or proportion of variance in temperature explained by the model)

```
plot(resid(m1)~dat$year) # resid(model object) generates the residual values
abline(h=0)
```



a6. a) The residuals appear to have a symmetric distribution around zero, which means they look okay and the mean error is 0.

```
beta0 = as.numeric(m1$coefficients[1])
beta1 = as.numeric(m1$coefficients[2])
predID2000 = beta0 + beta1*2000
predID2001 = beta0 + beta1*2001
predID2000
```

```
[1] -0.8559502
```

```
predID2001
```

```
[1] -0.837048
```

```
predID2001 - predID2000
```

```
[1] 0.01890226
```



a6 b). The predicted temperature for the year 2000 and 2001 are -0.8559502 and -0.837048 respectively. The difference between these data is 0.01890226 which makes sense this is the slope from the linear model.

\answerbox{a6 c). 22.62% of the variance in temperature is described by the model.}

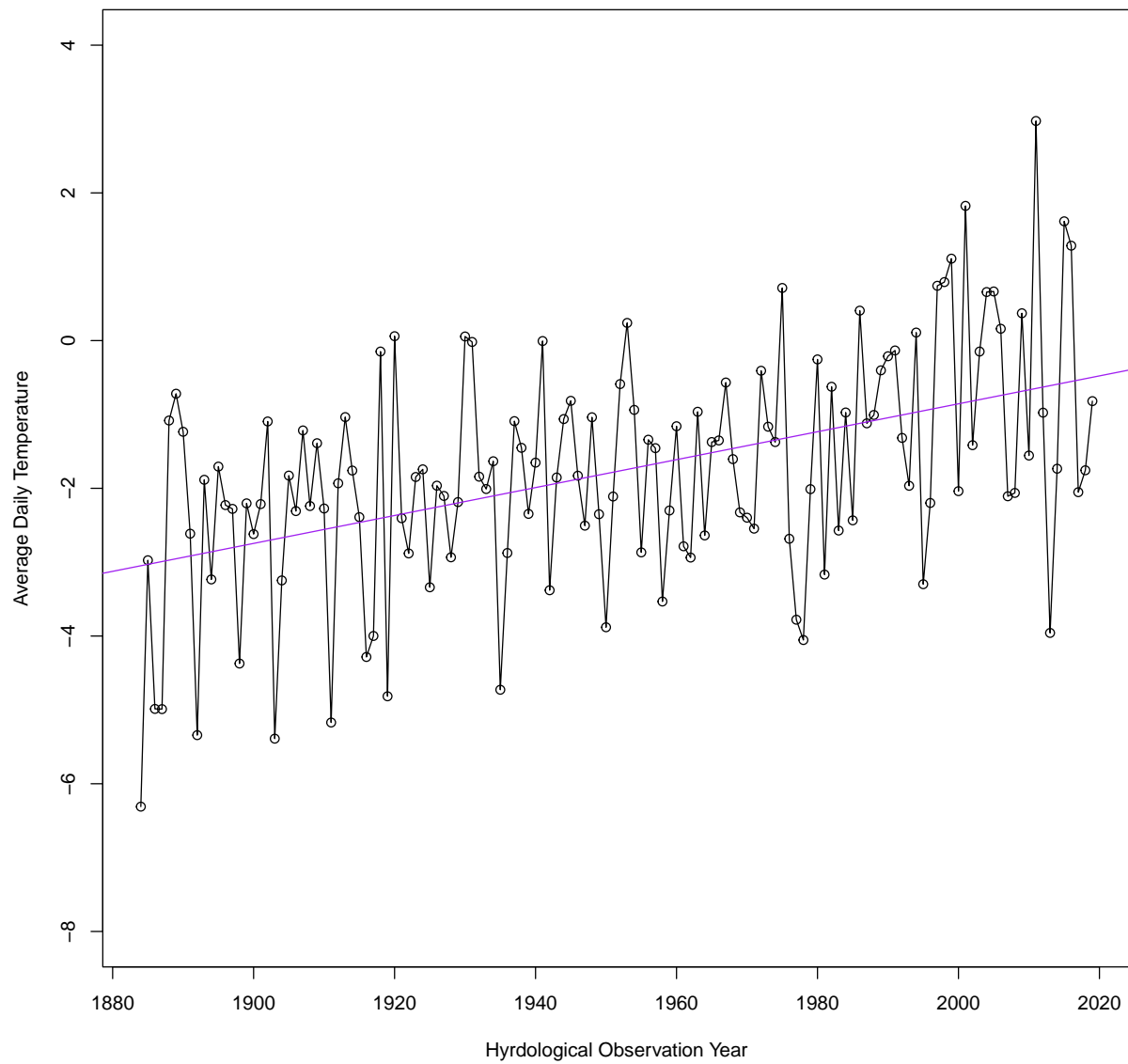


Having a hard time interpreting residuals? Ask Me, Anchal, or Brian!

#### 4.3.1 Plotting a line

The best way to present a linear model, is using a plot. Use `abline` to plot a line. You can also change the line type and line width using `lty` and `lwd` (see: <https://www.statmethods.net/advgraphs/parameters.html>). Try to run the following code, and make sure you like the plot.

```
plot(temp~year, ylim=c(-8,4), xlab='Hyrdological Observation Year', ylab= 'Average Daily Temperature',  
lines(temp~year, dat)  
abline(m1, col='purple')
```



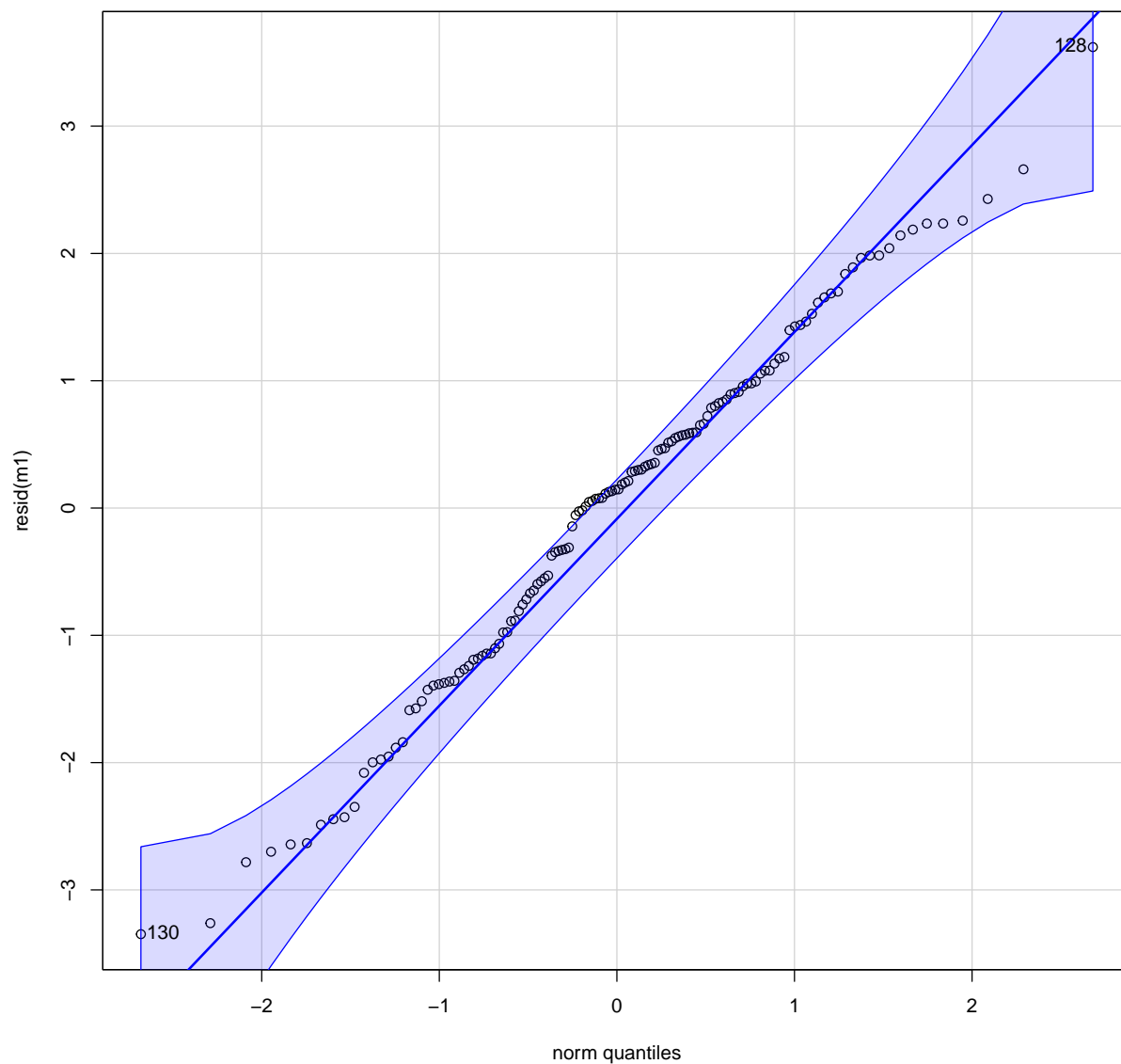
## 4.4 Assumptions

Now, during class we saw the assumptions of a linear model. While we can do formal tests for each assumption, most often that's not necessary if we run some visual diagnostics.

### 4.4.1 Normal distribution

First, for the normal distribution we can plot the residuals against the quantiles. We use a Q-Q plot for this:

```
library(car)
qqPlot(resid(m1))
```



[1] 128 130



If you can't load a package using library, you need to install the package first!

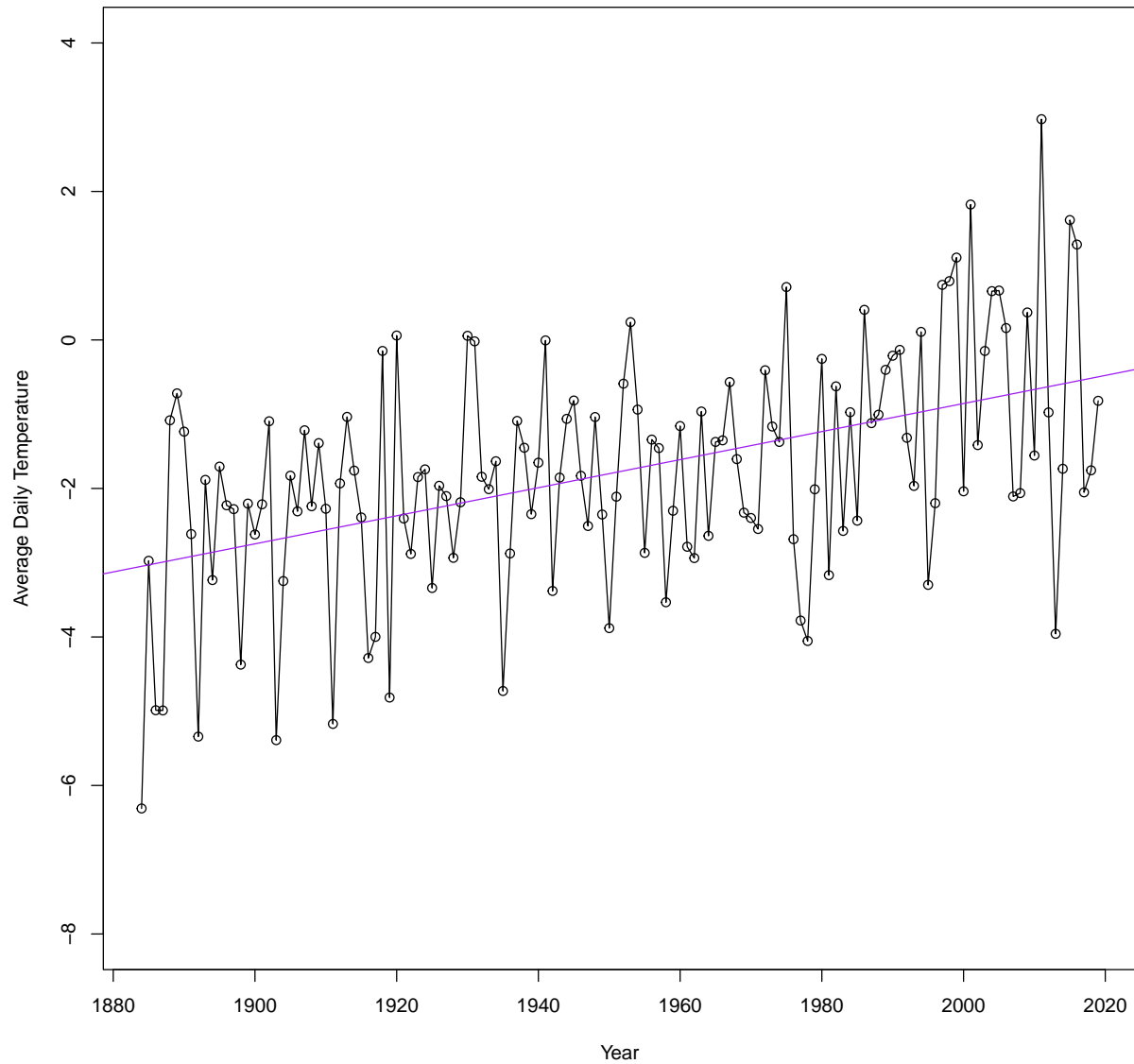
Explore the Q-Q plot.

#### 4.4.2 Linearity

Explore the plot with the regression line:

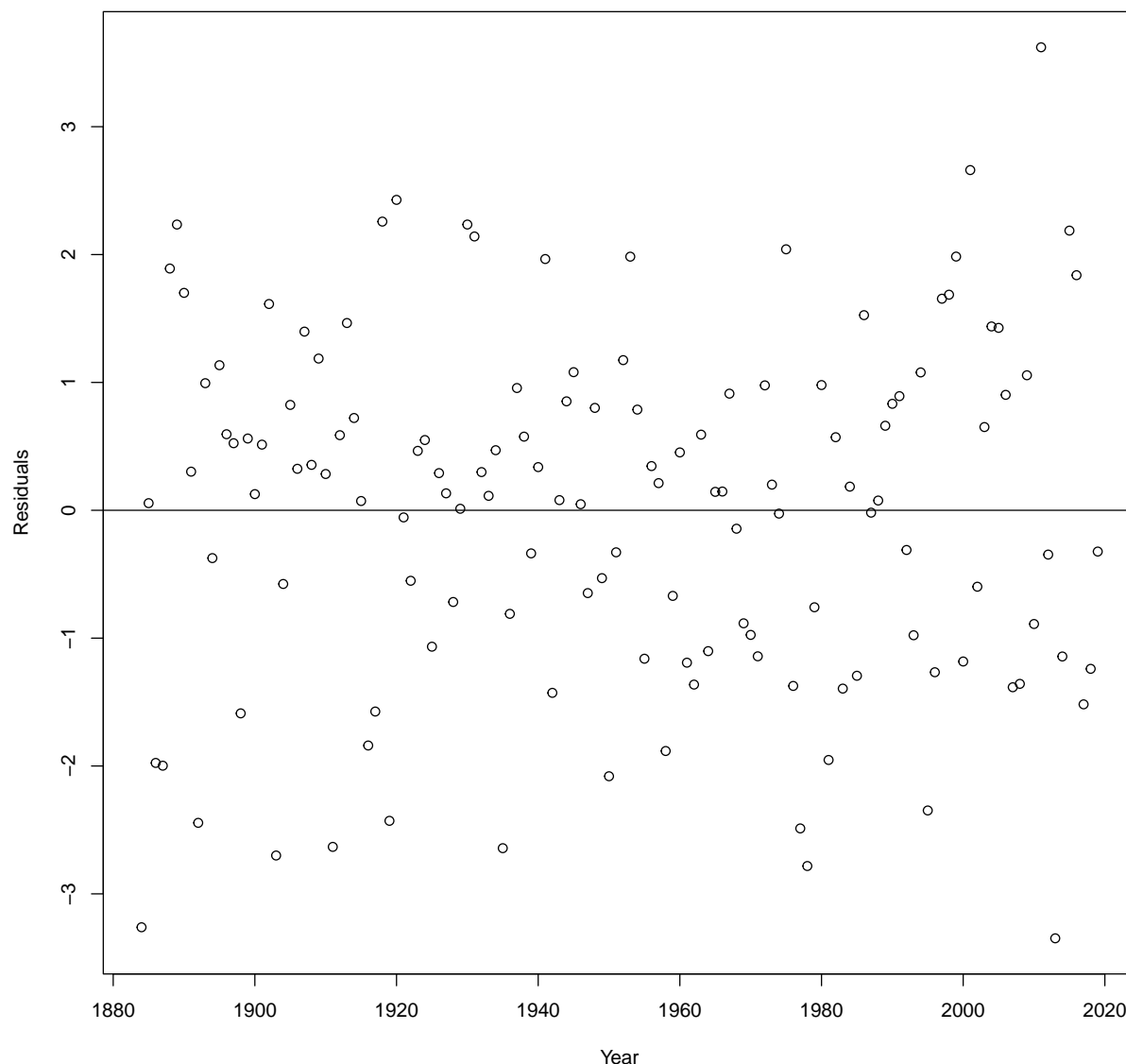
```
plot(temp~year, ylim=c(-8,4), xlab='Year', ylab= 'Average Daily Temperature', dat)
lines(temp~year, dat)
```

```
abline(m1, col='purple')
```



You can see whether the purple fitted line is centered. You can also plot the residual against the predictor using:

```
plot(resid(m1)~dat$year, xlab="Year", ylab="Residuals") # resid(model object) generates the residual va  
abline(h=0)
```



Remember that errors  $\epsilon_i$  has a mean of 0, so we should see the scatter of residuals to be roughly centered at 0 across the entire range of values of year. If the scatter forms a curved band

#### 4.4.3 Equal variance

Residuals should have an equal variance across the range of the predictor variable. We can use the residual vs. predictor response plots to evaluate this assumption. You have already run this plot!

#### 4.4.4 Independence of observations

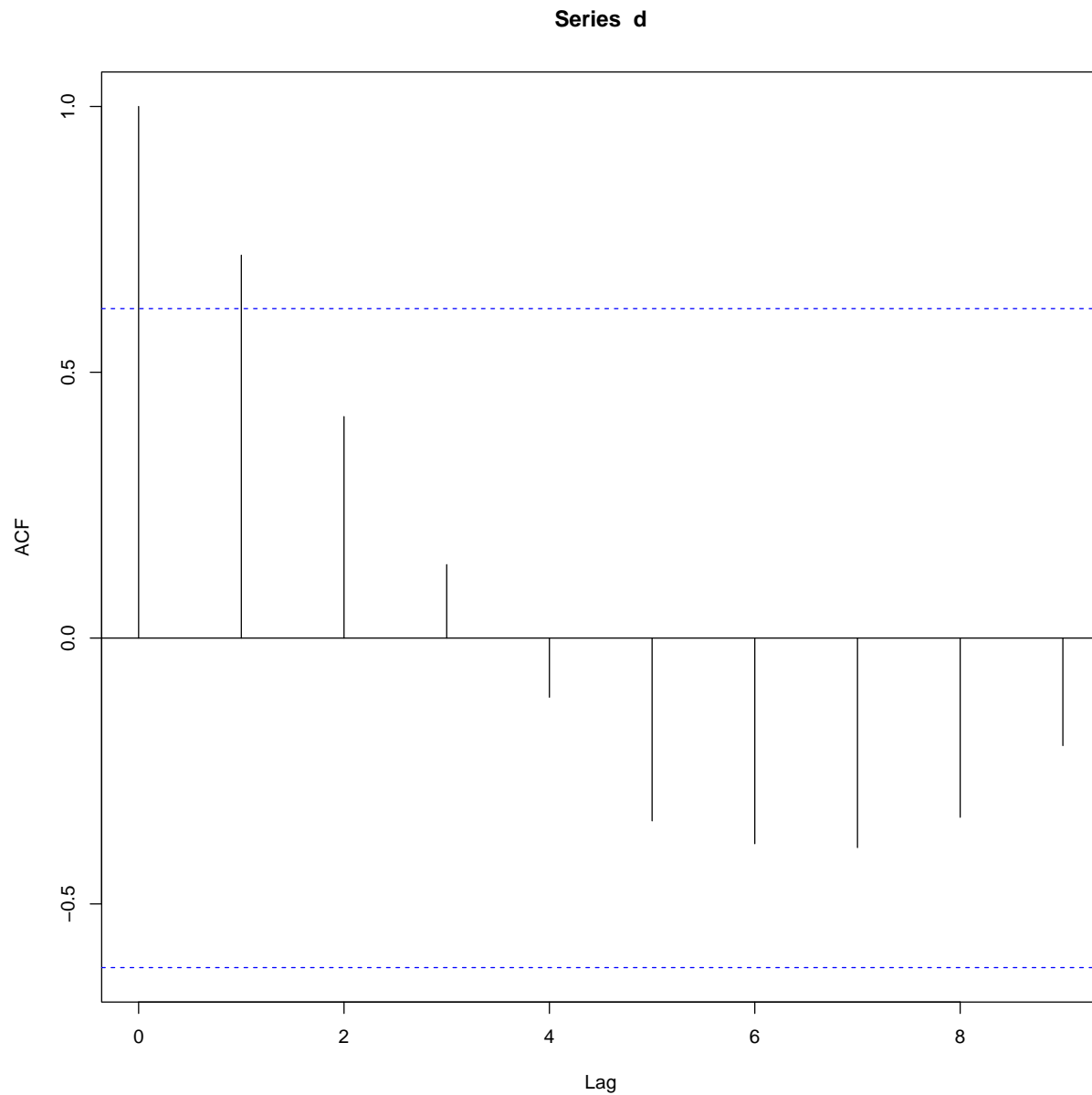
Often times, we know if our observations are independent or not based on the experimental design. However, spatial and temporal non-independence are pretty common. We get a lot more information about *how* data was collected than about *how* the data looks regarding data independence

In this case, the observations cannot be considered as independent because observations in successive years



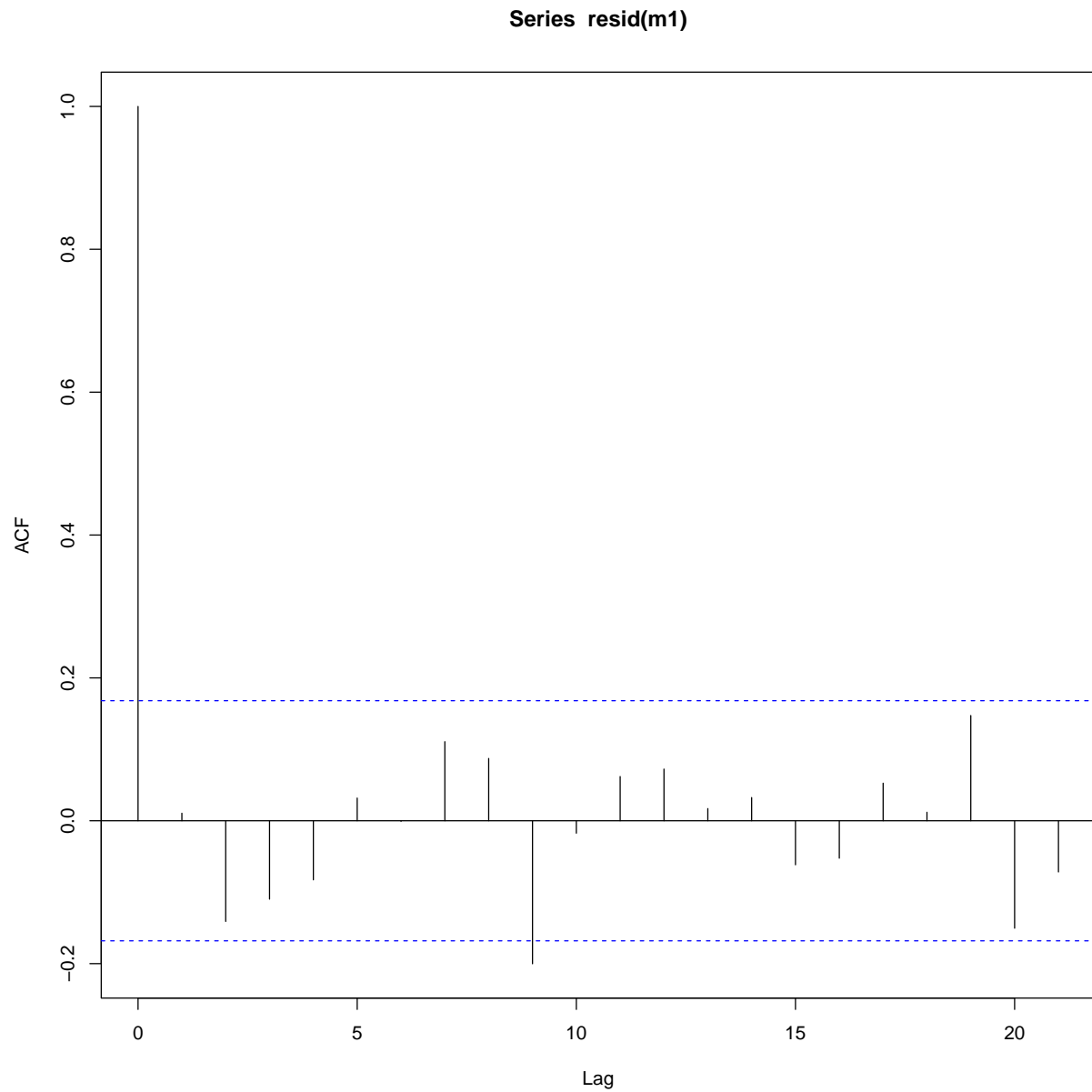
(say 2000 and 2001) would likely be more similar to each other compared to observations decades apart (say, 2000 and 1950). In a time-series, observations are almost certainly non-independent by definition.

This is a very complex topic, and we do not have much time to dig deep in this topic. However, just know that we can test spatial and temporal autocorrelation using the `acf(resid(ml))` function. If the plot looks like this:



It is auto-correlated. (Lines decreasing as lag increases).

```
acf(resid(m1))
```



In this case **you can copy this to answer a question** Our data is not independent (understand why it's not independent), but it is not auto-correlated. Because the lines behave randomly.



Q7. 2 pts. Test the four assumptions of a linear model, and explain whether this model meets all four assumptions



a7. (A1). Since the residuals lie along the QQ-line, we can conclude that SLR assumption 1 can be considered supported, that is errors are normally distributed.

(A2). Since the linear model is centered among the data, this support the assumption of linearity.

(A3). Since the residual plot is evenly distributed and symmetric around 0, this supports the assumption of equal variance.

(A4). The auto-correlation plot shows that the data are not correlated, which supports the assumption of independence. Our data is not independent, but it is not auto correlated.

## 4.5 Ice cover

We have answered one question: Is there climate warming over the past century or so around Lake Mendota? We saw that yes, there is, we also found the rate of change.

However, if you explore the dataset, you see that we have information on ice-cover. We can answer two more questions using linear models:

- 8) Is there a decline in ice cover duration over this time? If so, what is the rate of decline in ice cover duration
- 9) Is warming linked to a decline in ice cover duration? If so, what is the decline in ice cover duration per 1 deg C increase in air temperature?

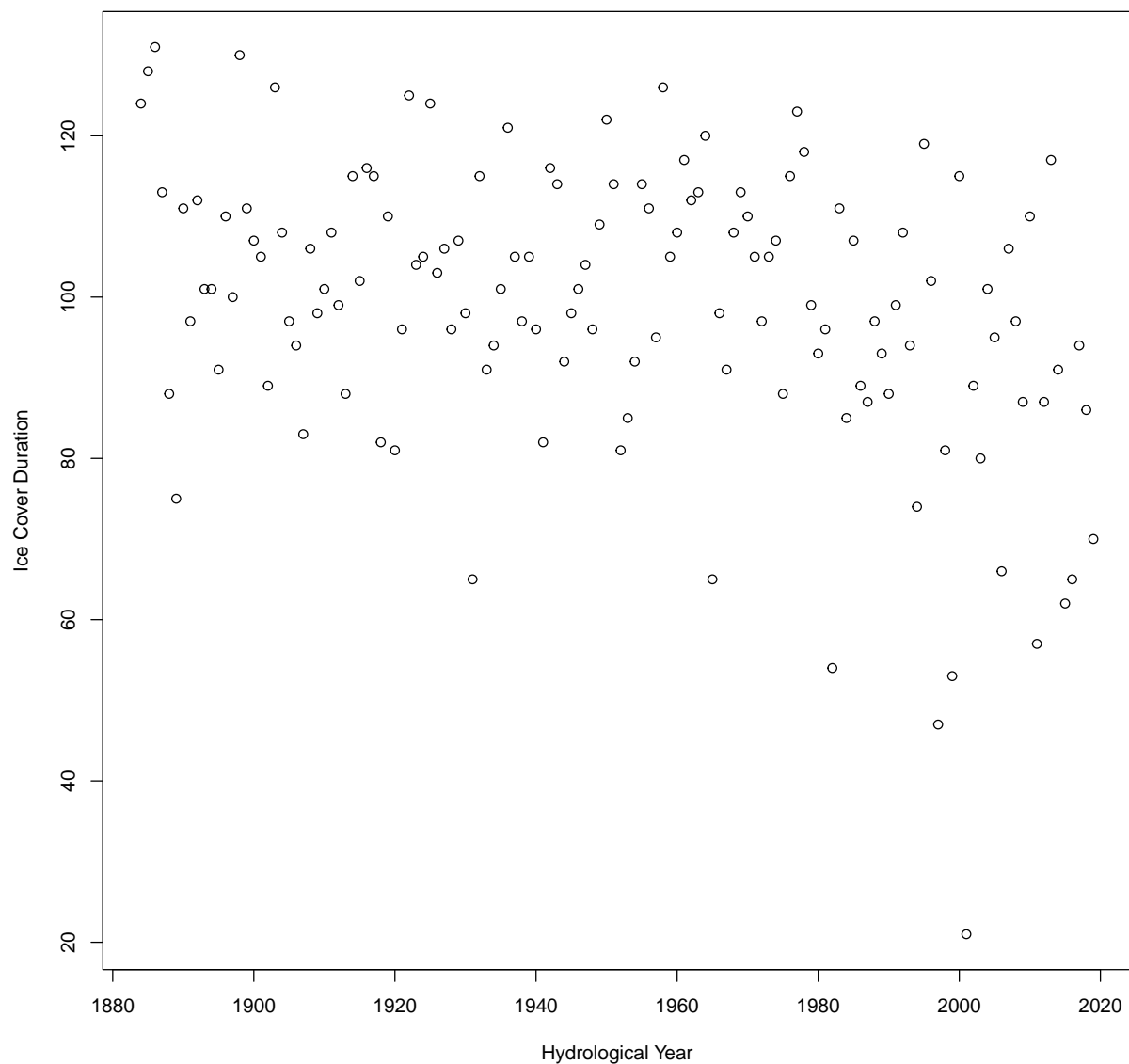


Q8. 6 pts. Answer the question. "Is there a decline in ice cover duration over this time? If so, what is the rate of decline in ice cover duration." You'll need to fit the regression model (1 pt). You then need to interpret the 6 main outputs of regression model; outputs are printed via `summary()` (2 pts). Further, you will need to test the 4 main assumptions of a SLR model (2 pts). And lastly, include a sentence or two to conclude your findings associated with each research question (1 pts).

```
head(dat)
```

```
year lakeid ice_on ice_off ice_duration temp 1 1884 Lake Mendota 1884-12-17 1885-04-20 124 -6.309945
2 1885 Lake Mendota 1885-12-12 1886-04-19 128 -2.974033 3 1886 Lake Mendota 1886-12-05 1887-04-15
131 -4.986740 4 1887 Lake Mendota 1887-12-24 1888-04-15 113 -4.989560 5 1888 Lake Mendota 1889-01-02
1889-03-31 88 -1.082320 6 1889 Lake Mendota 1890-01-14 1890-03-30 75 -0.719337
```

```
plot(dat$year, dat$ice_duration, xlab="Hydrological Year", ylab="Ice Cover Duration" )
```



```
m2<-lm(ice_duration~year,data=dat) #fit SLR model of temp against year, place in object m1
summary(m2)
```

Call: `lm(formula = ice_duration ~ year, data = dat)`

Residuals: Min 1Q Median 3Q Max -69.643 -7.846 1.526 11.859 28.359

Coefficients: Estimate Std. Error t value Pr(>|t|)

(Intercept) 424.01402 70.70278 5.997 1.76e-08 **year -0.16660 0.03622 -4.599 9.70e-06** — Signif. codes:  
0 ‘’ **0.001** ’’ 0.01 ’’ 0.05 ‘.’ 0.1 ’’ 1

Residual standard error: 16.58 on 134 degrees of freedom Multiple R-squared: 0.1363, Adjusted R-squared:  
0.1299 F-statistic: 21.15 on 1 and 134 DF, p-value: 9.7e-06



a8. a) The rate of decline of ice cover is 0.16660 units per year.

b) From the summary, we find the model formula is a linear model. We are modeling ice cover duration with respect to year.

The residuals (errors) tell us the difference between the actual data and predicted ice cover duration. We are also given summary statistics of including min/max, median, and quartiles. The median is 1.526 is close to 0 and the Q1 and Q3 are roughly symmetric around the median.

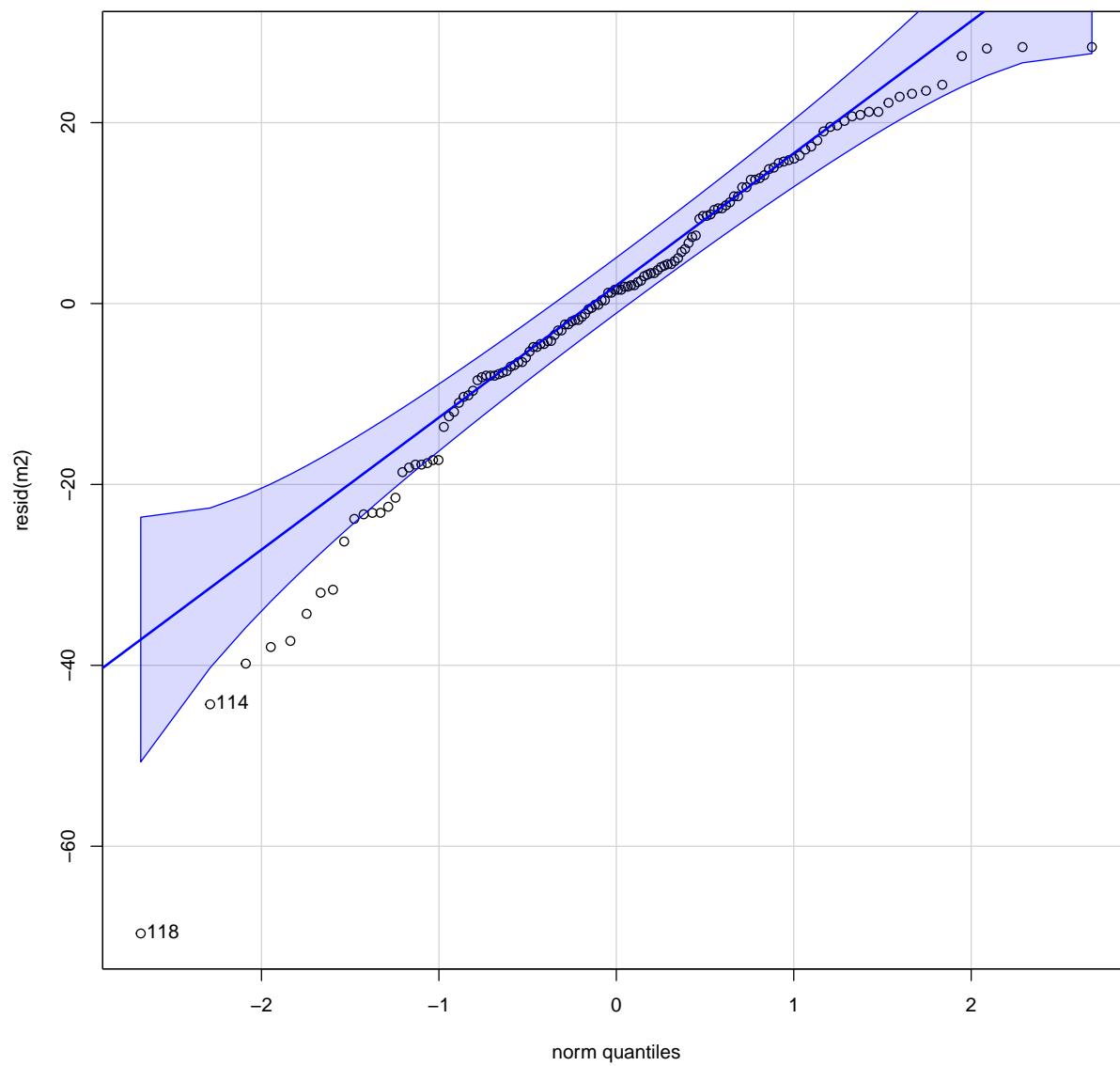
The coefficient given are the y-intercept and the slope, as providing standard error and p-value for each. The p-value tests whether each coefficient is significantly different from 0. Since the p-value is below 0.05 for both values, we are able to reject the null model, means that we have support that neither coefficient is 0.

The residual standard error is the standard deviation of the errors (residuals) which is 16.38. This indicates the spread of the errors.

$R^2$  represent the percentage of variance explained by the model. So, 13.63% of the variance is explained by the model.

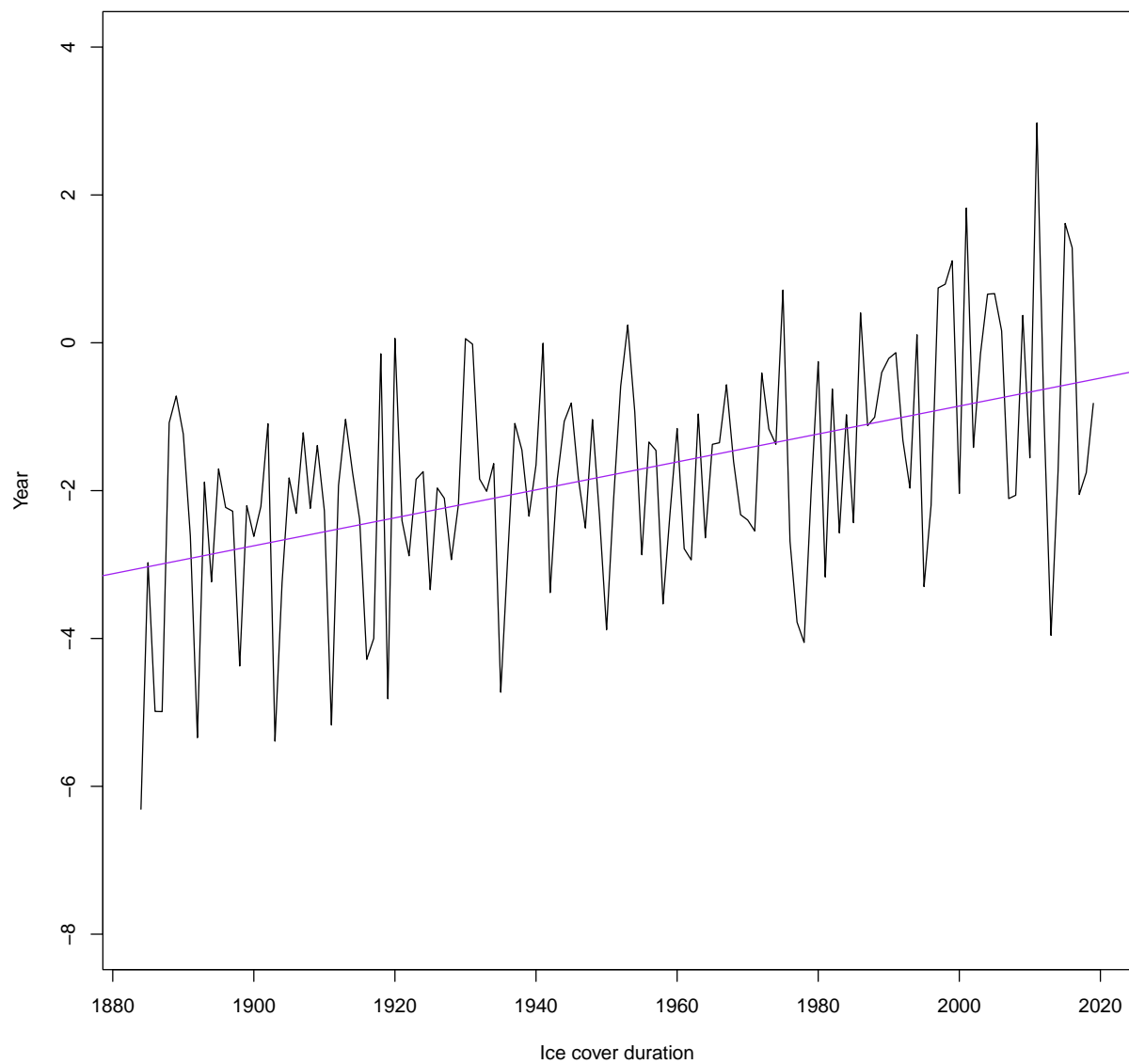
The last item indicates model significance and gives the f-statistic and a p-value. This low p-value means that there is little evidence for support for the null model over the the current model.

```
#a1 - normality  
qqPlot(resid(m2))
```

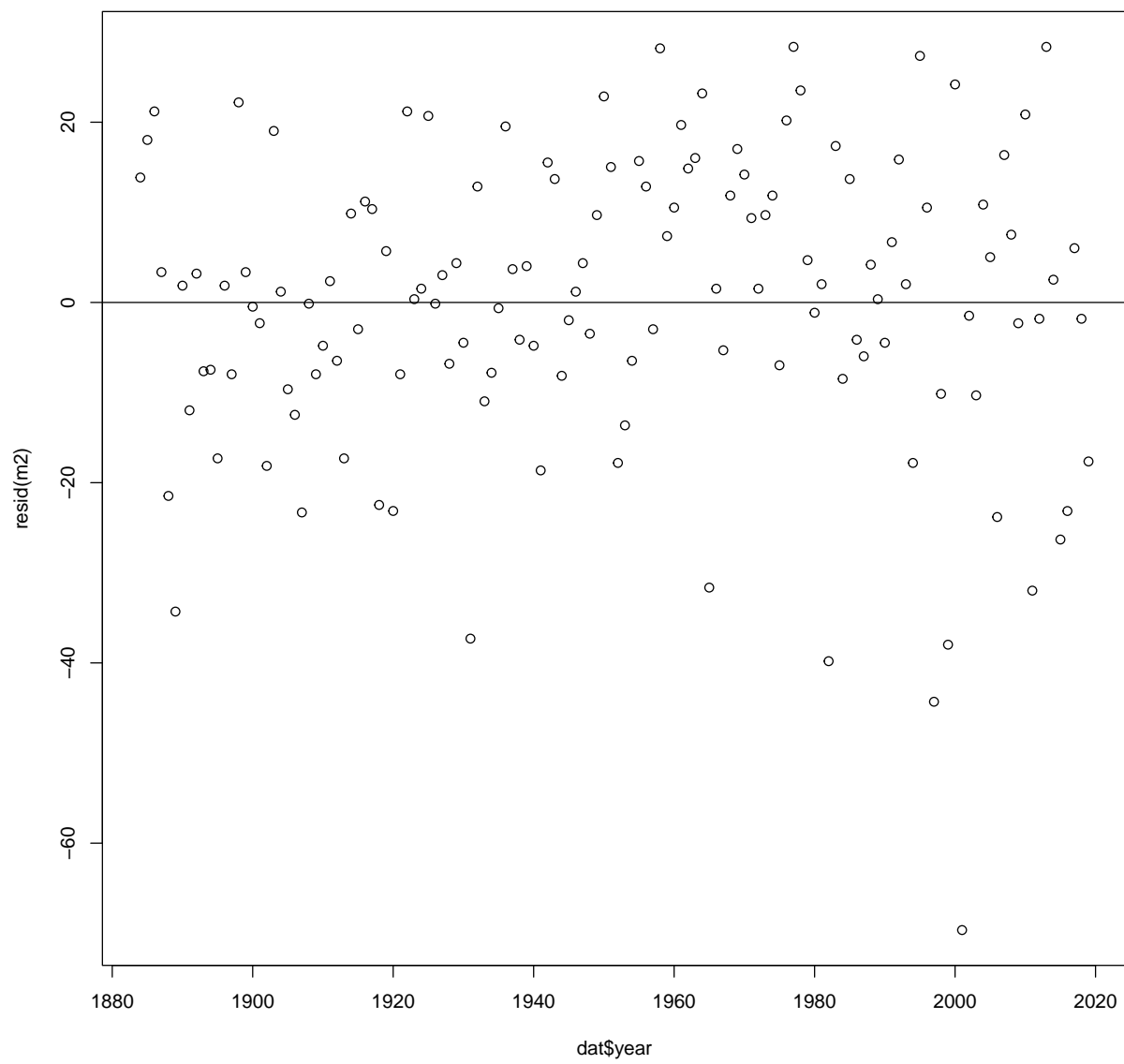


[1] 118 114

```
#a2 - linearity
plot(ice_duration~year, ylim=c(-8,4), xlab='Ice cover duration', ylab= 'Year', dat)
lines(temp~year, dat)
abline(m1, col='purple')
```

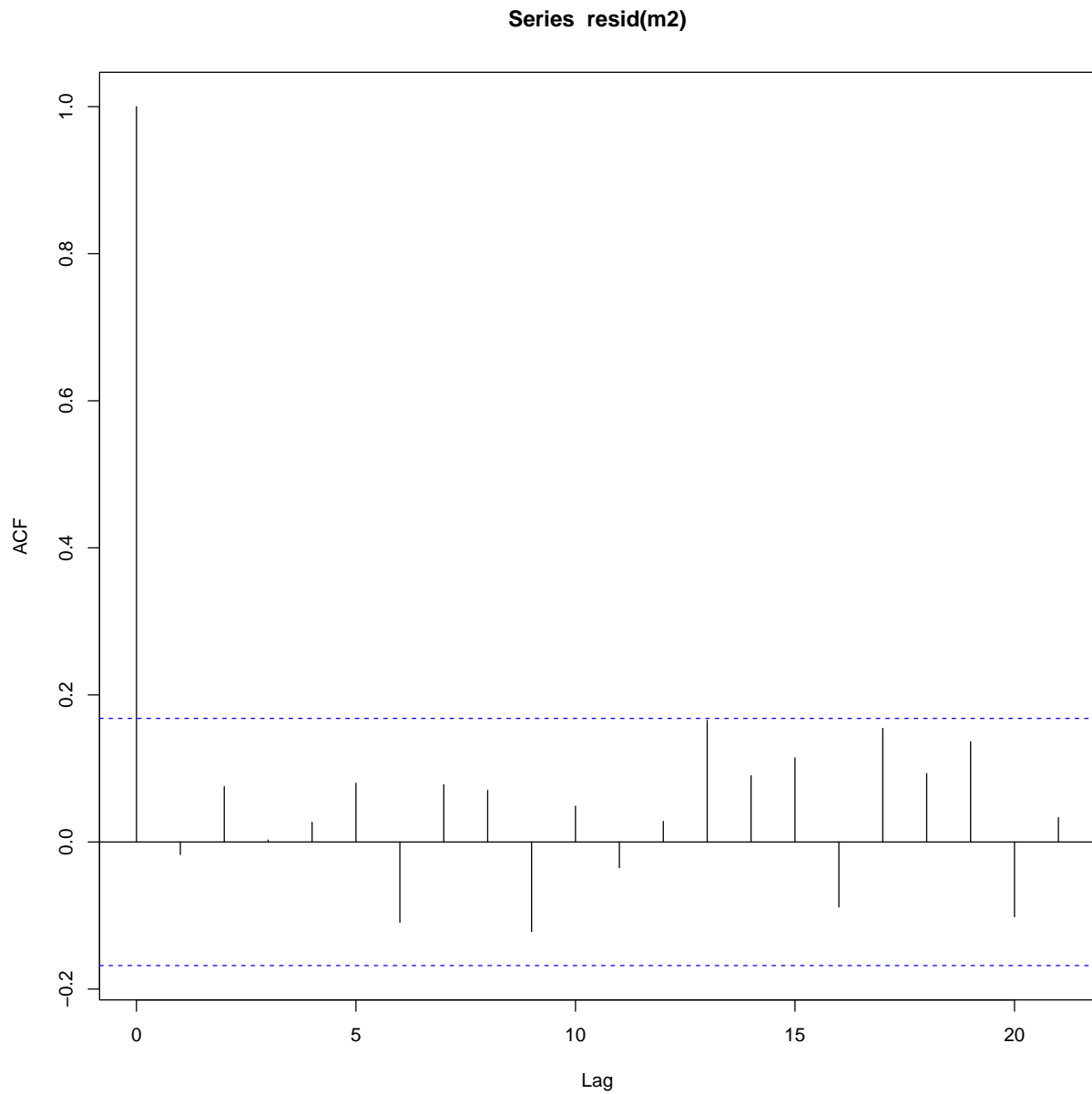


```
#a3 - equal variance  
plot(resid(m2)~dat$year) # resid(model object) generates the residual values  
abline(h=0)
```



```
#a4 - independence  
acf(resid(m2))
```





a8. c) (A1). Since some of the residuals lie outside of the shaded region along the QQ-line, we have less support that the errors are normally distributed.

(A2). Since the linear model is centered among the data, this supports the assumption of linearity.

(A3). Since the residual plot appears to be evenly distributed and symmetric around 0, this supports the assumption of equal variance.

(A4). Since the data is time related, the data cannot be independent. The auto-correlation plot shows that the data are not correlated.



a8 d. Is there a decline in ice cover duration over this time? Yes, the appears to be a linear trend for decline in ice cover duration over time. Since the p-value is below 0.05 for both values, we are able to reject the null model and have support for our alternative model. The rate of decline is 0.16660 units of ice cover per year.

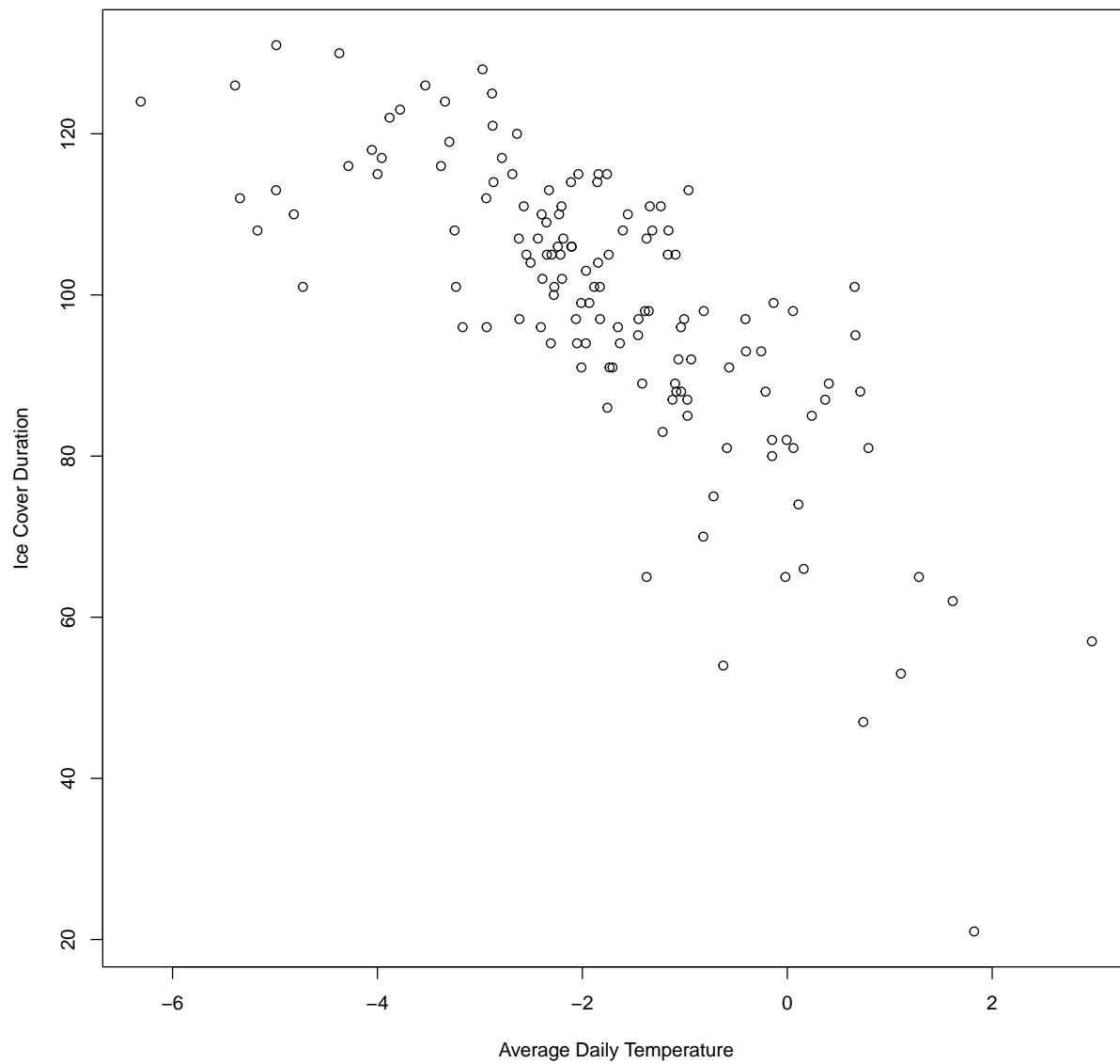


Q9. 6 pts. Answer the question. "Is warming linked to a decline in ice cover duration? If so, what is the decline in ice cover duration per 1 deg C increase in air temperature?" You'll need to fit the regression model (1 pt). You then need to interpret the 6 main outputs of regression model; outputs are printed via summary() (2 pts). Further, you will need to test the 4 main assumptions of a SLR model (2 pts). And lastly, include a sentence or two to conclude your findings associated with each research question (1 pts).

```
head(dat)
```

```
year lakeid ice_on ice_off ice_duration temp 1 1884 Lake Mendota 1884-12-17 1885-04-20 124 -6.309945
2 1885 Lake Mendota 1885-12-12 1886-04-19 128 -2.974033 3 1886 Lake Mendota 1886-12-05 1887-04-15
131 -4.986740 4 1887 Lake Mendota 1887-12-24 1888-04-15 113 -4.989560 5 1888 Lake Mendota 1889-01-02
1889-03-31 88 -1.082320 6 1889 Lake Mendota 1890-01-14 1890-03-30 75 -0.719337
```

```
plot(dat$temp, dat$ice_duration, xlab="Average Daily Temperature", ylab="Ice Cover Duration" )
```



```
m3<-lm(ice_duration~temp,data=dat) #fit SLR model of temp against year, place in object m1
summary(m3)
```

Call: `lm(formula = ice_duration ~ temp, data = dat)`

Residuals: Min 1Q Median 3Q Max -46.744 -6.320 1.143 7.292 23.156

Coefficients: Estimate Std. Error t value Pr(>|t|)

(Intercept) 83.5380 1.4975 55.79 <2e-16 **temp -8.6600 0.6341 -13.66 <2e-16** — Signif. codes: 0 ‘**’**  
**0.001** ‘**’** 0.01 ‘**’** 0.05 ‘**’** 0.1 ‘**’** 1

Residual standard error: 11.54 on 134 degrees of freedom Multiple R-squared: 0.5819, Adjusted R-squared:  
0.5788 F-statistic: 186.5 on 1 and 134 DF, p-value: < 2.2e-16



a9. a) The rate of decline of in duration of ice cover is 8.6600 units per degree Celsius.

b) From the summary, we find the model formula is linear model ice cover duration with respect to temperature.

The residuals (errors) tell us the difference between the actual data and predicted ice cover duration for this model. We are also given summary statistics of including min/max, median, and quartiles. The median is 1.143, which is close to zero. We also notice that the Q1 and Q3 are roughly symmetric.

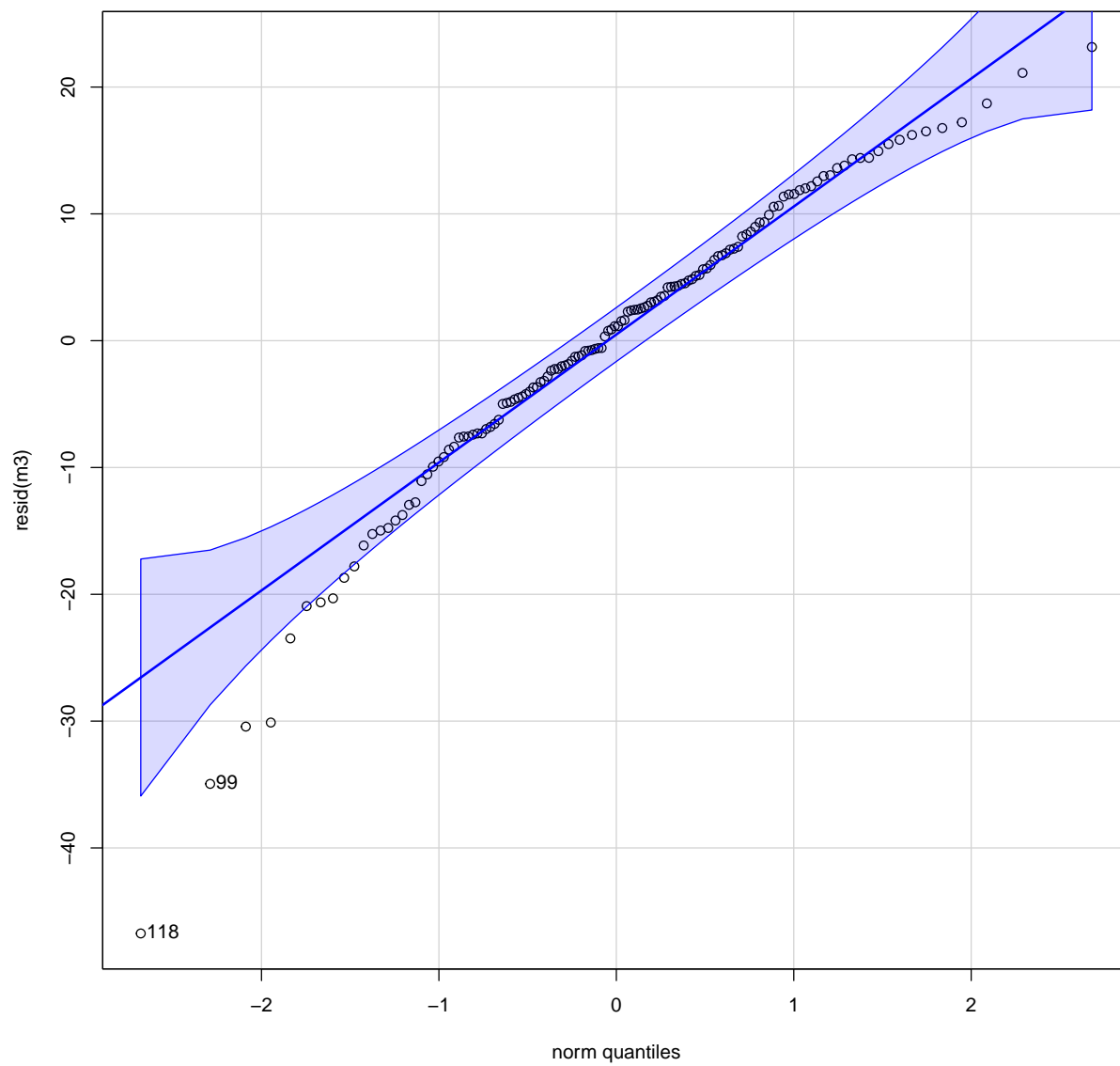
The coefficient givens are the y-intercept and the slope, as providing standard error and p-value for each. The p-value tests whether each coefficient is significantly different from 0. Since the p-value is below 0.05 for both values, we are able to reject the null model, which means that we have support that neither coefficient is 0.

The residual standard error is the standard deviation of the errors (residuals) which is 11.54.

$R^2$  represent the percentage of variance explained by the model. So, 58.19% of the variance is explained by the model.

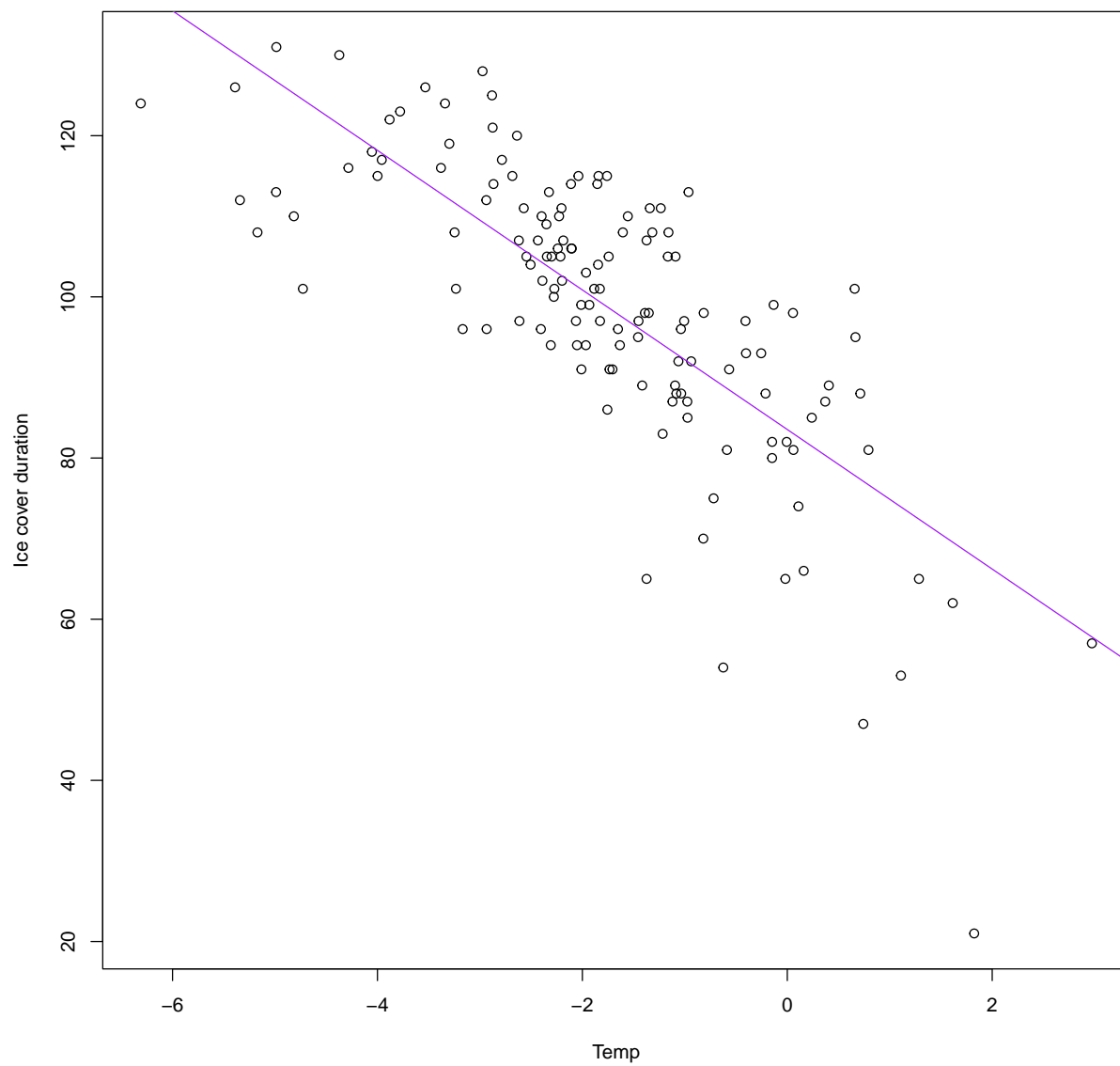
The last item indicates model significance and gives the f-statistic and p-value. This low p-value means that there is little evidence for support for the null model over the the current model.

```
#a1 - normality  
qqPlot(resid(m3))
```

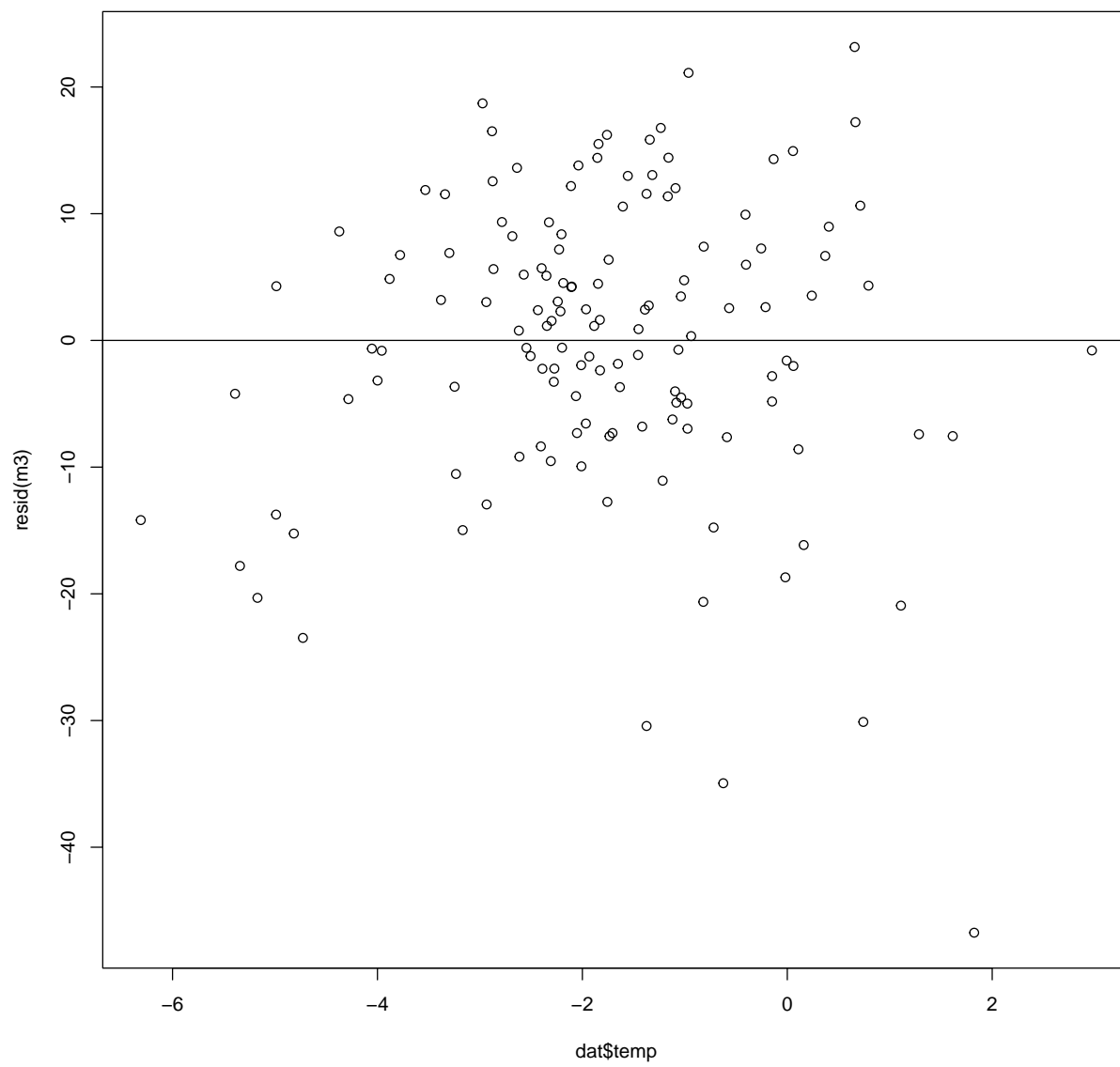


[1] 118 99

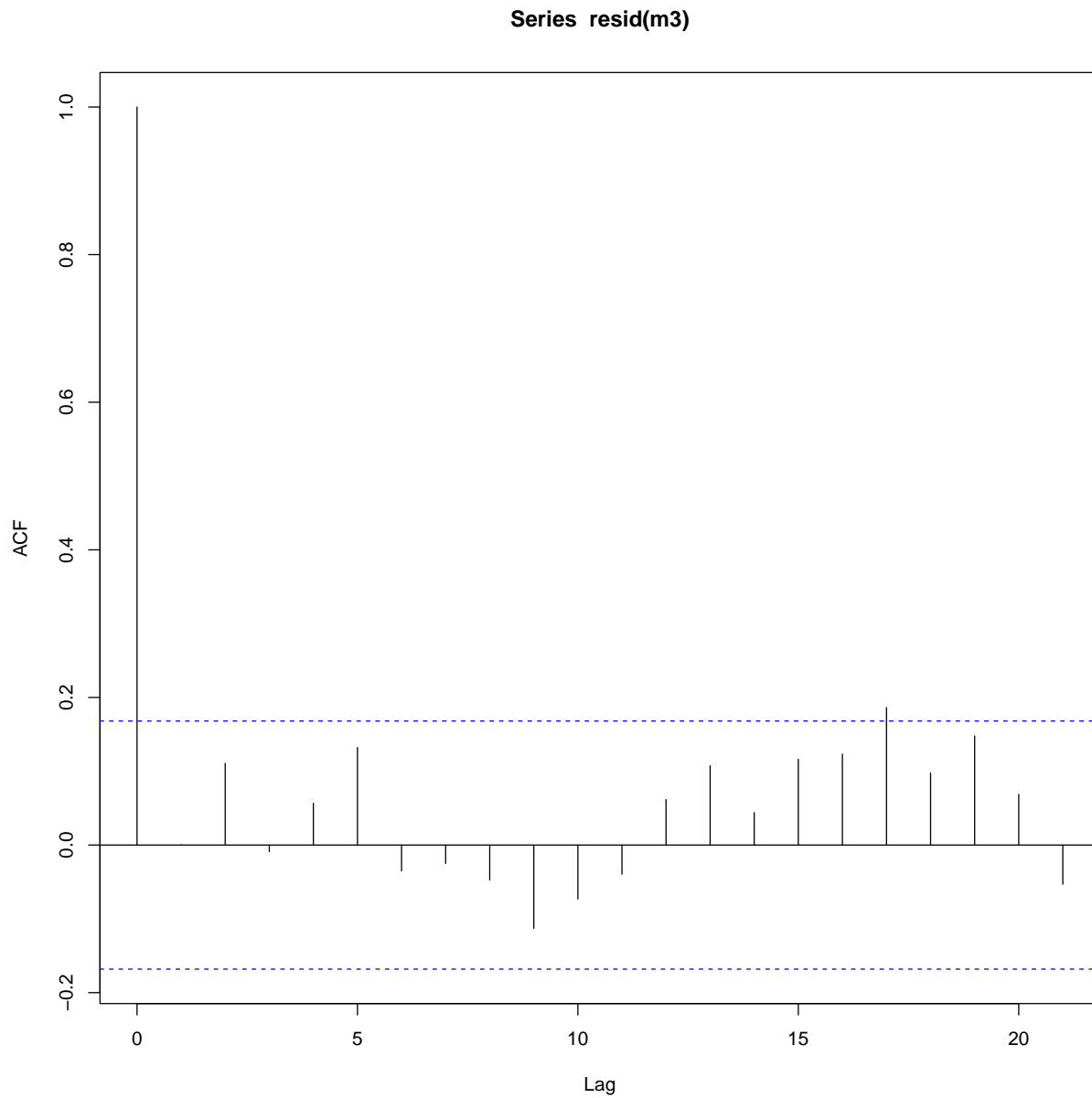
```
#a2 - linearity
plot(ice_duration~temp, xlab='Temp', ylab= 'Ice cover duration', dat)
#lines(ice_duration~temp, dat)
abline(m3, col='purple')
```



```
#a3 - equal variance  
plot(resid(m3)~dat$temp) # resid(model object) generates the residual values  
abline(h=0)
```



```
#a4 - independence  
acf(resid(m3))
```



a9. c) **\*\*Update\*\*** (A1). Since some of the residuals lie outside of the shaded region along the QQ-line, we have less support that the errors are normally distributed, but we can still proceed with caution.

(A2). Since the linear model is centered among the data, this supports the assumption of linearity.

(A3). Since the residual plot appears to be evenly distributed and symmetric around 0, this supports the assumption of equal variance.

(A4). Since the data is related by temperature, the data are independent. The auto-correlation plot shows that the data are not correlated.





a9 d. Is warming linked to a decline in ice cover duration? **\*\*Update\*\*** Yes, there appears to be a linear trend for decline in ice cover duration with respect to temperature. Since the p-value is below 0.05 for both values, we are able to reject the null model and have support for our alternative model. The ice cover duration decreases 8.6600 per increase in degree celsius.

If you are eager for more work, you can get some extra credit. Check the Resources, extra credit and sample code section of the lab.

## Categorical linear model and Anova

You are almost done with this lab! :) Just a couple more questions.



These two questions are regarding a topic we will see on Friday! Check that presentation if you have any questions

Download the `mthreepops.csv` file and load it in an object called `dat2`.

```
dat2 = read.csv("mthreepops.csv")
```

Explore the dataframe (using `head` or `summary`) so you understand it. We are back to working with mussels. In this case you have samples from **3** populations, and you wonder if there are differences in sizes.

To do this, first run a linear model, and check the summary

```
m4<-lm(Length~Site, data=dat2)
summary(m4)
```

Call: `lm(formula = Length ~ Site, data = dat2)`

Residuals: Min 1Q Median 3Q Max -10.7996 -3.7647 -0.8647 2.9495 14.4167

Coefficients: Estimate Std. Error t value Pr(>|t|)

(Intercept) 38.1053 0.9831 38.760 <2e-16 \*\* *SiteStream2* 3.4817 1.3903 2.504 0.0141

*SiteStream3* -0.2183 1.3903 -0.157 0.8756

— Signif. codes: 0 ‘**0.001**’ ‘0.01’ ‘0.05’ ‘0.1’ ‘1’

Residual standard error: 5.385 on 87 degrees of freedom Multiple R-squared: 0.09298, Adjusted R-squared: 0.07213 F-statistic: 4.459 on 2 and 87 DF, p-value: 0.01433



Q10. 2 pts. Look at the model summary, and interpret the results. Is any population different? Which one has the largest mean size and which one the smallest length size?



a10. Looking at the model summary, we have the model

$$y = 38.1053 + 3.4817x - 0.2183x$$

Now, let's run an ANOVA. This is very simple in R

```
anova(m2)
```

Analysis of Variance Table

Response: ice\_duration Df Sum Sq Mean Sq F value Pr(>F)  
year 1 5818 5818 21.154 9.7e-06 \*\*\* Residuals 134 36853 275  
— Signif. codes: 0 ‘‘**0.001**’’ 0.01 ‘‘0.05’’ 0.1 ‘‘1’’



Q11. 2 pts. State the null and alternative hypotheses. Decide whether to reject or fail to reject your null hypothesis, and explain what this means biologically



a11. The null hypothesis is that the mean of all 3 populations are equal. The alternative hypothesis is that at least one of the means are different from the others. Since the p-value is small (below 0.05), we have reason to reject the null hypothesis. Biologically, this means that all three sites do not have the same mean size of mussels.

Lab total: 36 pts

end of lab

### Resources, extra credit and sample code

*These are hyperlinks*

The base R cheat-sheet

# An introduction to R

[{Research notes}%20are%20shorter%20academic,notes%20undergo%20regular%20peer%2Dreview">{Research notes}%20shorter%20academic,notes%20undergo%20regular%20peer%2Dreview">{Research notes}%20academic,notes%20undergo%20regular%20peer%2Dreview">{Research notes}%20regular%20peer%2Dreview">{Research notes}%20peer%2Dreview">{Research notes}%2Dreview">{Research notes}%20notes%20are%20shorter%20academic,notes%20undergo%20regular%20peer%2Dreview">{Research notes}%20shorter%20academic,notes%20undergo%20regular%20peer%2Dreview">{Research notes}%20academic,notes%20undergo%20regular%20peer%2Dreview">{Research notes}%20regular%20peer%2Dreview">{Research notes}%20peer%2Dreview">{Research notes}%2Dreview">https://www.paulmusgrave.info/journal-lists/research-notes/#::~text=Research%20notes%20are%20shorter%20academic%20notes%20are%20shorter%20academic,notes%20undergo%20regular%20peer%2Dreview](https://www.paulmusgrave.info/journal-lists/research-notes/#::~text=Research%20notes%20are%20shorter%20academic%20notes%20are%20shorter%20academic,notes%20undergo%20regular%20peer%2Dreview)



EXTRA CREDIT: 10 pts Due before first partial exam. For the three research questions explored in the second half of this lab: (1) Is there climate warming over the past century or so around Lake Mendota? If so, what is the rate of warming?; (2) Is there a decline in ice cover duration over this time? If so, what is the rate of decline in ice cover duration; (3) Is warming linked to a decline in ice cover duration? If so, what is the decline in ice cover duration per 1 deg C increase in air temperature? Write a RESEARCH NOTE (AKA, A VERY SHORT research paper, no more than 4 pages), with introduction, methods, and results, and discussion. Include plots, tables, references or anything you think it's important. Upload it to Canvas. There will be an assignment called Extra Credit 1

## Sample Code

This is the data processing code used to obtain the data for this lab

```
## data processing steps for L19
## author: Xingli Giam (xgiam@utk.edu)
```

```

    ## roughly following approach here: https://lter.github.io/lterdatasampler/articles/ntl\_icecover\_vignette.html

install.packages('lterdatasampler')
library(lterdatasampler)

## explore the two datasets # use print() instead of head() because this is a tidyverse tibble dataframe
head(ntl_icecover, n=50)
head(ntl_airtemp, n=50)

## subset observations of only Lake Mendota
DurationMendota<- ntl_icecover[ntl_icecover$lakeid %in% 'Lake Mendota',]

## the lterdatasampler vignette states:
# 'according to the original metadata: "Daily temperature data prior to 1884 were estimated from 3
# we therefore use data from 1884

## only choose observations of lake ice duration starting from 1884
DurationMendota_1 <- DurationMendota[DurationMendota$year >= 1884,]

# check if the subset is correct
print(DurationMendota_1,n=20) # yes, first observation is 1884

# notice that year in the ice cover duration dataset refers to hydrological year, i.e., the year of
# therefore if ice starts to form in December of 2022, this (hydrological) year is taken as 2022
# say for the next cold season, ice only starts to form in Jan of 2024, this (hydrological) year is
# let me know if this doesn't make sense to you...

## we want to match the air temperature of the cold season of a given hydrological year to the lake ice
# e.g., for hydrological year 2022, we want to average the daily air temps of the cold season from 11/1/2021 to 10/31/2022

# install lubridate package to help us easily handle dates
install.packages('lubridate')
library(lubridate)

# ntl_airtemp - each row is the air temperature of a given day
# use month() to extract the month of a given day, put this information into new column called 'month'
ntl_airtemp$month <- month(ntl_airtemp$sampldate)

# check if the above code is doing what we want it to do - always check
head(ntl_airtemp, n=50) # yes, we see a new column called month, comparing this to sampldate, we can see

# create new column 'hydroyear', populate it with the data from the column 'year'
ntl_airtemp$hydroyear <- ntl_airtemp$year

# this part is the more complicated part, but you will gain the skills the more you work with data in R
# because the cold season is the start of a hydroyear, i.e., 2022 hydroyear is from November 2021 to October 2022
# hydroyear = year for months Nov and Dec, but hydroyear = year - 1 for months Jan to October

# what this does is telling R to pick out the rows for which months are from 1 (jan) to 10 (oct), and for those rows
ntl_airtemp$hydroyear[ntl_airtemp$month<=10] <- ntl_airtemp$year[ntl_airtemp$month<=10] - 1

# see if it does what we want it to do
head(ntl_airtemp, n=100) # looks good

```

```

# next, for each hydroyear, we want to retain only days falling between November and April (the 6 months)
ntl_airtemp_NovToApr <- ntl_airtemp[ntl_airtemp$month %in% c(11,12,1,2,3,4),]

# see if it does what we want it to do
head(ntl_airtemp_NovToApr, n=100) # looks good

# subset this dataset to start from hydroyear 1884
ntl_airtemp_NovToApr_1 <- ntl_airtemp_NovToApr[ntl_airtemp_NovToApr$hydroyear >= 1884,]

# see if it does what we want it to do
head(ntl_airtemp_NovToApr_1, n=100) # looks good

# now calculate average daily temperature from Nov to Apr for each hydroyear using the aggregate() function
AvgTempDat<-aggregate(ave_air_temp_adjusted~hydroyear, FUN="mean", data=ntl_airtemp_NovToApr_1)

# see if it does what we want it to do # use head() because the data frame resulting from aggregate() is large
head(AvgTempDat, 50) ## looks good! # each row represents the mean daily cold season temp for a given hydroyear

# now, let's combine the temp and ice duration dataset, making sure that each row refers to the ice duration for the same hydroyear
DurationMendota_2<- merge(DurationMendota_1, AvgTempDat, by.x="year", by.y="hydroyear", all.x=T)
# refer to this (https://stackoverflow.com/questions/20374459/merge-two-dataframes-with-repeated-columns)

# see if it does what we want it to do
head(DurationMendota_2, 50) ## awesome!

# check summary() to see if there are any NAs or missing data in any one of the rows
summary(DurationMendota_2) ## ok!

# last thing, rename the last column to make things easier for students
names(DurationMendota_2)[6] <- 'temp'

# save as csv

write.csv(DurationMendota_2, 'LakeMendota.csv', row.names=F)

## end

```