Lab 4: Null Hypothesis Significance Testing and Linear Models

Dr. Alejandro Molina-Moctezuma;

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

> ℹ 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 $X \sim N(\mu = 40, \sigma = 6)$

```r
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 [1] 35.16110 41.86010 41.02979 46.91927 35.67037 35.84521 29.19490 48.95783 [9] 34.76085 35.56883 30.45121 39.53396 29.68375 40.00724 23.14135

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

```r
X_bar<-
s<-
n<-
```

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

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

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 size is represented $X \sim N(\mu = 41.6, \sigma = 6)$

First, let's simulate this second population

```r
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.

```
[1] 28.97077 41.61400 49.58230 39.47187 37.36413 39.95577 40.55664 37.06439 [9] 45.56977 41.76724 43.36342
40.21923 33.03839 45.74475 23.60037
```

> 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 confidecen intervals, would you say these populations are different in size? More importatly, would you say that the mussels from the population 2 are 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 o 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?

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?

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!

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

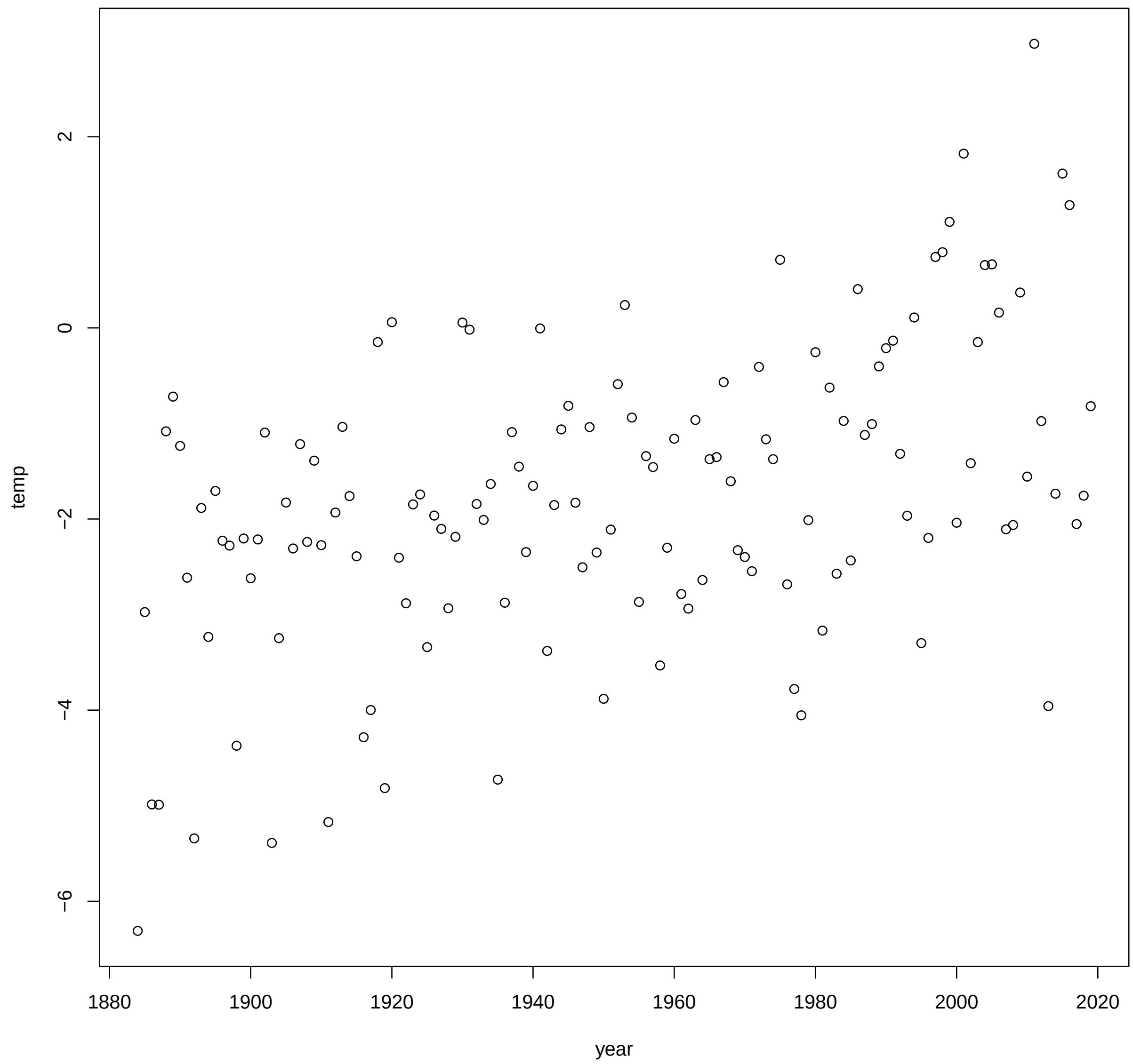
We are working on three research questions in this lab: (1) Is there climate warming over the past century or so around Lake Mendota? If so, what is the rate of warming?; (2) Is there a decline in ice cover duration over this time? If so, what is the rate of decline in ice cover duration; (3) Is warming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

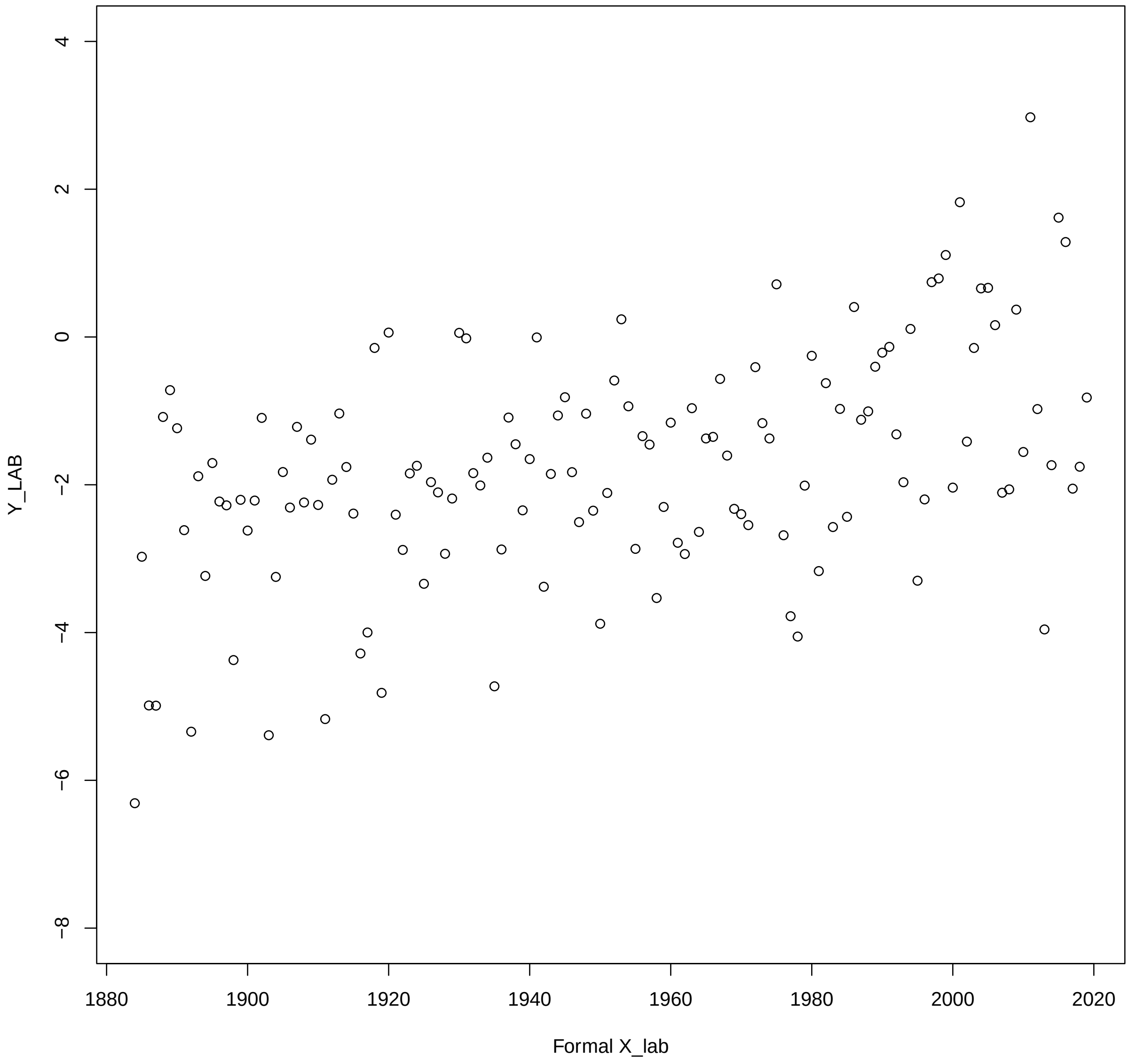
If the climate is warming over time, we would see an increase in air temperature over time. Let's plot air temperature on the y-axis and year on the x-axis.

```r
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

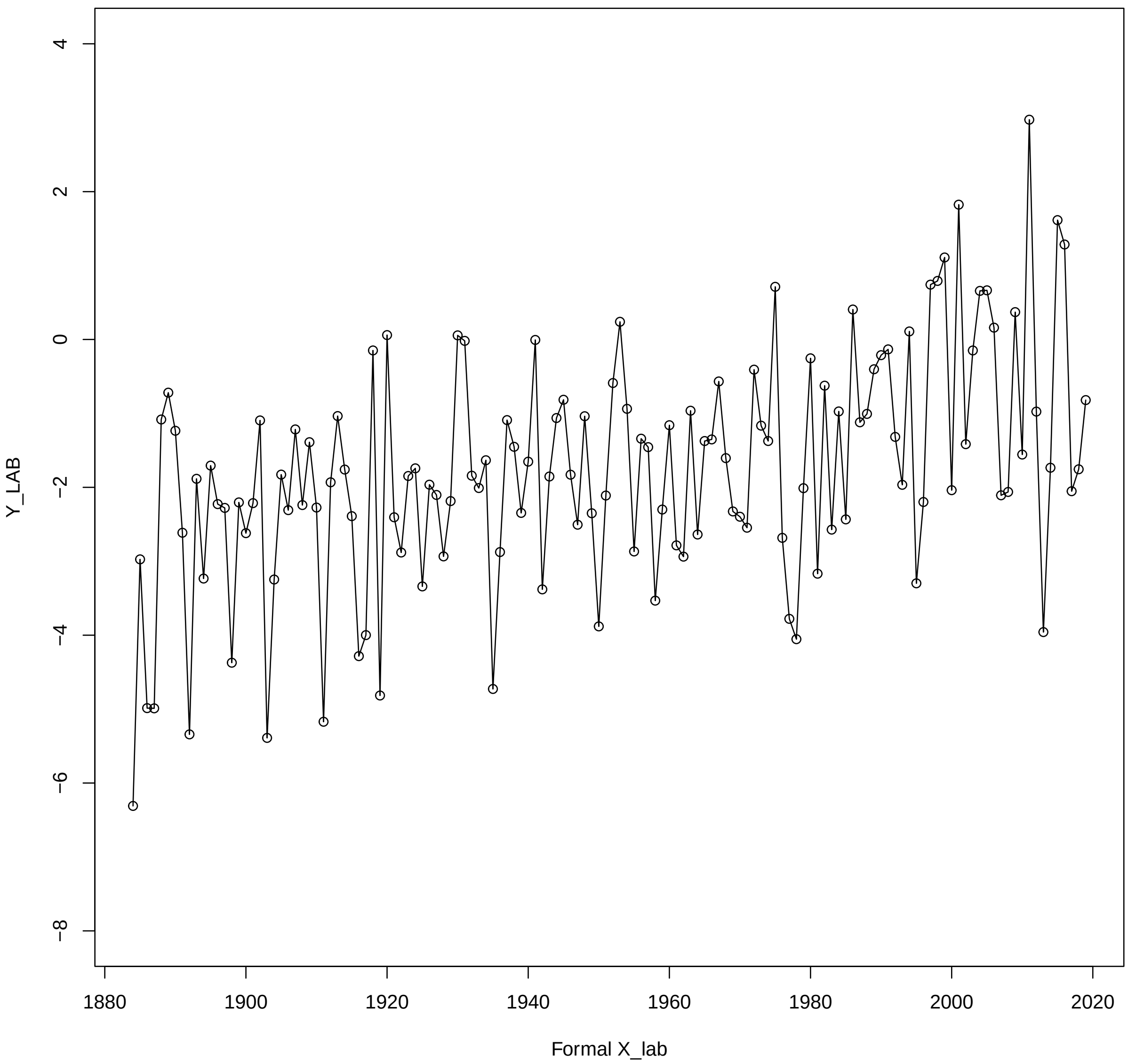
```
plot(temp~year, ylim=c(-8,4), xlab='Formal X_lab', ylab= 'Y_LAB', dat)
```



> 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='Formal X_lab', ylab= 'Y_LAB', dat)
lines(temp~year, dat)
```



4.3.1 Running a linear model

Running a linear model is very simple. We can use lm(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 β_0 and your β_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 β_0 and the slope (effect of year) is β_1. The larger β_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 ϵ 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.

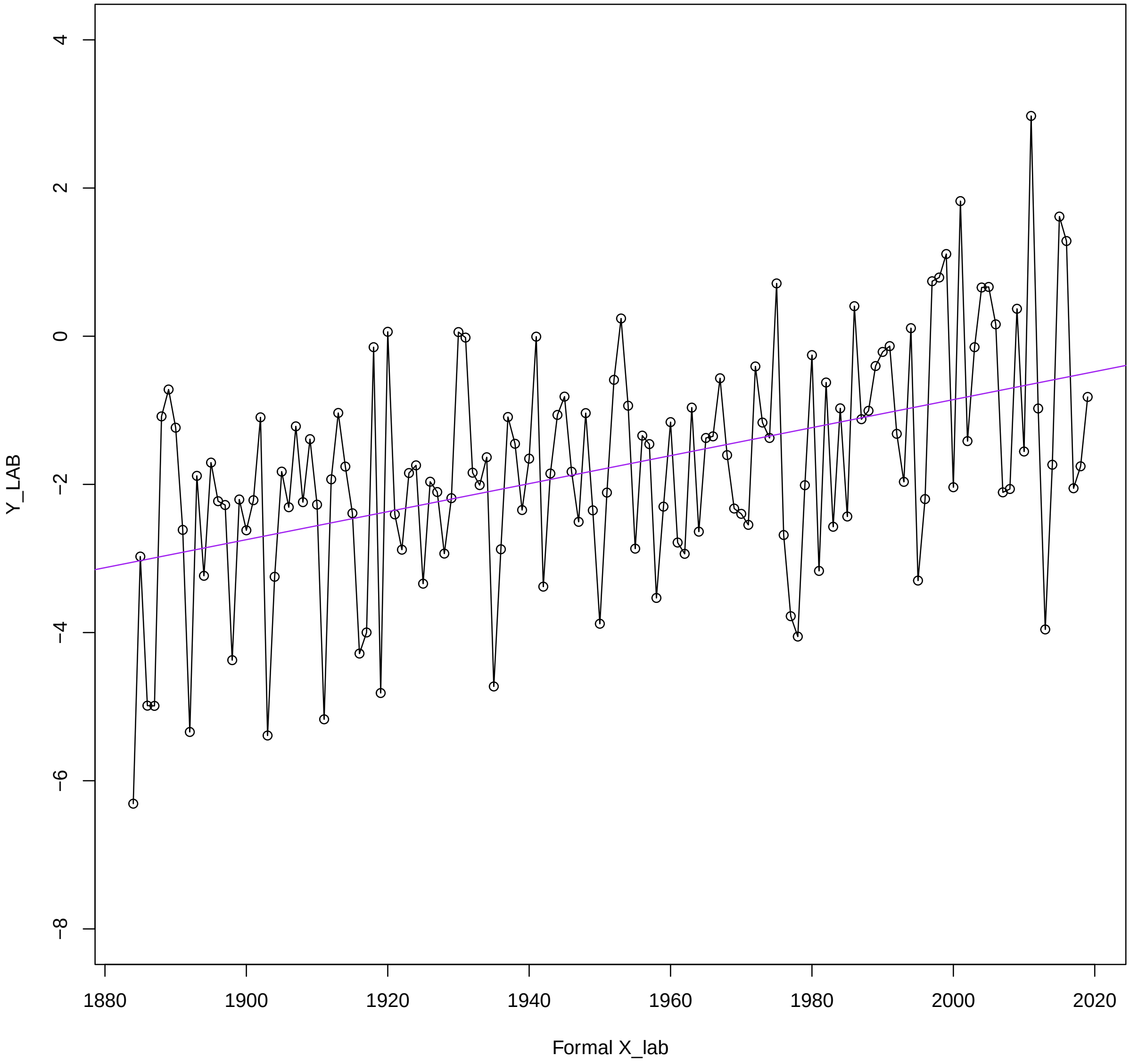
> Q6. 3 pts. a) Looking at the residuals, explain whether they look OK, or if there seems to be a problem with their distribution, b) substitute the coefficient values, and using R, obtain and report the predicted temperature for year 2000, and for year 2001. What is the difference in predicted temperature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

> 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 widht 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.

```r
plot(temp~year, ylim=c(-8,4), xlab='Formal X_lab', ylab= 'Y_LAB', dat)
lines(temp~year, dat)
abline(m1, col='purple')
```



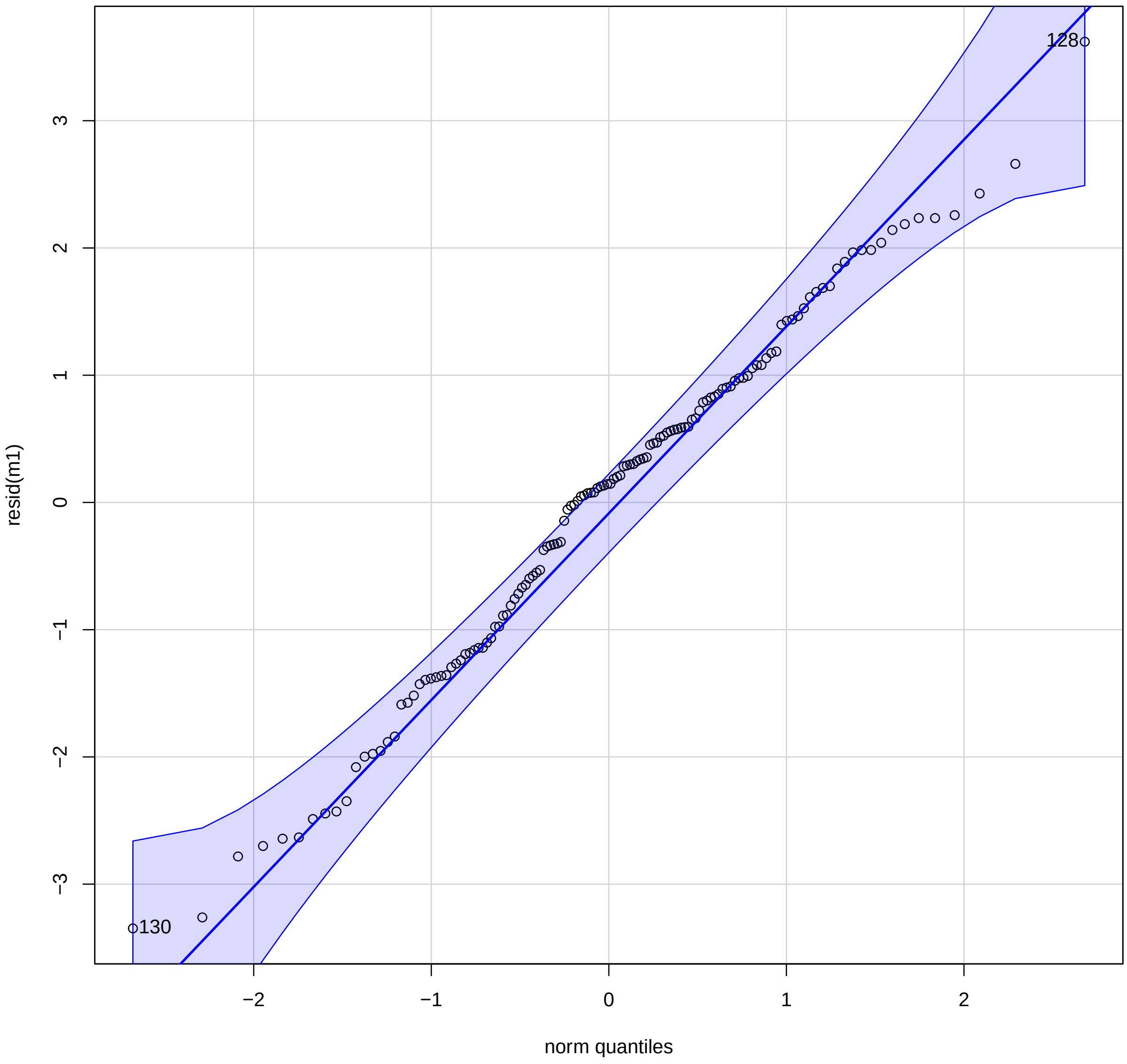
4.4 Assumptions

Now, during class we saw he 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
```

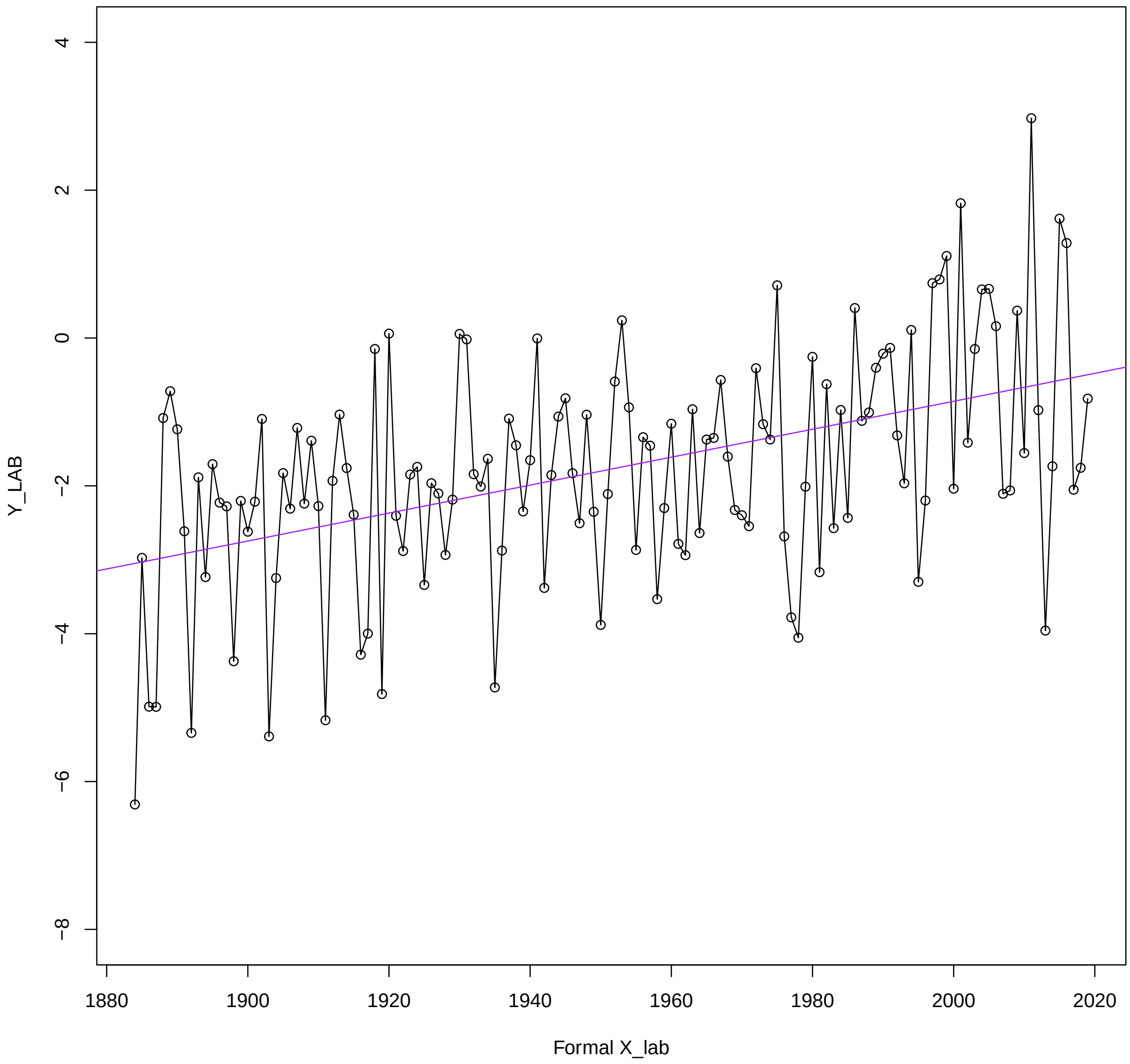
If you can't load a package using library, you need to install the package first!

Explore the Q-Q plot.

4.4.2 Linearity

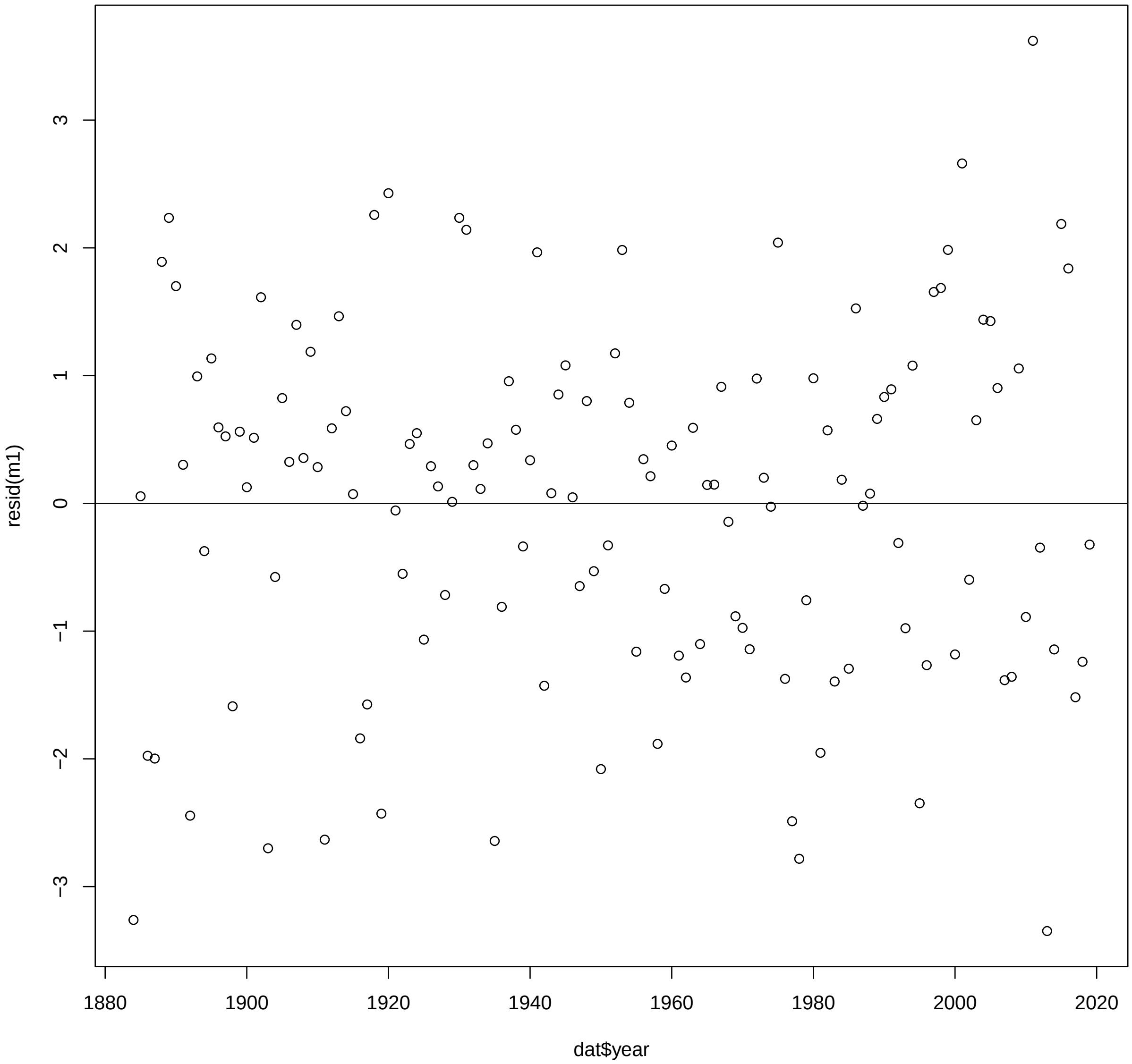
Explore the plot with the regression line:

```r
plot(temp~year, ylim=c(-8,4), xlab='Formal X_lab', ylab= 'Y_LAB', 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 agains the predictor using:

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



Remember that errors ϵ_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

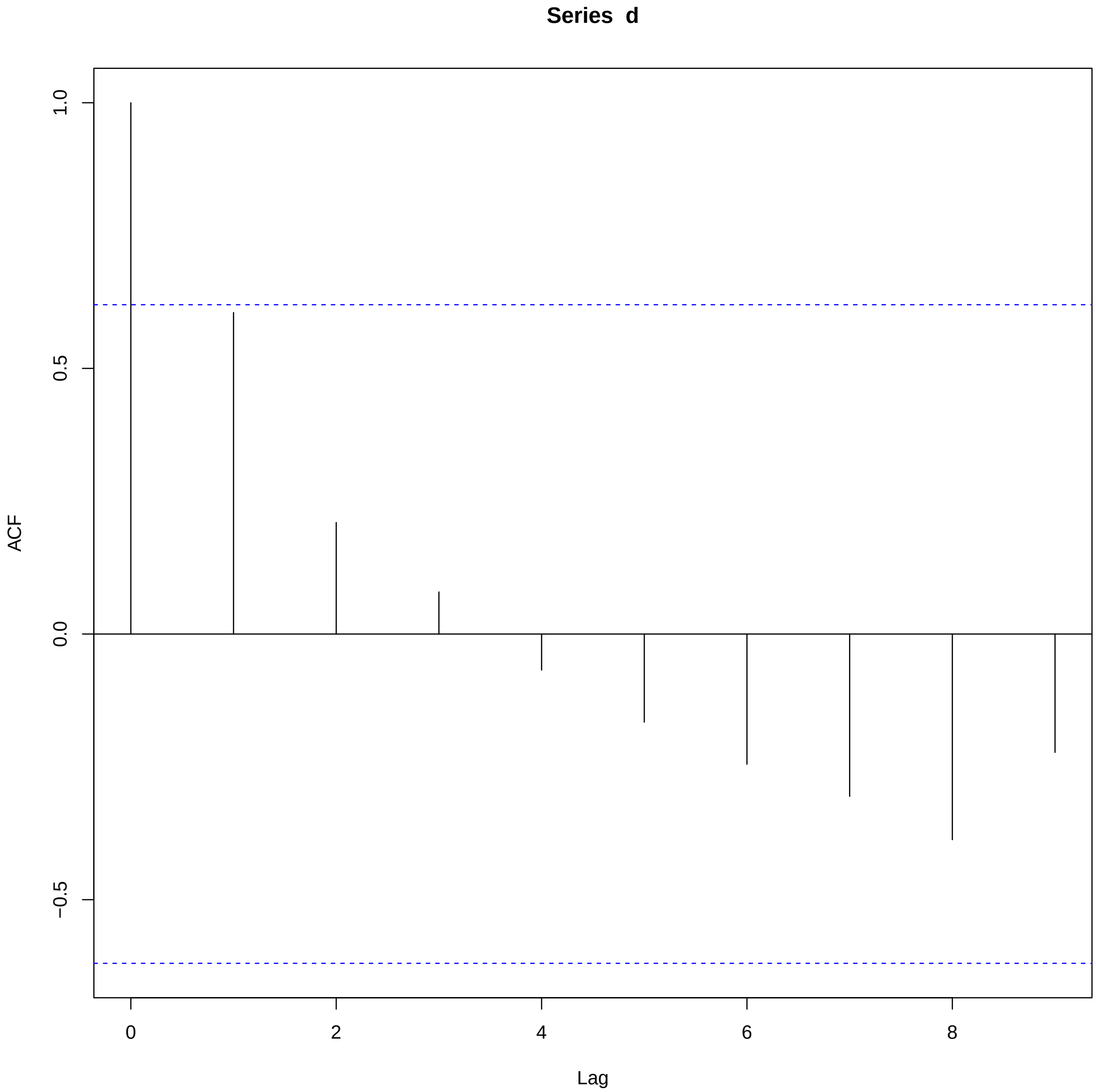
Residuals should have an equal variance across the range of the predictor variable. We can use the residual vs. predictor response plots to evaluate this assumption. You already run this plot!

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

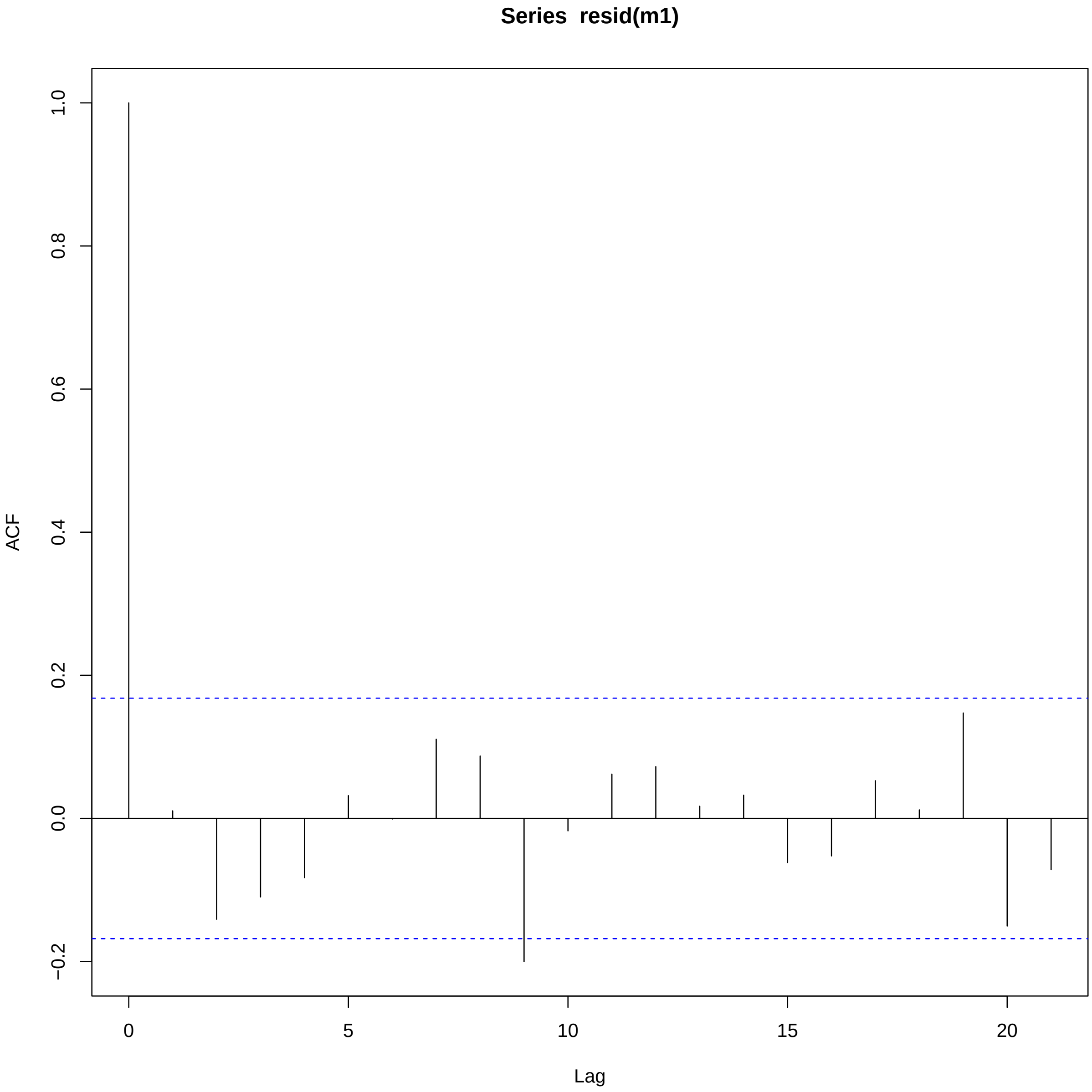
In this case, the observations cannot be considered as independent because observations in successive years (say 2000 and 2001) would likely be more similar to each other compared to observations decades apart (say, 2000 and 1950). In a time-series, observations are almost certainly non-independent by 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(lm)) function. If the plot looks like this:



It is autocorrelated. (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 autocorrelated. 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

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).

> 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).

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.

Lab total: 32 pts

end of lab

Resources, extra credit and sample code

These are hyperlinks

The base R cheat-sheet

An introduction to R

Research notes

> 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

```r
## data processing steps for L19
## author: Xingli Giam (xgiam@utk.edu)
    ## roughly following approach here: https://lter.github.io/lterdatasampler/articles/ntl_icecover_vig

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

## explore the two datasets # use print() instead of head() because this is a tidyverse tibble datafram
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

# 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$month <- month(ntl_airtemp$sampledate)

    # 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 sampledate, we c

# 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 2022 t
    # 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 f
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 month
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() func
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
    head(AvgTempDat, 50) ## looks good! # each row represents the mean daily cold season temp for a giv

# now, let's combine the temp and ice duration dataset, making sure that each row refers to the ice dur
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-co

    # 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
```