Linear Regression Final Report

------- Estimating the relationship between temperature anomaly and pollution using the linear regression method

Icey

December 2, 2022

Statement of Purpose

Global temperature has undergone a tremendous change as more and more human activities are involved. According to NASA statistics, extreme events, including wildfires, floods, and hurricanes, have become a frightening new normal. Hotter temperatures, air pollution, and violent storms are leading to immediate, life-threatening dangers for people, including difficulty breathing, malnutrition, and a higher risk of infectious diseases. Our group aims to analyze the relationship between temperature anomaly and pollution to lot further mitigate the extreme weather problem.

Introduction

Jes Fenger (2009) explicitly explained how air pollution could drive a huge temperature change in the last 50 years. Frederica Perera (2022) studied the relationship between fossil-fuel pollution and climate temperature. However, most recent literature only focused on one type of contamination. Our group considers that there will be some interactions among different types of pollution. In this article, our group will consider the collective impact of water pollution, air pollution, and land pollution on climate change.

Data collection

Year	Median temperature anomaly from 1961-1990 average	C02 emission	Forest area	Plastic production	oil leak	Ferterlize use	living planet index	O3 layer
1979	0.057	19610017729	#N/A	71000000	636000	111581366	83.34748	225
1980	0.092	19492613515	#N/A	70000000	206000	115808264	81.5080934	203
1981	0.14	19026910179	#N/A	72000000	48000	115094406	79.8368748	209.5
1982	0.011	18875760275	#N/A	73000000	12000	115399888	77.2813851	185
1983	0.194	18998031557	#N/A	80000000	384000	123657138	74.5181848	172.9
1984	-0.014	19656716874	#N/A	86000000	29000	130461720	71.8320243	163.6
1985	-0.03	20317483523	#N/A	90000000	85000	129244448	69.7531144	146.5
1986	0.045	20619623658	#N/A	96000000	19000	132786820	68.1256656	157.8
1987	0.192	21266194534	#N/A	104000000	38000	138463906	66.322487	123
1988	0.198	22101399490	#N/A	110000000	190000	144297750	64.5467006	171
1989	0.118	22406832518	#N/A	114000000	164000	141854332	63.5743437	127
1990	0.296	22749574360	4236433020	120000000	61000	136971688	62.6811227	124.2
1991	0.254	23238102148	4228594630	124000000	431000	133526594	61.6091396	119
1992	0.105	22569624280	4220756240	132000000	167000	124213148	60.1198509	114.3
1993	0.148	22803495349	4212917850	137000000	140000	119011616	58.1092963	112.6
1994	0.208	22964583359	4205079460	151000000	130000	121890670	56.2008977	92.3
1995	0.325	23453944453	4197241070	156000000	12000	129114058	54.0360964	#N/A
1996	0.183	24154961206	4189402680	168000000	80000	134127396	53.0421358	108.8
1997	0.39	24300550431	4181564290	180000000	72000	136832796	51.7500764	108.8
1998	0.539	24206271270	4173725900	188000000	13000	137925312	50.4664511	98.8

Figure 1

Our data contains 41 rows dated from 1979 to 2021, and 9 rows including one target response that is temperature anomaly, and 7 candidate variables which are CO2 emission, Ozone layer (air pollution), Forest area, Plastic production, Living planet index (land pollution), Oil

leak, Fertilizer use (water pollution). We obtained the temperature anomaly data from climate.gov.com and received water pollution data from Kaggle. Land pollution data and air pollution data are all collected from the University of OXFORD world data library.

Data analysis and results

1. Data cleaning

As we can see from figure 1, our original data set contains multiple NA values. The data is missing mainly because of lacking records in the early year or missing values under specific situation. In the project, our group use the "random substitution" method to fulfill the NA values. For example, for NA values in the Forest area column, we randomly draw a value from later years to substitute the early years' value. Moreover, for some unknown-reason-missing data in the other columns, we randomly sample a value from this column and set them as the missing values.

2. Data scaling

Our data set contains different types of variables, and each of them has a different unit and is not measured in the same way. Therefore, we normalized our dataset after data cleaning using the mean value and standard deviation.

3. Scatter plot matrix of response and variables.

From figure 2, we can see an obvious linear relationship between some of the variables and the response, especially in CO2 emission, Plastic production, and living planet index. Moreover, we can also notice some possible interactions between variables, for example, Forest area is highly likely to have an inverse relationship with Plastic production, and the living planet seems to increase with the increase of forest area. Therefore, our group is going to figure out the factors that may contribute to temperature anomaly.

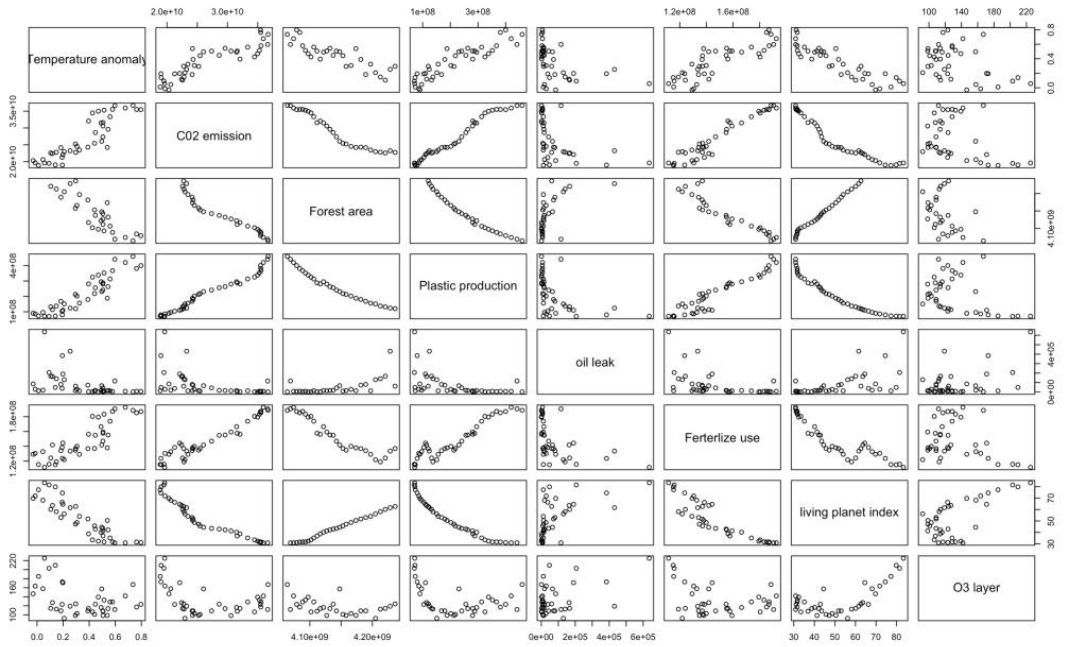

Figure 2

4. Residual checking

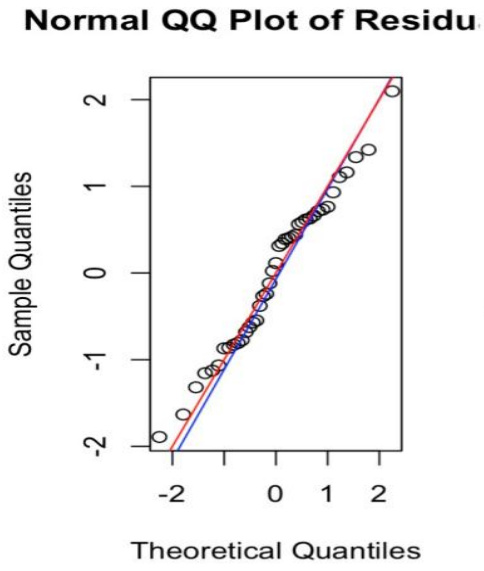

Figure 3

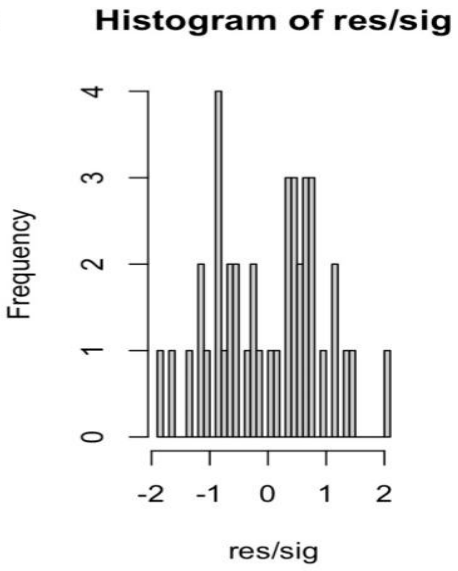

Figure 4

First, we use the QQ plot to check the residuals. In the QQ plot (Figure 3), the red line is a straight line that has intercept zero and slope 1, and the blue line is the residual line. From the

QQ plot, we can find that the blue line and the red line are almost overlapped. What is more, we also drew the frequency histogram of a standardized residual (Figure 4). From the plot, we can find that the standardized residual of this regression model follows a normal distribution. From these two p, we can pass the normal assumption for our model.

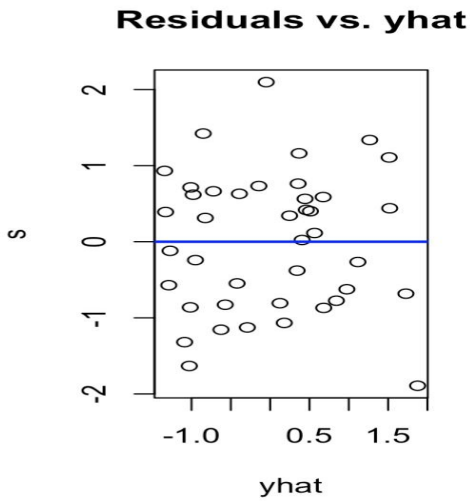

Figure 5

In Figure 5, we found that residuals are randomly distributed around 0 with no specific pattern within a fixed range, which means the residual has a constant variance. When both the assumption of linearity and homoscedasticity are met, the points in the residual plot will be randomly scattered. And model inferences like confidence intervals and model predictions should also be valid.

5. Model selection

According to previous research, seven factors may contribute to temperature anomaly. However, a high-dimensional linear regression model may result in overfitting and poor prediction. So we decide to do a variable selection to reduce the dimension of the linear model.

First, we we consider all possible subsets by comparing models adjusted R^2, Figure 6 and Figure 7 show the seven best models with different model sizes, and the model with five parameters which are carbon dioxide, forest area, plastic, living planet index, and ozone layers, has the largest adjusted R^2 over 0.84.

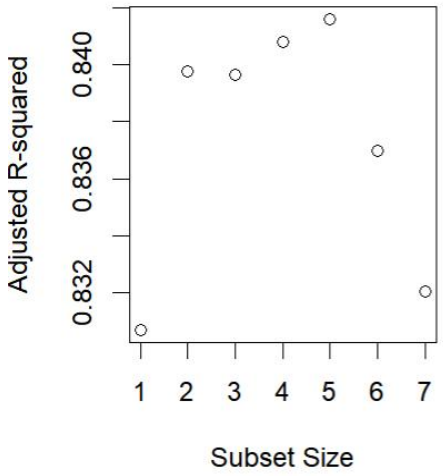

Figure 6

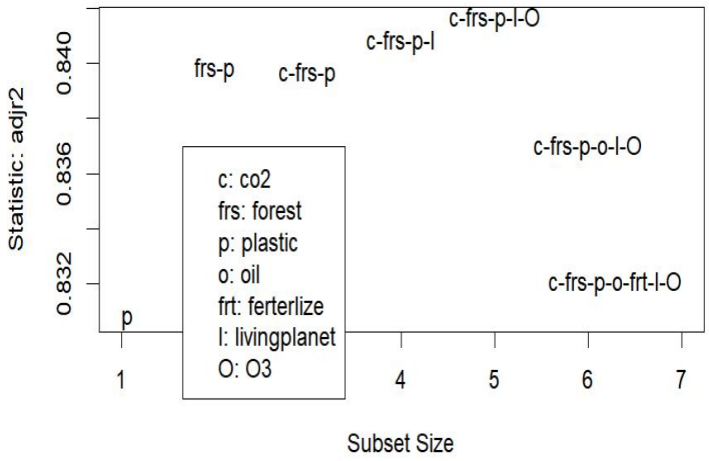

Figure 7

And then we compare the AIC, AICc, and BIC of these models, the model that has the minimum AIC and minimum AICc is model 2 which includes forest area and plastic. Model 3 has the minimum BIC which only includes plastic production.

Finally, we compare the parameter significance of model 1, model 2, and model 5 but all of them only have one statistically significant variable. Therefore, we select the model with the lowest residual standard error, model 5. The residual standard error of model 5 is 0.398 and the adjusted R square is 0.8416. In conclusion, we consider model 5 is best-fitted, which means the global temperature has a linear relationship with carbon dioxide, forest area, plastic, living planet, and ozone.

$$y = 5.246e-17 - 5.920e-01x_1 + 1.586e-01x_2 + 1.464x_3 - 2.110e01x_6 + 0.1198ex_7$$

6. Forward selection and backward selection

We also use forward selection to check whether linear model is well-fitted, and then the result obtained is that model 2 which includes plastic and forest area has the least AIC = -72.19.

$$y = 1.965e-16 + 9.813e-01x_3 + 1.319e-01x_2$$

And model included only plastic has the least BIC = -67.45.

$$y = -1.327e-16 + 9.138e-01\, x_3$$

And then we use backward selection and the result is same as forward selection.

7. LASSO and Ridge

7.1 LASSO

For variable selection and regularization, we also use the LASSO method to improve our model's prediction accuracy and interpretability. We choose the penalty parameter λ which gives the minimum mean cross-validated error.

We use two-thirds of the data as the training set, and the other as the test set, we fit the data by using LASSO regression, and the model we got:

$$y = 0.1744x_2 + 0.9564x_3 - 0.0754x_4 - 0.0463x_6$$

x_2, x_3, x_4, x_6denote forest area, plastic production, oil leak, and living planet index respectively, and other variables have coefficients equal to zero.

The fitted model has $R^2 = 0.85$ for the training set and has $R^2 = 0.839$ for the test set, which performed well.

7.2 Ridge

From Figure 2 we find that there are some correlations between variables, then we consider using ridge regression to fit the data. We also choose the penalty parameter λ which gives the minimum mean cross-validated error.

We also use two-thirds of the data as the training set, and the other as the test set. Then we fit the model as below:

$$y = 0.0064 + 0.1828x_1 + 0.1129x_2 + 0.4873x_3 - 0.048x_4 + 0.1975x_5 - 0.0751x_6 - 0.0441x_7$$

The fitted model has $R^2 = 0.83$ for the training set and has $R^2 = 0.7994$ for the test set.

8. Conclusion

By comparing the R^2 of the fitted LASSO and Ridge models, we can find that the model built by the LASSO method performs better than the Ridge model. Then we prefer to choose the model which is built by the LASSO method, which means the forest area, plastic production, oil leak, and living planet index have a stronger linear relationship with the anomaly temperature.

Furthermore, if we compare the LASSO model with the model we choose from the forward selection in section 5, we can find that both of these two models choose forest area, plastic production, and living planet index as variables that have linear relationships with the anomaly temperature. Then we can conclude that all of these three variables have the strongest linear influence on the target response which is anomaly temperature.

In conclusion, we prefer to choose the LASSO model or the model selected by the forward/backward selection methods to fit the dataset:

$$y = 1.965e-16 + 9.813e-01x_3 + 1.319e-01x_2$$

$$y = 0.1744x_2 + 0.9564x_3 - 0.0754x_4 - 0.0463x_6$$

Appendix

```
library("glmnet")
library(leaps)
library("ridge")

##data
data=read.csv("~/STAT605/data/615data.csv",na.strings = NA )
data=data[-1:-129,]
y=data$Median.temperature.anomaly.from.1961.1990.average
x=data.frame(co2=data$C02.emission,
             forest=data$Forest.area,
             plastic=data$Plastic.production,
             oil=data$oil.leak,
             ferterlize=data$Ferterlize.use,
             livingplanet=data$living.planet.index,
             O3=data$O3.layer)


##missing data
random.imp <- function (a){
  missing <- a=="#N/A"
  n.missing <- sum(missing)
  a.obs <- a[!missing]
  imputed <- a
```

```
  imputed[missing] <- sample (a.obs, n.missing, replace=TRUE)
  return (imputed)
}
for (i in 1:length(x[1,])) {
  x[,i]<-as.numeric(random.imp(x[,i]))
}

##scale data
yscale=scale(y)
xscale=scale(x)
n=length(x[,1])
p=length(x[1,])+1

df=data.frame(y=yscale,
  co2=x$co2,
            forest=x$forest,
            plastic=x$plastic,
            oil=x$oil,
            ferterlize=x$ferterlize,
            livingplanet=x$livingplanet,
            O3=x$O3)

plot(x, y)

##fit data
fit=lm(y ~ co2+forest+plastic+oil+ferterlize+livingplanet+O3,data=df)
```

```
sr=summary(fit)


##check residual
res=sr$residuals
sig=sr$sigma
hist(res/sig, breaks = 50)
qqnorm(res/sig, main = "Normal QQ Plot of Residuals") # use '?qqnorm' to see the help document
qqline(res/sig, col = "blue")
abline(a = 0, b = 1, col = 'red')
yhat=fit$fitted.values
s=res/sig
plot(yhat, s, main = "Residuals vs. yhat"); abline(h = 0, lwd = 2, col = "blue")


## variable selection
b <- regsubsets(xscale,yscale)
rs <- summary(b)
par(mfrow=c(1,2))
plot(1:7,rs$adjr2,xlab="Subset Size",ylab="Adjusted R-squared")
subsets(b,statistic=c("adjr2"))
om1 <- lm(yscale~xscale[,3])
om2 <- lm(yscale~xscale[,3]+xscale[,2])
om3 <- lm(yscale~xscale[,1]+xscale[,3]+xscale[,6])
om4 <- lm(yscale~xscale[,1]+xscale[,2]+xscale[,3]+xscale[,6])
om5 <- lm(yscale~xscale[,1]+xscale[,2]+xscale[,3]+xscale[,6]+xscale[,7])
om6 <- lm(yscale~xscale[,1]+xscale[,2]+xscale[,3]++xscale[,4]+xscale[,6]+xscale[,7])
```

```
om7 <- fit


om<-list(om1,om2,om3,om4,om5,om6,om7)
AIC<-c()
AICc<-c()
BIC<-c()
for (i in 1:7) {
  n <- length(om[[i]]$residuals)
  npar <- length(om[[i]]$coefficients) +1
  AIC<-rbind(AIC,extractAIC(om[[i]] ,k=2))
  AICc<-rbind(AICc,extractAIC(om[[i]],k=2)+2*npar*(npar+1)/(n-npar-1))
  BIC<-rbind(BIC,extractAIC(om[[i]],k=log(n)))
}
which(AIC[,2]==min(AIC[,2]))
which(AICc[,2]==min(AICc[,2]))
which(BIC[,2]==min(BIC[,2]))
backAIC  <- step(fit,direction="backward", data=df)
backBIC <- step(fit,direction="backward", data=df, k=log(n))
mint<- lm(y~1,data=df)
forwardAIC <- step(mint,scope=list(lower=~1,
                          upper=~co2+forest+plastic+oil+ferterlize+livingplanet+O3),
                direction="forward", data=df)
forwardBIC <- step(mint,scope=list(lower=~1,
                          upper=~co2+forest+plastic+oil+ferterlize+livingplanet+O3),
                direction="forward", data=df,k=log(n))
```

```
##lasso
##choose training set
t=sample(seq(1,n),round(n*2/3))
tx=xscale[t,]
ty=yscale[t,]

##choose test set
rx=xscale[-t,]
ry=yscale[-t,]

##fit lasso
aaa=cv.glmnet(tx,ty,alpha=1)
blam=aaa$lambda.min
bm=glmnet(tx,ty,alpha=1,lambda = blam)
summary(bm)
bm$dev.ratio
lcoef=as.numeric(coef(bm))

##compute R-square
xbind=cbind(1,rx)
reslasso=sum((ry-xbind%*%lcoef)^2)
sstot=sum((ry-mean(ry))^2)
1-reslasso/sstot


##ridge
```

```
##fit ridge
ridge=aaa=cv.glmnet(tx,ty,alpha=0)
rlam=ridge$lambda.min
rigr=glmnet(tx,ty,alpha=0,lambda = rlam)
rigr$dev.ratio
rcoef=as.numeric(coef(rigr))

##compute R-square
resridg=sum((ry-xbind%*%rcoef)^2)
sstot=sum((ry-mean(ry))^2)
1-resridg/sstot
```

Reference

[1] Water bodies and water quality world, 2020. (n.d.). University of OXFORD world data library. Retrieved November 2, 2022, from https://ourworldindata.org/grapher/water-bodies-good-water-quality?tab=chart&country=~OWID_WRL

[2] Kaggle. (n.d.). World cities air quality and water pollution. 2020. Retrieved November 2, 2022, from https://www.kaggle.com/datasets/cityapiio/world-cities-air-quality-and-water-polution

[3] NASA's Goddard Institute for Space Studies (GISS). (n.d.). GLOBAL LAND-OCEAN TEMPERATURE INDEX. Retrieved from https://climate.nasa.gov/vital-signs/global-temperature/

[4] How to Impute Missing Values in R, from https://www.geeksforgeeks.org/how-to-impute-missing-values-in-r/