Un estudi sobre l'efecte del monòxid de carboni en l'estat de benestar d'un país.

Universitat Autònoma de Barcelona
Anàlisi de Dades Complexes 2020-2021

Bernat Espinet Torrescassana

1564342

Figure 1: Air Pollution as Bad as Smoking for Human Health,engoo

Contents

1 Abstract

L'objectiu d'aquest anàlisi de dades és buscar quin tipus de relació hi ha entre la presència de monòxid de carboni (CO) amb la salut mental d'una regió. Per fer-ho, crearem diversos models i realitzarem varies tècniques de remostreig per tal d'arribar a una conclusió.

2 Motivació

El monòxid de carboni és un gas que s'allibera principalment en la combustió de diversos combustibles fòssils. El monòxid de carboni contribueix en la formació de gasos amb efecte hivernacle, ja que a través d'un procés lent d'oxidació, el monòxid de carboni es transforma en diòxid de carboni. També s'han donat casos d'intoxicació per una alta concentració del gas, ja que no permet l'absorció d'oxigen a la sang. En casos més moderats, el monòxid de carboni provoca marejos, fatiga i disfuncions nervioses.

En el nostre cas, no ens centrarem en l'impacte en la salut dels habitants, ens enfocarem en buscar si la presència d'aquest gas en diverses concentracions té un impacte en la qualitat de vida d'un país i en general, en el nivell de benestar públic, tenint en compte tant la salut fisica com la mental. Per tal de mesurar-ho, farem servir un estudi anomenat Life Ladder que ens proporciona una puntuació sobre 10, on 0 és un estat de desesperació i desesperança, i el 10 és de prosperitat màxima. Per fer-nos una idea, Espanya en el 2020 va puntuar 6.502.

En el 2020, el país amb més puntuació va ser Finlàndia amb una puntuació de 7.889 i per la cua tenim Zimbàbue amb una puntuació de 3.160.

ï..country	Life_Ladder
Finland	7.889
Iceland	7.575
Denmark	7.515
Switzerland	7.508
Netherlands	7.504
Sweden	7.314
Germany	7.312
Norway	7.290
New Zealand	7.257
Austria	7.213
Israel	7.195
Australia	7.137
Ireland	7.035
United States	7.028
Canada	7.025
Czech Republic	6.897
Belgium	6.839
United Kingdom	6.798
ıiwan Province of China	6.751
France	6.714
Saudi Arabia	6.560
Slovakia	6.519
Croatia	6.508
Spain	6.502
Italy	6.488
Slovenia	6.462
United Arab Emirates	6.458
Estonia	6.453
Lithuania	6.391
Uruguay	6.310
Kosovo	6.294
Cyprus	6.260
Kyrgyzstan	6.250
Latvia	6.229
Bahrain	6.173
Kazakhstan	6.168
Malta	6.157
Chile	6.151
Poland	6.139
Japan	6.118
Brazil	6.110
Serbia	6.042

Figure 2: Ranking de paisos segons Life Ladder (2020)

3 Dades

Per poder dur a terme aquest anàlisi, farem servir dues taules de dades. La primera taula que farem servir [1] publicada per OpenAQ i és actualitzat cada dia. OpenAQ és una plataforma que se centra a maximitzar la transparència de les dades i facilitant el tractament de la informació. Es tracta d'una taula que conté diverses mostres d'aire on s'ha analitzat la presència de diversos gasos arreu del món a partir del febrer del 2007.

També farem servir un segon conjunt de dades[2] publicat en el per Ajaypal Singh, un estudiant de la universitat Punjabi de la Índia. D'aquesta segona taula traurem la puntuació del test Life Ladder per a diversos anys i països.

Les dades i el codi en R el podeu trobar en el seguent enllaç de Github: https://github.com/beesto72/ADC

4 Anàlisi

Abans de crear hipòtesis, models i treballar amb els resultats, primer hem de manipular una mica les dades. Per començar, en el dataset de OpenAQ, trobem que les mostres d'aire es prenen en unitats diferents, per tant el primer que farem és transformar totes les files que tinguin unitats = PPM, aplicar el canvi de variables considerant que ens trobem en condicions estàndard per tenir totes les entrades amb les mateixes unitats $\mu g/m^3$.

També, d'aquesta taula de dades ens interessa extreure l'any de la columna de data, per això escorçarem els valors que es troben en la columna last Updated als 4 primers valors corresponent als anys. Per últim, en tractar amb dades de dos taules diferents, per evitar errors, buscarem quins països tenim en comú en les dues taules, i més tard, quins anys tenim en comú per aquest país.[3]

Per crear el nostre primer model, agafarem les mostres únicament d'Espanya, i en els casos de tenir més d'una mostra d'aire en el mateix any, ens guardarem la mitjana. Fent això aconseguim dos vectors, el primer que conté la puntuació d'Espanya en el test de Life Ladder en diversos anys i la seva concentració mitjana de cada any de CO. Per aquest primer model proposem la següent fórmula $Lifeladder\ log(CO) + intercept$ [4]:

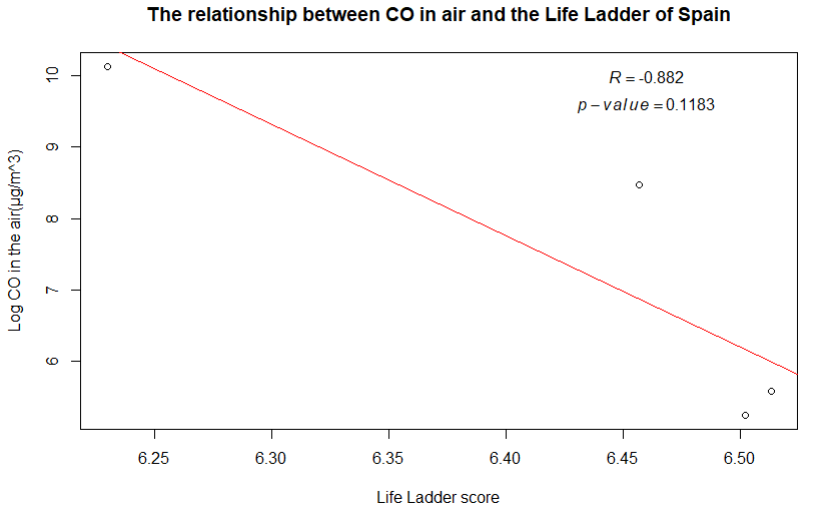

Figure 3: Representació de la relació entre el CO i la puntuació Life Ladder en Espanya

I obtenim el següent summary:

```
Call:
lm(formula = data_s[, 2] ~ data_s[, 1])

Residuals:
      1       2       3       4
-0.2826 -0.4076 -0.9183  1.6086

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  107.661     37.950   2.837    0.105
data_s[, 1]  -15.610      5.905  -2.643    0.118

Residual standard error: 1.356 on 2 degrees of freedom
Multiple R-squared:  0.7775,    Adjusted R-squared:  0.6662
F-statistic: 6.988 on 1 and 2 DF,  p-value: 0.1183
```

Figure 4: Resum del model obtingut per Espanya

Veiem que per aquest model, el R^2 és prou bo, però en contar únicament amb 4 punts, no ens en podem refiar. Per la poca quantitat de punts, també trobem un P-Value per sobre 0.05, el qual no ens permet refusar la H0.

Per tant ara anem a recollir dades de tots els països i anys que tenim en comú, i obtenim 60 punts que són combinacions de país i any que tenim disponibles. Si fem un simple plot dels punts obtinguts ara veiem el següent:

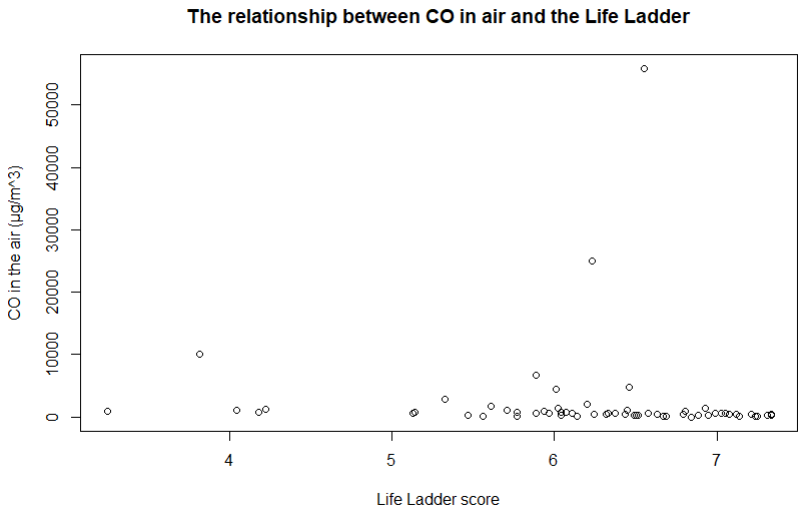

Figure 5: Representació gràfica dels punts obtinguts

Observant aquest scatterplot, veiem que tenim un parell d'anomalies que poden ser a causa de tenir poques mostres per aquella precisa combinació país/any o que el diagnòstic de l'aire fos en una regió molt industrialitzada. Per tant, plantegem el mateix model lineal que per únicament Espanya amb la fórmula $Lifeladder\ log(CO) + intercept$[5]:

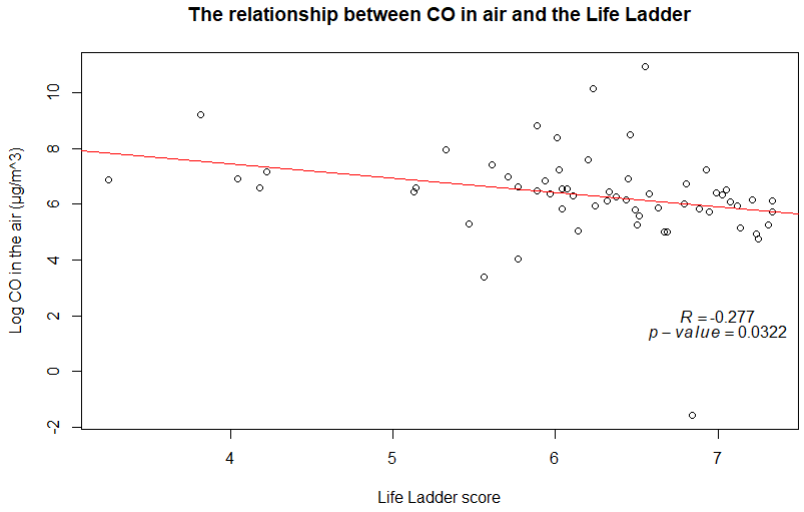

Figure 6: Representació de la relació entre el CO i la puntuació Life Ladder

I obtenim el següent summary:

```
Call:
lm(formula = data[, 2] ~ data[, 1])

Residuals:
    Min      1Q  Median      3Q     Max
-7.5537 -0.5220 -0.0058  0.4471  4.7893

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.4765     1.4539   6.518 1.88e-08 ***
data[, 1]    -0.5094     0.2321  -2.195   0.0322 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.608 on 58 degrees of freedom
Multiple R-squared:  0.07668,   Adjusted R-squared:  0.06076
F-statistic: 4.817 on 1 and 58 DF,  p-value: 0.0322
```

Figure 7: Resum del model obtingut

Veiem en aquest cas, que el p-valor es troba per sota el llindar 0.05, per tant refusem la H0 i podem afirmar que si que hi ha una correlació entre les dues variables. Però si ens fixem en el coeficient de Pearson, veiem que és prou baix (-0,277), és a dir que la relació és prou feble. També veiem que la nova recta de regressió difereix força en pendent del model d'Espanya.

Per acabar de perfilar el P-valor i afirmar amb certesa que podem refusar la H0, fem un bootstrap no paramètric i obtenim el següent p-valor[6]:

```
> experimental_pvalue
[1] 0.0354
```

Figure 8: P-valor obtingut del Bootstrap no paramètric

Per tant podem afirmar amb certesa que la H0 queda refusada.

Sabem que Espanya se'n va col·locar en el top 24 l'any 2020 i ara que tenim el model proposat, un experiment interessant és veure quina concentració de CO en l'aire hauríem de mantenir per tal d'arribar al top 10, és a dir puntuar 7,213 en el test. Per fer-ho, realitzarem un test permutacional per calcular un interval de confiança del 95% per una puntuació al test Life Ladder. Obtenim el següent Histograma[7]:

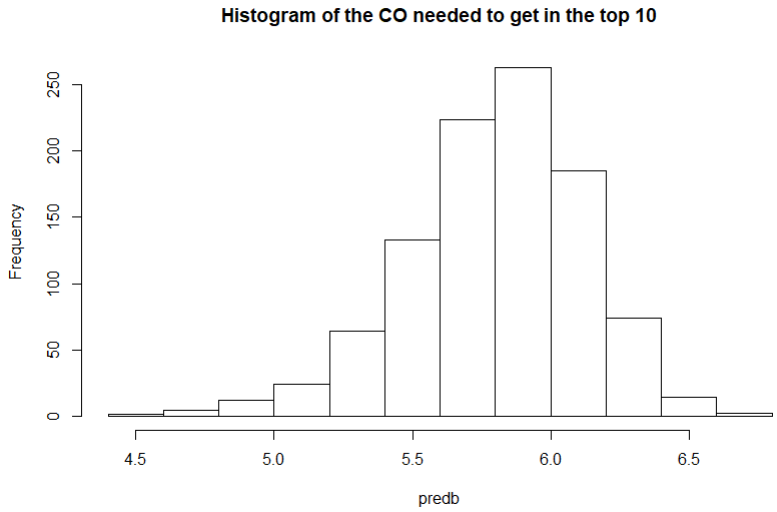

Figure 9: Histograma obtingut de les prediccions

I si calculem el Interval de confiança del 95%, obtenim el següents valors:

```
> Prediction_CO_to_top10Score
    2.5%    97.5%
5.127544 6.394568
```

Figure 10: Interval de confiança 95% per entrar en el top 10

Per tant, necessitaríem mantenir una concentració de monòxid de carboni entre 168.6018 µg/m^3 i 598.5798 µg/m^3.

Per tal d'aconseguir uns millors resultats, efectuem una modificació en el model inicial, com tenim quantitats de mostres molt dispars de combinacions any/país, el que podem fer és atorgar més pes a aquells punts on tenim més mostres, és a dir sobreposar el mateix punt N cops, on N és el nombre de mostres que tenim. Tornem a proposar un model lineal, però en aquest cas, tindrem 613 mostres, i fent servir la fórmula $Life ladder\ log(CO) + intercept$ Obtenim el següent model amb pesos normalitzats[8]:

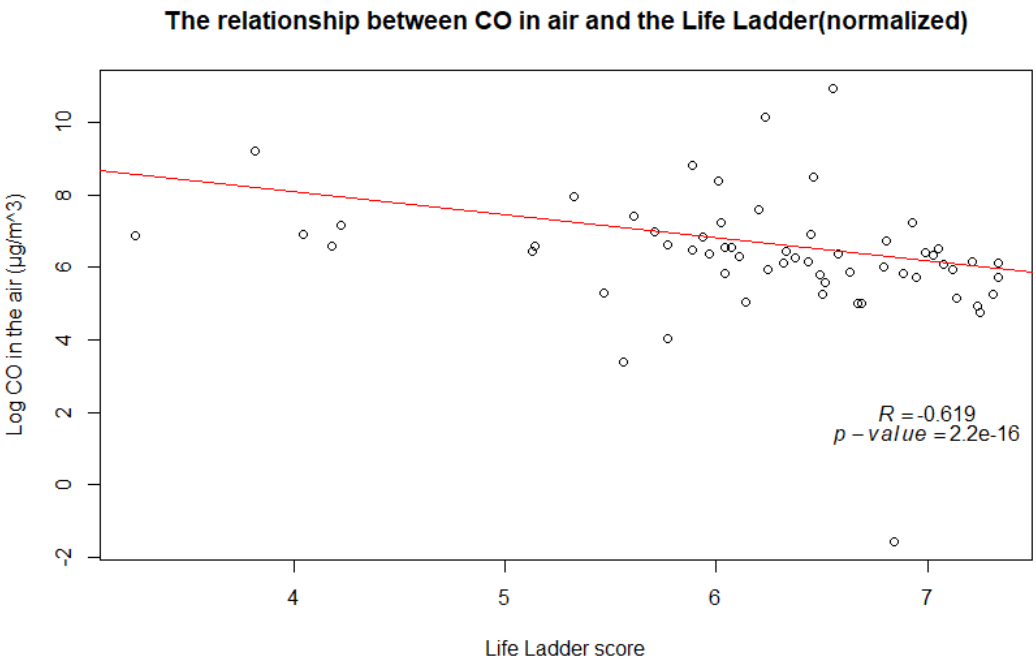

Figure 11: Representació de la relació entre el CO i la puntuació Life Ladder amb els pesos ajustats

I obtenim el següent summary:

```
Call:
lm(formula = data[, 2] ~ data[, 1])

Residuals:
    Min      1Q  Median      3Q     Max
-7.8355 -0.3210  0.1702  0.1702  4.4692

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.66325    0.20918   50.98   <2e-16 ***
data[, 1]   -0.64167    0.03291  -19.50   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9043 on 611 degrees of freedom
Multiple R-squared:  0.3835,    Adjusted R-squared:  0.3825
F-statistic: 380.1 on 1 and 611 DF,  p-value: < 2.2e-16
```

Figure 12: Resum del model obtingut

En aquest cas, veiem que el reassignat de pesos ha afavorit al model, ja que el coeficient de Pearson arriba a -0.619, que ja es considera una correlació forta. També veiem que amb aquest nou model, no te sentit buscar el p-value a través d'un test permutacional boostrap, ja que es troba molt lluny del llindar 0.05. També veiem que la recta de regressió queda lleugerament diferent al model inicial.

Per tant, anem a repetir la predicció per arribar al top 10 amb el model millorat, en aquest cas ens dóna l'histograma següent[9]:

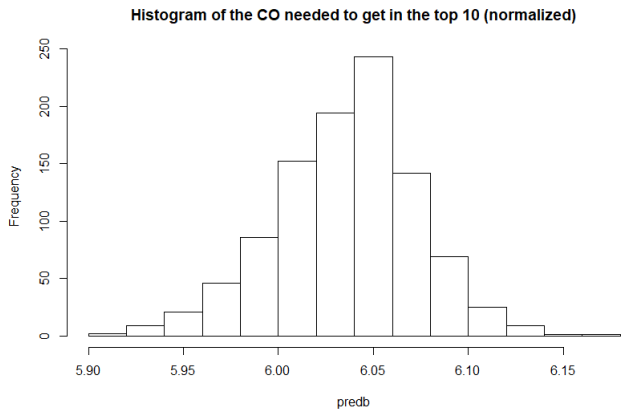

Figure 13: Histograma obtingut de les prediccions amb el nou model

Veiem que les prediccions del test permutacional es troben molt més compactes.

El nou interval de confiança del 95% obtingut és el següent:

```
> Prediction_CO_to_top10Score_norm
    2.5%    97.5%
5.955993 6.105340
```

Figure 14: Interval de confiança 95% per entrar en el top 10 amb el nou model

Per tant, fent servir el nou model, necessitaríem mantenir una concentració de monòxid de carboni entre 387.4977 µg/m^3 i 448.0765 µg/m^3, .

Com només tenim dades per Espanya de 4 anys, és difícil dir quina ha de ser la reducció percentual de contaminació. També, al només agafar un sol país, ens trobem amb el problema de què no tenim en compte la importància dels valors en funció del nombre de mostres. Però tot hi això, podem afirmar que una reducció de monòxid de carboni permetria augmentar la puntuació en el test Life Ladder en anys per venir.

5 Conclusió

Un cop finalitzat l'anàlisi, en podem treure les conclusions. Primer de tot, hem après que la presència de monòxid de carboni en l'aire no només afecta la salut física, sinó que també està força correlacionada amb l'estat de benestar de la població, és a dir, amb la salut mental.

Si passem a analitzar els resultats obtinguts, trobem que tot hi millorar força el model proposat amb la redistribució de pesos, segueix havent-hi molta variància en les dades. Això és causat principalment per dues coses, primer la no homogeneïtat de les mostres preses de l'aire, i el fet de no tenir en compte variables externes com podria ser l'impacte que hi ha hagut tant en la població com amb la contaminació l'arribada de la Covid-19. Per això mateix, en fer el bootstrap permutacional per calcular els Intervals de confiança del 95% per una certa predicció, obtenim un interval poc compacte. És a dir, que el nostre model no ens permet afilar prim.

Un cop hem vist que el monòxid de carboni afecta la població en més d'un aspecte, l'interessant seria intentar reduir la presència d'aquest gas dins de les zones urbanitzades. Aquest gas s'allibera en la combustió de combustibles a base de carboni, per tant, per part del govern es podria aplicar mesures per combatreho. Aquestes mesures podrien ser per exemple separar més les fàbriques de les grans ciutats, fomentar l'ús de la bicicleta com a mitjà de transport urbà i augmentar l'oferta de transport públic, ja que és una manera molt més neta de moure's que fer l'ús d'un cotxe particular.

6 Referències

- Monòxido de carbono,Ministerio para la transición ecológica y el reto demográfico, https://www.miteco.gob.es/es/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/calidad-del-aire/salud/monoxido-carbono.aspx

-[1]: OpenAQ,2007,Opendatasoft, https://public.opendatasoft.com/explore/dataset/openaq

-[2]: Ajaypal Singh, 2021, kaggle, https://www.kaggle.com/ajaypalsinghlo/world-happiness-report-2021

-[3]:

```
#First we are going to import both datasets
setwd("/Users/berna/Desktop/ADC treball/")
df_happiness <- read.table("/Users/berna/Desktop/ADC treball/world-happiness-report.csv",header = TRUE, sep = ",")
df_P <- read.table("/Users/berna/Desktop/ADC treball/openaq.csv",header = TRUE, sep = ";",fill=TRUE,stringsAsFactors = FALSE)
scalar1 <- function(x) {x / sqrt(sum(x^2))}
#Get the year by date of openaq dataset
df_P$Last_Updated <- substr(df_P$Last_Updated, 1,4)
#Get all data of CO pollution
to_purge=which(df_P$Pollutant == "CO")
df_Pollution=df_P[c(to_purge),]

#Set all values to same unit (µg/m^3)

for(row in 1:nrow(df_Pollution)){
  if(df_Pollution$Unit[row] == "ppm"){
    df_Pollution$Value[row]= as.character(as.numeric(as.character(df_Pollution$Value[row]))*1145.609)
    df_Pollution$Unit[row] = "Âµg/mÂ³"}
  }

#Get the concurrent countries with data
Countries <- unique(df_happiness$i..country)
Countries2 <- unique(df_Pollution$Country_Label)
Countries_in_Common=Reduce(intersect, list(Countries,Countries2))

```

-[4]:

```
#Get the data from spain
CO_s=c()
Life_Ladder_s=c()
index_countryH=which(df_happiness$i..country == "Spain")
index_country1=which(df_Pollution$Country_Label == "Spain")
index_country2=which(df_Pollution$Pollutant == "CO")
index_countryP=Reduce(intersect, list(index_country1,index_country2))

#We reduce the dataset to Spain and CO
df_per_countryH=df_happiness[c(index_countryH),]
df_per_countryP=df_Pollution[c(index_countryP),]
#We calculate the common years
common_years=Reduce(intersect, list(unique(df_per_countryP[ , 9]),unique(df_per_countryH[ , 2])))
for(year in common_years){
  #For each year, we will find the mean of CO pollution and pair it with the Life Ladder
  group1=which(df_per_countryH$year == year)
  group2=which(df_per_countryP$Last_Updated == year)
  mean_CO=sum(as.numeric(as.vector(df_per_countryP[group2,]$Value)))/length(df_per_countryP[group2,]$Value)
  #We will filter out negative pollution values
  if(mean_CO>0){
    CO_s=c(CO_s,mean_CO)
    Life_Ladder_s=c(Life_Ladder_s,df_per_countryH[group1,]$Life_Ladder)
  }
}
#Log regression fit
log_CO_s=c(log(CO_s))
plot(Life_Ladder_s,log_CO_s,main="The relationship between CO in air and the Life Ladder of Spain",xlab="Life Ladder score", ylab="Log CO in the air(µg/m^3)")
data_s<-cbind(Life_Ladder_s,log_CO_s)
fit_S<-lm(data_s[,2]~data_s[,1])
abline(summary(fit_S)$coefficients[1], summary(fit_S)$coefficients[2],col="red")
mylabel = bquote(italic(R) == .(format(cor(log_CO_s, Life_Ladder_s, method="pearson"), digits = 3)))
text(x = 6.46, y = 10, labels = mylabel)
mylabel = bquote(italic(p-value) == .(0.1183, digits = 3))
text(x = 6.46, y = 9.6, labels = mylabel)
```

-[5]:

```r
#Get data from all concurrent countries
Life_Ladder=c()
CO=c()
for(c in Countries_in_Common){
  #For each country, we find what years we have concurrences of both datasets
  index_countryH=which(df_happiness$i..country == c)

  index_country1=which(df_Pollution$Country_Label == c)
  index_country2=which(df_Pollution$Pollutant == "CO")
  index_countryP=Reduce(intersect, list(index_country1,index_country2))

  #We reduce the dataset to country and CO
  df_per_countryH=df_happiness[c(index_countryH),]
  df_per_countryP=df_Pollution[c(index_countryP),]
  #We calculate the common years
  common_years=Reduce(intersect, list(unique(df_per_countryP[ , 9]),unique(df_per_countryH[ , 2])))
  for(year in common_years){
    #For each country and year pair, we will find the mean of CO pollution and pair it with the Life Ladder
    group1=which(df_per_countryH$year == year)
    group2=which(df_per_countryP$Last_Updated == year)
    mean_CO=sum(as.numeric(as.vector(df_per_countryP[group2,]$value)))/length(df_per_countryP[group2,]$value)
    #We will filter out negative pollution values
    if(mean_CO>0){
      CO=c(CO,mean_CO)
      Life_Ladder=c(Life_Ladder,df_per_countryH[group1,]$Life_Ladder)
    }
  }
}
plot(Life_Ladder,CO,main="The relationship between CO in air and the Life Ladder",xlab="Life Ladder score", ylab="CO in the air (µg/m^3)")
log_CO=c(log(CO))
plot(Life_Ladder,log_CO,main="The relationship between CO in air and the Life Ladder",xlab="Life Ladder score", ylab="Log CO in the air (µg/m^3)")
data<-cbind(Life_Ladder,log_CO)
fit<-lm(data[,2]~data[,1])
summary(fit)
abline(summary(fit)$coefficients[1], summary(fit)$coefficients[2],col="red")
mylabel = bquote(italic(R) == .(format(cor(log_CO, Life_Ladder, method="pearson"), digits = 3)))
text(x = 7, y = 2, labels = mylabel)
mylabel = bquote(italic(p-value) == .(0.0322, digits = 3))
text(x = 7, y = 1.4, labels = mylabel)
```

-[6]:

```r
#We are going to make a permutation test on the F-statistic to ensure the pvalue is under 0.05
b<-anova(fit)
Ftrue<-b$"F value"[1];
nr=10000 #number of rearrangements to be examined
F=numeric(nr);
for (i in 1:nr){
  newx<- sample(data, 60)
  fit_p<-lm(newx ~ data[,1])
  b<-anova(fit_p)
  F[i]<-b$"F value"[1]}
experimental_pvalue=length(F[F >= Ftrue])/nr
```

-[7]:

```r
#Get CI for CO to get to top 10
data=data.frame(Life_Ladder,log_CO)
n<-length(Life_Ladder); data<-cbind(Life_Ladder,log_CO)
nb<-10000; z<-seq(1,n);predb<-numeric(nb)
thetab<-numeric(nb)
for(i in 1:nb){
  zb<-sample(z,n,replace=T)
  ajustb<- lm(data[zb,2]~ data[zb,1])
  predb[i]<-summary(ajustb)$coefficients[1]+summary(ajustb)$coefficients[2]*7.213}

hist(predb,main="Histogram of the CO needed to get in the top 10")
Prediction_CO_to_top10Score=quantile(predb,c(0.025,0.975))
```

-[8]:

```
#Get data from all concurrent countries
Life_Ladder_norm=c()
CO_norm=c()
for(c in Countries_in_Common){
  #For each country, we find what years we have concurrences of both datasets
  index_countryH=which(df_happiness$i..country == c)

  index_country1=which(df_Pollution$Country_Label == c)
  index_country2=which(df_Pollution$Pollutant == "CO")
  index_countryP=Reduce(intersect, list(index_country1,index_country2))

  #We reduce the dataset to country and CO
  df_per_countryH=df_happiness[c(index_countryH),]
  df_per_countryP=df_Pollution[c(index_countryP),]
  #We calculate the common years
  common_years=Reduce(intersect, list(unique(df_per_countryP[ , 9]),unique(df_per_countryH[ , 2])))
  for(year in common_years){
    #For each country and year pair, we will find the mean of CO pollution and pair it with the Life Ladder
    group1=which(df_per_countryH$year == year)
    group2=which(df_per_countryP$Last_Updated == year)
    mean_CO=sum(as.numeric(as.vector(df_per_countryP[group2,]$Value)))/length(df_per_countryP[group2,]$Value)
    sample=length(df_per_countryP[group2,]$Value)
    if(mean_CO>0){
      for (i in 1:sample){
        CO_norm=c(CO_norm,mean_CO)
        Life_Ladder_norm=c(Life_Ladder_norm,df_per_countryH[group1,]$Life_Ladder)
      }
    }
  }
}
log_CO_norm=c(log(CO_norm))
plot(Life_Ladder_norm,log_CO_norm,main="The relationship between CO in air and the Life Ladder(normalized)",xlab="Life Ladder score", ylab="Log CO in the
data<-cbind(Life_Ladder_norm,log_CO_norm)
fit_norm<-lm(data[,2]~data[,1])
summary(fit_norm)
abline(summary(fit_norm)$coefficients[1], summary(fit_norm)$coefficients[2],col="red")
mylabel = bquote(italic(R) == .(format(cor(log_CO_norm, Life_Ladder_norm, method="pearson"), digits = 3)))
text(x = 7, y = 2, labels = mylabel)
mylabel = bquote(italic(p-value) == .(2.2e-16, digits = 3))
text(x = 7, y = 1.4, labels = mylabel)
```

-[9]:

```
#Get CI for CO to get to top 10 (normalized)
data=data.frame(Life_Ladder_norm,log_CO_norm)
n<-length(Life_Ladder_norm); data<-cbind(Life_Ladder_norm,log_CO_norm)
nb<-10000; z<-seq(1,n);predb<-numeric(nb)
for(i in 1:nb){
  zb<-sample(z,n,replace=T)
  ajustb<- lm(data[zb,2]~ data[zb,1])
  predb[i]<-summary(ajustb)$coefficients[1]+summary(ajustb)$coefficients[2]*7.213}

hist(predb,main="Histogram of the CO needed to get in the top 10 (normalized)")
Prediction_CO_to_top10Score_norm=quantile(predb,c(0.025,0.975))
```