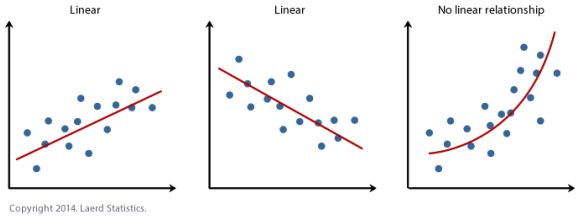
ENV 710

linear models

Duke | NICHOLAS SCHOOL OF THE ENVIRONMENT

roadmap

- exam results by next week
- comment on grading
- download `tad.usgab.csv` & covid.csv



linear models

↓

continuous dependent variable, continuous independent variable

⇣

continuous dependent variable, categorical independent variable

group projects

Name	Group
Benaka, Isaac	1
Go, Li Jia	1
Owens, Katie	1
Bernaus, Katrina	2
Gulino, Justin	2
Palia, Sophia	2
Bi, Yuntian	3
Haber, Jordan	3
Pang, Miaojun	3
Bliska, Hanna	4
Harvey, Marla	4
Pike, Rachel	4
Brentjens, Emma	5
Hays, Brandon	5
Price, Noah	5

Name	Group
Campos, Gabriel	6
Healey, Liam	6
Rowley, Caroline	6
Carlson, Maria	7
Hyun, Jiwon	7
Satagopan, Nanditha Ram	7
Davidson, Kelly	8
Jackson, Rachel	8
Seagle, Jenna	8
Diaz, Danae	9
Kuhlmann, Emily	9
Sepe, Stevie	9
Dye, Logan	10
Li, Jiahuan	10
Zungailia, Isabel	10
Franzetti, Tristan	11
Martinez, Laura	11
White, Libby	11
Freedman, Jacob	12
Merritt, Melissa	12
Wong, Richard	12

I – cars

Does the distance a car can travel in a set period of time increase with its speed?

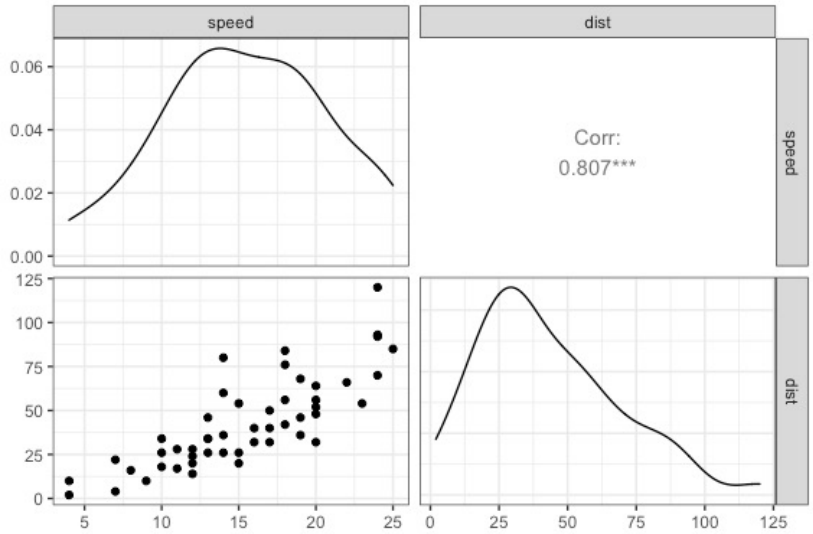
1. download the data: `data(cars)`
2. hypotheses? response variable? explanatory variable?
3. build the model, validate the model assumptions, interpret model fit and parameters

1 – cars

Cars is a standard built-in dataset, in R – it consists of 50 observations (rows) and 2 variables (columns) – dist (distance) and speed (speed).

1. what is your hypothesis?
2. which is the response variable? explanatory variable? write the model…
3. build the model, validate the model assumptions, interpret model fit and parameters

```
GGally::ggpairs(cars)
```

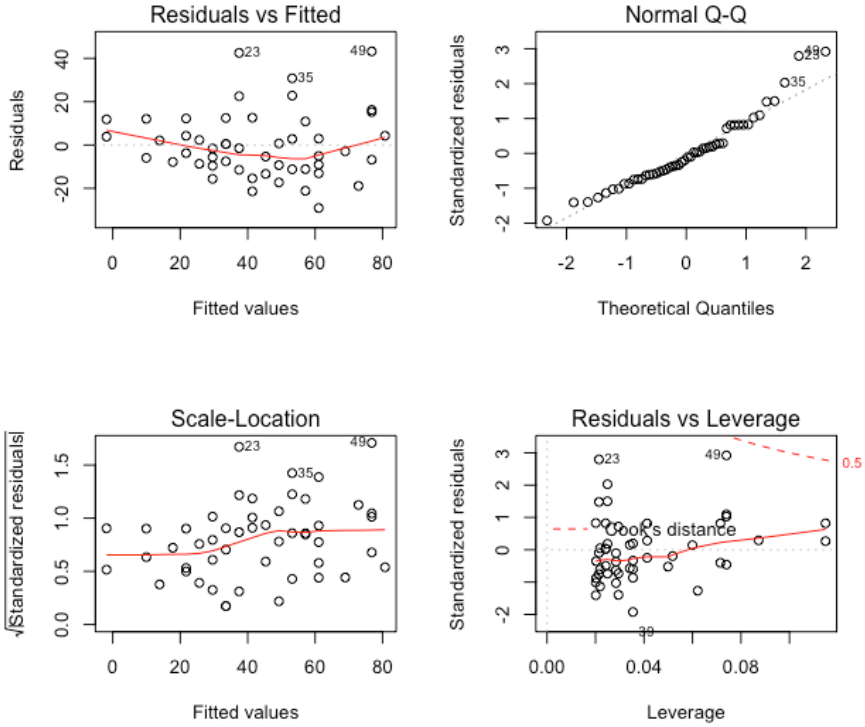


I – cars

```
c1 <- lm(dist ~ speed, data = cars)
summary(c1)
```

```
Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

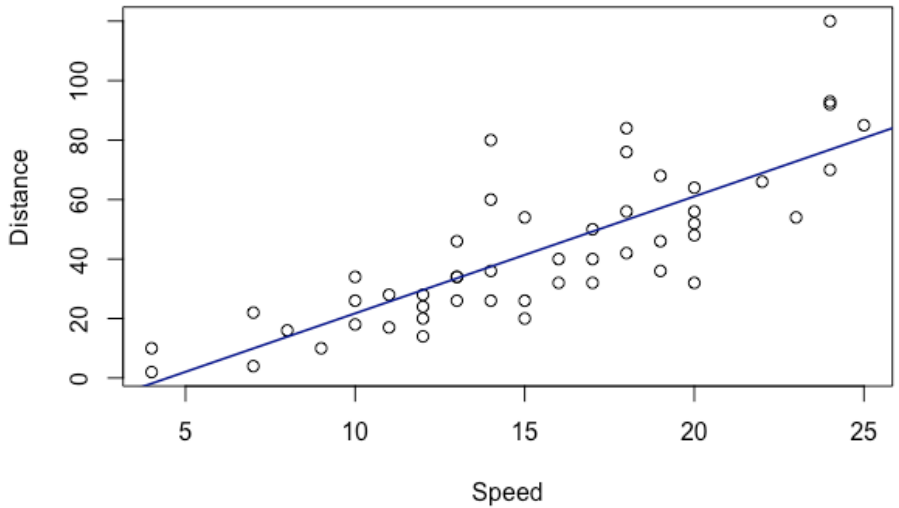
```
plot(c1)
```



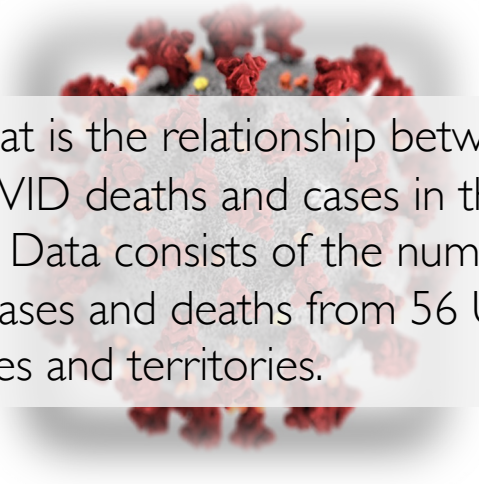
I – cars

```
plot(x = cars$speed, y = cars$dist,
     ylab = "Distance", xlab = "Speed")

abline(c1, col = "darkblue", lwd = 1.5)
```

```
summary(c1)

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```



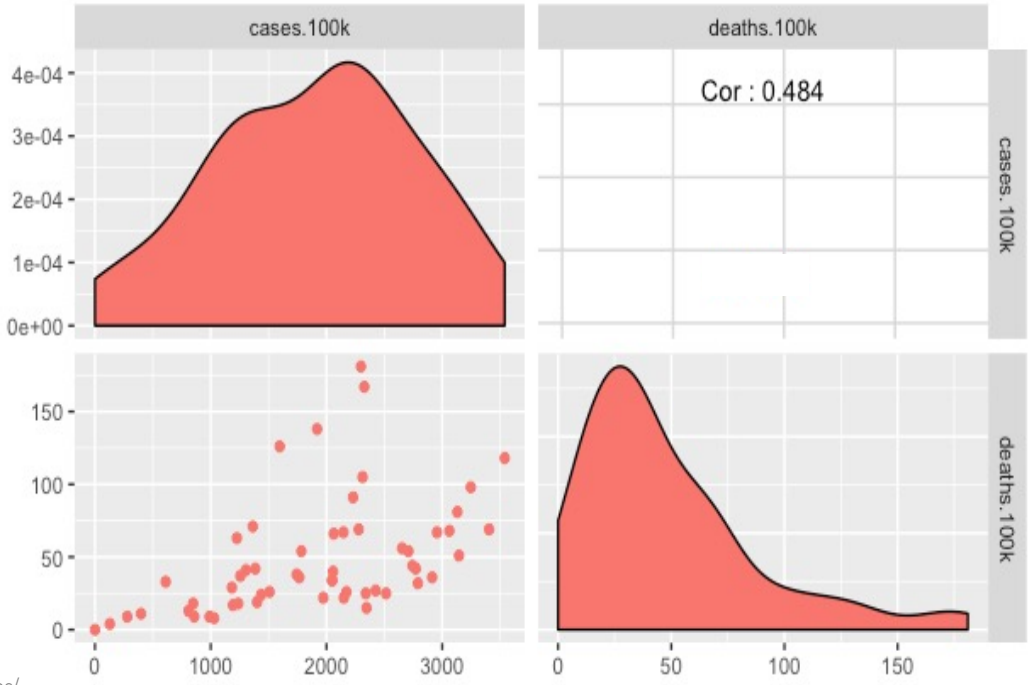
2 – COVID



What is the relationship between COVID deaths and cases in the US? Data consists of the number of cases and deaths from 56 US states and territories.

1. what is the null hypothesis?
2. what is the alternative hypothesis?
3. what is the model?
4. how do you feel about the data?

```
> summary(covid)
    place              cases.100k     deaths.100k
 Length:56          Min.   :   0   Min.   :  0.00
 Class :character   1st Qu.:1252   1st Qu.: 22.00
 Mode  :character   Median :2049   Median : 36.50
                    Mean   :1894   Mean   : 48.66
                    3rd Qu.:2448   3rd Qu.: 67.00
                    Max.   :3540   Max.   :181.00

> GGally::ggpairs(covid, columns = 3:4,
                  ggplot2::aes(colour = factor(1))
```

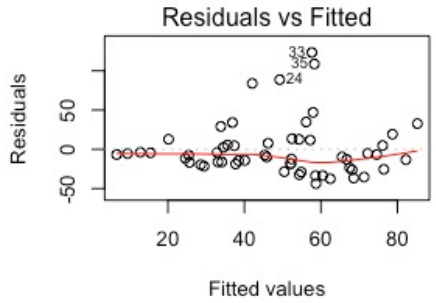


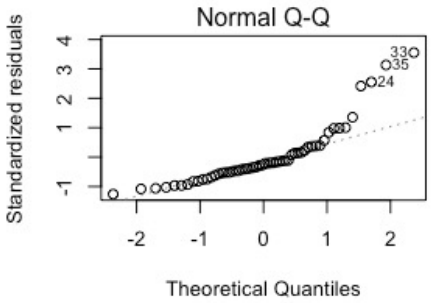
https://www.washingtonpost.com/graphics/2020/national/coronavirus-us-cases-deaths/

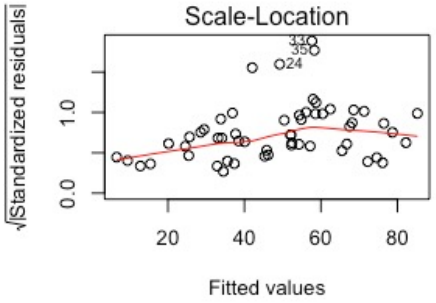
2 – COVID

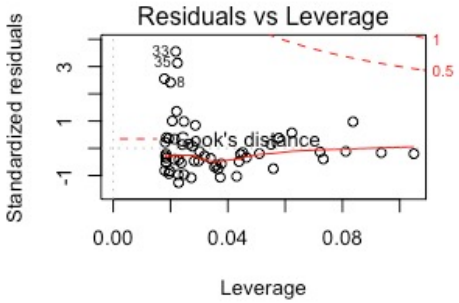
```
cvd <- with(covid, lm(deaths.100k ~ cases.100k))
summary(cvd)

Call:
lm(formula = deaths.100k ~ cases.100k)

Residuals:
    Min      1Q  Median      3Q     Max
-43.710 -18.914  -8.482   8.797 123.356

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.59726   11.36961   0.580 0.564157
cases.100k   0.02221    0.00547   4.061 0.000159 ***
---

Residual standard error: 35.1 on 54 degrees of freedom
Multiple R-squared:  0.234, Adjusted R-squared:  0.2198
F-statistic: 16.49 on 1 and 54 DF,  p-value: 0.0001589
```

```
summary.aov(cvd)

            Df Sum Sq Mean Sq F value    Pr(>F)
cases.100k   1  20322   20322    16.5 0.000159 ***
Residuals   54  66529    1232
```









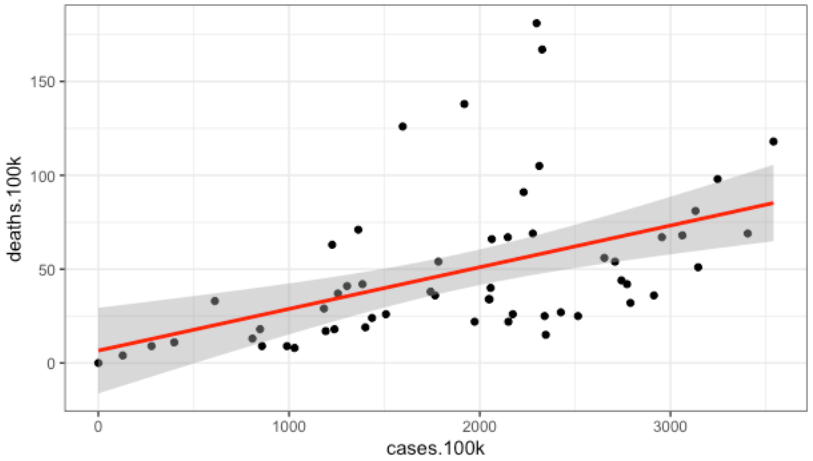
2 – COVID

```
cvd <- with(covid, lm(deaths.100k ~ cases.100k))
summary(cvd)
```

```
Call:
lm(formula = deaths.100k ~ cases.100k)

Residuals:
    Min      1Q  Median      3Q     Max
-43.710 -18.914  -8.482   8.797 123.356

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.59726   11.36961   0.580 0.564157
cases.100k   0.02221    0.00547   4.061 0.000159 ***
---

Residual standard error: 35.1 on 54 degrees of freedom
Multiple R-squared:  0.234, Adjusted R-squared:  0.2198
F-statistic: 16.49 on 1 and 54 DF,  p-value: 0.0001589
```

```
summary.aov(cvd)
```

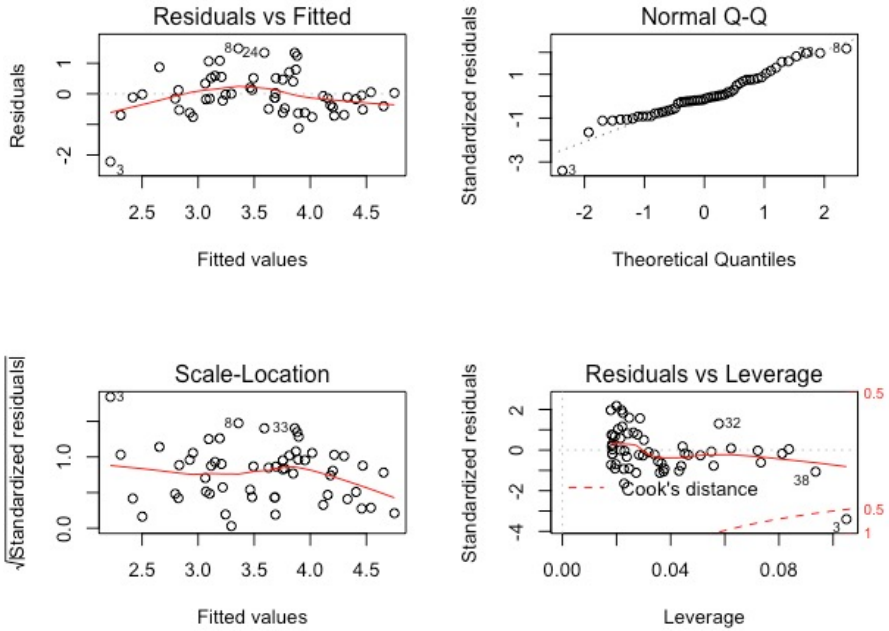
```
            Df Sum Sq Mean Sq F value    Pr(>F)
cases.100k   1  20322   20322    16.5 0.000159 ***
Residuals   54  66529    1232
```




```
ggplot(covid, aes(x = cases.100k, y = deaths.100k)) +
    geom_point() + stat_smooth(method = "lm", col = "red") +
     theme_bw()
```

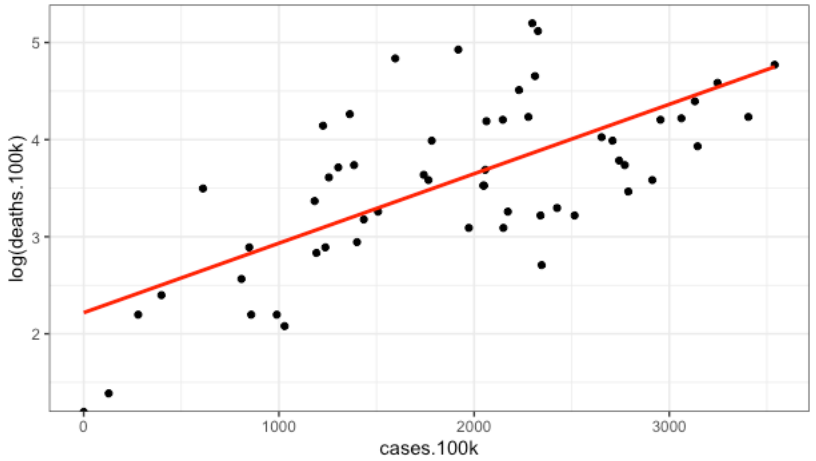
2 – COVID

```
cvd1 <- with(covid, lm(log(deaths.100k + 1) ~ cases.100k))
summary(cvd1)

Call:
lm(formula = log(deaths.100k + 1) ~ cases.100k)

Residuals:
     Min       1Q   Median       3Q      Max
-2.21777 -0.48060 -0.09219  0.47731  1.48480

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.2177726  0.2234578   9.925 8.93e-14 ***
cases.100k  0.0007153  0.0001075   6.654 1.50e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6899 on 54 degrees of freedom
Multiple R-squared:  0.4505,  Adjusted R-squared:  0.4404
F-statistic: 44.28 on 1 and 54 DF,  p-value: 1.501e-08
```



2 – COVID

```
cvd1 <- with(covid, lm(log(deaths.100k + 1) ~ cases.100k))
summary(cvd1)

Call:
lm(formula = log(deaths.100k + 1) ~ cases.100k)

Residuals:
     Min       1Q   Median       3Q      Max
-2.21777 -0.48060 -0.09219  0.47731  1.48480

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.2177726  0.2234578   9.925 8.93e-14 ***
cases.100k  0.0007153  0.0001075   6.654 1.50e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6899 on 54 degrees of freedom
Multiple R-squared:  0.4505,  Adjusted R-squared:  0.4404
F-statistic: 44.28 on 1 and 54 DF,  p-value: 1.501e-08
```

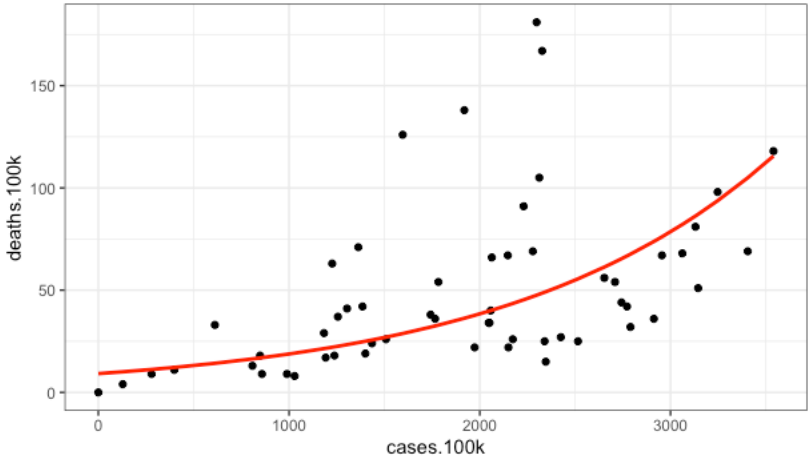



```
ggplot(covid, aes(x = cases.100k, y = log(deaths.100k)))+
 geom_point() +
 geom_line(data = fortify(cvd1), aes(x = cases.100k,
           y = .fitted), colour = "red", lwd = 1) +
  theme_bw()
```

2 – COVID

```
cvd1 <- with(covid, lm(log(deaths.100k + 1) ~ cases.100k))
summary(cvd1)

Call:
lm(formula = log(deaths.100k + 1) ~ cases.100k)

Residuals:
     Min       1Q   Median       3Q      Max
-2.21777 -0.48060 -0.09219  0.47731  1.48480

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.2177726  0.2234578   9.925 8.93e-14 ***
cases.100k  0.0007153  0.0001075   6.654 1.50e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6899 on 54 degrees of freedom
Multiple R-squared:  0.4505,   Adjusted R-squared:  0.4404
F-statistic: 44.28 on 1 and 54 DF,  p-value: 1.501e-08
```

```
exp(coef(cvd1))
(Intercept)  cases.100k
   9.186846    1.000716
```

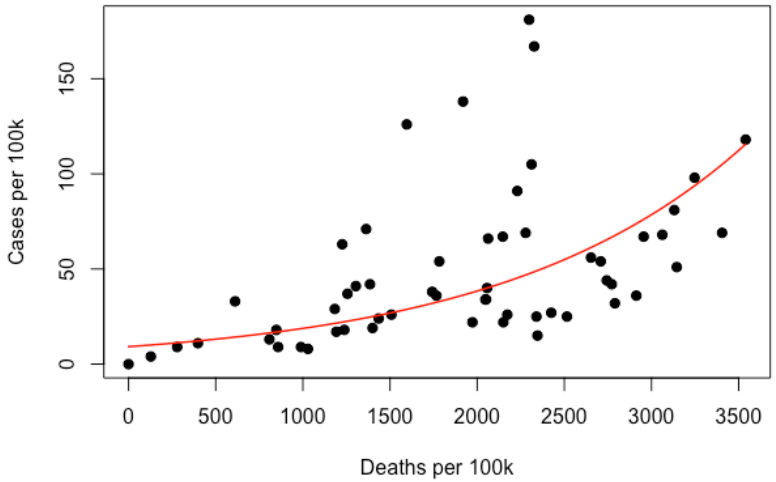



```
ggplot(covid, aes(x = cases.100k, y = deaths.100k))+
  geom_point() +
  geom_line(data = fortify(cvd1), aes(x = cases.100k,
            y = exp(.fitted)),
            colour = "red", lwd = 1) +
  theme_bw()
```

2 – COVID

```
cvd1 <- with(covid, lm(log(deaths.100k + 1) ~ cases.100k))
summary(cvd1)

Call:
lm(formula = log(deaths.100k + 1) ~ cases.100k)

Residuals:
     Min       1Q   Median       3Q      Max
-2.21777 -0.48060 -0.09219  0.47731  1.48480

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.2177726  0.2234578   9.925 8.93e-14 ***
cases.100k  0.0007153  0.0001075   6.654 1.50e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6899 on 54 degrees of freedom
Multiple R-squared:  0.4505,   Adjusted R-squared:  0.4404
F-statistic: 44.28 on 1 and 54 DF,  p-value: 1.501e-08
```

```
exp(coef(cvd1))
(Intercept)  cases.100k
   9.186846    1.000716
```




```
with(covid, plot(x = cases.100k, y = deaths.100k,
     pch = 19, xlab = "Deaths per 100k",
     ylab = "Cases per 100k"))
curve(expr = exp(cvd1$coefficients[1] +
      cvd1$coefficients[2]*x), add = T, col = "red",
      lwd = 1.5)
```

3 – GHG emissions

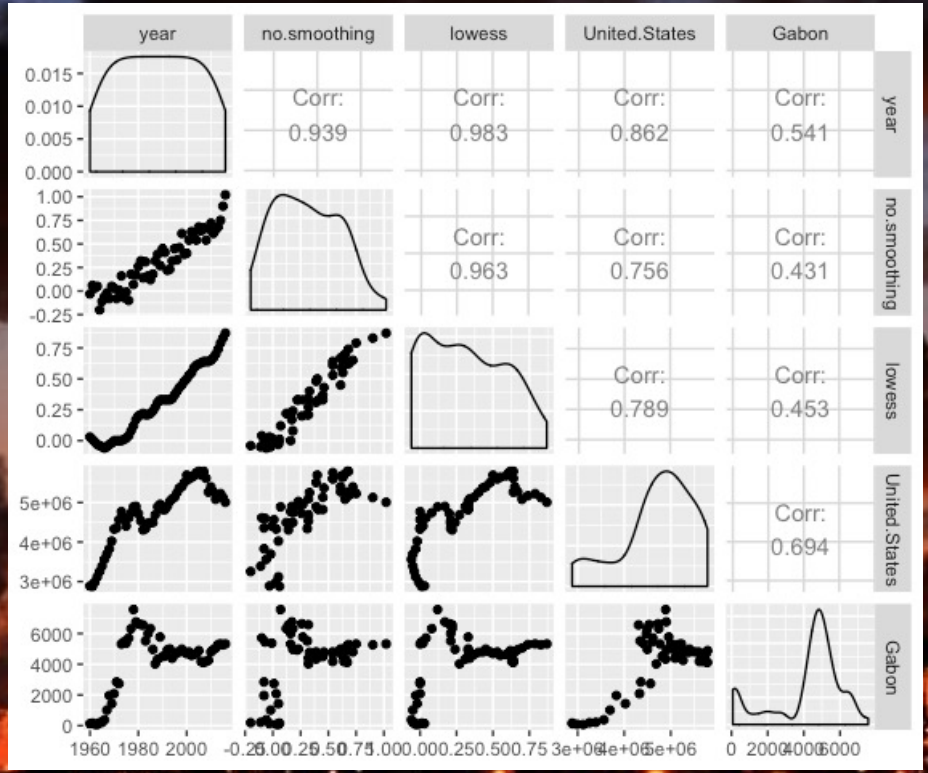
Is the global temperature anomaly related to greenhouse gas emissions in the United States?

https://data.worldbank.org/indicator/EN.ATM.CO2E.KT

https://data.giss.nasa.gov/gistemp/graphs/graph_data/Global_Mean_Estimates_based_on_Land_and_Ocean_Data/graph.txt

3 – GHG emissions

	year	no.smoothing	lowess	United.States	Gabon
1	1960	-0.03	0.03	2890696	132.0
2	1961	0.06	0.01	2880506	165.0
3	1962	0.03	-0.01	2987208	88.0
4	1963	0.05	-0.03	3119231	73.3
5	1964	-0.20	-0.04	3255995	190.7
6	1965	-0.11	-0.05	3390923	216.4



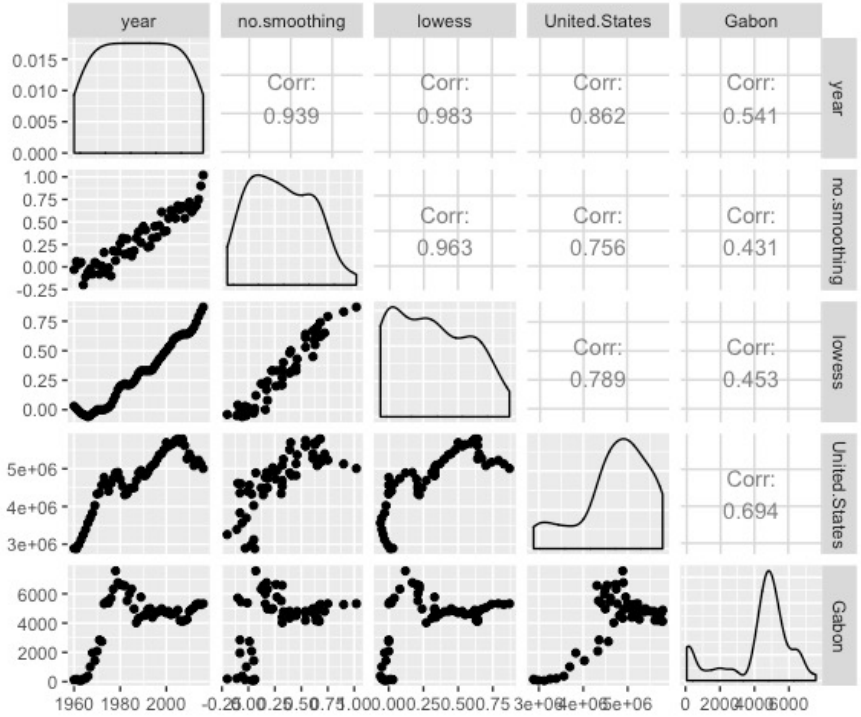
3 – GHG emissions

is the global temperature anomaly related to greenhouse gas emissions in the United States?

temperature anomalies (no.smoothing)

CO2 emissions (kt - kiloton) from the US and Gabon (an African country)

1. download the data
2. response variable = no.smooth, explanatory variable = United.States
3. build the model, validate the model assumptions, interpret model fit and parameters

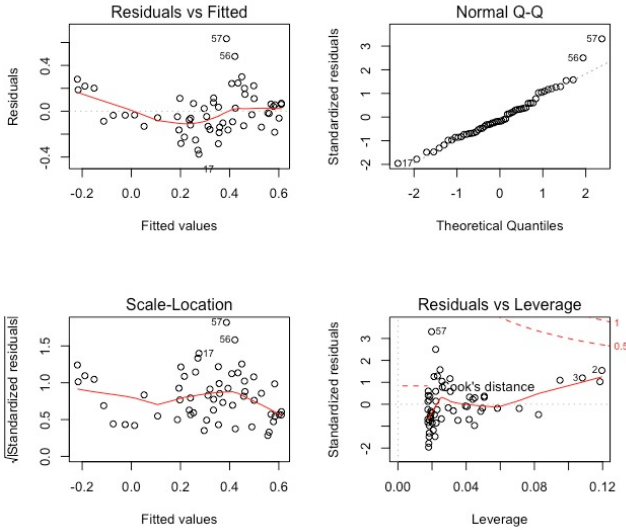


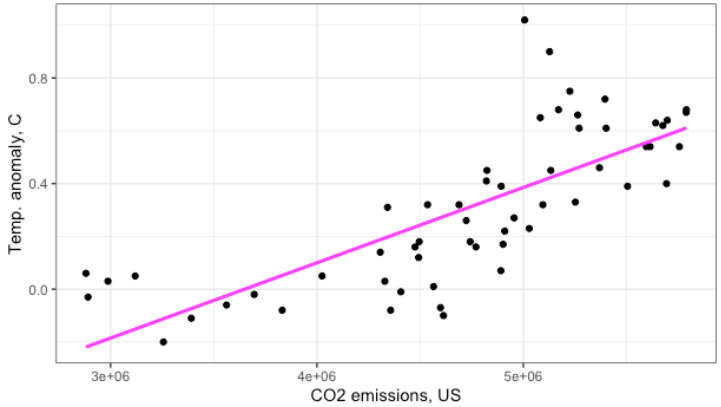
3 – GHG emissions

is the global temperature anomaly related to greenhouse gas emissions in the United States?

```
us.0 <- with(dat2, lm(no.smoothing ~ United.States))
summary(us.0)
```

```
Call:
lm(formula = no.smoothing ~ United.States)

Residuals:
     Min       1Q   Median       3Q      Max
-0.3749 -0.1318 -0.0332  0.1103  0.6329

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.041e+00  1.597e-01  -6.519 2.31e-08 ***
United.States  2.853e-07  3.330e-08   8.567 1.05e-11 ***
---
Residual standard error: 0.1933 on 55 degrees of freedom
Multiple R-squared:  0.5716, Adjusted R-squared:  0.5639
F-statistic:  73.4 on 1 and 55 DF,  p-value: 1.048e-11
```



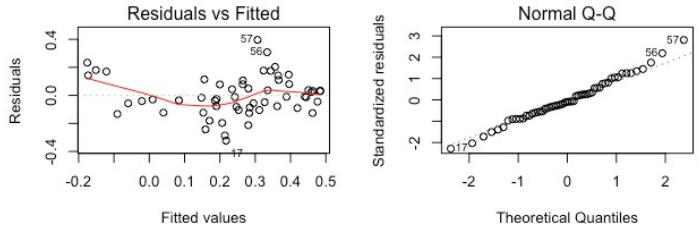


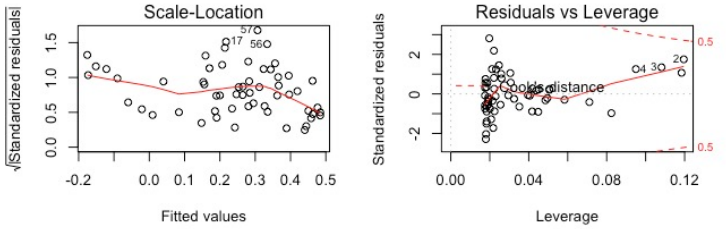
3 – GHG emissions

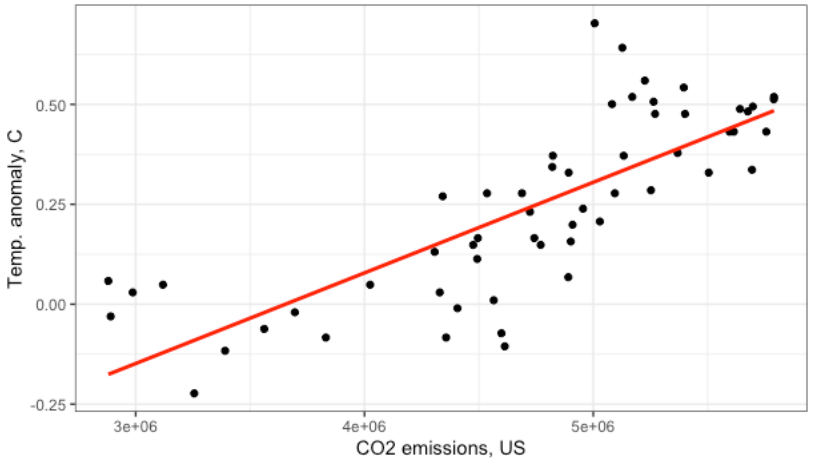
Is the global temperature anomaly related to greenhouse gas emissions in the United States?

```
us.1 <- lm(log(no.smoothing + 1) ~ United.States),
          data = dat2)
```

```
lm(formula = log(no.smoothing + 1) ~ United.States)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32277 -0.08579 -0.01265  0.08013  0.39651

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -8.288e-01  1.176e-01  -7.045 3.18e-09 ***
United.States  2.268e-07  2.453e-08   9.246 8.55e-13 ***
---

Residual standard error: 0.1423 on 55 degrees of freedom
Multiple R-squared:  0.6085,  Adjusted R-squared:  0.6014
F-statistic:  85.5 on 1 and 55 DF,  p-value: 8.552e-13
```







Questions?