CSI 2300: Intro to Data Science

In-Class Exercise 19: Modeling – Simple Linear Regression

For this lecture, we'll return to a dataset we've used before; the COVID dataset from the Colorado Department of Public Health and Environment (CDPHE)[1]. If you want to refer back, we used this dataset in both lectures 3 and 15.

1. Load the data.

    - Load in the dataset from the CSV file `CDPHE_COVID19_Wastewater_Dashboard_Data.csv`.
    - Rename the column `SARS_CoV_2_copies_L` to `sars_rna`, and rename the column `Number_of_New_COVID19_Cases_by_` to `cases`.
    - Remove the observations that have value `NA` for `sars_rna`, that have the value 0 for `sars_rna` or that have the value 0 for `cases`.
    - Add two columns that represent the logarithm of `sars_rna` and the logarithm of `cases`. ($\log(0) = -\infty$, which is why we removed the zero-cases in the previous step.)

```
library(tidyverse)
# -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
# v dplyr     1.1.4     v readr     2.1.5
# v forcats   1.0.0     v stringr   1.5.1
# v ggplot2   3.5.1     v tibble    3.2.1
# v lubridate 1.9.4     v tidyr     1.3.1
# v purrr     1.0.2
# -- Conflicts ------------------------------------------ tidyverse_conflicts() --
# x dplyr::filter() masks stats::filter()
# x dplyr::lag()    masks stats::lag()
# i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts

covid_data <- read_csv("CDPHE_COVID19_Wastewater_Dashboard_Data.csv")
# Rows: 3498 Columns: 5
# -- Column specification --------------------------------------------------------
# Delimiter: ","
# chr (2): Date, Utility
# dbl (3): SARS_CoV_2_copies_L, Number_of_New_COVID19_Cases_by_, ObjectId
#
# i Use `spec()` to retrieve the full column specification for this data.
# i Specify the column types or set `show_col_types = FALSE` to quiet this message.

covid_clean <- covid_data %>%
```

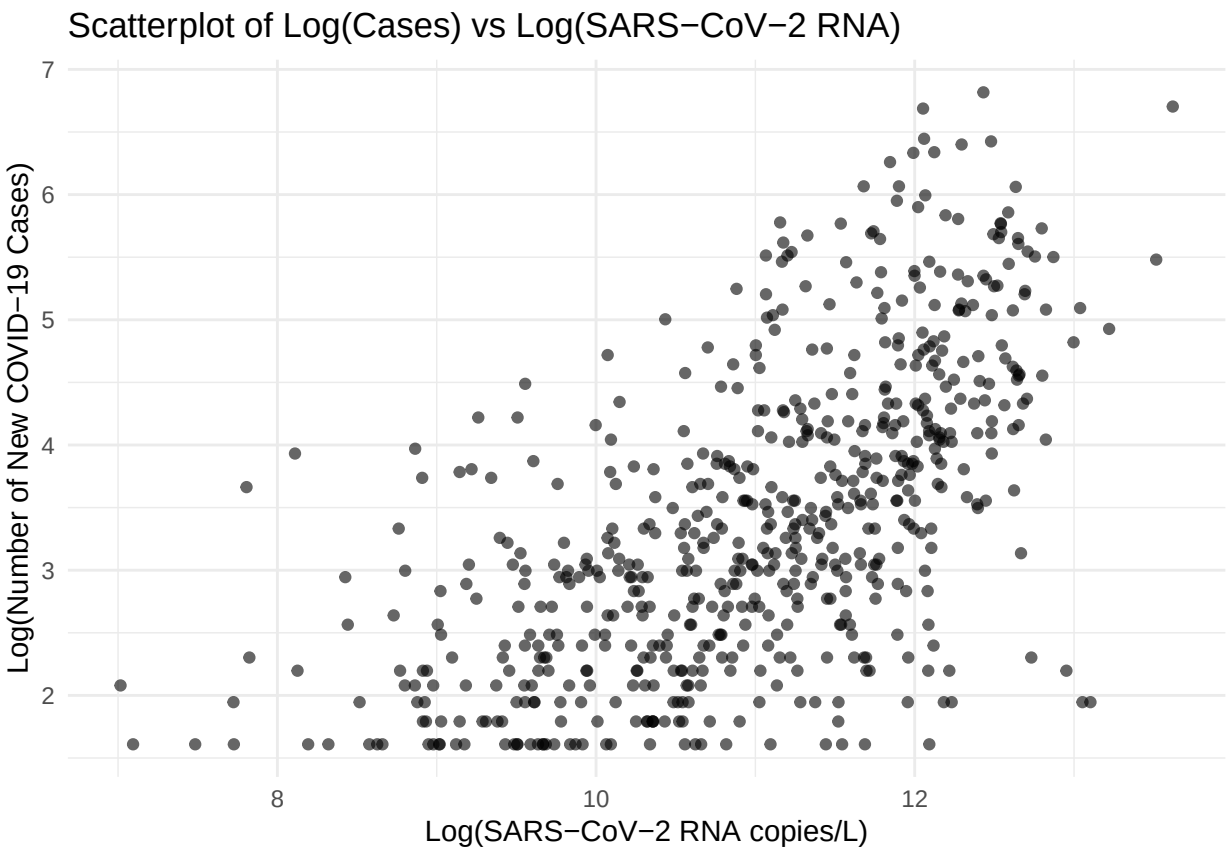
[1] https://cdphe.maps.arcgis.com/apps/opsdashboard/index.html#/d79cf93c3938470ca4bcc4823328946b

```r
  rename(
    sars_rna = SARS_CoV_2_copies_L,
    cases = Number_of_New_COVID19_Cases_by_
  ) %>%
  filter(
    !is.na(sars_rna),
    sars_rna != 0,
    cases != 0
  ) %>%
  mutate(
    log_sars_rna = log(sars_rna),
    log_cases = log(cases)
  )

head(covid_clean)
```

```
# # A tibble: 6 x 7
#   Date       Utility                sars_rna cases ObjectId log_sars_rna log_cases
#   <chr>      <chr>                     <dbl> <dbl>    <dbl>        <dbl>     <dbl>
# 1 08/06/2020 Metro Wastewater RW~   17309.     40       16         9.76      3.69
# 2 08/06/2020 Metro Wastewater RW~    7078.     53       20         8.86      3.97
# 3 08/02/2020 CO Springs - JD Phi~   24177.      5       23        10.1       1.61
# 4 08/02/2020 CO Springs - Las Ve~   50394.     15       29        10.8       2.71
# 5 08/06/2020 Pueblo                 17678.      6       31         9.78      1.79
# 6 08/06/2020 SPR                    15334.     11       54         9.64      2.40
```

2. Make a scatter plot with the log of the number of cases and the log of the number of SARS CoV-2 RNA copies. Use good labels. Which do you think should be the independent variable, and which should be the dependent variable? Why do you think that?

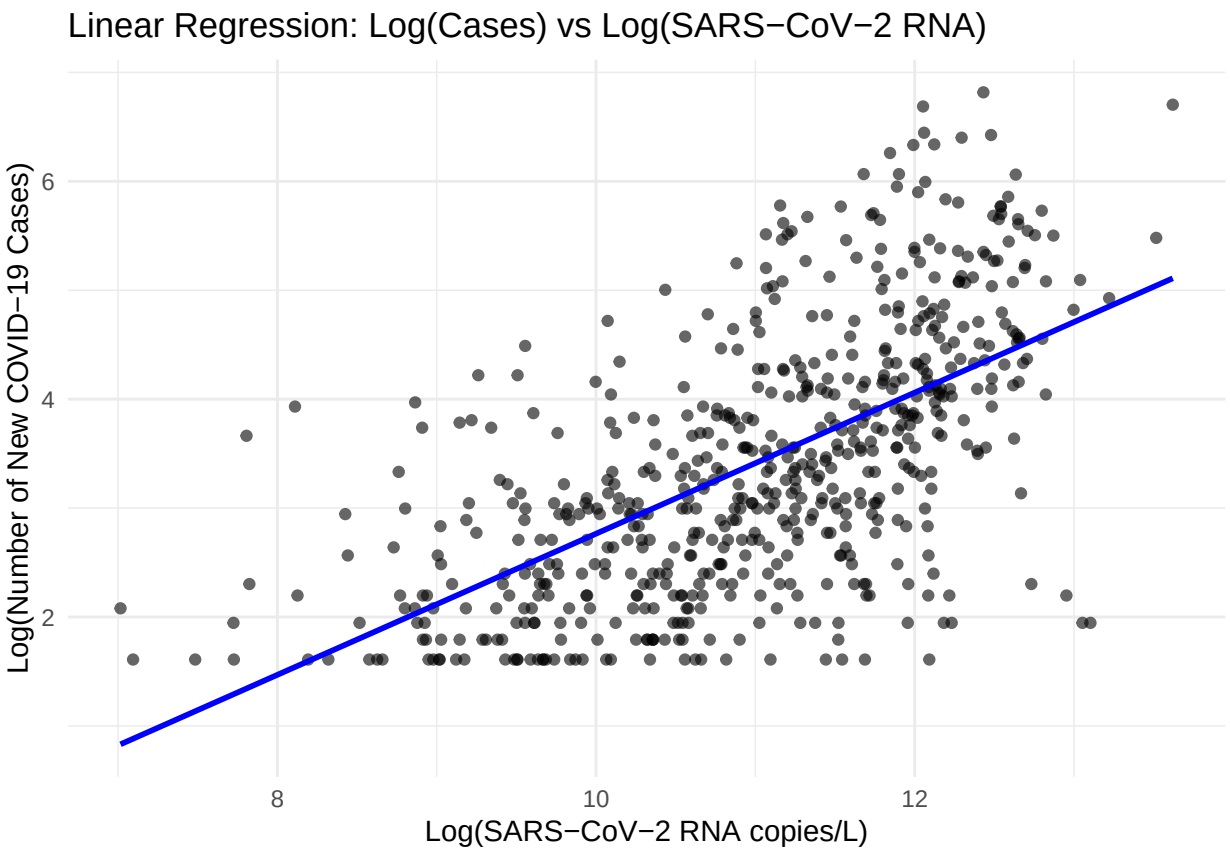
```r
ggplot(covid_clean, aes(x = log_sars_rna, y = log_cases)) +
  geom_point(alpha = 0.6) +
  labs(
    x = "Log(SARS-CoV-2 RNA copies/L)",
    y = "Log(Number of New COVID-19 Cases)",
    title = "Scatterplot of Log(Cases) vs Log(SARS-CoV-2 RNA)"
  ) +
  theme_minimal()
```



The independent variable should be log(SARS-CoV-2 RNA copies/L), and the dependent variable should be log(Number of New COVID-19 Cases). The amount of SARS-CoV-2 RNA in wastewater shows the number of new COVID-19 cases in the community so it makes sense to use the log of SARS-CoV-2 RNA as the predictor (independent variable) and the log of new cases as the response (dependent variable).

3. Fit a linear model using `lm()` for the same data you just plotted. Save the model returned by `lm()` in a variable so you can investigate it further. Overlay the fitted model's line using `abline()`.

```
model <- lm(log_cases ~ log_sars_rna, data = covid_clean)

ggplot(covid_clean, aes(x = log_sars_rna, y = log_cases)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(
    x = "Log(SARS-CoV-2 RNA copies/L)",
    y = "Log(Number of New COVID-19 Cases)",
    title = "Linear Regression: Log(Cases) vs Log(SARS-CoV-2 RNA)"
  ) +
  theme_minimal()
# `geom_smooth()` using formula = 'y ~ x'
```



4. Investigate the fit of the model.

    - Does it *look* like a good fit?
    - What is the value of R^2 (you can use the one called "multiple R^2")? Is that good or bad (or can you tell)?
    - Are the two coefficients of the model significant?
    - Explain what the two coefficients mean (including their sign – positive or negative).

```
# Model summary
summary(model)
#
# Call:
# lm(formula = log_cases ~ log_sars_rna, data = covid_clean)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -2.82894 -0.63304 -0.07786  0.58101  2.59289
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
```

```
# (Intercept)   -3.7141     0.3705  -10.02   <2e-16 ***
# log_sars_rna   0.6479     0.0335   19.34   <2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.9526 on 620 degrees of freedom
# Multiple R-squared:  0.3762,  Adjusted R-squared:  0.3752
# F-statistic:   374 on 1 and 620 DF,  p-value: < 2.2e-16
```

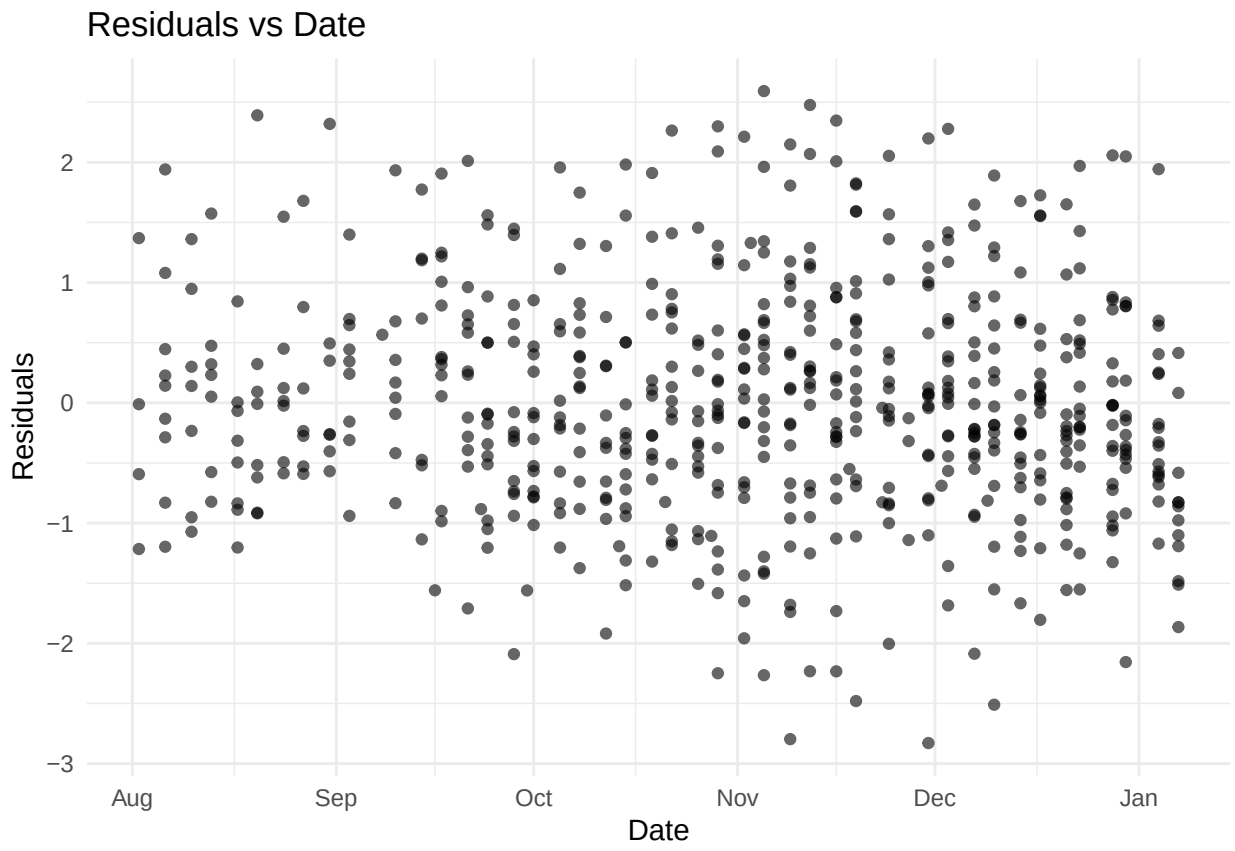
The model shows a moderate fit, with an R2R2 value of 0.38, meaning a quarter of the variation in new COVID-19 cases is explained by the viral RNA in wastewater. Both the intercept and the slope are highly significant, indicating a real relationship between the variables. The positive slope means that as SARS-CoV-2 RNA in wastewater increases, the number of new COVID-19 cases also tends to increase.

5. Calculate the residuals of your model. Then analyze them:

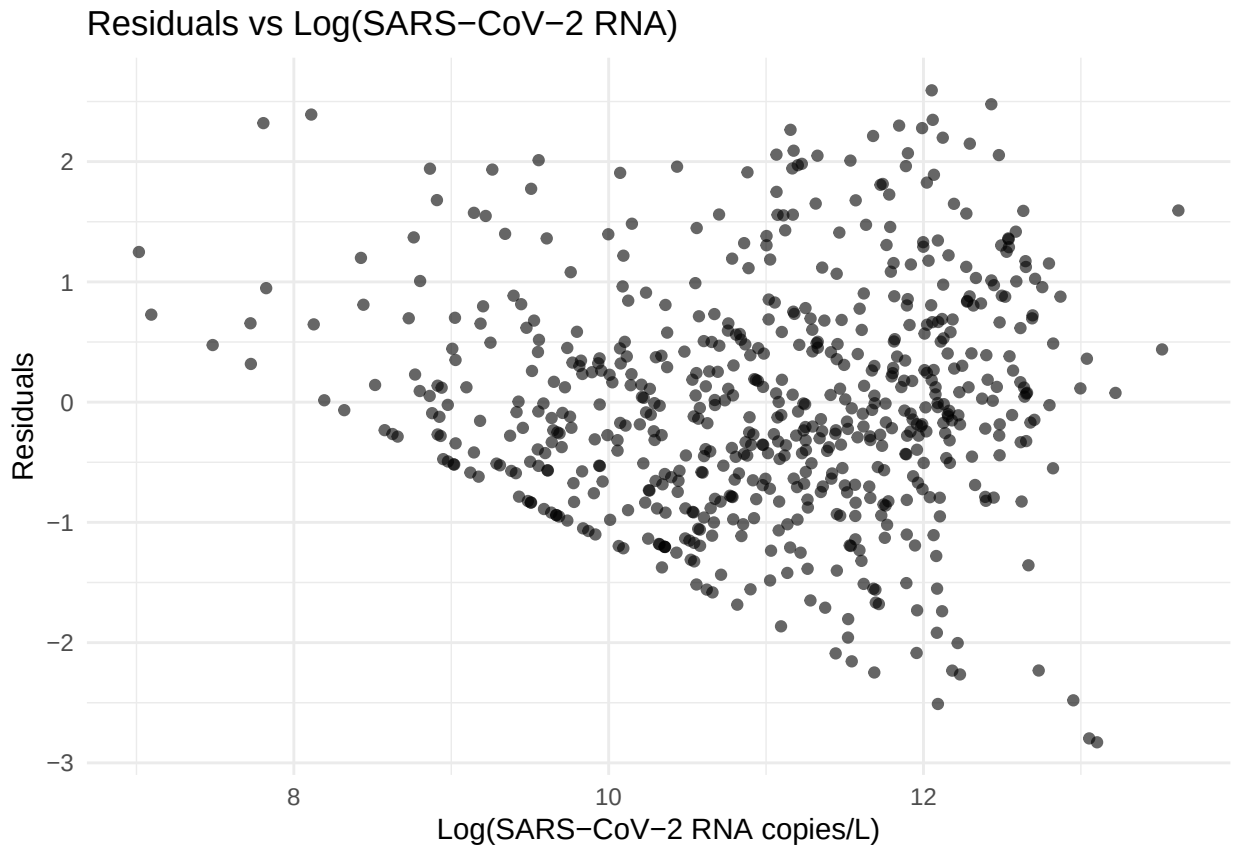
    - Compute the mean (average) of the residuals. What is it close to? What that imply?
    - Show a scatter plot of residuals (y-axis) and date (x-axis). What can you learn from this plot? (Hint: don't forget to use `lubridate` for the dates.)
    - Show a scatter plot of residuals (y-axis) and log of rna (x-axis). What can you learn from this plot?

```
library(lubridate)
library(tidyverse)

# Add residuals to the data
covid_clean <- covid_clean %>%
  mutate(
    residuals = resid(model)
  )

mean_resid <- mean(covid_clean$residuals)
mean_resid
# [1] 9.995777e-18

covid_clean$Date <- mdy(covid_clean$Date)

num_na_dates <- sum(is.na(covid_clean$Date))
num_na_dates
# [1] 0
```

```r
covid_clean_no_na <- covid_clean %>% filter(!is.na(Date))

ggplot(covid_clean_no_na, aes(x = Date, y = residuals)) +
  geom_point(alpha = 0.6) +
  labs(
    x = "Date",
    y = "Residuals",
    title = "Residuals vs Date"
  ) +
  theme_minimal()
```




```r
ggplot(covid_clean, aes(x = log_sars_rna, y = residuals)) +
  geom_point(alpha = 0.6) +
  labs(
    x = "Log(SARS-CoV-2 RNA copies/L)",
    y = "Residuals",
    title = "Residuals vs Log(SARS-CoV-2 RNA)"
  ) +
  theme_minimal()
```



The residuals appear to be spread out unevenly, with a funnel shape, meaning that the number of the errors increases as the log of SARS-CoV-2 RNA increases.

6. Explain whether this model is useful. If you think it is useful: for what purpose is it useful, and for whom? If you do not think it is useful: why not?

This model is somewhat useful because it shows a moderate relationship between SARS-CoV-2 RNA in wastewater and new COVID-19 cases. Public health officials could use it as an early warning tool to help predict trends in case numbers. However, since the model does not explain all the variation, it should be used with caution and improved with more complex modeling.