BIOSTAT/STAT 570: Coursework 1

To be submitted to the course canvas site by 1:30pm Friday 7th October, 2022.

The data we analyze were collected to investigate the determinants of pollution for 41 cities in the United States. In this study, the SO2 level is considered as the dependent variable and six variables are considered as potential explanatory variables:

- x_{i1}: average annual temperature in degrees F
- x_{i2}: number of manufacturers employing > 20 workers
- x_{i3}: population size in thousands
- x_{i4}: average annual wind speed in miles per hour
- x_{i5}: average annual rainfall in inches
- x_{i6}: average number of days rainfall per year

We let Y_i represent SO2 concentration (in micrograms per cubic meter), and $(x_{i1}, x_{i2}, \ldots, x_{i6})$, the covariates, for city i, $i = 1, \ldots, n = 41$. We fit the model

$$y_i = \beta_0 + \sum_{j=1}^{6} x_{ij}\beta_j + \epsilon_i,$$

$i = 1, \ldots, n$, in `R`, using least squares, and the output below (which has been edited slightly) was produced.

The computation part

1. Using `R`, reproduce every number in the handout using matrix and arithmetic operations.

The interpretation part: imagine this part will be read by a non-statistician

1. Based on the fitted model, provide an informative plot that summarizes the association between the SO2 level and the 6 covariates.
2. Give interpretations of each of the parameters β_j, $0 = 1, \ldots, 6$.

The assumptions part

State the assumptions that are required valid for the following (in all cases $j = 0, \ldots, 6$):

1. An unbiased estimate of β_j.
2. An accurate estimate of the standard error of β_j, .
3. Accurate coverage probabilities for 100$(1-\alpha)$% confidence intervals of the form

$$\widehat{\beta}_j \pm \widehat{\mathsf{var}}(\widehat{\beta}_j)^{1/2} \times z_{1-\alpha/2},$$

   where $z_{1-\alpha/2}$ represents the $(1-\alpha/2)$ quantile of an $N(0,1)$ random variable.
4. Accurate coverage probabilities for 100$(1-\alpha)$% confidence intervals of the form

$$\widehat{\beta}_j \pm \widehat{\mathsf{var}}(\widehat{\beta}_j)^{1/2} \times t_{n-4}(1-\alpha/2),$$

   where $t_{n-7}(1-\alpha/2)$ represents the $(1-\alpha/2)$ quantile of a standard Student's t random variable on $n-4$ degrees of freedom.
5. An accurate prediction for an *observed* outcome at $\boldsymbol{x} = \boldsymbol{x}_0$.

```
library(gamlss.data)
data(usair)
lmod <- lm(y~x1+x2+x3+x4+x5+x6,data=usair)
summary(lmod)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 111.72848   47.31810   2.361 0.024087 *
x1           -1.26794    0.62118  -2.041 0.049056 *
x2            0.06492    0.01575   4.122 0.000228 ***
x3           -0.03928    0.01513  -2.595 0.013846 *
x4           -3.18137    1.81502  -1.753 0.088650 .
x5            0.51236    0.36276   1.412 0.166918
x6           -0.05205    0.16201  -0.321 0.749972

Residual standard error: 14.64 on 34 degrees of freedom
```