ES/STT 7140: Homework 3 Solution

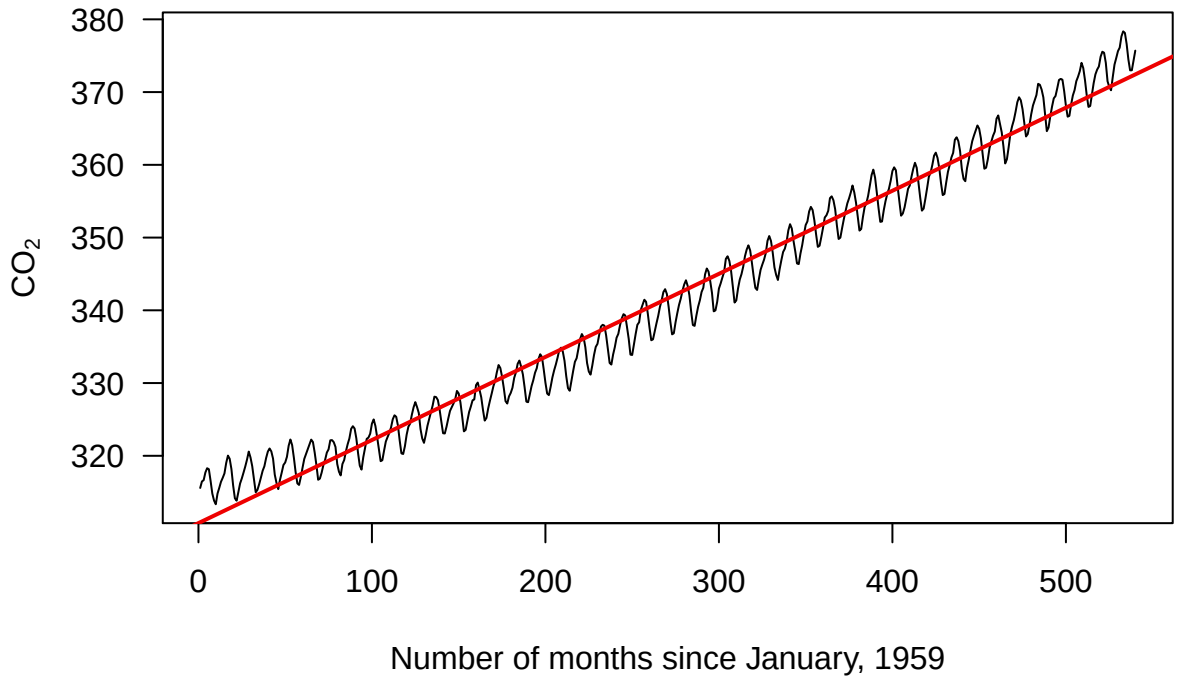
Brandon M. Greenwell

Question 1 (autocorrelation) The data set `co2.csv` contains monthly mean atmospheric CO_2 concentrations measured at Mauna Loa, Hawaii, from January 1959 to December 2003. Atmospheric CO_2 concentrations show a distinct seasonal pattern, reflecting the annual cycle of plant activities. The data set has four columns: `CO2` (monthly CO_2 concentrations in ppm), `Month` (calendar month), `Year`, and `Months` (months since January, 1959). The data set can be loaded using the following lines of code:

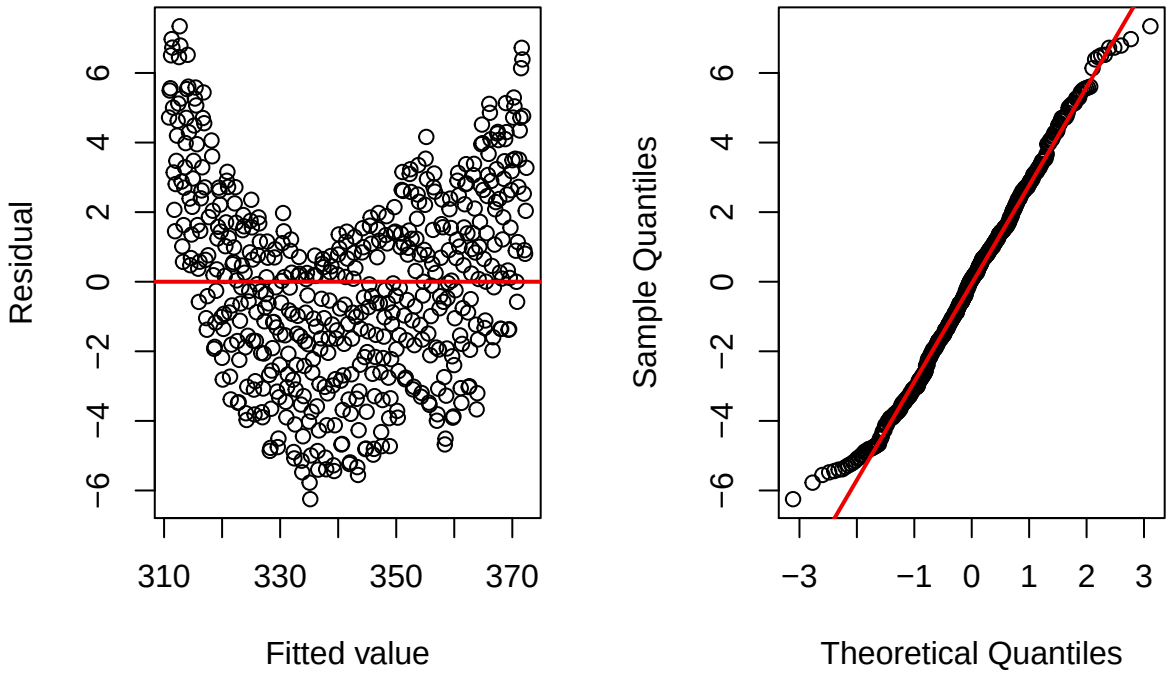
```
# Load the data
url <- paste0("https://raw.githubusercontent.com/bgreenwell/",
              "stt7140-env/master/data/co2.csv")
co2 <- read.csv(url, header = TRUE)

# Scatterplot
plot(
  CO2 ~ Months,
  data = co2,
  type = "l",
  xlab = "Number of months since January, 1959",
  ylab = expression(CO[2]),
  las = 1
)

# Fit a simple linear regression model
fit <- lm(CO2 ~ Months, data = co2)
summary(fit)  # print model summary
```

```
##
## Call:
## lm(formula = CO2 ~ Months, data = co2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2497 -1.9573  0.0778  1.8497  7.3421
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.107e+02  2.424e-01  1281.9   <2e-16 ***
## Months      1.142e-01  7.744e-04   147.5   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.791 on 535 degrees of freedom
## Multiple R-squared:  0.976,  Adjusted R-squared:  0.976
## F-statistic: 2.176e+04 on 1 and 535 DF,  p-value: < 2.2e-16
```

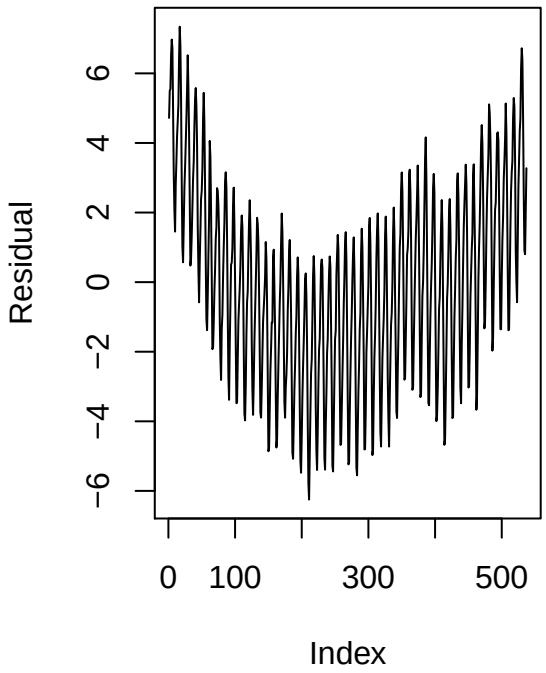
```
abline(fit, lwd = 2, col = "red2")  # add regression line to previous plot
```




```r
# Residual plots
par(mfrow = c(1, 2))
res <- residuals(fit)
plot(fitted(fit), res, xlab = "Fitted value", ylab = "Residual")
abline(h = 0, lwd = 2, col = "red2")  # reference line
qqnorm(res, main = "")
qqline(res, lwd = 2, col = "red2")  # reference line
```




```r
plot(res, xlab = "Index", ylab = "Residual", type = "l")
```



a. The above scatterplot displays the relationship between atmospheric CO_2 concentrations (`CO2`) and the number of months since January, 1959 (`Months`). The plot indicates an overall increasing relationship between `CO2` and `Months`. However, the "sawtooth" pattern in the plot indicates a strong seasonal effect as well.

b. The slope from the SLR model is 0.1142224 which implies that for each additiaonal month since January, 1959, we expect an average increase of CO_2 of 0.1142224 ppm. However, the plots displayed below indicate that the SLR is not adequate due to the obvious nonlinear trend in the data. Furthermore, the figure on the bottom right suggests that the we do not have independent observations (the pattern in the residuals is obviously not random!). Hence, the SLR model does not seem appropriate for describing these data.

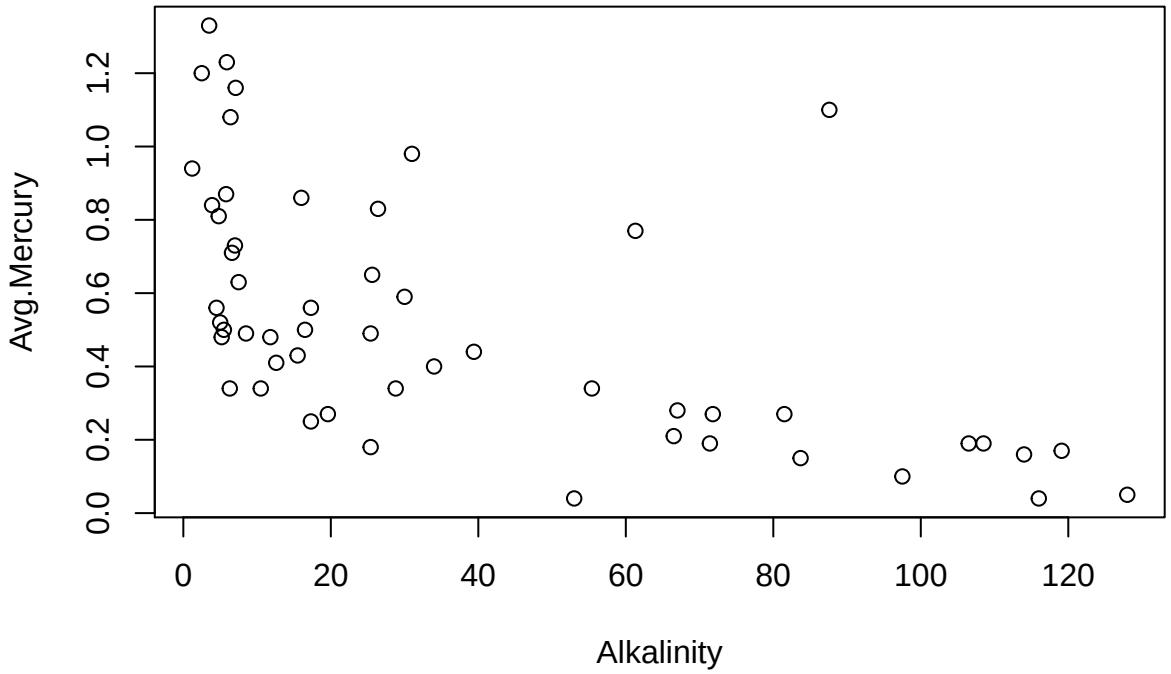
Question 2 (transformations) Mercury contamination of edible freshwater fish poses a direct threat to our health. Large mouth bass were studied in 53 Florida lakes to examine the factors that influence the level of mercury contamination. Water samples were collected from the surface of the middle of each lake in August 1990 and then again in March 1991. The pH level, the amount of chlorophyll, calcium, and alkalinity were measured in each sample. The average of the August and March values were used in the analysis Next, a sample of fish was taken from each lake with sample sizes ranging from 4–44 fish, each used to measure a mercury concentration and the average concentration is reported. The authors of the study [Lange et. al., 1993] indicated that alkalinity is the best predictor of the average mercury concentration and a linear model was suggested. The data can be loaded using the following lines of code:

```
url <- paste0("https://raw.githubusercontent.com/bgreenwell/",
              "eesR/master/R/Data/HgBass.txt")
hg_bass <- read.table(url, header = TRUE)
```

a. Discuss an appropriate sampling plan that could have been used to sample the fish from each lake (be sure to explain why you think this sampling plan is appropriate).

b. Fit a simple linear regression model using the average mercury concentration (`Avg.Mercury`) as the response and alkalinity (`Alkalinity`) as the predictor. Discuss the model fit and interpret the estimated slope parameter and R^2 value in the context of the problem. Be sure to discuss any potential problems with the model (**Hint:** look at the residual plots!). The following lines of code may be useful:

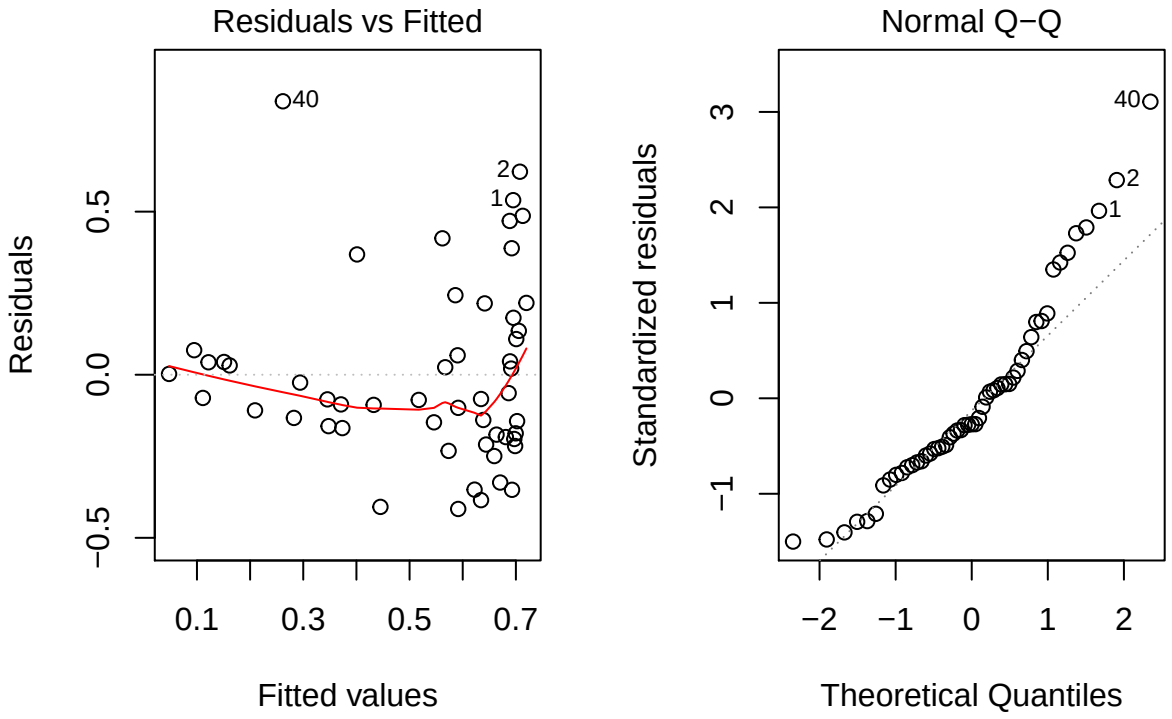
```r
# PLOT YOUR DATA FIRST!
plot(Avg.Mercury ~ Alkalinity, data = hg_bass)
```




```r
# Fit a simple linear regression model
mod1 <- lm(Avg.Mercury ~ Alkalinity, data = hg_bass)
summary(mod1)  # print model summary
```

```
##
## Call:
## lm(formula = Avg.Mercury ~ Alkalinity, data = hg_bass)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41148 -0.17963 -0.07442  0.10931  0.83828
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.726140   0.053599  13.548  < 2e-16 ***
## Alkalinity  -0.005302   0.001006  -5.272 2.76e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2771 on 51 degrees of freedom
## Multiple R-squared:  0.3527, Adjusted R-squared:   0.34
## F-statistic: 27.79 on 1 and 51 DF,  p-value: 2.763e-06
```

```r
# Residual plots
par(mfrow = c(1, 2))
plot(mod1, which = 1L:2L)
```



c. Use a log-transformation on one or both of the variables to see if the model in (b) can be improved. (You should try all three and select the model you think fits best.) Discuss the "best" fitting model (including why you think it's the best). Interpret the slope of your chosen model and provide a 95% confidence interval for it (**Hint:** use the `confint()` function). Be sure to interpret the confidence interval. The following lines of code may be useful:

```
mod2 <- lm(log(Avg.Mercury) ~ Alkalinity, data = hg_bass)
mod3 <- lm(Avg.Mercury ~ log(Alkalinity), data = hg_bass)
mod4 <- lm(log(Avg.Mercury) ~ log(Alkalinity), data = hg_bass)
```