MTH 3220 Lab 8

Due Thu., Nov. 7

1 Part A: Linear Regression

1.1 Yellowstone Lake Data Set

Yellowstone Lake is located in the southeastern part of Yellowstone National Park and covers an area about 136 mi^2 (352 km^2) depending on the level of water in the lake.

The water level in Yellowstone Lake varies each year in response to differences in the winter's snowpack accumulation, spring precipitation, and air temperatures. Restriction at the outlet of the lake retards the outflow, and water backs up during periods of high inflows. The U.S. Geological Survey started publishing Yellowstone Lake elevations in 1922 and outflows in 1926.

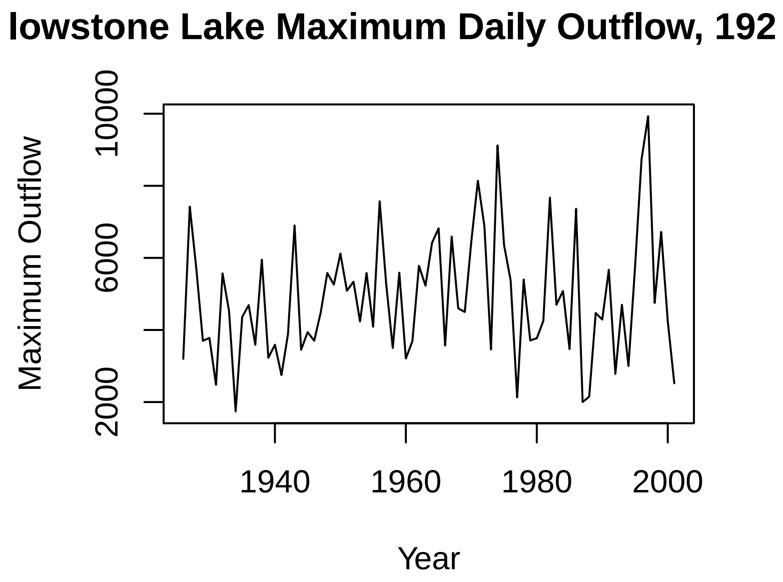
The file **yellowstone_lake.txt** contains data on the **maximum daily outflow** (ft^3/sec) and **maximum daily elevation** (ft) for each of the **years 1926 - 2001**.

1. Use `read.table()` to read the data into a data frame in R.
2. If we make a scatterplot of **maximum daily outflow** (y) versus **year** (x) by typing:

```
plot(x = my.data$Year, y = my.data$MaxOut,
     xlab = "Year", ylab = "Maximum Outflow",
     main = "Yellowstone Lake Maximum Daily Outflow, 1926-2001")
```

   Is there a ***trend*** in the lake's **outflow**? Since the X variable is *time*, it's better to make a ***time series plot*** by specifying `type = "l"` (for ”line”) in the call to `plot()`:

```
plot(x = my.data$Year, y = my.data$MaxOut,
     xlab = "Year", ylab = "Maximum Outflow",
     main = "Yellowstone Lake Maximum Daily Outflow, 1926-2001",
     type = "l")
```



3. We want to fit the ***linear regression model***

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

to the data, with **maximum daily outflow** as the response variable and **year** as the predictor, and carry out a t test of the hypotheses

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_a : \beta_1 &\neq 0 \end{aligned}$$

The function `lm()` will carry out the ***linear regression analysis***. Among its arguments are:

`formula`	A formula specifying the regression model
`data`	A data frame in which the variables specified in the formula will be found.

Use `lm()` to fit the regression model and carry out the hypothesis test, saving the results in an object called, say, `my.reg`, for example by typing:

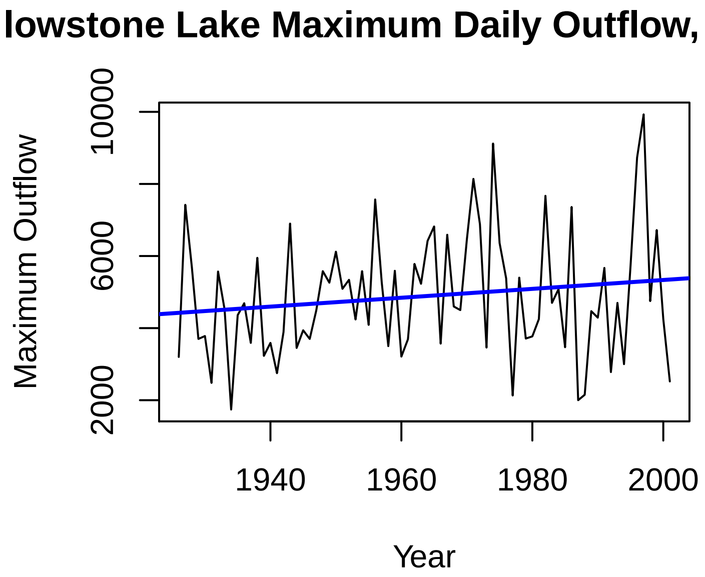
```
my.reg <- lm(MaxOut ~ Year, data = my.data)
```

4. Now use `summary()` to obtain the estimates of the coefficients β_0 and β_1 and to look at the results of the t test for the slope.

5. Re-create the ***time series plot*** of Step 2, then **add** the **fitted regression line** to the plot by typing:

```r
abline(my.reg)
```

Your plot should look something like this:



6. The object `my.reg` is a *list* (type `is.list(my.reg)`). Typing:

```r
names(my.reg)
```

will show the names of the objects stored in `my.reg`, and the `$` operator can be used to extract specific objects from `my.reg`.

Check the **normality assumption** for the error term ϵ in the regression model by making a histogram (use `hist()`) of the residuals (`my.reg$residuals`).

7. Now check the **constant standard deviation assumption** for ϵ by making a plot (use `plot()`) of the residuals (y-axis) versus the fitted values (`my.reg$fitted.values`, x-axis).

Add a horizontal line to the plot at $y = 0$ by typing:

```r
abline(h = 0)
```

8. Look at the ***regression ANOVA table*** by typing:

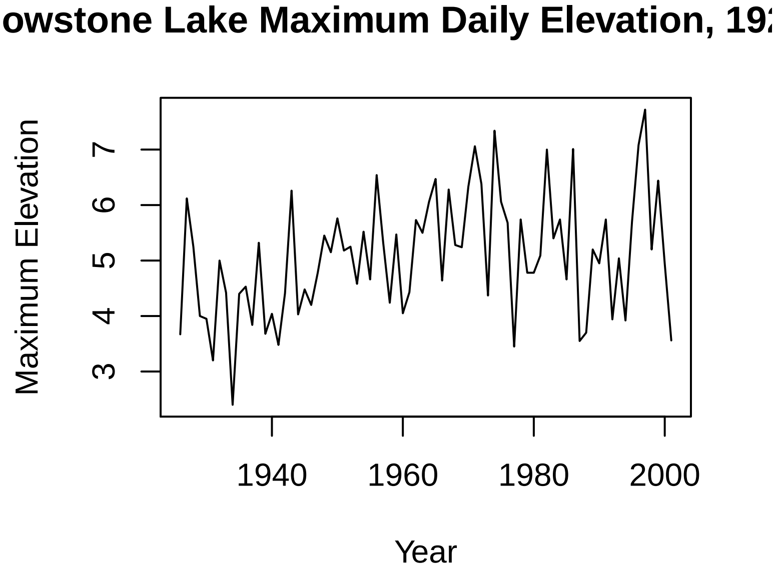
```r
anova(my.reg)
```

2 Part B: More Linear Regression

2.1 Yellowstone Lake Data Set (Continued)

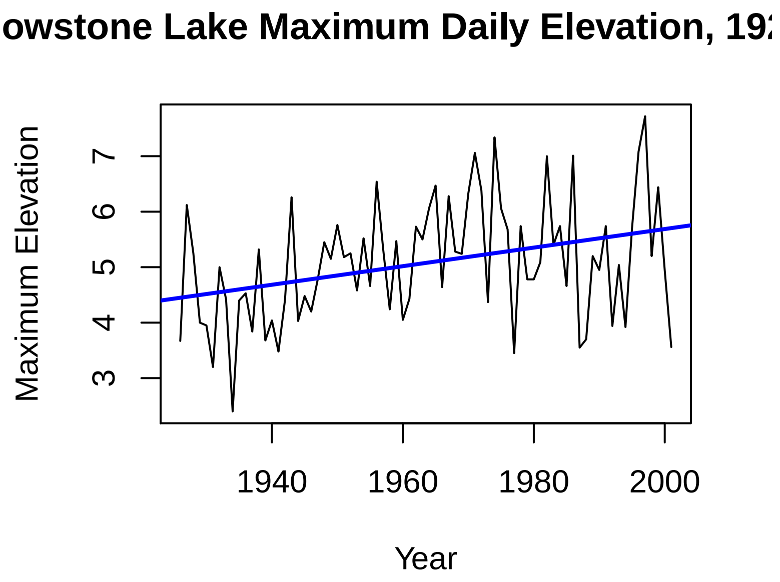
Now we'll analyze the **maximum daily elevations** of Yellowstone Lake.

1. Make a ***time series plot*** of the **maximum daily elevation** (y) versus **year** (x). Include an appropriate title and x and y axis labels. Your plot should look something like this:



2. Use `lm()` to carry out a linear regression analysis, with **maximum daily elevation** as the response variable and **year** as the predictor. Then use `summary()` to look at the results.

3. Add the regression line to the plot from Step 1. Your plot should look something like this:



4. Check the **normality assumption** for the error term ϵ in the regression model by making a histogram of the residuals.
5. Now check the **constant standard deviation assumption** for ϵ by plotting the residuals (y-axis) versus the fitted values (x-axis). Add a horizontal line to the plot (`abline(h = 0)`).
6. Look at the ***regression ANOVA table*** (`anova(my.reg)`).

3 Part C: Correlation

3.1 Yellowstone Lake Data Set (Continued)

We want to determine how closely the **maximum daily outflow** is related to the **maximum daily elevation** of Yellowstone Lake.

1. Make a scatterplot of **maximum daily outflow** (y) versus **maximum daily elevation** (x) by typing:

```
plot(x = my.data$MaxElev, y = my.data$MaxOut, pch = 19)
```

2. Now compute the ***correlation*** r between these two variables by typing:

```
cor(x = my.data$MaxElev, y = my.data$MaxOut)
```