MA117 - Worksheet 12
Review
March 19, 2021 - Week 3, Friday

Problem 1. Load in the diamonds data set again:

```
diamonds <- read.csv("https://sagrawalx.github.io/teaching/sp21-b6_ma117/wksht/
diamonds.csv"
```

Recall that this data set records lots of information about lots of diamonds, including the following.

- `price`: price in US dollars
- `x`: length in mm
- `y`: width in mm
- `z`: depth in mm
- `depth`: total depth percentage (ie, `z / mean(x, y)`)

(a) Make a plot of a diamond's `price` against its *depth* using: `plot(diamonds$depth, diamonds$price)`. Describe the relationship. Does this seem like a linear relationship?
(b) Run a least squares regression using: `lsline <- lm(diamonds$price ~ diamonds$depth)`.
(c) Plot the least squares line using: `abline(coefficeints(lsline))`.
(d) Run `summary(lsline)` to get some numerical output about the least squares line. Look through this output to answer the following questions.
(e) What is the (adjusted) R^2 value? What does that tell you about the line?
(f) What is the slope of the best fit line?
(g) What is the t-value associated to the slope of this best fit line? At a significance level of $\alpha = 0.05$, would you reject the hypothesis that the best fit line has slope zero? What about if $\alpha = 0.01$?

Problem 2. Use the `read.csv` function to load the following dataset about vocabulary into R:
`https://sagrawalx.github.io/teaching/sp21-b6_ma117/wksht/vocab.csv`
This data matrix contains 30351 observations of 5 variables:

- `X` is an ID of the respondent.
- `year` of the survey.
- `sex` of the respondent (`Female` or `Male`)
- `education` in years.
- `vocabulary` test score (out of 10).

(a) During what time period was this data collected?

(b) Is there a *statistically* significant difference between the vocabularies of men and women? If so, is there a *practically* significant difference? Do you clearly understand the difference between these two questions?

(c) Are `education` and `vocabulary` associated? If so, is it a positive association or a negative one, and how practically significant does this association seem?

Problem 3. Use the `read.csv` function to load the following dataset about the number of pages in a textbook and the price of a textbook into R.
`https://sagrawalx.github.io/teaching/sp21-b6_ma117/wksht/textbooks.csv`
Is there a relationship between the number of pages and the price of a textbook? If you had to use this data to predict the price of a 500 page textbook, what would your estimate be?

Problem 4. Use the `read.csv` function to load in the following dataset about land area and the number of mammal species on several islands in Southeast Asia.
`https://sagrawalx.github.io/teaching/sp21-b6_ma117/wksht/speciesdiversity.csv`
There are five variables.

- `Name` of the island
- `Area` of the island in sq km
- `Species`: the number of mammal species
- `logArea`: the natural log (base e) of `Area`
- `logSpecies`: the natural log (base e) of `Species`

(a) Do `Area` and `Species` seem like they have a linear relationship? What about `Area` and `logSpecies`? `logArea` and `Species`? `logArea` and `logSpecies`?

(b) (Challenging?) Elaborate on your answers to the above questions and derive a formula that predicts the number of mammal species based on the area of an island.

Problem 5. Use the `read.csv` function to load in the following dataset from a survey conducted among students at Grinnell College in 1992.
`https://sagrawalx.github.io/teaching/sp21-b6_ma117/wksht/grinnell.csv`
There are 59 observations on the following variables.

- `Year`: Class year (`1` to `4`)
- `Sex`: `0`=male, `1`=female
- `Vote`: Voting status: `0`=not eligible, `1`=eligible/not registered, `2`=registered/didn't vote, `3`=voted
- `Paper`: Read news (per week): `0`=never, `1`=less than once, `2`=once, `3`=2 or 3 times, `4`=daily
- `Edit`: Read editorial page? `0`=no or `1`=yes
- `TV`: Watch TV news: `0`=never, `1`=less than once, `2`=once, `3`=2 or 3 times, `4`=daily
- `Ethics`: Politics should be ruled by: `1`=ethical considerations to `5`=practical power

- `Inform`: How informed are you about politics? `1`=uninformed to `5`=very well informed
- `Participate`: Missing if `Vote`=`0`, `0` if `Vote`=`1` or `2`, `1` if `Vote`=`3`

Ask any question you want about this data, and then answer it!