

MA117 - WORKSHEET 13
REVIEW
Week 3, Friday

Problem 1. Load in the `diamonds.csv` data set again:

https://sagrawalx.github.io/teaching/fa21-b1_ma117/class/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, $z / \text{mean}(x, y)$)

(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)
abline(coefficients(lsline))
summary(lsline)
```

The first command calculates the least squares line and stores it in `lsline`. The second command plots the line, and the third displays some useful information about `lsline`. Look through this output to answer the following questions.

- (c) What is the (adjusted) R^2 value? What does that tell you about the line?
- (d) What is the slope of the best fit line?
- (e) What is the t-value associated to the slope of this best fit line?
- (f) At a significance level of $\alpha = 0.05$, would you reject the hypothesis that the best fit line has slope zero?

Problem 2. Here's a data set about vocabulary:

https://sagrawalx.github.io/teaching/fa21-b1_ma117/class/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.
- **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? Run an appropriate hypothesis test to answer this question. If there is an association, is it a positive association or a negative association, and how practically significant does this association seem?

Problem 3. Here's a data set about textbooks:

https://sagrawalx.github.io/teaching/fa21-b1_ma117/class/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. Here's a data set about species diversity on several Southeast Asian islands:

https://sagrawalx.github.io/teaching/fa21-b1_ma117/class/speciesdiversity.csv

There are five variables.

- **Name** of the island
 - **Area** of the island in sq km
 - **Species** is the number of mammal species
 - **logArea** is the natural log (base e) of **Area**
 - **logSpecies** is 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) Use your answers to part (a) to derive a formula that predicts the number of mammal species on an island based on the area of that island.