HW5

Please submit the solution in the form of R Markdown report, knitted into either of the available formats (HTML, pdf or Word). Provide all relevant code and output. Goal of this homework is to have you familiarized with 1) statistical inference for both simple and multiple linear regression (confidence intervals, t-test, F-test, prediction/confidence bands); 2) practice your R coding.

Problem #1

For *Advertisement.csv* data set, proceed to fit simple linear regressions of *sales* onto *radio*.

1. Interpret the hypothesis test results. Is there a statistically significant relationship between predictor and response? Why?
2. Interpret the confidence intervals for both the intercept and the slope.
3. For a $20,000$\$ investment into radio ads, provide and interpret
    a. A single prediction.
    b. 95% confidence bands.
    c. 95% predictions bands.

    Which ones are wider - confidence or prediction? Why do you think that is?

Problem #2

For the *FL_crime.csv* data, proceed to fit

1. For simple linear regression $crime \sim education$,
    a. Write down the **full modeling equation**, with all **error assumptions**.
    b. Fit the model, provide the **fitted equation**. Provide a plot of the fitted line.
    c. Is there a statistically significant relationship? If yes, interpret the effect of education on crime.
2. For multiple linear regression $crime \sim education + urbanization$,
    a. Write down the **full modeling equation**, with all **error assumptions**.
    b. Fit the model, provide the **fitted equation**. Provide a plot of the **fitted plane**.
    c. Provide interpretation of *education* effect on *crime*. Did it change compared to part 1? Why?
    d. Provide interpretation of the intercept. By the sound of it, does it make sense to interpret it here?
3. For multiple linear regression $crime \sim education + urbanization + income$,
    a. Write down the **full modeling equation**, with all **error assumptions**.
    b. Having fitted the model from part $3(a)$, provide the **fitted equation**.
    c. Write down the hypotheses (in terms of parameters of the model in part $3(a)$) and make conclusions for t-tests on significance of each **individual** predictor.
    d. Interpret the effect of the only statistically significant predictor from part $3(c)$.
    e. Formulate the hypotheses (in terms of parameters of the model in part $3(a)$ for testing the overall model significance. Provide the conclusion of the respective test.

Problem #3 (Why need F-statistic?)

1. Generate a data example where you have a response variable Y and a predictor variable X that are **unrelated** to each other (make sure to use a **random** generation mechanism). How would you do that? How would you demonstrate that they're unrelated (think of basic visualizations)?

2. Having settled on a method of generating such unrelated variables in part 1, proceed to:

    a. Generate response variable vector Y (e.g. of length 200).
    b. Generate 50 predictor variable vectors $X_1, X_2, \ldots, X_{50}$ according to your method from part 1. **Record them**.

3. Fit a **multiple** linear regression model, regressing response Y from part 2(a) on all 50 X's you've generated in part 2(b).

    a. Report the # of individual t-tests that resulted into a significant p-value, hence rejecting H_0 at $\alpha = 0.05$ level. Were those rejections correct decisions or not? Why? (**Hint**: Remember what's the true relationship between X and Y, given how you generated the data.) If not, what type of error do they correspond to? Why?
    b. Given that the individual significant t-test aren't necessarily indicative of at least one predictor having a true relationship with the response, what would be the appropriate testing procedure to address that question? Conduct that test, report its p-value, interpret its result.