Homework 3

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 1) familiarized χ^2-test of independence for contingency tables; 2) familiarized with permutation test for contingency tables; 3) interpretation of linear regression; 4) practice your R coding.

Problem #1.

1. Code up your own *my.permutation.test*() function to conduct permutations tests on contingency tables.

As inputs, it should take

- data frame with two categorical variables as columns (first one - explanatory, second one - response),
- # of randomly generated permutations to be executed.

As output, it should provide:

- contingency table for the data frame,
- permutation p-value,
- plot the histogram of permutation distribution for X^2 statistic.

NOTE: To obtain X^2 values for each permutation, you are allowed to use *R*'s *chisq.test*() function.

2. Proceed to apply the *my.permutation.test*() function (and subsequently interpret the results) to:

   a. Snowden data (from the lecture), with $10,000$ permutations. What's the conclusion? Compare the resulting histogram with the one in the slides (they should be roughly similar).
   b. Airbnb data (from previous HW), with **just** $1,000$ **permutations**. What's the conclusion? Compare the shape of resulting histogram with the density of χ^2 distribution with appropriate degrees of freedom. What does it tell us about whether χ^2-test results from previous HW were appropriate for Airbnb data?

Problem #2

In *Advertisement.csv* data set, proceed to study the relationship between the sales and radio advertising expenses. In particular, proceed to

1. Plot their relationship. Does linear regression appear as appropriate model here?

2. Regardless of the answer to Part 1, proceed to fit the linear regression and write down the **fitted model equation**.

3. Interpret both the slope and the intercept.
4. Provide and **interpret** the prediction for a $50,000\$$ investment into this advertisement media.
5. Report and interpret the Residual Standard Error (RSE).
6. Report and interpret the R^2 statistic.

Problem #3

1. When one obtains the fitted simple linear regression formula

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X,$$

the actual mathematical formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$ are:

$$\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{\mathbf{X}})(Y_i - \bar{\mathbf{Y}})}{\sum_i (X_i - \bar{\mathbf{X}})^2}, \qquad \hat{\beta}_0 = \bar{\mathbf{Y}} - \hat{\beta}\bar{\mathbf{X}},$$

where $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ - vector of explanatory variable values, and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ - vector of response values.

   a. Proceed to write your own function which will calculate these estimates. Specifically, your function should

      - Take vectors $\mathbf{X}, \mathbf{Y}$ as inputs.
      - Output estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

      Hint: You're better off avoiding loops. Conduct calculations on full vectors (AKA "vectorized calculation").

      Note: **Your function shouldn't "cheat" by using R's built-in "lm()" function. You need to explicitly implement above-mentioned formulas (by applying basic vectorized operations on inputs $\mathbf{X}, \mathbf{Y}$).**

   b. **"Sanity check"**: Demonstrate that your function works properly on *sales* ~ *TV* regression example (*Advertising.csv* data), by supplying it vector of *TV* ad expense values as $\mathbf{X}$ argument, vector of *Sales* values as $\mathbf{Y}$ argument, and comparing your calculated estimates with the ones yielded by *lm*() function.

2. a. Write your own function that, for a simple linear regression $Y = \beta_0 + \beta_1 X + \epsilon$, will

      - Take vectors $\mathbf{X} = (X_1, X_2, \ldots, X_n), \mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ as inputs.
      - Output Residual Standard Error (RSE) and R^2 statistic.

      In your function, you're allowed to use *lm*() function, but only for purposes of obtaining $\hat{Y}$ values. You are NOT allowed to extract *RSE* or R^2 values directly from the *lm* object. Instead, you will need to explicitly implement *RSE* & R^2 formulas from slide deck #3.

   b. **"Sanity check"**: Demonstrate that your function works properly by supplying it vector of ad expense values for arbitrary media from *Advertising.csv* data (*TV*, *radio*, *newspaper*, whichever) as $\mathbf{X}$ argument, vector of *sales* values as $\mathbf{Y}$ argument, and comparing your calculated RSE and R^2 values with the ones yielded by *summary*() function applied to the *lm* object.