Homework 4
Due Tuesday, October 8

Show all your work. Data files available on class website.

1. Consider the following data

$X1$	1	2	3	4	5
$X2$	10.1	9.0	8.9	8.2	12.6
Y	1.9	2.2	2.9	3.2	5.2

   (a) Set up and run the model $Y = X1 + X2$ by using `lm()` function in R, and report the summary.

   (b) Write down a vector $\mathbf{Y}$ and a matrix $\mathbf{X}$, in terms of $\mathbf{Y} = \{Y_i\}$ and $\mathbf{X} = \{X_{ij}\}$ (include intercept term in $\mathbf{X}$), and in terms of actual numerical entries in $\mathbf{Y}$ and $\mathbf{X}$. Then, use these and R matrix computation to confirm the numbers in part (a).

   (c) Repeat parts (a) and (b) with the model $Y = X2$. Compare this model with both the model $Y = X1$ (done in previous HWs) and the model $Y = X1 + X2$ above. Comment.

2. Refer to the `wateruse` data in Example 3 of the lecture. Set up and use R matrix computation to confirm the numbers from `lm()` functions (follow the parts of Example 2 in the lecture; may skip the Prediction part), for both models.

3. Please see the dataset `cofreeway.txt`. Here is the description:

```
Hour -hour of the day, from midnight to midnight
CO   -average summer weekday CO concentration (parts per million)
TD   -average weekday traffic density
WS   -average perpendicular wind-speed component
```

   We are interested in the effects of `Hour`, `TD` and `WS` have on `CO` .

   (a) Provide a summary statistics, scatter plots (using pairs plot), and the correlation for all variables.

   (b) Compute the regression of `CO` on `Hour`, `TD` and `WS`, using `lm()` function in R, and report the summary.

   (c) Verify the numbers in part (b) by matrix computation in R.

   (d) Provide a residual plot.

   (e) Interpret the results and comment on any interesting observations.

4. Consider the `water` dataset from `library(alr4)` (see `?water` for the description of the dataset). We are interested in the effects of `OPBPC`, `OPRC`, `OPSLAKE` variables have on `BSAAM`. Use these information to repeat all parts of Problem 3 (for part (a), use only the variables of interest).

5. Consider the `BGSgirls` dataset from `library(alr4)` (see `?BGSgirls` for the description of the dataset). We are interested in the effects of `HT2`, `HT9`, `WT2`, `WT9`,`ST9` variables have on `BMI18`. Use these information to repeat all parts of Problem 3 (for part (a), use only the variables of interest).

6. Refer to the `physics` data from Example 4 of the lecture. First, write down $\mathbf{Y}$ and $\mathbf{X}$, in terms of $\mathbf{Y} = \{Y_i\}$ and $\mathbf{X} = \{X_{ij}\}$ (include intercept term in $\mathbf{X}$), and in terms of actual numerical entries in $\mathbf{Y}$ and $\mathbf{X}$. Then, please confirm all numbers from `summary(physics2.lm)` and `summary(physics3.lm)` by matrix computation in R.

7. Please find the dataset `un.txt`. Let `fertility` be the predictor and `ppgdp` be the response.

   (a) Fit a (simple) linear regression model. Can you explain the discrepancy between t-test and R^2 value?

   (b) Fit a quadratic (polynomial of degree 2) regression. Does this help the model fit?

   (c) Fit a cubic (polynomial of degree 3) regression. What happens now?

   (d) Please create a scatterplot and superimpose regression lines (linear, quadratic, and cubic) on it. Does the plot help explain the results in the previous parts?

   (e) What is your overall conclusion about the analysis?