550.413 Assignment 1

Fall 2016

Instruction: This assignment consists of 4 problems. The assignment is due on **Wednesday, September 28, 2016** at 3pm, in class. If you cannot make it to class, please leave the assignment under the door at Whitehead 306E and email the course instructor. If possible, please type up your assignments, preferably using LaTeX. For problems with R programming, please also attach a print-out of the **R** code. When asked to perform hypothesis testing, you are free to use any (reasonable) choice of significance level α.

Problem 1: (10pts)

Suppose we are given n data points $\{(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)\}$

(a) We are interested in fitting the linear regression model $Y_i = \beta_1 X_i + \epsilon_i$ where the ϵ_i are independent and identically distributed $N(0, \sigma^2)$. Derive the least square estimate $\hat{\beta}_1$ of β. Find the distribution of $\hat{\beta}_1$ and propose an estimate for its variance.

(b) We are also interested in fitting the linear regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where the ϵ_i are again independent and identically distributed $N(0, \sigma^2)$ variables. It turns out, incidentally, that the $\{X_i\}$ satisfies $\sum_{i=1}^{n} X_i = 0$. What are the least squares estimates of β_0 and β_1 in this case ? Do you observe any interesting aspect of the least square estimation due to the fact that $\sum_{i=1}^{n} X_i = 0$? When doing simple linear regression, can you always assume, without loss of generality, that $\sum_{i=1}^{n} X_i = 0$?

Problem 2: (10pts)

Suppose we are given n data points $\{(X_1, Y_1, Z_1), (X_2, Y_2, Z_2), \dots, (X_n, Y_n, Z_n)\}$. We are interested in fitting the linear regression model $Y_i = \alpha + \beta X_i + \epsilon_i$ and $Z_i = \gamma + \beta X_i + \eta_i$ for $i = 1, 2, \dots, n$ where the $\{\epsilon_i\}$ and the $\{\eta_i\}$ are independent random variables with zero mean and common variance σ^2. Derive the least squares estimates $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ of α, β and γ algebraically. Note that we require the linear coefficient β in both the regression model for Y_i on X_i and Z_i on X_i to be the same.

Hint: The least square objective function can be written as

$$Q = \sum_{i=1}^{n}(Y_i - \alpha - \beta X_i)^2 + \sum_{i=1}^{n}(Z_i - \gamma - \beta X_i)^2$$

We can then estimate α, β and γ by taking the partial derivatives of Q with respect to α, β, and γ, set the resulting partial derivatives to 0 and solve for the estimates $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$.

Problem 3: (20pts)

Install the R package SemiPar to get access to the sausage data set for the calories vs sodium content among several sausage types. Using R but not the lm command in R, do the following.

(a) Perform a simple linear regression using calories as the response variable and sodium level as the predictor variable. Find the least square estimate for the coefficients.

(b) Setup a hypothesis test to test whether calories count is associated with the sodium content. What is your conclusion ?

(c) Predict the calories count (that is, obtain the prediction intervals) when the sodium content is 350 mg, 520 mg, and 441 mg.

(d) Find the equation defining the 95% confidence band for the regression line using the Working-Hotelling approach.

Problem 4: (20pts)

The data for this problem is available from the link `https://us.sagepub.com/sites/default/files/upm-binaries/26929_lordex.txt`

The data records a study about misinformation and facilitation effects in children. The data consists of 51 observations, with each observation corresponding to a child between the age of 4 to 9 years old. The children saw a magic show and were then asked questions – in two separate sessions – about the events that happened during the show. The first session takes place one week after the show, while the second session takes place roughly 10 months after the show. The children was scored in each session based on how much they managed to recall the events of the magic show. The study is described in more detail in the paper "Post-Event Information Affects Children's Autobiographical Memory after One Year" by K. London, M. Bruck and L. Melnyk, *Law and Human Behavior*, Volume 33, 2009. A snippet of the data is given below; here AGEMOS refers to the age of the child in months at the start of the first session, Initial and Final are the scores of the child in the first and second session, respectively.

	AGEMOS	Final	Initial
1	55	0	0
2	82	6	8
3	81	3	6
4	71	0	3
5	84	2	15
6	76	2	8

Once you download the above data file, you can read it into **R** using the following command [1]

```r
df <- read.table("26929_lordex.txt", sep = "", header = T)
## Now make a new column called age.binarize
df$age.binarize <- (df$AGEMOS <= 78)
df$age.binarize <- factor(df$age.binarize, levels = c(T, F), labels = c("younger", "older"))
## Now make a new column called score.difference
df$score.difference <- df$Final - df$Initial
```

We have decided to binarize the age of the children into two categories, namely those for which the child is 78 months or *younger*, and those for which the child is 79 months or *older*. After adding the above columns, the above snippet of data becomes

	AGEMOS	Final	Initial	age.binarize	score.difference
1	55	0	0	younger	0
2	82	6	8	older	-2
3	81	3	6	older	-3
4	71	0	3	younger	-3
5	84	2	15	older	-13
6	76	2	8	younger	-6

Using the above data, answer the following questions.

(a) A scientist wants to inquire whether or not older children remember events longer than younger children. He thinks that the way to do this is by performing a regression with score.difference as the response variable and age.binarize as the predictor variable, i.e., he consider the model

$$\text{score.difference}_i = \beta_0 + \beta_1 \times \mathbf{1}\{\text{age.binarize}_i = \text{“older”}\} + \epsilon_i$$

where $\mathbf{1}\{\text{age.binarize}_i = \text{“older”}\}$ is 1 if the i-th child is older than 79 months and 0 otherwise. Without using the lm command in **R**, find the least square estimate for β_0 and β_1 under this model. What is the estimated coefficient $\hat{\beta}_1$? Assuming the normal error regression model,

[1] **R** might warns you about EOF in the downloaded file, but you can safely ignore this warning.

comment on the output of this regression, e.g., is the estimated coefficient $\hat{\beta}_1$ statistically significant ? Under this model, what does the estimated coefficient $\hat{\beta}_1$ say about the scores of the older children compared to the scores of the younger children ?

(b) Another scientists also wants to inquire whether or not older children remember events longer than younger children. She thinks that the way to do this is by performing a regression with Final as the response variable and Initial and age.binarize as the predictor variables, i.e., she consider the model

$$\text{Final}_i = \beta_0 + \beta_2 \times \text{Initial}_i + \beta_1 \times \mathbf{1}\{\text{age.binarize}_i = \text{``older''}\} + \epsilon_i$$

Without using the lm command, compute the least square estimate for β_1 under this model [2]. Under this model, what does the estimated coefficient $\hat{\beta}_1$ say about the scores of the older children compared to the scores of the younger children ?

(c) (Bonus: 10pts) Comment on the discrepancy in the estimate for β_1 between the above two regression models. What do you think is the reason behind this discrepancy ?

[2] Once again, write down the least square objective function in terms of the parameters β_0, β_1 and β_2 and then find $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ by setting the partial derivatives with respect to β_0, β_1 and β_2 to 0. For simplicity, you can also assume that the least square estimate for β_2 is known to be $\hat{\beta}_2 = 0.12$