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 $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$

(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 $\sum_{i=1}^n X_i = 0$?

Solution

For part (a), the least square formulation is

$$\arg\min_{\beta} Q = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - \beta_1 X_i)^2$$

Taking the partial derivatives of Q with respect to β_1 and setting the resulting expression to 0 yields the following estimate $\hat{\beta}_1$ of β_1

$$\hat{\beta}_1 = \sum_i \frac{X_i Y_i}{\sum X_i^2}$$

As $\hat{\beta}_1$ is a linear combination of the Y_i, it is normally distributed. It is also an unbiased estimator of β_1, and its variance is easily computed to be

$$\text{Var}[\hat{\beta}_1] = \text{Var}\Big[\sum_i \frac{X_i Y_i}{\sum X_i^2}\Big] = \sum_i \Big(\frac{X_i}{\sum X_i^2}\Big)^2 \sigma^2 = \frac{\sigma^2}{\sum_i X_i^2}$$

As σ^2 is unknown, we can replace it with an estimate based on the residuals $e_i = Y_i - \hat{\beta}_1 X_i$. An example of such an estimate for σ^2 is

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} e_i^2$$

One can then show that s^2 is an unbiased estimator of σ^2. Note that other estimators for σ^2 is possible provided that some justification is given.

For part (b), the least square estimator for $\hat{\beta}_0$ and $\hat{\beta}_1$ is given by

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \bar{Y}$$
$$\hat{\beta}_1 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \sum_{i=1}^{n} \frac{X_i Y_i}{\sum X_i^2}$$

Thus, the estimator $\hat{\beta}_0$ for β_0 can be computed independently of the estimator $\hat{\beta}_1$ for β_1. That is to say, to estimate β_0, we can perform a regression without the linear term β_1 and to estimate β_1, we can perform a regression without the intercept term β_0.

We can reparametrize X_i so that $\sum X_i = 0$ when performing a simple linear regression. Indeed, by letting $X_i' = X_i - \bar{X}$, we have $\sum X_i' = 0$. That is to say, the regression problem $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ can always be reparametrized as

$$Y_i = \beta_0 + \beta_1 \bar{X} + \beta_1 (X_i - \bar{X}) + \epsilon_i = \beta_0' + \beta_1 X_i' + \epsilon_i$$

with $\beta_0' = \beta_0 + \beta_1 \bar{X}$ and $X_i' = X_i - \bar{X}$.

Problem 2: (10pts)

Suppose we are given n data points $\{(X_1, Y_1, Z_1), (X_2, Y_2, Z_2), \ldots, (X_n, Y_n, Z_n)\}$. 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 $i = 1, 2, \ldots, n$ 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 $\hat{\gamma}$.

Solution:

The sum of square formulation for the above problem is as follows.

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

Thus, the least square estimates $\hat{\alpha}, \hat{\beta}$ and $\hat{\gamma}$ for α, β, γ is given by

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

Taking the partial derivatives of Q with respect to α, β and γ gives

$$\frac{\partial Q}{\partial \alpha} = -2\sum_{i=1}^{n}(Y_i - \alpha - \beta X_i)$$

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

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

Setting the partial derivatives of Q with respect to α, β and γ to all be 0 (and denote $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$ as the solutions of the resulting equations), we have

$$n\hat{\alpha} = \sum_{i=1}^{n}(Y_i - \hat{\beta} X_i)$$

$$n\hat{\gamma} = \sum_{i=1}^{n}(Z_i - \hat{\beta} X_i)$$

or equivalently

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$
$$\hat{\gamma} = \bar{Z} - \hat{\beta}\bar{X}$$

If we now set the partial derivative of Q with respect to β to be 0 and use the above values of $\hat{\alpha}$ and $\hat{\gamma}$ as the estimate for α and γ, we get,

$$\sum_{i=1}^{n} X_i(Y_i - \bar{Y} + \hat{\beta}\bar{X} - \hat{\beta}X_i) + \sum_{i=1}^{n} X_i(Z_i - \bar{Z} + \hat{\beta}\bar{Z} - \hat{\beta}X_i) = 0$$

Therefore,

$$2\hat{\beta}\sum_{i} X_i(X_i - \bar{X}) = \sum_{i=1}^{n} X_i(Y_i - \bar{Y} + Z_i - \bar{Z})$$

Further simplifications then yield

$$\hat{\beta} = \frac{1}{2}\sum_{i=1}^{n} \frac{(Y_i + Z_i - \bar{Y} - \bar{Z})(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}.$$

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.

Solution

(a) The following code computes the least square estimates for β_0 and β_1. The mean squared error MSE is also reported.

```r
library("SemiPar")
data(sausage)
Y <- sausage$calories
```

```r
X <- sausage$sodium

n <- length(X)
deg.freedom <- n - 2
txy <- sum((X - mean(X)) * (Y - mean(Y)))
txx <- sum((X - mean(X))^2)
(beta1.hat <- txy/txx)

## [1] 0.1431946

(beta0.hat <- mean(Y) - beta1.hat * mean(X))

## [1] 84.61059

Yhat <- beta0.hat + beta1.hat * X
residuals <- Y - Yhat

SSE <- sum(residuals^2)
(MSE <- SSE/deg.freedom)

## [1] 687.9577
```

(b) To test the hypothesis that calories count is associated with the sodium content of the sausage, we can test whether the linear coefficient β_1 is zero or non-zero. That is, we test the hypothesis

$$H_0\colon \beta_1 = 0 \quad \text{against} \quad H_A\colon \beta_1 \neq 0$$

If we reject the null hypothesis, then we conclude that there is an association between sodium content and calories count. We note that under the null hypothesis that $\beta_1 = 0$, one has

$$\frac{\hat{\beta}_1}{s\{\hat{\beta}_1\}}$$

follows a Student t distribution with 52 degrees of freedom. We now perform some quick computation.

```r
alpha <- 0.05
(s.beta1.hat <- sqrt(MSE/txx))

## [1] 0.03758564

(T <- beta1.hat/s.beta1.hat)

## [1] 3.809823

(reject <- abs(T) > qt(1 - alpha/2, df = deg.freedom))

## [1] TRUE

(p.value <- 2 * pt(abs(T), df = deg.freedom, lower.tail = FALSE))

## [1] 0.0003693124
```

From the above computation, we have that the test statistic $T = \frac{\hat{\beta}_1}{s\{\hat{\beta}_1\}} = 3.810$. The 97.5% quantile level for the Student t distribution with 52 degrees of freedom is 2.007 and so we reject the null hypothesis. That is, we conclude that there is an association between the sodium content and calories count. The p-value for this hypothesis test is 3.7×10^{-4}.

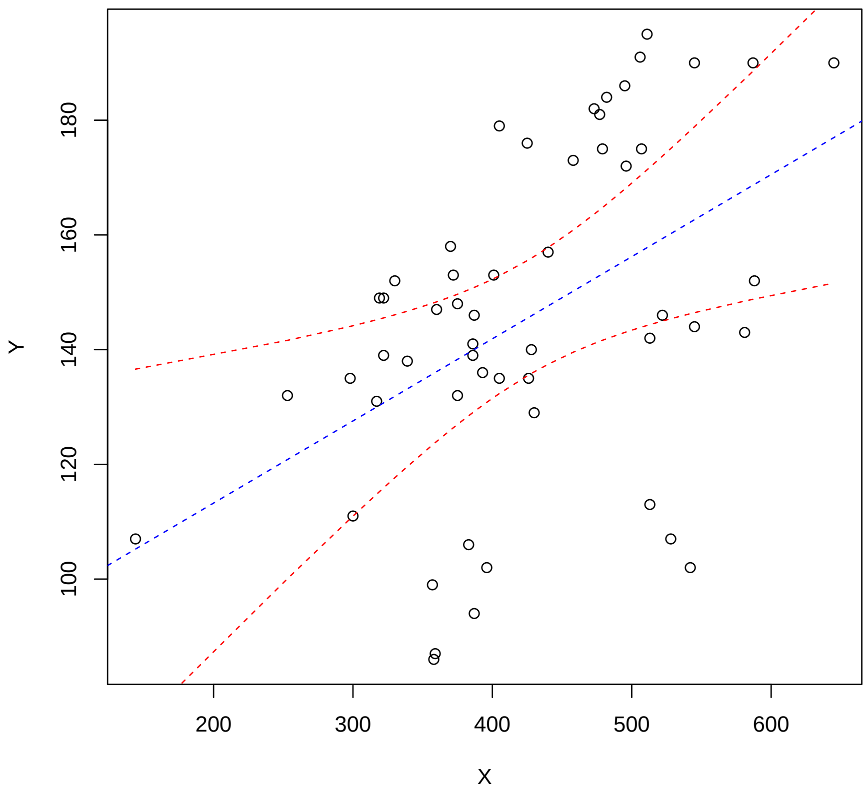
(c) We now perform some quick computations to obtain the prediction intervals for the calorie count when the sodium content is 350 mg, 441 mg and 520 mg.

```
x.pred <- c(350, 441, 520)
(yhat.pred <- beta0.hat + beta1.hat * x.pred)

## [1] 134.7287 147.7594 159.0718

(s.pred <- sqrt(MSE * (1 + 1/n + (x.pred - mean(X))^2/txx)))

## [1] 26.61970 26.47767 26.71127

lower.pred <- yhat.pred - qt(1 - alpha/2, df = deg.freedom) * s.pred
upper.pred <- yhat.pred + qt(1 - alpha/2, df = deg.freedom) * s.pred
pred.intervals <- data.frame(lower.bound = lower.pred, upper.bound = upper.pred)
pred.intervals

##   lower.bound upper.bound
## 1    81.31237    188.1451
## 2    94.62810    200.8907
## 3   105.47172    212.6719
```

For example, for a sausage whose sodium content is 441 mg, the 95% prediction interval for the calorie count is $(81.312, 188.145)$.

(d) We now compute the Working-Hotelling 95% Working-Hotelling confidence band.

```
(W <- sqrt(2 * qf(1 - alpha/2, 2, deg.freedom)))

## [1] 2.815456

s.WH <- sqrt(MSE * (1/n + (X - mean(X))^2/txx))

lower.WH <- Yhat - W * s.WH
upper.WH <- Yhat + W * s.WH

plot(X, Y)
abline(beta0.hat, beta1.hat, lty = 2, col = "blue")
## We want to first sort the X. This is so that the plotting of the confidence bands and
## confidence region is easier.
X.ordering <- order(X)

lines(X[X.ordering], lower.WH[X.ordering], lty = 2, col = "red")
lines(X[X.ordering], upper.WH[X.ordering], lty = 2, col = "red")
```



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]

```
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(TRUE, FALSE), 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

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

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, 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

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

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 ?

Solution:

(a) We compute the least square estimates for β_0 and β_1 under the model $\mathrm{score.difference}_i = \beta_0 + \beta_1 \times \mathbf{1}\{\mathrm{age.binarize}_i = \text{``older''}\} + \epsilon_i$. We also compute the estimate of the variance of $\hat{\beta}_1$. We first (re)-convert the `age.binarize` variable into a $\{0,1\}$ valued variable before computing the estimates [3].

```
Y <- df$score.difference
X <- numeric(length(df$age.binarize))
X[df$age.binarize == "older"] <- 1
X[df$age.binarize == "younger"] <- 0
txy <- sum((Y - mean(Y)) * (X - mean(X)))
txx <- sum((X - mean(X))^2)
(beta1.hat <- txy/txx)

## [1] -3.939145

(beta0.hat <- mean(Y) - beta1.hat * mean(X))

## [1] -1.842105
```

[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 $\hat{\beta}_2 = 0.12$

[3] Reconverting a factor into an integer is quite clunky, but ...

```r
Yhat <- beta0.hat + beta1.hat * X
n <- length(Y)
(MSE <- 1/(n - 2) * sum((Y - Yhat)^2))

## [1] 15.95908

(s2.beta1.hat <- MSE/txx)

## [1] 1.338673
```

From the above quantities, assuming the normal error regression model, the estimated coefficient $\hat{\beta}_1$ is statistically significant, i.e., we reject the null hypothesis $\mathbb{H}_0\colon \beta_1 = 0$ in favor of the alternative hypothesis $\mathbb{H}_A\colon \beta_1 \neq 0$ at a significance level of $\alpha = 0.01$. The value of the test statistic $T = \frac{\hat{\beta}_1}{s\{\hat{\beta}_1\}} = -3.405$. As $\hat{\beta}_1 = -3.939$ is negative, this model seems to suggests that older children tend to forget more than younger children.

(b) Let $\mathcal{I}_0$ and $\mathcal{I}_1$ be the subset of indices corresponding to children 78 months or younger and 79 months or older, respectively. Also let Y_i and X_i denote the initial score and the final score of the i-th child, respectively.

Then the least square objective function is

$$Q = \sum_{i \in \mathcal{I}_1} (Y_i - \beta_0 - \beta_2 X_i)^2 + \sum_{i \in \mathcal{I}_2} (Y_i - \beta_0 - \beta_1 - \beta_2 X_i)^2.$$

.

Then taking the partial derivatives with respect to β_0, β_1 and β_2 and setting these partial derivatives to 0, we obtain the following system of equations.

$$2 \sum_{i \in \mathcal{I}_0} (Y_i - \hat{\beta}_0 - \hat{\beta}_2 X_i) + 2 \sum_{i \in \mathcal{I}_1} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$
$$2 \sum_{i \in \mathcal{I}_1} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$
$$2 \sum_{i \in \mathcal{I}_0} X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_2 X_i) + 2 \sum_{i \in \mathcal{I}_1} X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$

Adding the first equation to the second equation in the above systems, we obtain

$$\sum_{i \in \mathcal{I}_0} (Y_i - \hat{\beta}_0 - \hat{\beta}_2 X_i) = 0$$

and hence

$$\hat{\beta}_0 = \bar{Y}^{(0)} - \hat{\beta}_2 \bar{X}^{(0)}$$

where $\bar{X}^{(0)}$ and $\bar{Y}^{(0)}$ are the sample means of the initial and final scores among all the children 78 months or younger. Substitution of $\hat{\beta}_0$ into the second equation in the above system then yields

$$\hat{\beta}_1 = \bar{Y}^{(1)} - \bar{Y}^{(0)} + \hat{\beta}_2 \bar{X}^{(0)} - \hat{\beta}_2 \bar{X}^{(1)}.$$

Once again, $\bar{X}^{(1)}$ and $\bar{Y}^{(1)}$ are the sample means of the initial and final scores among all the children 79 months or older, respectively. Substitution of $\hat{\beta}_0$ and $\hat{\beta}_1$ into the third equation in the above system finally yields

$$2\sum_{i \in \mathcal{I}_0} X_i(Y_i - \bar{Y}^{(0)} - \hat{\beta}_2(X_i - \bar{X}^{(0)})) + 2\sum_{i \in \mathcal{I}_1} X_i(Y_i - \bar{Y}^{(1)} - \hat{\beta}_2(X_i - \bar{X}^{(1)})) = 0$$

and hence[4]

$$\hat{\beta}_2 = \frac{\sum_{i \in \mathcal{I}_0}(Y_i - \bar{Y}^{(0)})(X_i - \bar{X}^{(0)}) + \sum_{i \in \mathcal{I}_1}(Y_i - \bar{Y}^{(1)})(X_i - \bar{X}^{(1)})}{\sum_{i \in \mathcal{I}_0}(X_i - \bar{X}^{(0)})^2 + \sum_{i \in \mathcal{I}_1}(X_i - \bar{X}^{(1)})^2}.$$

We then have the following computation (once again, very clunky, but ...)

```
Y <- df$Final
X <- df$Initial
W <- numeric(length(df$age.binarize))
W[df$age.binarize == "older"] <- 1
W[df$age.binarize == "younger"] <- 0
Y.subset0 <- Y[which(W == 0)]
Y.subset1 <- Y[which(W == 1)]
X.subset0 <- X[which(W == 0)]
X.subset1 <- X[which(W == 1)]
txy.subset0 <- sum((Y.subset0 - mean(Y.subset0)) * (X.subset0 - mean(X.subset0)))
txx.subset0 <- sum((X.subset0 - mean(X.subset0))^2)
txy.subset1 <- sum((Y.subset1 - mean(Y.subset1)) * (X.subset1 - mean(X.subset1)))
txx.subset1 <- sum((X.subset1 - mean(X.subset1))^2)
(beta2.hat <- (txy.subset0 + txy.subset1)/(txx.subset0 + txx.subset1))

## [1] 0.1206255

(beta0.hat <- mean(Y.subset0) - beta2.hat * mean(X.subset0))

## [1] 0.9811496

(beta1.hat <- mean(Y.subset1) - mean(Y.subset0) - beta2.hat * (mean(X.subset1) - mean(X.subset0)))

## [1] 1.619138
```

The estimate $\hat{\beta}_1 = 1.619$ in this model is positive, thus suggesting that older children tends to forget less than younger children.

(c) This phenomenon (the reversal of the sign for the estimate $\hat{\beta}_1$ between the two model) is known as *Lords paradox* and is first presented in the paper F. Lord "A paradox in the interpretation of group comparisons", Psychology Bulletin, Vol. 68, 1967. The explanation for this phenomenon is that while it may appear that the two scientists are trying to answer the same question (the question of whether older children remembers longer than younger children or not), they are in fact addressing different questions. The first scientist is

[4] Note the similarity between the expression for $\hat{\beta}_2$ in the current problem and the expression for the least square estimate $\hat{\beta}$ in Problem 2

performing a test of unconditional comparison between the score difference of the two groups. The second scientist, meanwhile, is perfoming a test of average conditional comparison, conditioning on the initial score. That is to say, the second scientist is quantifying, given a group of children with the same initial score, what is the effect of the age of the child on his/her final score. The following histograms show that in general, younger children has much lower initial score then older children, and since $\hat{\beta}_2 = 0.121$ in the model for the second scientist, it is not unexpected that the two scientists reach different conclusion. For more on Lord's paradox, see D. Hand paper "Deconstructing statistical questions", Journal of the Royal Statistical Society, 1994 and H. Wainer and L. Brown book chapter "Three statistical paradoxes in the interpretation of group differences", Handbook of Statistics, Vol. 26, 2007. The dataset for this problem is from D. Wright and K. London book "Modern regression techniques using R: a practical guide", Sage Publishers, 2009.