Perspectives on computational analysis - PS2

Maria Adelaida Martinez Cabrera

10/17/2018

Question 1

Imputing age and gender (3 points). You have a dataset called BestIncome.txt that has 10,000 observations on four variables: labor income,capital income, height, weight. You have another dataset from a government survey called SurveyIncome.txt that has 1,000 observations on four variables: total income, weight, age, and gender. You want to use the BestIncome.txt data, but you need age and gender variables.

a)Propose a strategy for imputing age and gender variables into the BestIncome.txt data by using information from the SurveyIncome.txt data. Describe your proposed method, including equations.

One possible strategy to impute age and gender from SurveyIncome.txt to BestIncome.txt is using a linear model. In this case, the variables that we can use in the model are weight and total income that are the ones that we have in both data sets. We define total income as the sum of capital and labor income in the BestIncome data set.

More specifically, the model that we will run for age is:

$age_i = \beta_0 + \beta_1 tot\_income_i + \beta_2 weight_i + \epsilon$

and we are going to predict the age using the estimated coefficients on the BestIncome data:

$\hat{age}_i = \hat{\beta}_0 + \hat{\beta}_1 tot_i ncome_i + \beta_2 \hat{weight}_i$

The model for gender is:

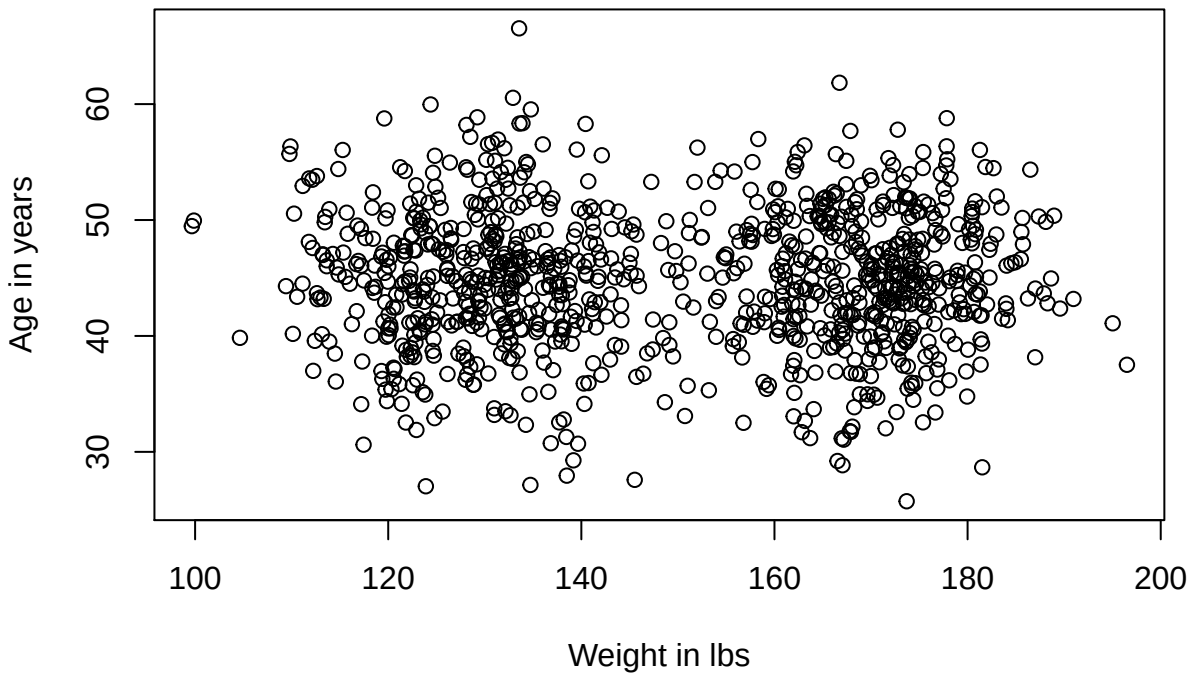
$female_i = \alpha_0 + \alpha_1 tot\_income_i + \alpha_2 weight_i + \epsilon$

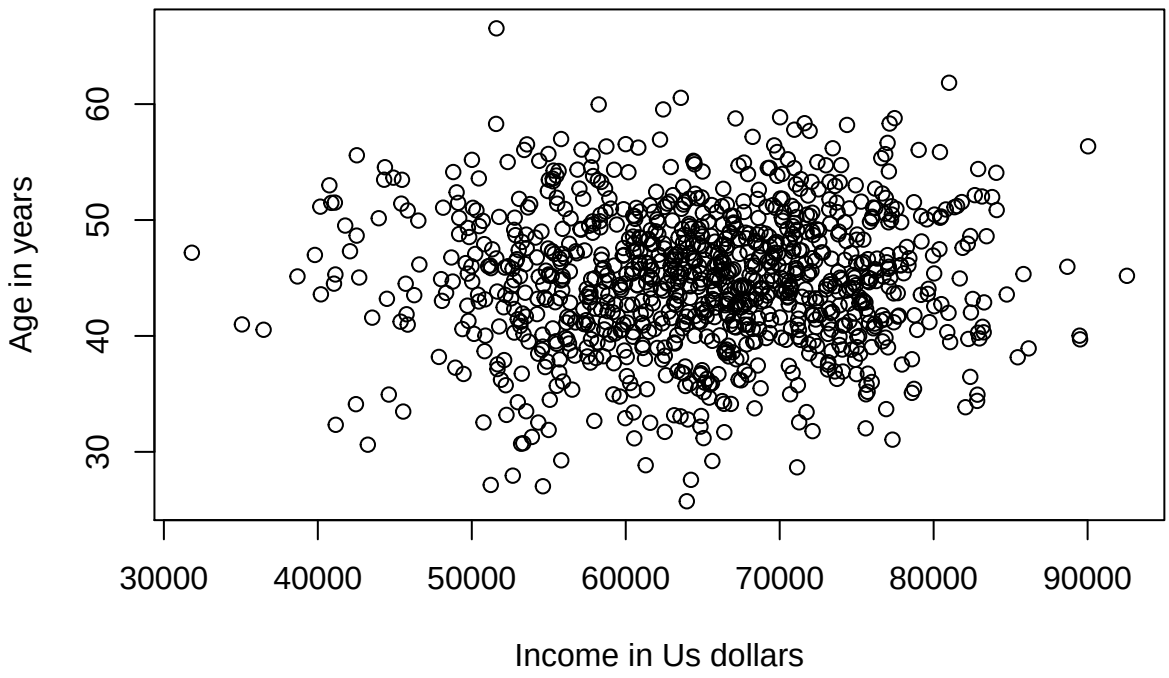
Given that gender is a binary variable, the second equation is going to be a linear probability model where the predicted values are continuous probabilities of being female given a certain level of total income and weight. To transform this predicted values into a binary variable, we are going to assign female if the predicted probability is higher than 0.5 (male otherwise).

$\mathbf{P}[\hat{female}_i] = \hat{\alpha_0} + \hat{\alpha_1} tot\_income_i + \alpha_2 \hat{weight}_i$

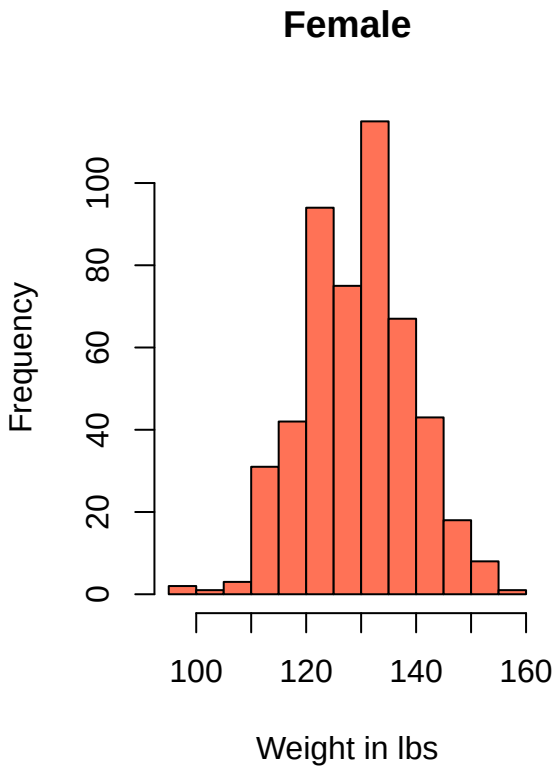
$\hat{female}_i = 1$ if $\mathbf{P}[\hat{female}_i] > 0.5$

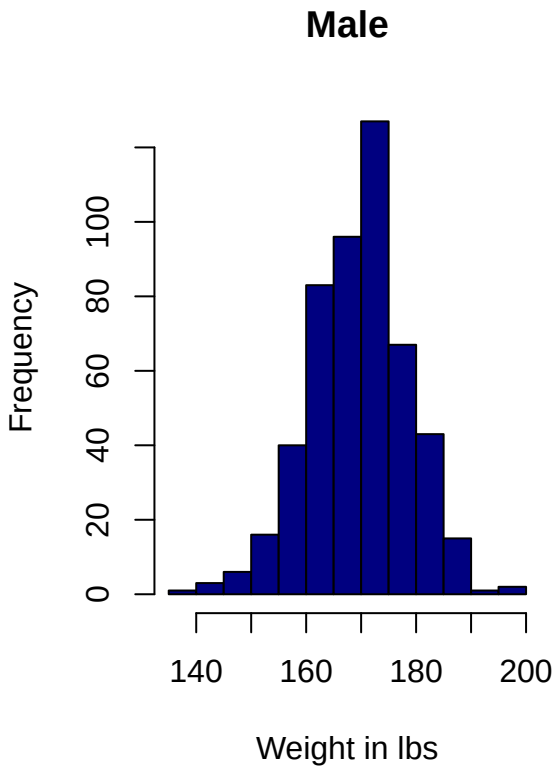
Before we use this strategy is useful to visualize the data and see the basic statistics.First, we did plots of age on income and weight. In this case, we don't see any clear pattern or functional form. Because of that, the model we are going to use is going to be linear in all parameters. Second, we did histograms of weight and income by gender. As we can see in the histograms, there are ranks of weight and income where we can perfectly predict the gender. More specifically: all the individuals with weight lower than 139.6lbs or with income lower than 49,743.3 are women, and individuals with weight higher than 155lbs or income higher than 88,686.3 are men.


Age in years
30
40
50
60
100
120
140
160
180
200
Weight in lbs

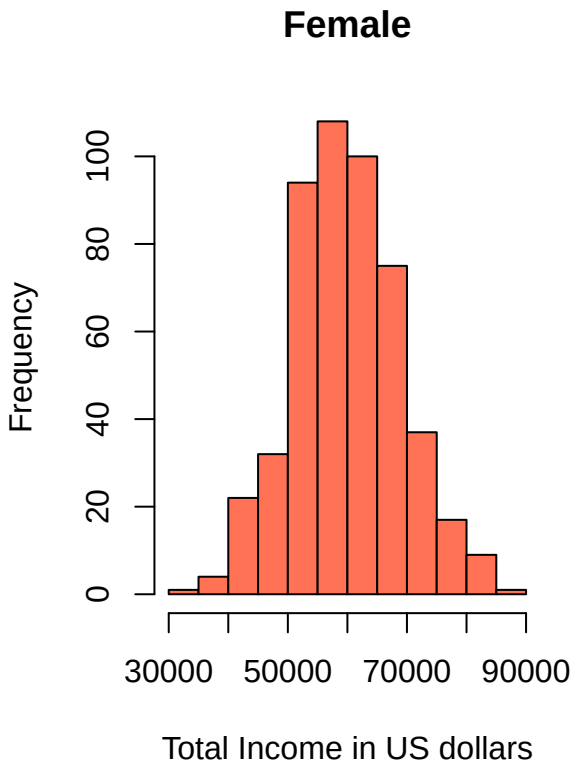


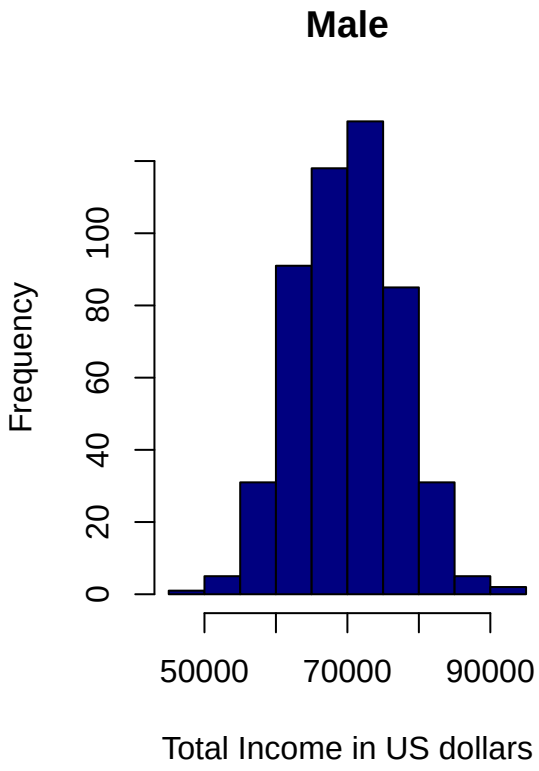

```
##
## Call:
## lm(formula = SurvIncome$age ~ SurvIncome$tot_inc + SurvIncome$weight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.9129  -3.7610   0.0717   4.0397  21.9223
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.421e+01  1.490e+00  29.666   <2e-16 ***
## SurvIncome$tot_inc  2.520e-05  2.263e-05   1.114    0.266
## SurvIncome$weight  -6.722e-03  9.803e-03  -0.686    0.493
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.941 on 997 degrees of freedom
## Multiple R-squared:  0.001267,   Adjusted R-squared:  -0.0007361
## F-statistic: 0.6326 on 2 and 997 DF,  p-value: 0.5314
```






```
##       mean       min      max
## 0 169.5635 139.60751 196.5033
## 1 129.5209  99.66247 155.0075
```






```
##       mean      min      max
## 0 69864.05 49743.27 92556.14
## 1 59878.37 31816.28 88686.26
```

```
##
## Call:
## lm(formula = SurvIncome$female ~ SurvIncome$tot_inc + SurvIncome$weight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70371 -0.13714 -0.00253  0.13815  0.59659
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.761e+00  5.110e-02  73.600  < 2e-16 ***
## SurvIncome$tot_inc -5.250e-06  7.760e-07  -6.765 2.28e-11 ***
## SurvIncome$weight  -1.953e-02  3.362e-04 -58.098  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2037 on 997 degrees of freedom
## Multiple R-squared:  0.8345, Adjusted R-squared:  0.8341
## F-statistic:  2513 on 2 and 997 DF,  p-value: < 2.2e-16
```

b)Using your proposed method from part (a), impute the variables age and gender into the BestIncome.txt data.

Code

c) Report the mean, standard deviation, minimum, maximum and number of observations for your imputed age and gender variables.

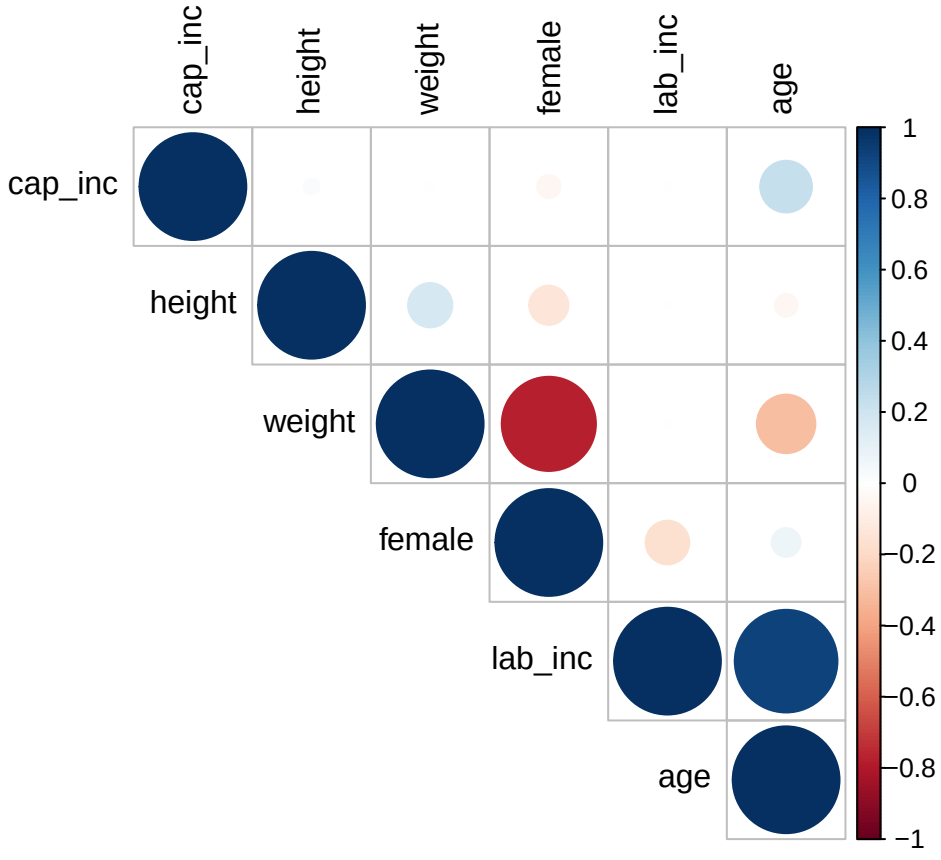
```
##      mean_age lenght_age  min_age  max_age  sd_age
## [1,] 44.89083      10000 43.97649 45.70382 0.21915
```

```
##      mean_female lenght_female min_female max_female sd_female
## [1,]      0.4616         10000          0          1 0.4985482
```

d)Report the correlation matrix for the now six variables: labor income, capital income cap inci, height, weight, age, and gender in the BestIncome.txt data.

```
##         lab_inc cap_inc  height  weight  female     age
## lab_inc  1.0000  0.0053  0.0028  0.0045 -0.1670  0.9241
## cap_inc  0.0053  1.0000  0.0216  0.0063 -0.0471  0.2342
## height   0.0028  0.0216  1.0000  0.1721 -0.1348 -0.0451
## weight   0.0045  0.0063  0.1721  1.0000 -0.7774 -0.3003
## female  -0.1670 -0.0471 -0.1348 -0.7774  1.0000  0.0726
## age      0.9241  0.2342 -0.0451 -0.3003  0.0726  1.0000
```

```
## corrplot 0.84 loaded
```



Question 2

Stationarity and data drift (4 points). Suppose you are interested in question that Salganik (2018) brings up in Chapter 2, namely, "Is higher intelligence associated with higher income?" Suppose that you wanted to test the hypothesis that higher intelligence is associated with higher income using two of the variables in the dataset IncomeIntel.txt. This dataset consists of 1,000 observations of university students who applied to graduate school in the United States over the time period 2001 to 2013. The dataset contains three variables on each observation: year of graduation, GRE quantitative score, and income 4 years after graduation. It is worth noting that the GRE quantitative scoring scale changed in 2011. You want to perform a simple linear regression of the following form to test this hypothesis,

$$salaryp4_i = \beta_0 + \beta_1 gre\_qnt_i + \varepsilon_i$$

where β_0 and β_1 are regression coefficients and ε_i is an error term that is assumed to be normally distributed.

a)Estimate the coefficients in the regression above by ordinary least squares without making any changes to the data. Report your estimated coeffi- cients and standard errors on those coefficients.

```
##
## Call:
## lm(formula = IncomeIntel$salary ~ IncomeIntel$gre_qnt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -28761  -7049   -293   6549  37666
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          89541.293    878.764  101.89   <2e-16 ***
## IncomeIntel$gre_qnt    -25.763      1.365  -18.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10460 on 998 degrees of freedom
## Multiple R-squared:  0.2631, Adjusted R-squared:  0.2623
## F-statistic: 356.3 on 1 and 998 DF,  p-value: < 2.2e-16
```

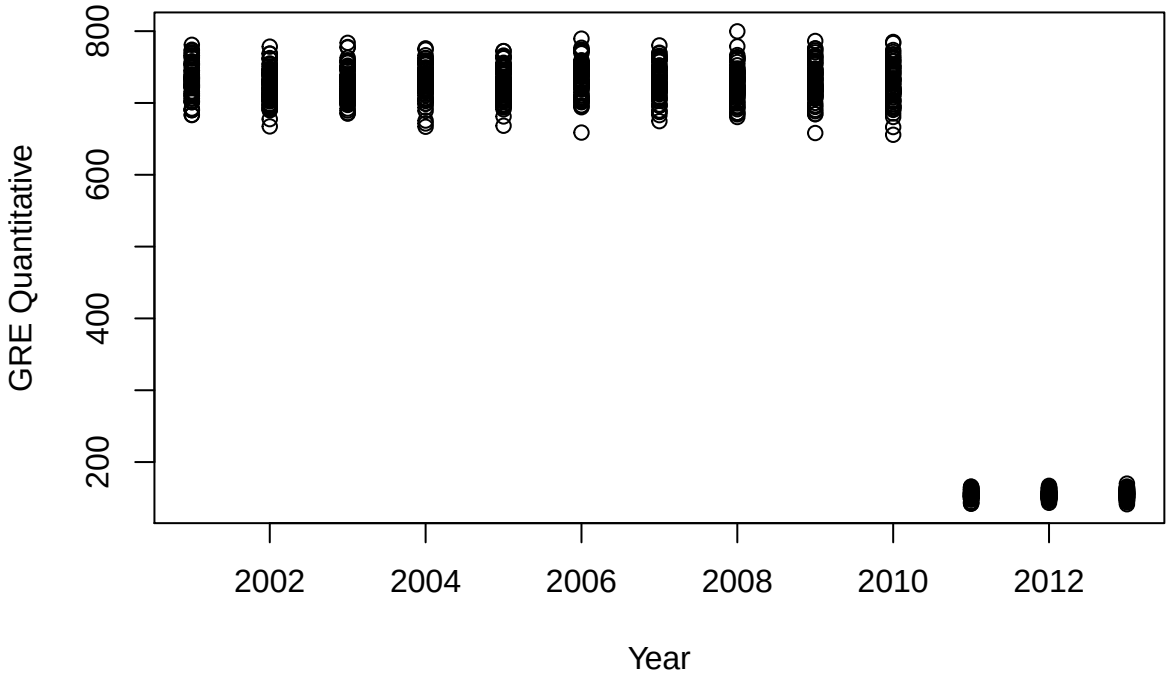
b)Create a scatter plot of GRE quantitative score on the y-axis and graduation year on the x-axis. Do any problems jump out in this variable and your ability to use it in testing your hypothesis? Propose and implement a solution for using this variable in your regression.

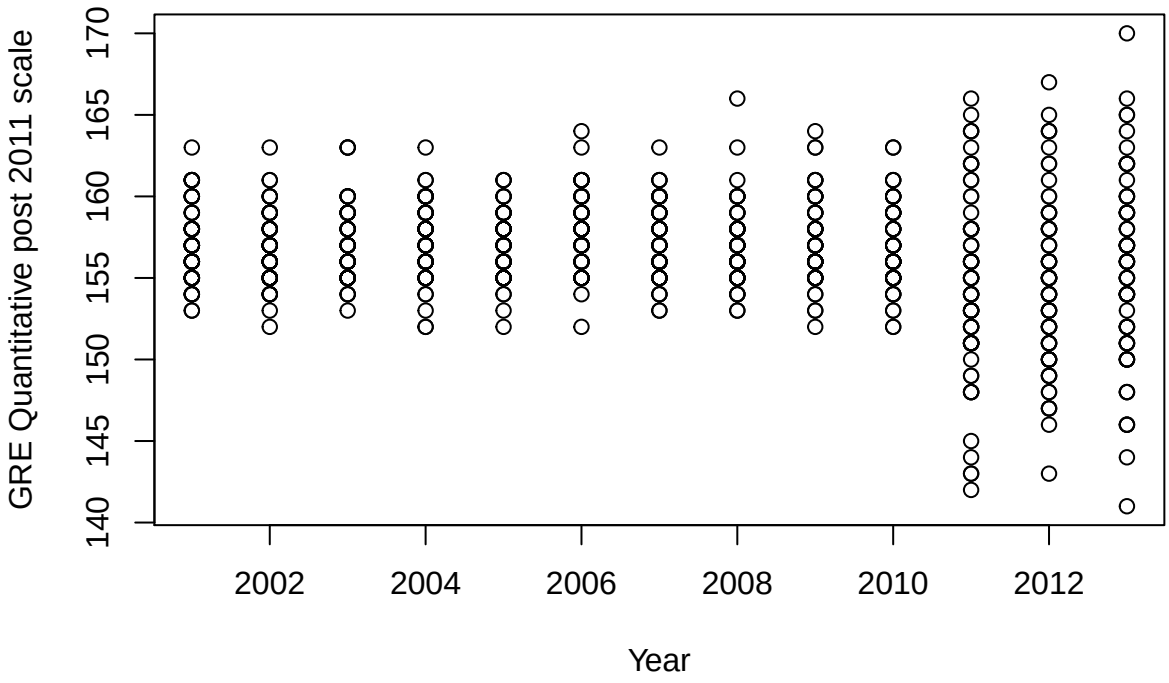
As we can see in the plot of GRE on graduation year, the GRE scored scale change in 2011. To fix this, we download the table that converts the GRE score from pre-2011 to post-2011 scale (GreConv.csv) and merges it to IncomeIntel.txt. The steps we followed were these:

1) Given that the GRE scores are in continuous and with decimals in IncomeIntel.txt, and in multiples of 10 in GreConv.csv, we round the gre_qnt values of IncomeIntel.txt to the closest multiple of 10.

2) Merge both data sets using this variable in IncomeInte.txt and the old_gre variable in the GreConv.

3) With both data sets merged through the pre-2011 scores, we create a new variable with post-2011 scores for everyone. This variable is equal to the merged values of post-2011 gre scores for graduation years previous to 2011 and the value of the original variable without decimal points for the years after 2011. As we can see in the new plot, the problem was solved because all years are now on the same scale. However rounding the scores we clearly lost variance.


GRE Quantitative
800
600
400
200
2002
2004
2006
2008
2010
2012
Year

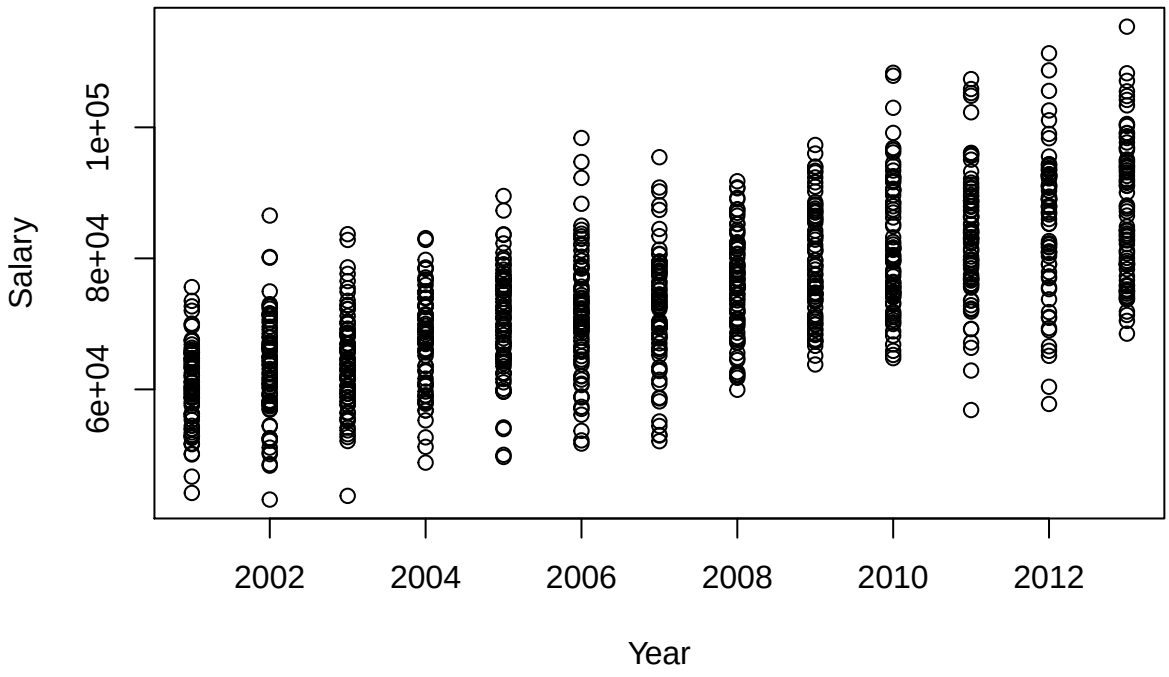


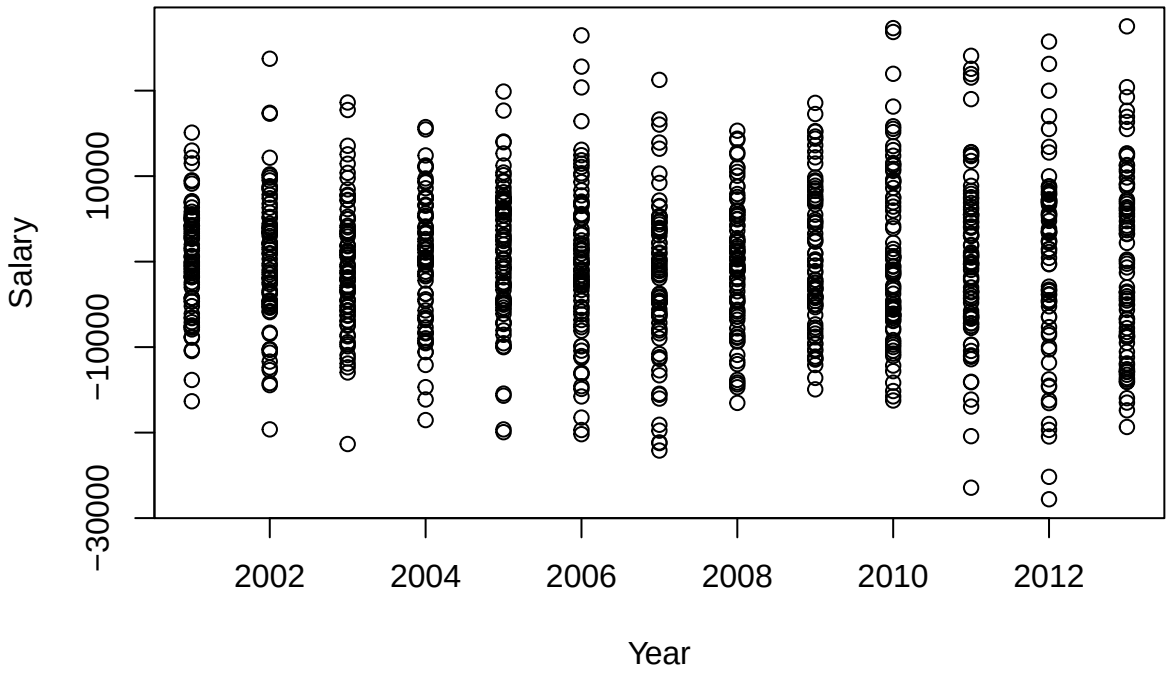

```
##
## Call:
## lm(formula = IncomeIntel$salary ~ IncomeIntel$correct_gre)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -31878  -8285   -599   7369  36828
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              164202.6    18315.5   8.965  < 2e-16 ***
## IncomeIntel$correct_gre   -575.1      117.0  -4.917 1.03e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12030 on 998 degrees of freedom
## Multiple R-squared:  0.02365,    Adjusted R-squared:  0.02267
## F-statistic: 24.17 on 1 and 998 DF,  p-value: 1.029e-06
```

c)Create a scatter plot of income 4 years after graduation on the y-axis and graduation year on the x-axis. Do any problems jump out in this variable and your ability to use it in testing your hypothesis? Propose and implement a solution for using this variable in your regression.

Is clear from the scatter plot that the salary has an increasing tendency. The basic solution for this problem is to detrend the salary variable. One way is to add to the ols regresion the year variable or, as is done in this problem set detrend the variable and use it in the model. To detrend the variable we run an OLS of salary on the year variable, the residuals of these regression are the detrend salaries. Bassically what we are

doing here is to purge the data from any time trend.






```
##
## Call:
## lm(formula = IncomeIntel$det_salary ~ IncomeIntel$gre_qnt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27879.8  -5714.1    -29.9   5597.4  27433.6
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         114.3979   731.4394   0.156    0.876
## IncomeIntel$gre_qnt  -0.1918     1.1361  -0.169    0.866
##
## Residual standard error: 8703 on 998 degrees of freedom
## Multiple R-squared:  2.855e-05,  Adjusted R-squared:  -0.0009734
## F-statistic: 0.0285 on 1 and 998 DF,  p-value: 0.866
```

d)Using the changes you proposed in parts (b) and (c), re-estimate the regression coefficients with your updated salary and gre variables.Report your new estimated coefficients and standard errors on those coeffcients. How do these coefficients differ from those in part (a)? Interpret why your changes from parts (b) and (c) resulted in those changes in coefficient values? What does this suggest about the answer to the question?

After we correct by the system drift and the stationarity of the data the coefficient on GRE loses its significance. The change in the scale and the positive trend of the salaries were driving the significant results in a). Both where confounding factors of the data that where showing a negative relationship between intelligence, measured by the GRE quantitative score, on future salary. This new result suggests that there

is no relationship between the GRE quantitative score and the salary. This doesn't mean that there is no relationship between intelligence and income, we might think that in this case, it might be that the GRE quantitative score is not measuring intelligence as a whole, but just a particular part of it.

```
##
## Call:
## lm(formula = IncomeIntel$det_salary ~ IncomeIntel$correct_gre)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27971.2  -5763.1     39.9   5703.2  27205.3
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              10872.71   13240.16   0.821    0.412
## IncomeIntel$correct_gre    -69.46      84.56  -0.821    0.412
##
## Residual standard error: 8700 on 998 degrees of freedom
## Multiple R-squared:  0.0006755,  Adjusted R-squared:  -0.0003258
## F-statistic: 0.6746 on 1 and 998 DF,  p-value: 0.4116
```

Question 3: Assessment of Kossinets and Watts (2009)

This paper studies the origins of homophily in a particular U.S. university using e-mail data along with individual characteristics. More specifically, they ask, what are the relative roles of similarity and structural proximity on new tie formation? In other words, what is the relative change in the probability of a new tie in the network, corresponding to a unit change in similarity; and in what proportion the proximity of the individuals explains this change. The authors measured similarity regarding gender, age, status, field and year, as an aggregate measure that counts the number of matches that two individuals have over these attributes. Similarly, they measured structural proximity in network distance and number of shared classes.

To answer this question, the authors used three different data sources: 1) The logs of e-mail interactions within the university over one academic year, 2) a database of individual attributes 3) records of course registration recorded separately by semester. As is described in page 410, the authors categorized the variables in four groups: personal characteristics, organizational affiliations, course-related variables, and email related variables. The descriptions and definitions of each of these variables are in Appendix A.

The period for which the data spans is one academic year composed by fall and spring semesters. The total number of observations used this paper is 30,396 out of 43,553 email users in the university network. The individuals were selected into the sample if they were active e-mail users during both semesters and if they exchanged e-mails with others during the academic year. For those 30,396 individuals, the authors collected 7,156,162 exchanged messages during 270 days of observation. The distribution of these observations between the different groups in the university is as follows: 21% undergraduate students, 27% graduate and professional students, 13% faculty members, 13.4% administrators, and staff, and 25% affiliates.

Throughout the data cleaning process, the authors drop 13,157 individuals that where part of the entire sample of users, this represents approximately 30% of this sample. What this means is that they are doing all the analysis with only 70% percent of the users. Additionally, given that this sample is retrained only to e-mail accounts on the central university server this could be even less. Dropping 30% of the total number of users might be in part the reason why they find that tie formation is a rare event on their network. It diminishes their ability to answer the research question because they have less variation in the specific event of study that is the formation of new ties. Along these lines the authors don't explain well why they are only using one academic year, it is possible to find more new ties if they take into account more than one academic year, and relax the conditions for being an active e-mail user.

The underlying theoretical construct on the paper are the social relations between individuals. However, is difficult to think about e-mail logs as very personal social interactions between individuals. Daily we send and receive dozens of emails from people we don't know, and we can think about it as one of the less personal types of communication channel these days. To address this weakness of the data the authors try to select the sample of emails that, on their characteristics, look more personal than others. The authors choose the sample of emails with only one recipient and show how the results are robust using e-mails with up to five recipients.

Nonetheless, the authors acknowledge that most of the interactions between individuals with a close relationship happen to be through different channels. A clear example is shown in the results where on average the undergraduate students had fewer contacts than the faculty. This doesn't mean that they have less social relations, it shows how the email measure of social relations is not accurate for that specific population. As the author states, this pattern might be explained by the popularity of other communication channels among undergraduates.