Perspectives on computational analysis HW2

Angela Zorro Medina

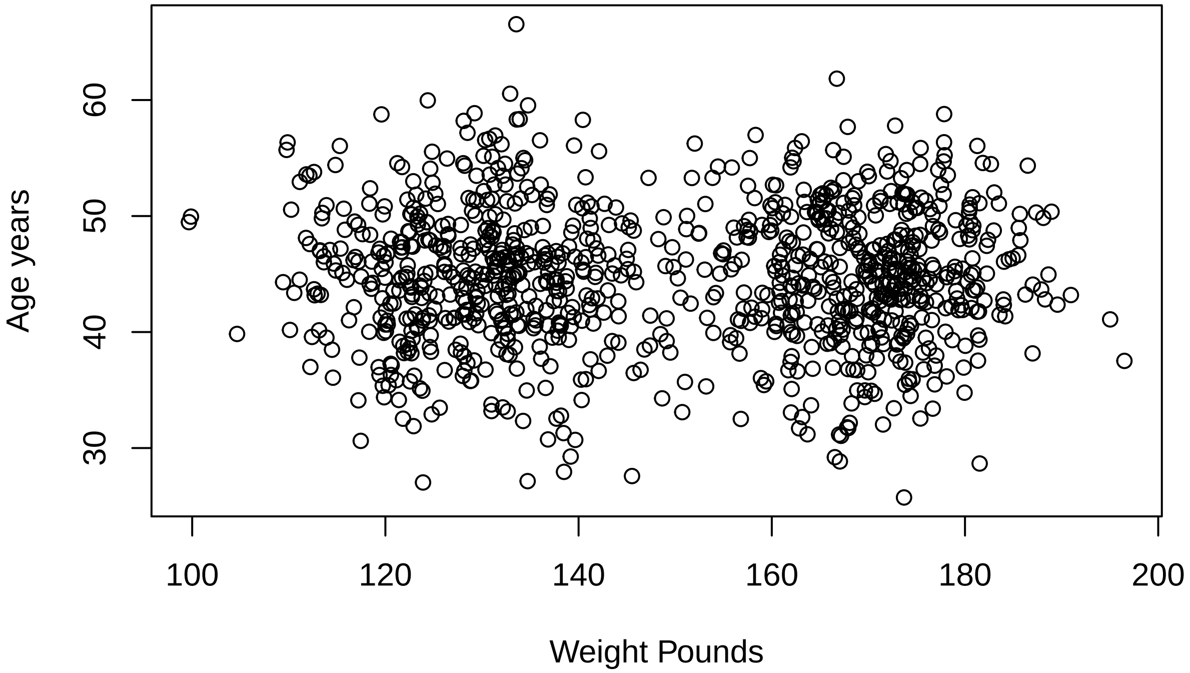
10/17/2018

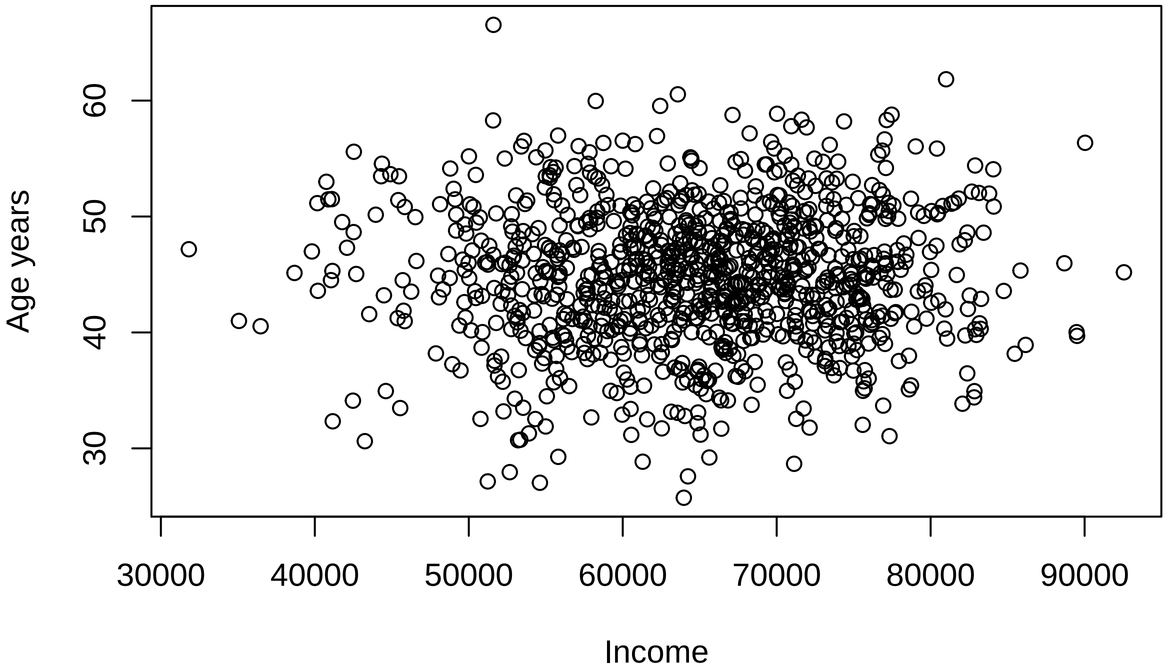
Question 1

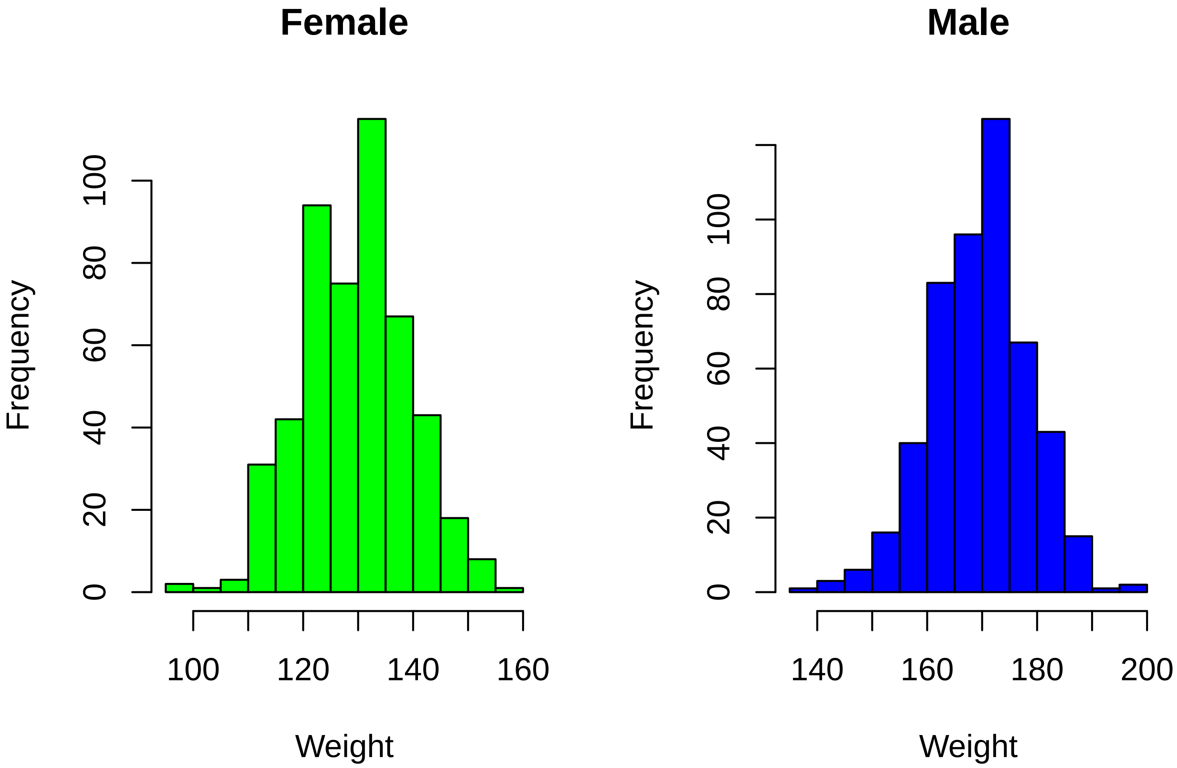
Imputing age and gender (3 points). You have a dataset called BestIncome.txt that has 10,000 observations on four variables: labor income (lab inci, dollars), capital income (cap inci, dollars), height (hgti, inches), weight (wgti, lbs.). You have another dataset from a government survey called SurveyIncome.txt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

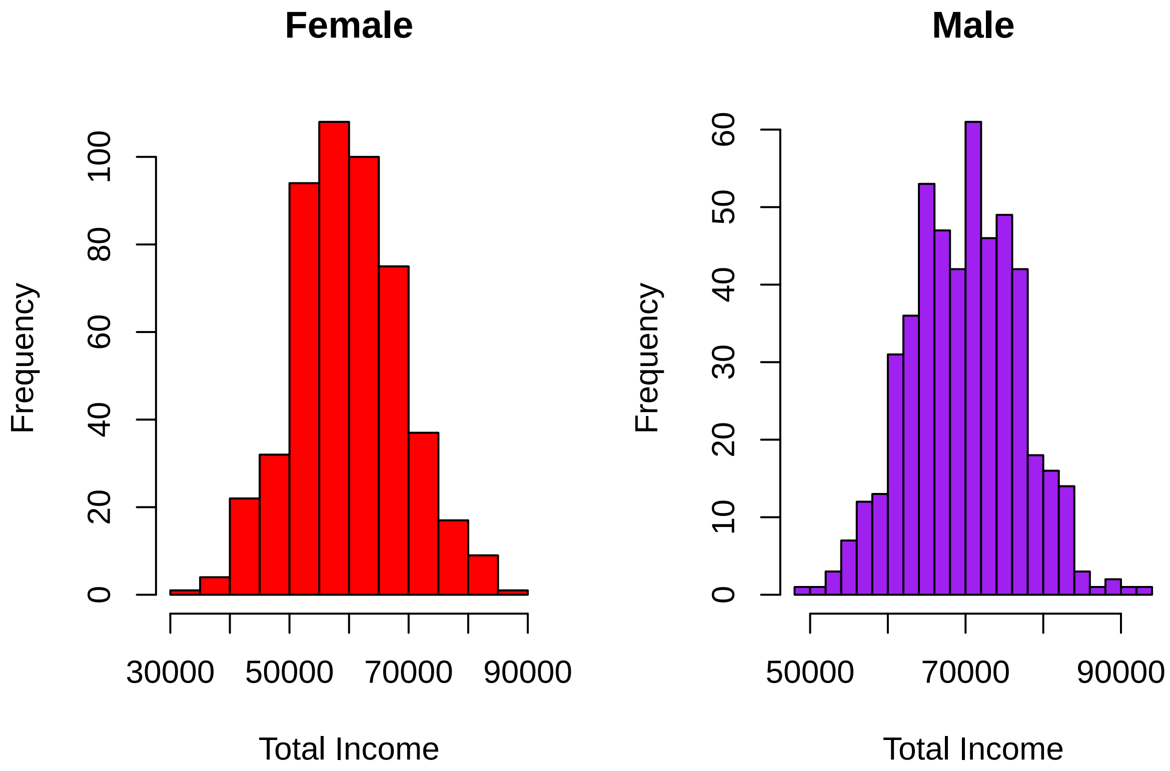
a)Propose a strategy for imputing age(agei) and gender (femalei) variables into the BestIncome.txt data by using information from the SurveyIncome.txt data. Describe your proposed method, including equations.

Our first step consisted in understanding our databases, the variables we had and how they were distrubuted To do this we took into account the main descriptives statistics of these variables and the distribution of our variables. We did plots of age on income and weight, and histograms of female and wieght and income. In the first plot we were not able to identify a clear pattern. In oursecond set of graphs, we identified weight ranges between it is possible to identify the sex of the person we are observing. For example, if a person weights moren than 155lbs then it has to be a man, or is person weights less than 139.6 lbs she has to be a woman. A similar pattern was identified in the histogram of female and income. While an observation with income higher than 88,700 has to be a man and observation with income lower than 49,700 has to be a woman.




Age years
30
40
50
60
30000
40000
50000
60000
70000
80000
90000
Income


Female
Frequency
0
20
40
60
80
100
100
120
140
160
Weight
Male
Frequency
0
20
40
60
80
100
140
160
180
200
Weight



```
##        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
```

Before choosing our functional form, we did some tests with other functional forms based on the inverse U-shape that is usually observed between income and age, however the R-square was really low at those other functional forms we evaluate. Therefore, we estimate the following linear model:

$$age_i = \beta_0 + \beta_1 TIncome_i + \beta_2 Weight_i + \epsilon_i$$

To input age, after estimating the model presented above, we will use the coefficients to predict the age based on the total income variable created at BestIncome, and the weight information we have in that dataset.

After visualizing our female variable, we propose input each of the variables in the following way:

$$\mathbf{P}[female_i] = \beta_0 + \beta_1 TIncome_i + \beta_2 Weight_i + \epsilon_i$$

After estimating the model presented using the dataset SurvIncome. We will use the coefficients $beta_0$,β_1 and β_2 to calculate the probability that a specific observation in the dataset BestIncome is female. After doing that, if the probability is higher than 0.5, we will assume that that observation is a woman. Moreover, if the probability is lower or equal to 0.5, we will assume that observation belongs to a man.

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

```
##
## Call:
## lm(formula = SurvIncome$Age ~ SurvIncome$TIncome + SurvIncome$Weight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.9129  -3.7610   0.0717   4.0397  21.9223
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.421e+01  1.490e+00  29.666   <2e-16 ***
## SurvIncome$TIncome  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
```

```
##
## Call:
## lm(formula = SurvIncome$Female ~ SurvIncome$TIncome + SurvIncome$Weight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70371 -0.13714 -0.00253  0.13815  0.59659
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.761e+00  5.110e-02  73.600  < 2e-16 ***
## SurvIncome$TIncome -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
```

Code

```
##        (Intercept) SurvIncome$TIncome  SurvIncome$Weight
##       3.761142e+00      -5.249560e-06      -1.953025e-02
```

```
##        (Intercept) SurvIncome$TIncome  SurvIncome$Weight
##      44.2096668124       0.0000252022      -0.0067221442
```

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

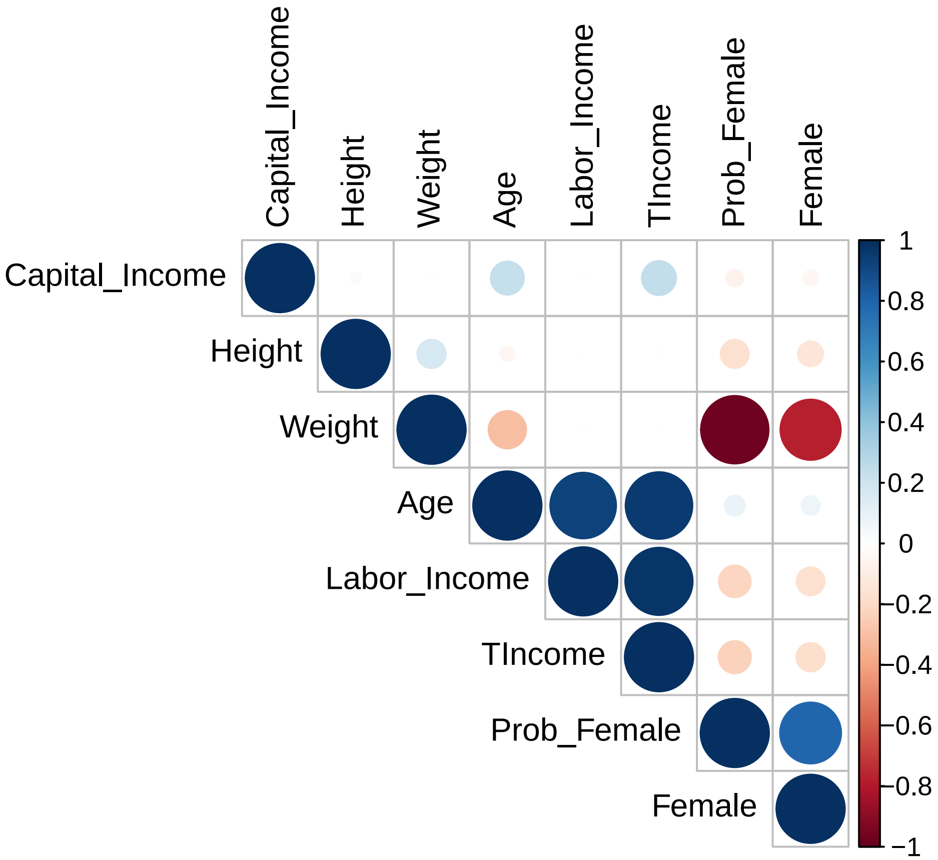
```
##      mean_age lenght_age  min_age  max_age  sd_age
## [1,] 44.89083      10000 43.97649 45.70382 0.21915
```

```
##      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 (lab inci), capital income cap inci, height (hgti), weight (wgti), age (agei), and gender (femalei) |in the BestIncome.txt data.

```
##                Labor_Income Capital_Income  Height  Weight TIncome
## Labor_Income         1.0000         0.0053  0.0028  0.0045  0.9702
## Capital_Income       0.0053         1.0000  0.0216  0.0063  0.2475
## Height               0.0028         0.0216  1.0000  0.1721  0.0079
## Weight               0.0045         0.0063  0.1721  1.0000  0.0059
## TIncome              0.9702         0.2475  0.0079  0.0059  1.0000
## Prob_Female         -0.2158        -0.0601 -0.1695 -0.9760 -0.2236
## Female              -0.1670        -0.0471 -0.1348 -0.7774 -0.1732
## Age                  0.9241         0.2342 -0.0451 -0.3003  0.9521
##                Prob_Female  Female     Age
## Labor_Income       -0.2158 -0.1670  0.9241
## Capital_Income     -0.0601 -0.0471  0.2342
## Height             -0.1695 -0.1348 -0.0451
## Weight             -0.9760 -0.7774 -0.3003
## TIncome            -0.2236 -0.1732  0.9521
## Prob_Female         1.0000  0.7954  0.0852
## Female              0.7954  1.0000  0.0726
## Age                 0.0852  0.0726  1.0000
```

```
## corrplot 0.84 loaded
```



Question 2

Stationarity and data drift (4 points). Suppose you are interested in a 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 (grad yeari), GRE quantitative score (gre qnti), and income 4 years after graduation (salary p4i). It is worth noting that the GRE quantitative scoring scale changed in 2011.[1] You want to perform a simple linear regression of the following form to test this hypothesis,

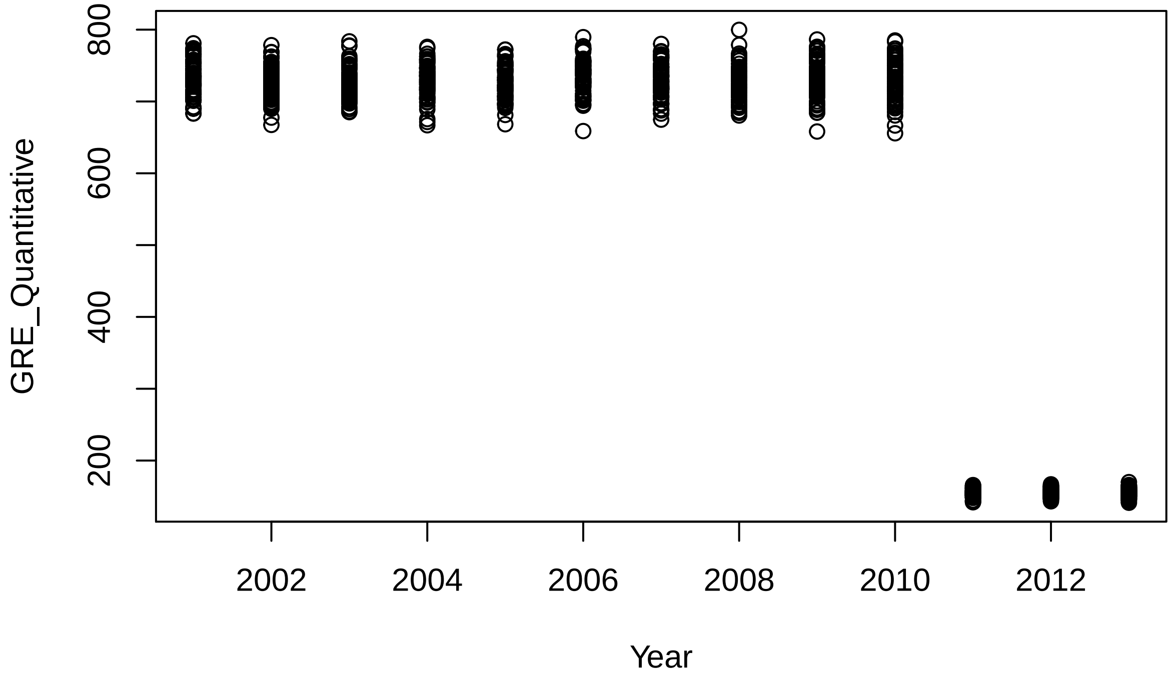
$$salaryp4_i = \beta_0 + \beta_1 greqnt_i + \epsilon_i$$

where β_0 and β_1 are regression coecients 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 coefficients and standard errors on those coefficients.

```
##
## Call:
## lm(formula = IncomeIntel$Salary ~ IncomeIntel$GRE)
##
## 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   -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
```

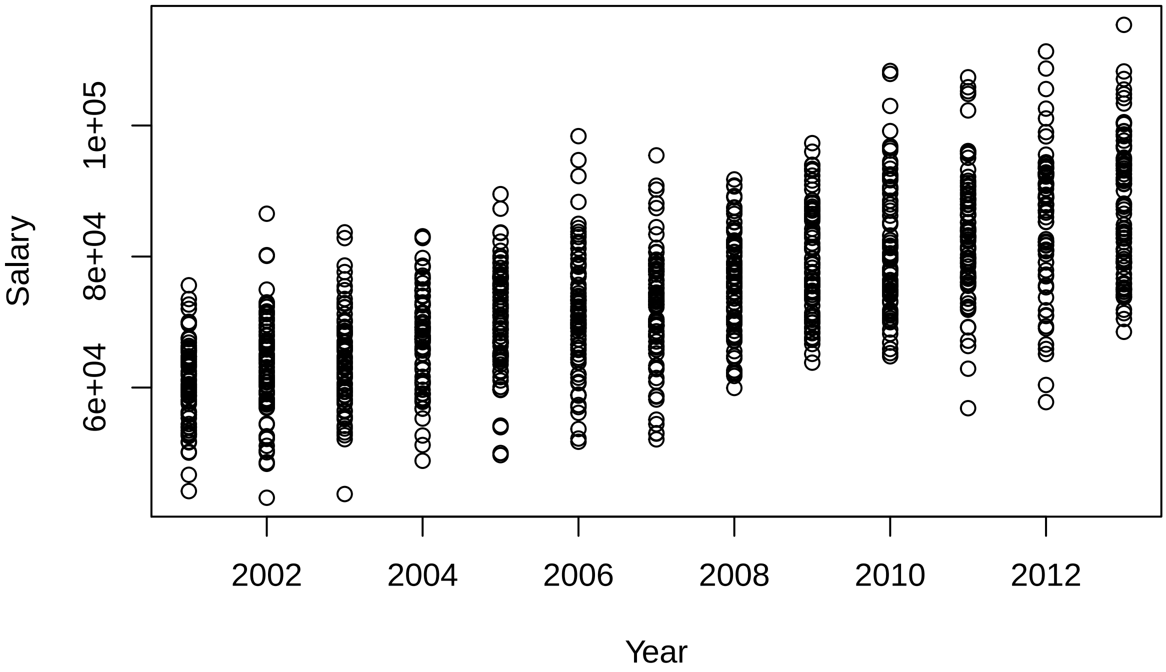
(b) Create a scatter plot of GRE quantitative score (gre qnti) on the y-axis and graduation year (grad yeari) on the x-axis. Do any problems jump out in this variable and your ability to use it in testing your hypothesis? Propose and implement a solution for using this variable in your regression.



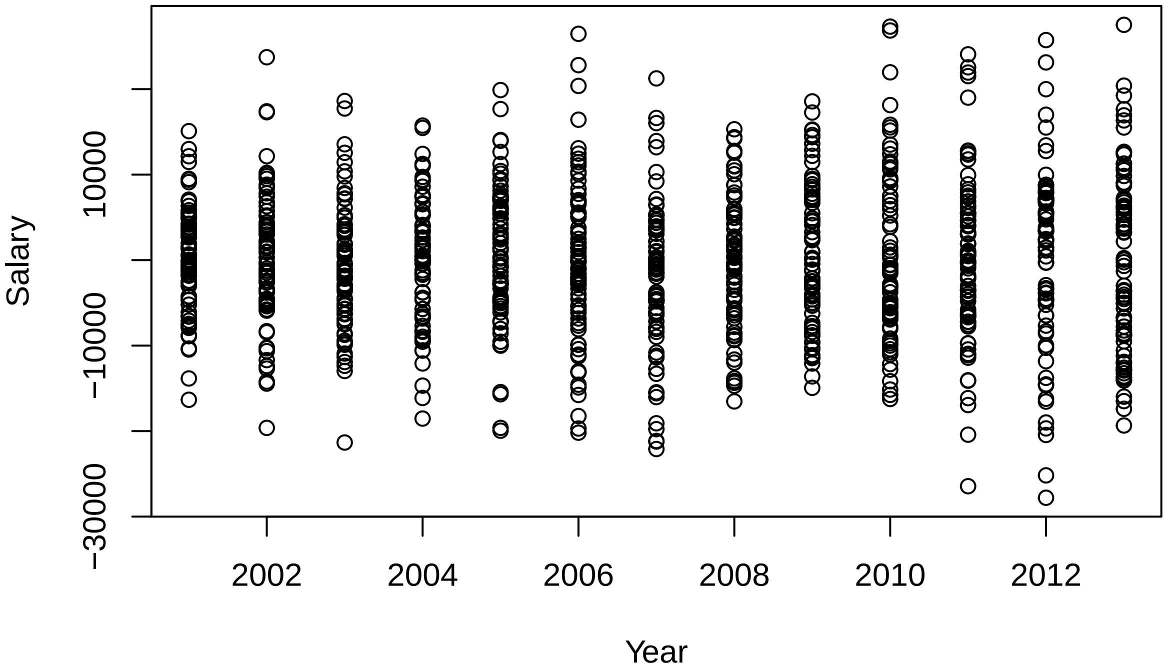
By looking at the plot, we observe a jump after 2010. This jump can be explained by the change in the way the GRE score was measured. To have all GRE scores under the same scale we used the conversion table and rounded the values of the GRE to the closest 10th number.For example: 652->650.

```
##
## Call:
## lm(formula = IncomeIntel$Salary ~ IncomeIntel$GRE_F)
##
## 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$GRE_F   -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 (salary p4i) on the y-axis and graduation year (grad yeari) 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.



By looking at the scatter plot we can identify that salary has an increasing trend. A common strategy used in the relevant literature is to detrend the variable by estimating a regression of salary against year and using the residuals of that regression as the detrend salary.



(d) Using the changes you proposed in parts (b) and (c), re-estimate the regression coeffcients with your updated salary p4i and gre qnti variables. Report your new estimated coeffcients and standard errors on those coefficients. How do these coeffcients differ from those in part (a)? Interpret why your changes from parts (b) and (c) resulted in those changes in co-effcient values? What does this suggest about the answer to the question (evidence for or against your hypothesis)?

```
##
## Call:
## lm(formula = IncomeIntel$DSalary ~ IncomeIntel$GRE_F)
##
## 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$GRE_F   -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
```

In the estimations obtained in part (a) we found that GRE was significant, which means that the evidence suggested that higher quantitative GRE scores caused lower salaries. However, after correcting for the change in the way the GRE was measured and detrending the salary we were able to indentify that our first results were biased. This final regression suggest that there is not empirical evidence supporting the fact that higher quantitative scores have an impact on income.If the GRE quantitative scores are actually capturing the intelligence of a person is different question. Therefore, these results are limited only to the relation between GRE quantitative scores and salary and cannot be extendend to the effect of intelligence in salary.

3. Assessment of Kossinets and Watts (2009) (3 points). Read the paper, Kossinets andWatts (2009). Write a one-to-two page response to the paper that answers the following questions. Make sure that your response is a single owing composition that follows the rules of spelling, grammar, and good writing.

During the last decades, the sociological research on social networks has been studying why individuals tend to associate with similar individuals. To explain this phenomenon, social scientists have developed two theoretical frameworks. On one side, social-psychology studies show that people form ties with similar individuals because they have a preference to do it. Potential explanations to those preferences are related to empirical findings that trust and solidarity are easier to establish within similar people than with dissimilar counterparts (Portes and Sensenbrenner 1993; Mollica, Gary, and Trevino 2003). If we accept the fact that trust and solidarity are easier to find within similar people, then individuals may choose to form ties to similar counterparts because the cost of maintaining ties with those persons is lower than if they do It with dissimilar individuals.

Despite the intuitive explanation offer by the theory presented above, some structuralist researchers have pointed out how the structure of society constrains our decisions to form ties with other persons. For example, Liben-Nowell et. al. (2005) show that geographical processes determined approximately 70% of friendships. This result supports the idea that the reason why individuals form ties with similar counterparts is that the social structure only allows individuals to interact with similar persons. Under this structural framework, individuals preferences are not causing homophily, on the contrary, the division, and organization of the society predetermined the existence of this phenomenon.

The empirical evidence does not resolve the dispute between the structuralistic and individualist origin of homophily. On the contrary, the research conducted in this area only allows us to conclude that a combination of individual and structural processes is causing the tendency to associate with similar individuals. However, the question of the origin itself continues to be unsolved. In the paper "Origins of Homophily in Evolving Social Network," Kossinets and Watts try to resolve this dispute by answering the question of how do individuals selectively make or break some ties over others. By understanding this process, the authors try to separate the factors associated with individual preferences to those factors that are work as social constraints to develop social networks.

To answer the question of the origin of homophily, Kossinets and Watts (2009) collected around of 7,156,162 e-mail messages exchanged during 270 days by 30,396 individuals in a university. They used this data to explain how the structural proximity and individual preferences for similarity interplay in this academic context, and how this interplay reveals the role of structural proximity in homophilic behavior. The authors used three different sources of data. First, they used the logs of e-mail interactions within the university over one academic year. The e-mail data is the center of their research. First, they defined the exchange of e-mails as a social association, which means that when one individual sends an e-mail to another person, she is deciding to connect with that individual. And second, they defined the decision to maintain a relationship as a repeated e-mail exchange process over time. Their second set of data allowed them to see the individual attributes (age, gender, home state, years in school, school, dormitory, academic field) and the status of each individual inside the university (formal status, primary department). They divided the university population into nine subgroups that they called "status division of individuals." The distinction of these groups is fundamental in their definition of structural proximity because they assume that individuals under the same status form the "risk set." Finally, their third dataset gives them information of course registration, which is an important structural constraint to develop social networks.

Using the data described above, Kossinets and Watts (2009) conclude that neither of the individual preferences or structural, theoretical views can adequately account for the striking levels of homophily observed in the population they studied. The authors concluded that both views play an important but partial role, where each reinforces the other. Moreover, they show that their findings are constant whether they consider homogeneity of ties themselves or homogeneity of social contexts.

Despite the effort of Kossinets and Watts (2009) to separate social structure and individual preferences, it is not clear how what they defined as preferences are not shaped by a larger social structure. In the first place, they decided to do the study at a university, even though they use the Ivy League university example as a way to show how the population in a specific setting already shares substantial similarities. For example, students at a university tend to have similar backgrounds regarding parents education, parents income, among others. In the same way, if we look at the faculty of the universities, we can find that they have a similar background: professors at elite U.S. schools probably are similar to the students at those elite universities. The point with this critic is that there is no evidence that what Kossinets and Watts called preferences is capturing exogenous individual preferences and that those preferences are not the result of a structural constraint.

An example of how individual preferences could not be independent of the social structure is language. Constantly we see that individuals prefer to engage in social interactions with other persons that speak their native language. However, individuals cannot decide which is their mother tongue, and that is a structural constraint that is present throughout the whole life of the person. Moreover, if a student, professor, the staff does not speak English as her first language, and there is a small number of people available to talk in that language, the social structure of that particular university is defining the people you are going to have a relationship. In the absence of that social structure the outcome could be different.

Finally, the way Kossinets and Watts (2009) solved the gaps and missing values result problematic in the academic context that they are studying. A common thing, especially at the graduate level, is the absence of students from the university. For example, students apply to different fellowships outside their home university and usually spend one year in another institution or doing fieldwork. Before leaving, the student used to have spaces to interact with a specific group inside the university (i.e., cohort). After the exchange, the student faced new constraints to connect to people. The initial group is not available anymore in most of the cases (some have graduated, or simply the student is one year behind). Because of this disruption, now this student could end up having more connections to younger people not because it is a preference but because they are the only people she can connect now. By treating the missing values of this student and interpolating her information in the way the authors do it, they are ignoring all these factors that affect the process of deciding to connect with some people over others.

References

Kossinets, G., & Watts, D. J. (2009). Origins of homophily in an evolving social network. American Journal of Sociology, 115(2), 405-450. Retrieved from http://dx.doi.org/10.1086/599247

Mollica, K., Gray, B., & Treviño, L. (2003). Racial Homophily and Its Persistence in Newcomers' Social Networks. Organization Science, 14(2), 123-136. Retrieved from http://www.jstor.org/stable/4135155

Portes, A., & Sensenbrenner, J. (1993). Embeddedness and Immigration: Notes on the Social Determinants of Economic Action. American Journal of Sociology, 98(6), 1320-1350. Retrieved from http://www.jstor.org/stable/2781823