MACS30000 Assignment 2

Dr. Richard Evans
Submitted by: Junho Choi
Due October 17, 2018 (11:30 AM)

Problem 1

(a) For the imputation of age and gender variables (age_i and $female_i$, respectively) in the data file `BestIncome.txt`, I propose the following strategy. Firstly, as `SurvIncome.txt` contains variables age_i and $female_i$, I propose to use these two variables as dependent variables (separately, for each model) where the explanatory variables are total income (tot_inc_i) and weight (wgt_i). That is, I will run two regression models, the first being OLS and the second being a logistic regression, in which the equations are as follows:

$$age_i = \beta_0 + \beta_1 tot\_inc_i + \beta_2 wgt_i + \varepsilon_i \tag{1}$$

$$p(female_i = 1) = (1 + e^{-BX})^{-1} \quad \text{where} \quad BX = b_0 + b_1 tot\_inc_i + b_2 wgt_i \tag{2}$$

Next, the said models will be used with the variables capital income (cap_inc_i), labor income (lab_inc_i), and weight (wgt_i) in the `BestIncome.txt` dataset to predict age_i and $female_i$. Note that the said dataset actually does not contain a variable called $totinc_i$. Therefore, the total income variable will too be "imputed" by adding labor and capital incomes (i.e. $tot\_inc_i = lab\_inc_i + cap\_inc_i$), with the assumption that those two are the only components to total income.

I note further that $female_i$ should be a binary variable (with values 0 or 1). Because the logistic regression returns the probability of a certain observation being female (i.e. $female_i = 1$), one would need to set some probability threshold for the predicted probabilities to be classified into binary values. I will describe in Problem 1-(b) of how will be implemented.

(b) For this part, I refer to the Jupyter Notebook file (`A2_junhoc.ipynb`) for coding details (see under the heading "Problem 1-(b)"). I note that I used Scikit-Learn's function `LogisticRegression` and its `.predict()` command to impute the binary values of $female_i$, instead of setting a probability threshold. This is because the said function and command automatically select the threshold from the (trained) model.[1]

(c) For this part, I refer to `A2_junhoc.ipynb` once again for coding details (see under the heading "Problem 1-(c)"). In the table below (Table 1), I report the descriptive

[1] Actually, I attempted to calculate the probabilities myself (after executing the logistic regression from `statsmodels.api`'s function `Logit`) using the log-odds and was going to set a probability threshold of 0.5. However, it seems that `math.exp()` cannot calculate some of the larger values of log-odds, which made me opt to using Scikit-Learn instead. This is also documented in `A2_junhoc.ipynb` as well.

statistics of the imputed age_i and $female_i$ in `BestIncome.txt`.

Table 1: Descriptive Statistics of the Imputed Age and Female Variables

	age_i (imputed)	$female_i$ (imputed)
Mean	44.89	0.47
SD	0.22	0.50
Minimum	43.98	0.00
Maximum	45.70	1.00
N	10,000	10,000

NOTE: SD refers to standard deviation and N refers to the number of observations. All values except for N was rounded to the nearest hundredth.

(d) For this part, I refer to `A2_junhoc.ipynb` once again for coding details (see under the heading "Problem 1-(d)"). In the table below (Table 2), I report the correlation matrix for the six variables in `BestIncome.txt` (i.e. lab_inc_i, cap_inc_i, hgt_i, wgt_i, age_i, and $female_i$).

Table 2: Correlation Matrix of the Six Variables in `BestIncome.txt`

	lab_inc_i	cap_inc_i	hgt_i	wgt_i	age_i	$female_i$
lab_inc_i	1.000000	-	-	-	-	-
cap_inc_i	0.005325	1.000000	-	-	-	-
hgt_i	0.002790	0.021572	1.000000	-	-	-
wgt_i	0.004507	0.006299	0.172103	1.000000	-	-
age_i	0.924053	0.234159	−0.045083	−0.300288	1.000000	-
$female_i$	0.677675	0.176901	−0.066972	−0.382659	0.784260	1.000000

NOTE: Each cell contains correlation coefficient between row and column variables. As the matrix is symmetric, redundant correlation coefficients were omitted (and "-"s are to represent this). age_i and $female_i$ are imputed variables.

Problem 2

(a) For this part, I refer to `A2_junhoc.ipynb` once again for coding details (see under the heading "Problem 2-(a)"). Firstly, I execute the OLS regression proposed in the problem without any alterations to the data. The below table (Table 3) reports the regression results, including the coefficients and the standard errors associated with them.

It is strange, however, to observe that there is a strong (statistically significant at $\alpha = 0.01$) negative relationship between GRE quantitative score (gre_qnt_i) and income four years after graduation ($salary\_p4_i$). At least in the field of economics, it is expected that income and one's abilities are positively correlated.[2] Since GRE

[2]For instance, one could refer to a canonical paper by Jacob Mincer (1958)

quantitative scores are meant to serve as a proxy for one's quantitative problem-solving skills, one would expect the two variables' relationship to be at least non-negative. This suggests that there may be something awry with the current data (in `IncomeIntel.txt`).

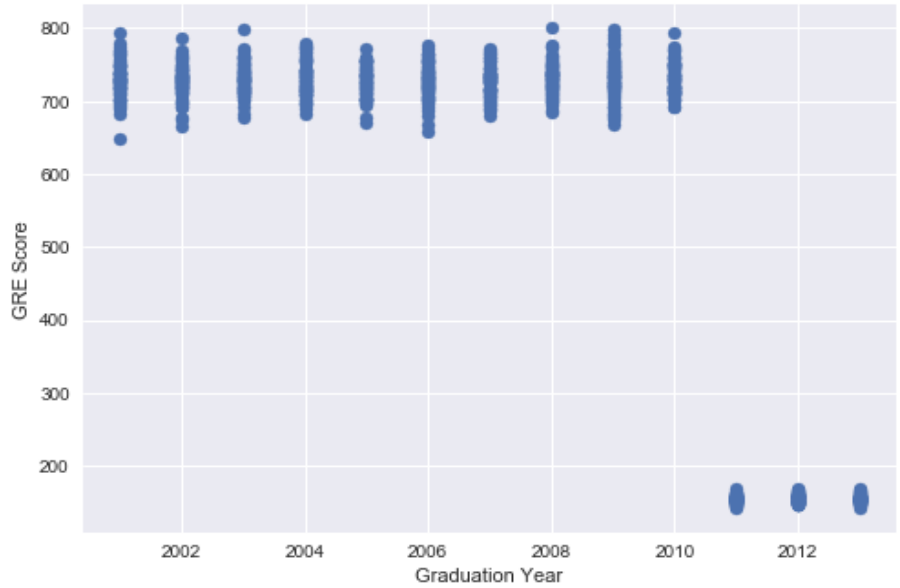
Table 3: Regression Results, gre_qnt_i and $salary\_p4_i$

	DV: $salary\_p4_i$
	(1)
Constant	-25.763^{***}
	(1.37)
gre_qnt_i	$8{,}954^{***}$
	(878.76)
N	1,000

NOTE: DV refers to dependent variable. Standard errors in parentheses.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

(b) Once again, I refer to `A2_junhoc.ipynb` for coding details (see under the heading "Problem 2-(b)"). Let us first view the scatterplot to make note of problems (if there are any). The below figure (Figure 1) is a scatterplot how the GRE quantitative scores change by graduation year.

Figure 1: GRE Scores by Graduation Year

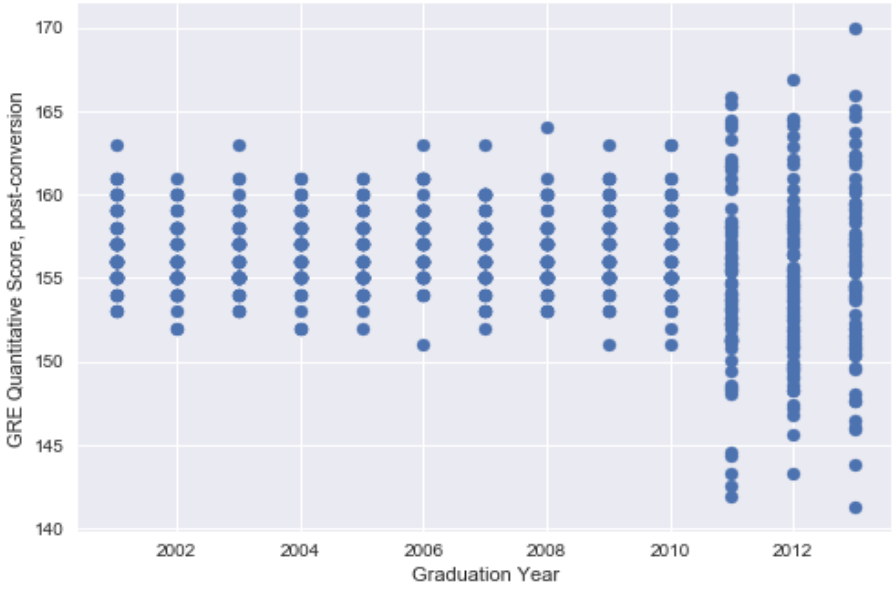


It is clear, from Figure 1, that the pre-2011 scores have not been converted according to the new score guideline. To an outsider who does not know about the changed score guideline, it would be as if people who graduated in 2011 and on just were (by a considerable degree) not good at taking the GRE quantitative section. This,

combined with the fact that $salary\_p4_i$ exhibits an increasing trend (as we will see in Problem 2-(c) below), would show why there is a strong (statistically significant) negative correlation between $salary\_p4_i$ and gre_qnt_i.

Therefore, one would need to convert the older score (i.e. pre-2011 scores) to a more modern scale. I reference the GRE Verbal and Quantitative Reasoning Concordance Tables on how to convert the older-guideline score to a newer-guideline one.[3] After this conversion, the scores are distributed as in the below figure (Figure 2).

Figure 2: GRE Scores, post-conversion, by Graduation Year



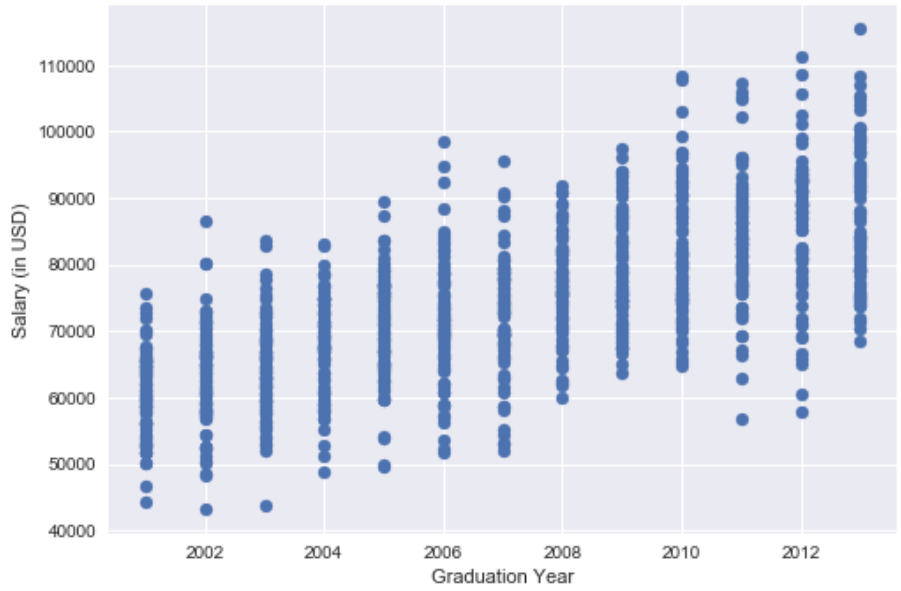
It can be noted, however, that the distribution of scores is more spread-out in the post-2011 scores. This is perhaps because the concordance table is imperfect; however, this is unlikely as it was provided by ETS. I believe that a more probable reason is, because there were alterations to the test in 2011, the test essentially became a different test in comparison to the one used prior to 2011. Therefore, I will consider dividing up the dataset to pre-2011 and in-and-post-2011 subgroups and run separate regressions accordingly.

(c) Again, I refer to `A2_junhoc.ipynb` for coding details (see under the heading "Problem 2-(c)"). Firstly, let us observe the scatterplot and see what may be problematic. The below figure (Figure 3) is a scatterplot showing how salary after 4 years have changed by graduation year.

However, it seems that the wages have some underlying trend, whether it be due to inflation, general growth in the economy, or some combination of two and/or other factors. Therefore, a need to de-trend the variable $salary\_p4_i$ arises. There could be multiple ways of implementing this, one of which is to directly implement a time trend variable in the new regression. Another is to directly calculate the

[3]These tables can be found at `https://www.ets.org/s/gre/pdf/concordance_information.pdf`.

Figure 3: Salary after 4 Years by Graduation Year



average growth rate over the years, and use this to calculate a de-trended version of the variable. I will describe the latter in detail for this question, and regarding the former I will implement it in Problem 2-(d).

Let yearly averages of $salary\_p4_i$ be denoted as m_t for year t. Then, the average of the (average) yearly growth rate (denoted by g) can be calculated as:

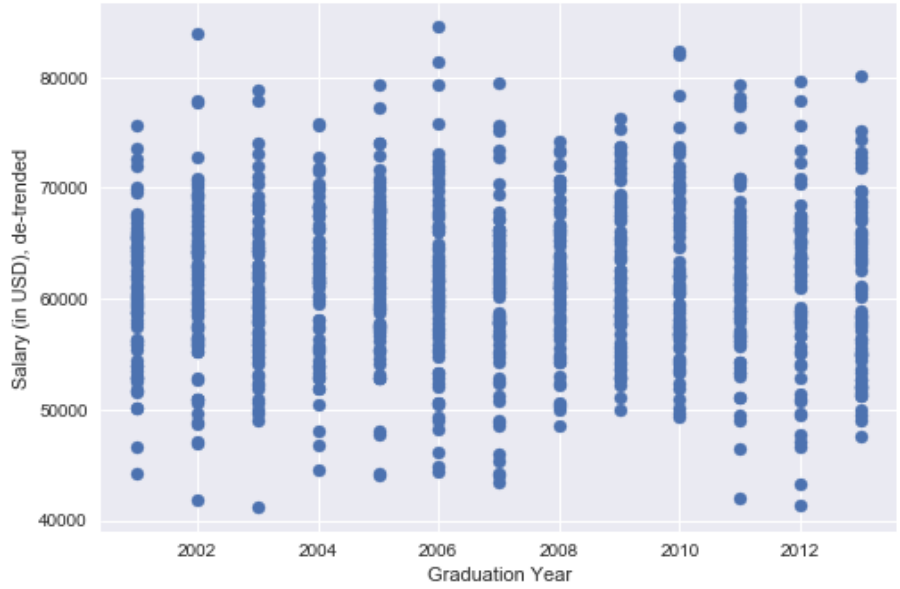
$$g = \left(\frac{1}{2013 - 2002 + 1}\right) \sum_{t=2002}^{2013} \left(\frac{m_t - m_{t-1}}{m_t}\right) \tag{3}$$

If we further denote $salary\_p4_i$ with a time (year) subscript t (i.e. $salary\_p4_{it}$), we can implement de-trending by applying the process. Let the de-trended value of $salary\_p4_{it}$ be denoted as ds_{it}. Then,

$$ds_{it} = \left(\frac{1}{1+g}\right)^{t-2001} salary\_p4_{it} \tag{4}$$

Using this process, a de-trended version of the variable $salary\_p4_i$ can be calculated. By plotting (scatterplot) this against graduation years, one can indeed confirm that the variable no longer exhibits an increasing trend. This is shown in the below figure (Figure 4).

Figure 4: Salary after 4 Years (de-trended) by Graduation Year



(d) Once again, the coding details can be found in `A2_junhoc.ipynb` (under the heading "Problem 2-(d)"). Using the converted variables, I will now produce regression results for the same regression in question. The regression results are written in Table 4's column (1). However, because there was worry that pre- and post-2011 GRE quantitative exams are quite different (for this I refer to Problem 2-(b)), I also split the dataset into pre- and post-2011 data and run the regressions. These results are written in Table 4's columns (pre) and (post).

In addition to this, I have discussed in Problem 2-(c) that a trend in the dependent variable can be accounted for by introducing a time trend variable. That is, we can run the regression below:

$$salary\_p4_i = \beta_0 + \beta_1 gre\_qnt_i + \beta_2 Time_i + \varepsilon_i \tag{5}$$

where $Time_i(t) = t - 2001$ and t is the graduation year (so that $t \in [2001, 2013] \cap \mathbb{N}$). Also, I note that in the above equation $salary\_p4_i$ has not been de-trended, while gre_qnt_i has been converted into a post-2011 score standard. The result for running this regression in the above equation (equation (5)) can be found in column (2) of Table 4.

While different approaches may have been utilized, one common feature that stands out is that the coefficients on gre_qnt_i (for each column) are no longer statistically significant. Despite retaining a negative value, this means that the GRE quantitative scores do not have much correlation with salary after 4 years. Clearly, the change from statistical significance (at $\alpha = 0.01$) to statistical insignificance is due to the de-trending and conversion that have been applied to the data. Prior to this data-cleaning procedure, the salaries were in an increasing trend in time and GRE

quantitative scores "dropped" significantly in 2011 and on. This would be the reason behind statistically significant and negative relationship between the two said variables.

One may further ask, then, why are the coefficients neither *nonnegative* nor *statistically significant*. If one goes back to the theory that income is highly associated with one's abilities, and if GRE quantitative scores are good proxies of one's abilities, then they should be positively correlated with the variable $salary\_p4_i$. One could argue that GRE scores are not a good proxy of one's abilities, but I do not think this is a good argument as many universities utilize the said scores to gauge potential students' abilities.

In light of this, I propose another explanation: that the sample (and the variable $salary\_p4_i$ associated with the sample) is unsuitable for the research question. The research question of this problem is linking intelligence to one's income earnings. However, consider the fact that all observations are *students who applied to graduate schools*; in four years, they are likely to be still in graduate school (for instance, in a PhD program). Even if one is not (due to reasons like attending a shorter program than a PhD program or discontinuing their graduate education), chances are they will have less experience in the job market than their peers. In addition, the very homogeneity (i.e. that they are all graduate school applicants) may mean that their income in four years is somewhat determined. These could be the potential reasons why the proxy for intelligence or ability (GRE quantitative score) is not strongly or positively correlated with salary after four years for this sample.[4]

Table 4: Regression Results, gre_qnt_i and $salary\_p4_i$ (Post-Conversion)

	DV: $salary\_p4_i$			
	(1)	(2)	(Pre)	(Post)
gre_qnt_i (PC)	−57.836	−86.998	−69.188	−63.758
	(71.25)	(87.93)	(122.30)	(93.85)
$Time_i$		2265.486***		
		(74.50)		
Constant	70,450***	74,180***	72,290***	71,180***
	(11,100)	(13,800)	(19,200)	(14,500)
N	1,000	1,000	770	230

NOTE: DV refers to dependent variable, and $salary\_p4_i$ is de-trended in all columns **except for column (2)**. (PC) stands for "post-conversion." Standard errors in parentheses. Columns (Pre) and (Post) divide the entire dataset into pre-2011 data and in-and-post-2011 data, respectively.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

[4]I am still puzzled that the (yearly) salaries of graduate applicants after four years are almost $60,000 on average because this should be much higher than regular stipends (or even regular yearly salaries at many jobs). I will just assume that in this hypothetical situation, this makes sense.

Problem 3

In Kossinets and Watts (2009), the authors explore the topic of how homophily arises; that is, the tendency to associate with those who are like oneself. They, in specific, seek to answer the following research question: *between structural opportunities and personal preferences, which one plays a more important role in generating homophilic tendencies (or do both have significant roles)?* Structural opportunities are tied to induced homophilies, where some environment or situation brings together similar people; on the other hand, personal preferences are tied to choice homophilies, where one's preferences directly lead to "liking the alike." While the two concepts are theoretically well-defined, it is hard to make the distinction in real life as homophily is often the culmination of combining structural opportunities and personal preferences (Kossinets and Watts 2009, pp.407–409).

To disentangle the relationship between the impacts of structural opportunities and those of personal preferences, the authors use a dataset with number of (usable) observations being 7,156,162, with each observation being associated with a single electronic mail. These messages are from 30,396 individuals who are faculty members, staffs, students (undergraduate and graduate), and so forth affiliated with a university and was collected within the span of one academic year (or 270 days to be exact). The variables associated with the dataset are from three different data sources: i) e-mail interaction logs; ii) database with the said individual's characteristics; iii) course registration records. Further information about the variables themselves (such as their definitions and further information) can be found in Appendix A of the paper (Kossinets and Watts 2009, pp.439–442).

In constructing the said 7,156,162 *usable* data, the authors have applied various data cleansing methods to overcome obstacles such as the missingness of data. These procedures are described in Appendix B of the paper in detail (Kossinets and Watts 2009, pp.442–443). Regarding these procedures, however, two potential problems can be identified. First is the problem of merging different datasets from different sources to create a single dataset, which is relatively a minor one for the study at hand. In general, combining different datasets can be dangerous as there is always the danger of distinct observtions being "matched" just because there are similar characteristics shared between the two. Or, if a certain theoretical model is used in merging different datasets, there can be the danger of data extrapolation. In this situation, however, the authors carefully track the individual and group identifiers in the data-merging process. This effort reduces the risk that invalid matchings are made.

In addition to the issue of merging datasets, there is also the issue of imputing missing data. The authors describe that the original, "dirty" dataset is filled with missing values as well as inconsistencies. They attempt to overcome this problem by using additional data. While the dataset reported in the main part of the paper covers the span of one academic year, it seems that the authors actually collected a 2-academic-year worth of data. By comparing the data from two different academic years, if certain information

is missing from one year, it can be supplemented by using the information from another year (given that the latter is not missing). The authors try to minimize the errors in this process by using various methods of *intrapolation*. However, this would not mean that all errors and problems associated with missing data are removed, as there could be no additional data for intrapolation of missing data to occur in the first place. If such were *not* missing at random (for instance, information about an entire department missing), the study could have problematic conclusions. However, it seems from descriptive statistics that such problems were not found (or at most very minimal).

Furthermore, the study may also have problems in its match between theoretical construct and actual data. Because the study wants to see how homophily emerges in social relationships in general, the most ideal data would comprise of various individuals' interactions regardless of types. But this would be nearly impossible as a researcher would have to track those individuals' every movement. Even if this were possible, it would be very costly to continue this data-collecting procedure for an elongated period. Therefore, the authors opt to using e-mails as proxies for social interactions. Yet this may be questionable, as e-mails may not be a dominating method of building social relationships. Also, as the authors suggest, the use of e-mails may be biased depending on group characteristics. The authors attempt to overcome this predicament by addressing that the number of e-mails represent only a fraction of the total, and supplement that other studies have also suggested methods to analyze social network via using e-mail data (Kossinets and Watts 2009, p.413).

In spite of the fact that the study is not completely free of potential weaknesses, the authors seem to have exerted great effort in reducing such by using optimal approaches such as data intrapolation. Even in the case of representing social relationships with e-mails, it can still be thought of as the most realistic (or second-best) approach to understanding how homophily arises in a certain society.