Assignment #2

Sanja Miklin

October 17th, 2018

1. Imputing age and gender.

Before I dive into the data, I look at some simple decriptives to compare the two data sets

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

variable	min	max	mean	sd
cap_inc	1495.19190	19882.32007	9985.79856	2010.123691
hgt	58.17615	72.80228	65.01402	1.999692
lab_inc	22917.60790	90059.89854	57052.92513	8036.544363
total_inc	33651.69181	98996.05376	67038.72370	8294.497996
wgt	114.51070	185.40828	150.00601	9.973001

variable	min	max	mean	sd
age	25.74133	66.53465	44.83932	5.9391855
female	0.00000	1.00000	0.50000	0.5002502
tot_inc	31816.28165	92556.13546	64871.21086	9542.4442143
wgt	99.66247	196.50327	149.54218	22.0288831

The mean of the shared variables, *tot_inc* and *wgt* are fairly similar for both data sets. Notably, SD of *wgt* is much larger in the second data set, which is a bit surprising.

a) Propose a strategy for imputing age (age_i) and gender ($female_i$) variables

To impute age and gender into BestIncome, I would use a regression imputation. I would use Surv.Income data to run regression models that would estimate age and gender using the remaining variables shared by the two data sets, that is total income (labor income + capital income) and weight.

Because female is a dummy variable taking only values 0 and 1, I would use logisic regression.

```
##
## Call:
## lm(formula = age ~ tot_inc + wgt, data = SurvIncome)
##
## Coefficients:
```

```
## (Intercept)      tot_inc          wgt
##  44.2096668    0.0000252   -0.0067221

##
## Call:  glm(formula = female ~ tot_inc + wgt, family = binomial, data = SurvIncome)
##
## Coefficients:
## (Intercept)      tot_inc          wgt
##  76.7929018   -0.0001555   -0.4460483
##
## Degrees of Freedom: 999 Total (i.e. Null);  997 Residual
## Null Deviance:       1386
## Residual Deviance: 72.1  AIC: 78.1
```

My formulas to impute age and gender are therefore the following:

$$age_i = 44.2097 + 0.0000252 \times tot\_inc_i - 0.0067221 \times wgt$$

$$female_i = \frac{\exp(76.7929 - 0.0001555 \times tot\_inc_i - 0.4460 \times wgt)}{1 + \exp(76.7929 - 0.0001555 \times tot\_inc_i - 0.4460 \times wgt)}$$

I however, do note that that the R^2 for the *age* model is really low and the coefficients are not significant. As coefficients are really small, the values produced through this imputation will all be very close to the mean. This is not ideal, but then again, simply using a mean value is a method of imputation as well. Still, I would not want to use this imputated variable in my data analysis.

Aditionally, I have to other options of calculating $female_i$, that is I can either round the value to 0 or 1 as to make it more meaningful (although it seems that practice is dicouraged, as discussed here, or I can do a single Bernoulli draw using the value as p (as suggested here).

b) Impute the variables age (age_i) and gender ($female_i$) into the BestIncome.txt data.

After testing all three approaches to imputing gender I mentioned above I saw they all produced a similar mean. I decided to go with the Bernoulli draw, as it looked fun, and it would produce meaningful values (0 and 1) without resorting to rouding, wich is apaprently an inferior method.

lab_inc	cap_inc	hgt	wgt	age	female
52655.61	9279.510	64.56814	152.9206	44.74252	1
70586.98	9451.017	65.72765	159.5344	45.15425	0
53738.01	8078.132	66.26880	152.5024	44.74233	1
55128.18	12692.670	62.91056	149.2182	44.91573	0
44482.79	9812.976	68.67830	152.7264	44.55131	0
55394.63	10769.461	67.37055	151.6027	44.85795	0

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

var	mean	sd	min	max	n
age	44.89072	0.2191325	43.97646	45.70364	10000
female	0.46040	0.4984543	0.00000	1.00000	10000

d) Report the correlation matrix for the now six variables in the BestIncome.txt data.

	lab_inc	cap_inc	hgt	wgt	age	female
lab_inc	1.0000000	0.0053253	0.0027898	0.0045069	0.9240460	-0.2021599
cap_inc	0.0053253	1.0000000	0.0215716	0.0062987	0.2341567	-0.0530370
hgt	0.0027898	0.0215716	1.0000000	0.1721027	-0.0450868	-0.1156436
wgt	0.0045069	0.0062987	0.1721027	1.0000000	-0.3003101	-0.7125505
age	0.9240460	0.2341567	-0.0450868	-0.3003101	1.0000000	0.0188964
female	-0.2021599	-0.0530370	-0.1156436	-0.7125505	0.0188964	1.0000000

2. Stationarity and data drift

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.

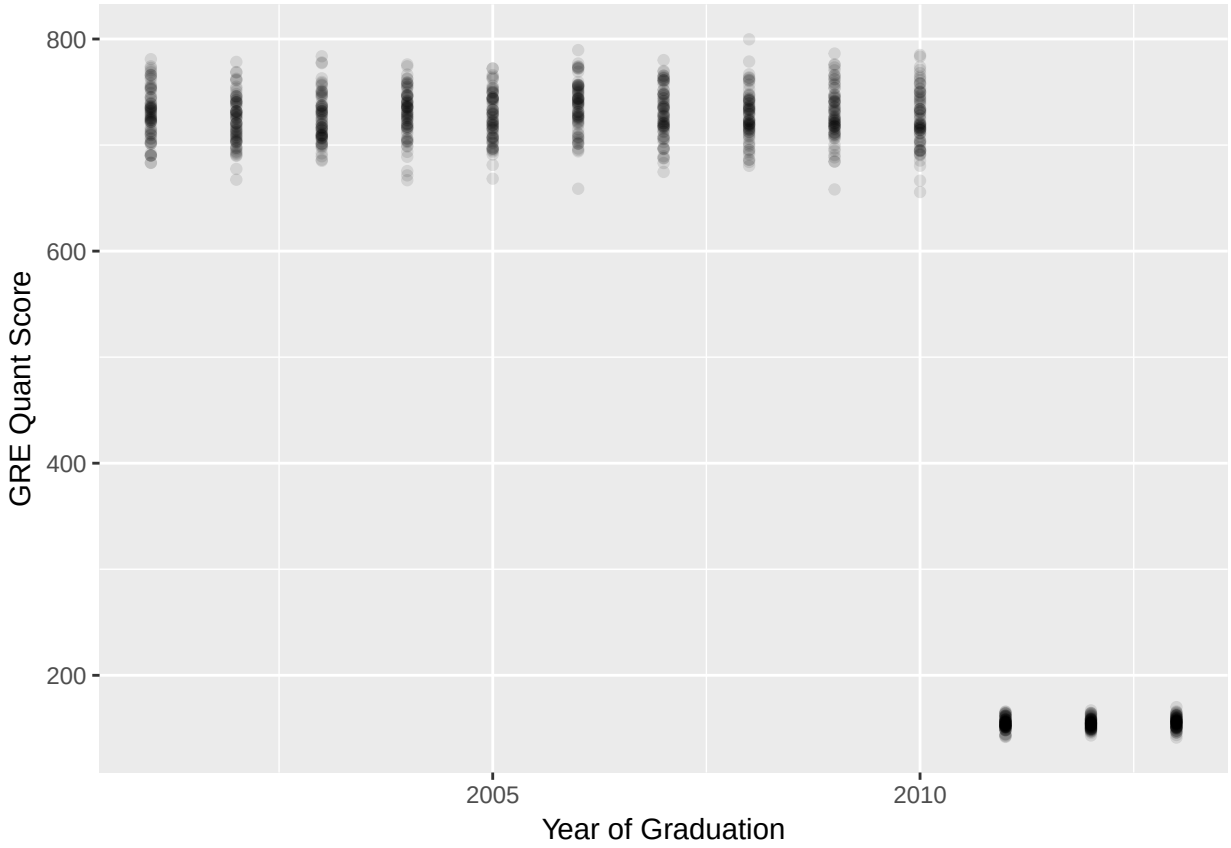
a) Estimate the coefficients in the regression above by ordinary least squares without making any changes to the data. Report your estimated coeffi- cients and standard errors on those coefficients.

```
##
## Call:
## glm(formula = salary_p4 ~ gre_qnt, data = IncomeIntel)
##
## Deviance 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 ***
## gre_qnt       -25.763      1.365  -18.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 109322312)
##
##     Null deviance: 1.4805e+11  on 999  degrees of freedom
## Residual deviance: 1.0910e+11  on 998  degrees of freedom
## AIC: 21352
##
## Number of Fisher Scoring iterations: 2
```

The estimated β_0 is 89541.293 with a standard error of 878.764. The estiamted β_1 is -25.763 with a standard error of 1.365, and **is** significant.

The regression model is $salary\_p4_i = 89541.293 - 25.763 \times gre\_qnt_i + \epsilon_i$

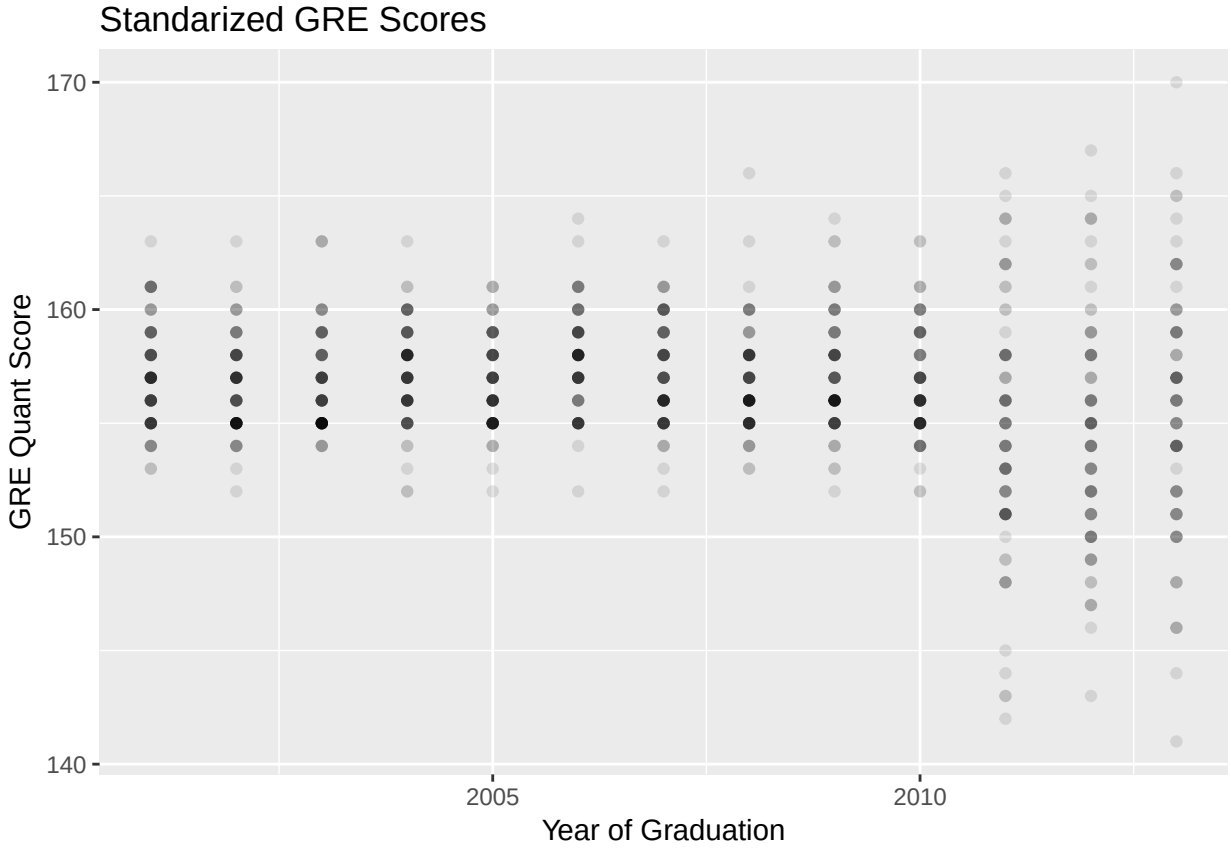
(b) Create a scatter plot of GRE quantitative score (gre_qnt_i) on the y-axis and graduation year ($grad\_year_i$) on the x-axis. Do any problems jump out in this variable and your ability to use it in testing your hypothesis? Propose and implement a solution for using this variable in your regression.



There seems to be a significant issue—the gre_qnt values are distributed between 600 and 800 up until 2010, and bellow 200 after 2010. This is because the GRE scale changed in 2011 from a [200,800] scale to a [130,170] scale.

To use this data, we should scale all the data to the same scale. GRE website provides a concordance table for the actual scores, so it would be the best to use that one. However, an issue that emerges is that the old scale and the old tests did not differentiate high scores very well, and the score of 800 maps onto the score of 166, not 170 on the new scale.

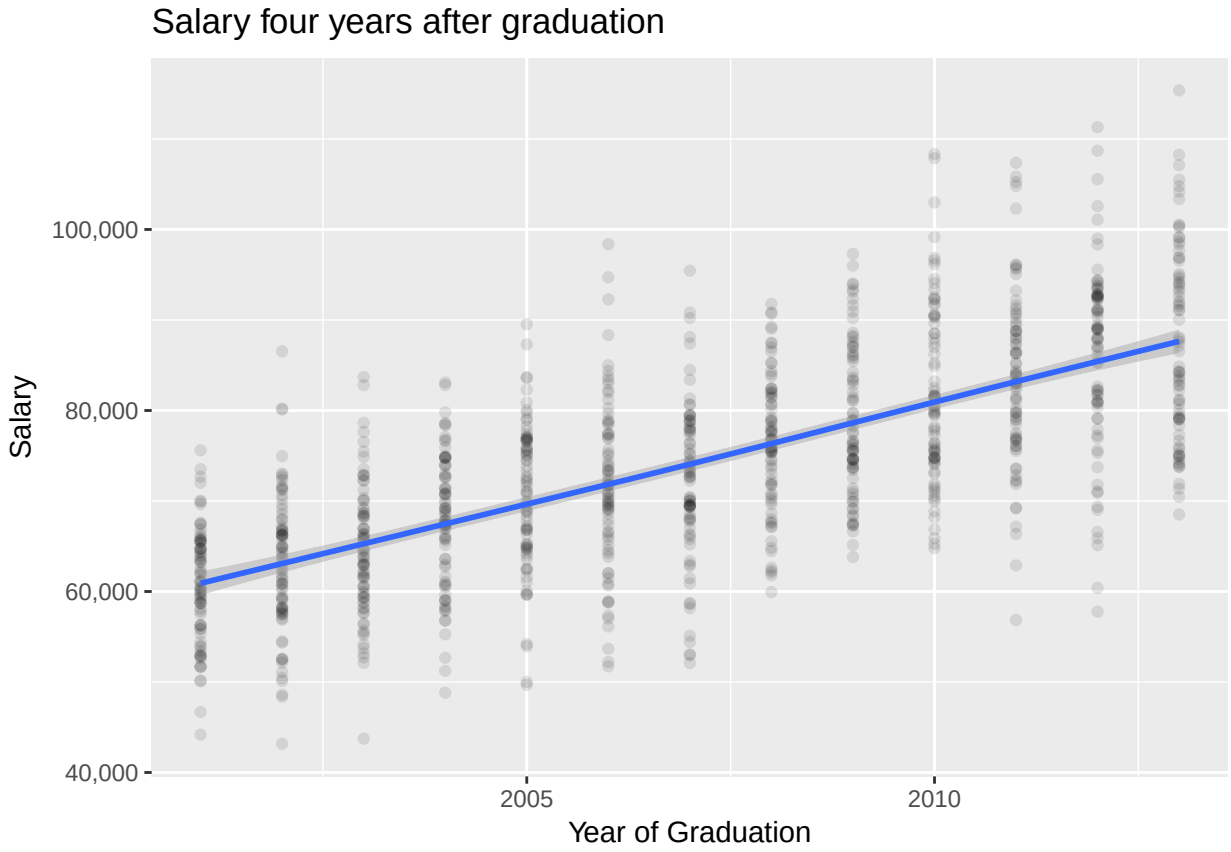
In order to implement this solution, I am converting the GRE scores in the data set to the actual GRE scores, that is rounding the old scale to the closest 10, and rounding the new scale to the closest integer.



This scatter plot looks much better. However, there is much more varriance in scores after 2010, likely because the new test is more difficult and spreads the high schores out more. This is not ideal, but it is what we've got. Ultiamtely, we can consider running analysis on the data up until 2010 only.

(c) Create a scatter plot of income 4 years after graduation ($salaryp4_i$) on the y-axis and graduation year ($gradyear_i$) 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.

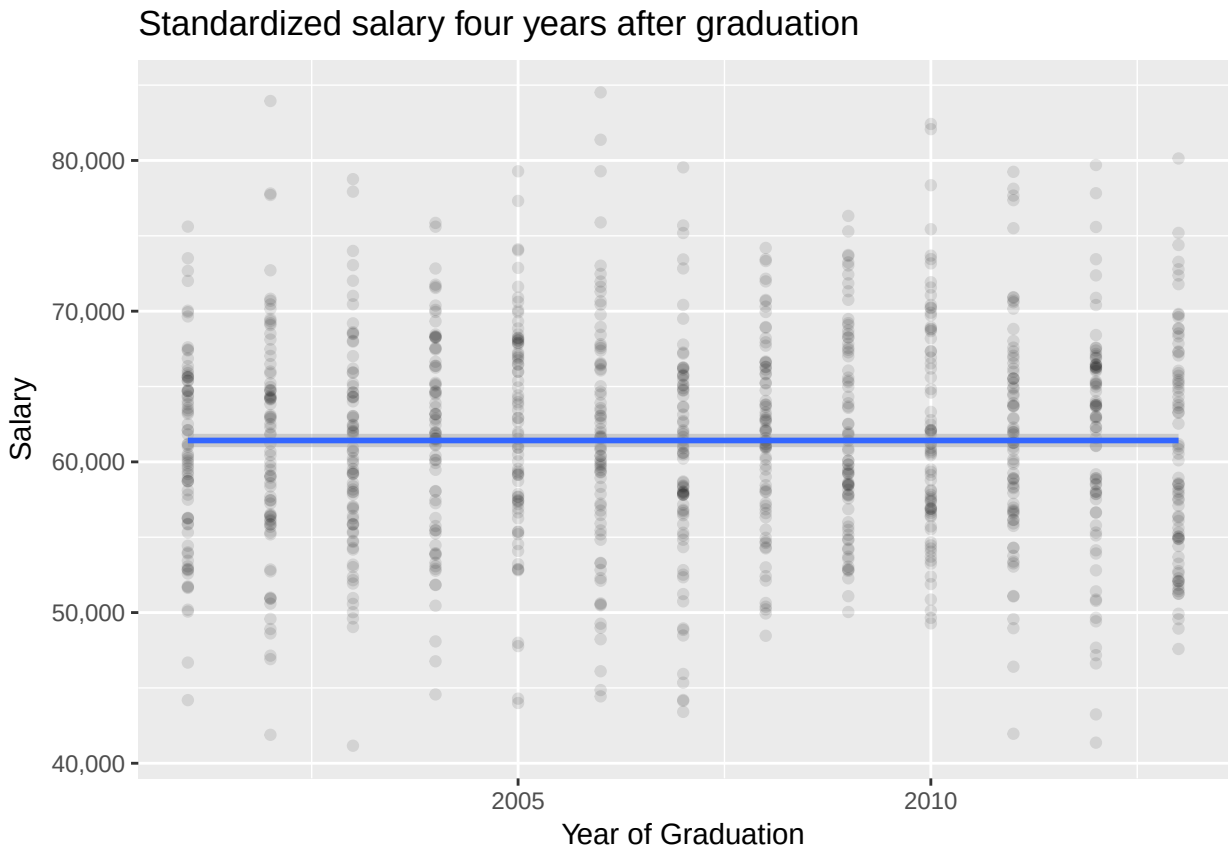
```
## `geom_smooth()` using method = 'gam'
```



There is a clear trend in the data—as the time increases, the mean income increases as well, likely due to inflation.

To correct this, we can calculate the anual growth rate w.r.t. 2001 average income and then use that growth rate to standardize all values to the 2001 level.

```
## `geom_smooth()` using method = 'gam'
```



This looks much better!

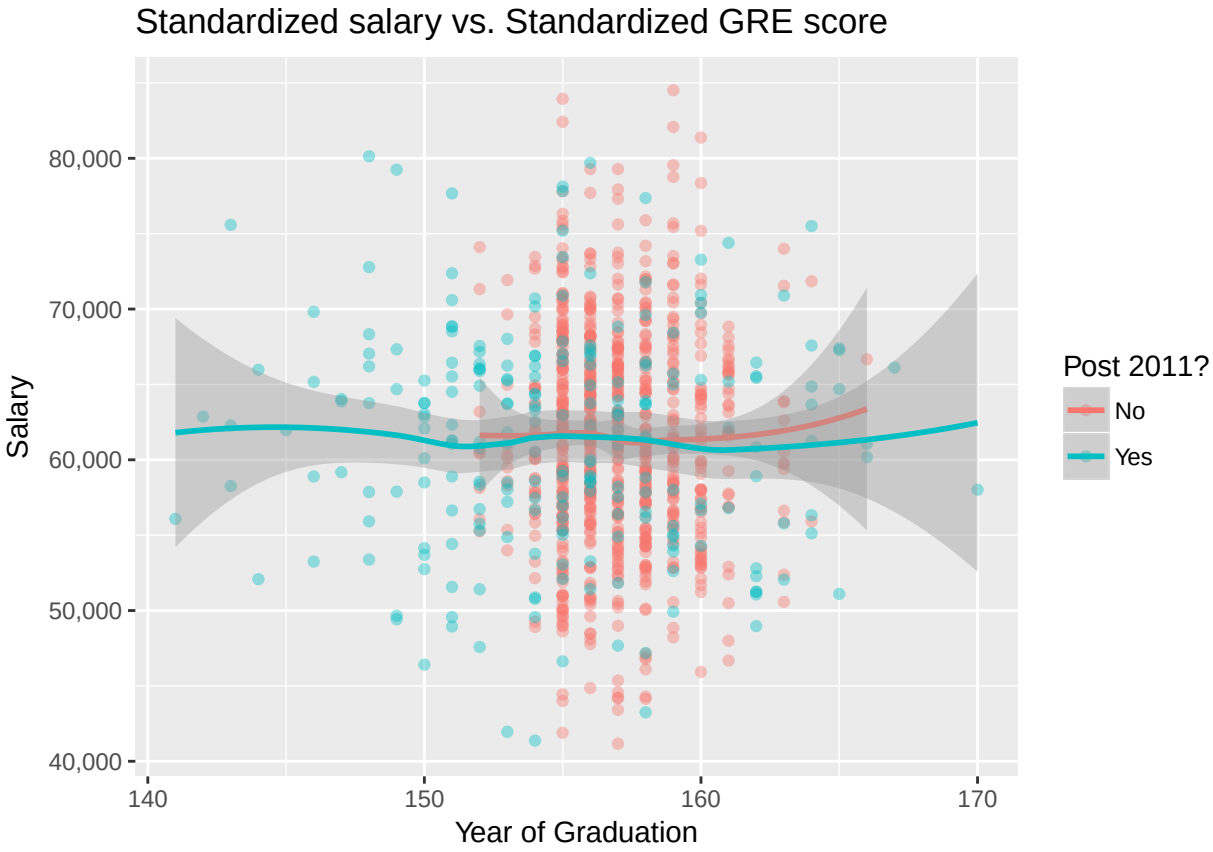
(d) Using the changes you proposed in parts (b) and (c), re-estimate the re- gression coefficients with your updated salary p4i and gre qnti variables. Report your new estimated coefficients and standard errors on those coef- ficients. How do these coefficients differ from those in part (a)? Interpret 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 (evidence for or against your hypothesis)?

```
##
## Call:
## glm(formula = salary_p4 ~ gre_qnt, data = IncomeIntel)
##
## Deviance 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 ***
## gre_qnt       -25.763      1.365  -18.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 109322312)
##
```

```
##     Null deviance: 1.4805e+11  on 999  degrees of freedom
## Residual deviance: 1.0910e+11  on 998  degrees of freedom
## AIC: 21352
##
## Number of Fisher Scoring iterations: 2
```

```
##
## Call:
## glm(formula = salary_p4 ~ gre_qnt, data = IncomeIntel_trans)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -20234.9   -4781.4     102.9    4784.0   23203.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 68190.79   10849.97   6.285 4.89e-10 ***
## gre_qnt       -43.26      69.30  -0.624    0.533
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 50948072)
##
##     Null deviance: 5.0866e+10  on 999  degrees of freedom
## Residual deviance: 5.0846e+10  on 998  degrees of freedom
## AIC: 20588
##
## Number of Fisher Scoring iterations: 2
```

```
## `geom_smooth()` using method = 'loess'
```



The estimated β_0 is 68190.79 with a standard error of 10849.97. The estiamted β_1 is -43.26 with a standard error of 69.30, and **is not** significant.

Although at first,the gre_qnt coefficient was significant, after the data was transformed this ceased to be the case. Most likely, time was the confounding factor: GRE scores were significantly higher before 2011, while salaries increased with later years of graduation.

After accounting for data drift, there doesn't seem to be a relationship between gre_qnt scores and income four yars after graduation. Thinking about our original hypothesis, the data offers no evidence that higher intelligence is associated with higher income.

3. Assessment of Kossinets and Watts (2009)

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

In their reseach, Kossinets and Watts (2009) set out to investigate the origins of homophily. More specifically, their research question is: **What are the roles of and the interplay between individual choice and structural constraints in the development of homophily in a social network?**

To answer their question, the researchers used data from a large US university, including students, faculty and staff, that spans the fall and the spring semester. Their data set draws on **three separate sources***: 1) the logs of university e-mail interactions (timestamp, sender ID and recipient IDs), 2) a database of individual attributes of individuals (e.g. gender, age, department, status etc.) and 3) course registration records. After data cleaning, their data set comprised of **30,296 nodes**, that is stable (present for both semesters) e-mail

users, and **7,156,162 tie observations**, that is the e-mails exchanged by the users over 270 days. The description and the definition of all variables can be found in Appendix A (pp. 439-42).

During the data cleaning, a **large number of users and e-mails was excluded from the data set**. For example, of the 43,553 users that used university e-mail during the period of interest, 8,979 (around 20%) were *not* active during both semesters—according to the authors, likely due to population turnover—and were excluded from the analysis. Further 4,178 (around 10%) were excluded because they were not exchanging e-mails with others in the university e-mail network, at it is unclear who those users are or why that is the case. Finally, an unknown number of e-mail accounts was excluded because they were not tied to the main university system (@university.edu) but rather the departmental system (@department.university.edu). Furthermore, although bulk e-mails (sent to more than one recipient) were used to establish 'implicit foci' in analysis, which were used to determine the 'strength of shared membership', these bulk-emails (accounting for about 18% of all e-mails) were not included as indicators of actual relationships (even though the only record of a friendship might be in their participation in a group e-mail). As a result, the final dataset omitted an unknown number of e-mail relationships (or social relationships evident in the e-mail communications), as well as some individuals (e.g. those present only for a semester) that might've nevertheless facilitated relationship formation, both of which would have an impact on the results of the analysis.

Most importantly, the use of e-mail logs to answer questions about "social relationships" is problematic in itself. On one hand, it is debatable to what extent an e-mail exchange (receiving and reciprocating an e-mail, which is taken as a tie within the author's analysis) indicates and actual social relationship. For example, an administrator might request that a student send over some documents, with a student responding , but it might not be appropriate to call this a "social relationship." On the other hand, social relationships transcend both e-mail exchanges and the university. Much communication between actors occurs through other channels, such as over text, phone calls and especially in person (two roommates, though highly involved in each other's social lives and networks, might never exchange an e-mail). At the same time, actors and groups outside of the university network, but within the community, might be playing a very important role in facilitating social relationship formation. The authors are aware, however, that *e-mail does not capture the whole relevant social network*, as they point out that they are interested "in the process of network evolution, rather than the network structure itself" (p. 417). From that standpoint, we can see the e-mail dataset as an *(accessible) subset of social relationships*, within which the process of relationship formation and dissolution can still be examined.