MACS 30000 PS 2, Fall 2018

Cosette L. Hampton

10/16/2018

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Imputing age and gender (3 points)

```
setwd("/Users/cosettelh/Documents/UChi_local/Grad_MPP/MACSS/persp-analysis_A18/Assignments/A2")

bestincome <- read.csv("BestIncome.txt", header = FALSE,
                         col.names = c("lab_inc", "cap_inc", "hgt", "wgt"))
incomeintel <- read.csv("IncomeIntel.txt", header = FALSE,
                          col.names = c("grad_year", "gre_qnt", "salary_p4"))
survincome <- read.csv("SurvIncome.txt", header = FALSE,
                         col.names = c("tot_inc", "wgt", "age", "female"))
```

```
print(bestincome)
as.data.frame(describe(bestincome, na.rm = TRUE))
stat.desc(bestincome)


print(survincome)
as.data.frame(describe(survincome, na.rm = TRUE))
stat.desc(survincome)

print(incomeintel)
as.data.frame(describe(incomeintel, na.rm = TRUE))
stat.desc(incomeintel)
```

(A)

OLS models help estimate an unknown outcome given several parameters. We can impute age and gender into BestIncome by using an OLS model based on an equation with SurveyIncome since both datasets have the variables weight and tot_inc can be computed with lab_inc and cap_inc.

```
#Adding tot_pop column
bestincome$tot_inc <- bestincome$lab_inc + bestincome$cap_inc
print(bestincome)
as.data.frame(describe(bestincome, na.rm = TRUE))
```

```
#Providing the OLS models
lm_age <- lm(age~tot_inc+wgt, data = survincome)
lm_female <- lm(female~tot_inc+wgt, data = survincome)

lm_age
```

```
##
## Call:
```

```
## lm(formula = age ~ tot_inc + wgt, data = survincome)
##
## Coefficients:
## (Intercept)      tot_inc          wgt
##  44.2096668    0.0000252   -0.0067221
```

```
lm_female
```

```
##
## Call:
## lm(formula = female ~ tot_inc + wgt, data = survincome)
##
## Coefficients:
## (Intercept)      tot_inc          wgt
##   3.761e+00   -5.250e-06   -1.953e-02
```

So the equations are: $44.2096668 + (totinc * 0.0000252) + (wgt * -0.0067221)$ $3.761 + (totinc * -0.000005250) + (wgt * -0.01953)$

(B)

```
age_form <- 44.2096668+(bestincome$tot_inc*0.0000252)+(bestincome$wgt*-0.0067221)
gen_form <- 3.761+(bestincome$tot_inc*-0.000005250)+(bestincome$wgt*-0.01953)

bestincome$age <- age_form
bestincome$female <- gen_form

head(bestincome, 10)
```

```
##     lab_inc   cap_inc      hgt      wgt  tot_inc      age    female
## 1  52655.61  9279.510 64.56814 152.9206 61935.12 44.74248 0.4493007
## 2  70586.98  9451.017 65.72765 159.5344 80038.00 45.15422 0.2250934
## 3  53738.01  8078.132 66.26880 152.5024 61816.14 44.74230 0.4580933
## 4  55128.18 12692.670 62.91056 149.2182 67820.85 44.91569 0.4907093
## 5  44482.79  9812.976 68.67830 152.7264 54295.77 44.55128 0.4932014
## 6  55394.63 10769.461 67.37055 151.6027 66164.09 44.85791 0.4528382
## 7  62627.90  9730.261 64.54769 151.4220 72358.16 45.01522 0.4238484
## 8  54936.56  8712.628 63.08035 153.9178 63649.18 44.77898 0.4208276
## 9  52730.25  9260.990 63.41790 147.3275 61991.24 44.78150 0.5582392
## 10 60525.27 10310.989 65.31023 154.1793 70836.26 44.95833 0.3779877
```

```
bestincome <- bestincome %>%
  mutate_at(vars(female), funs(round(.,0))) %>%
  select(-tot_inc)
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
head(bestincome, 10)
```

```
##     lab_inc   cap_inc      hgt      wgt      age female
## 1  52655.61  9279.510 64.56814 152.9206 44.74248      0
## 2  70586.98  9451.017 65.72765 159.5344 45.15422      0
## 3  53738.01  8078.132 66.26880 152.5024 44.74230      0
## 4  55128.18 12692.670 62.91056 149.2182 44.91569      0
## 5  44482.79  9812.976 68.67830 152.7264 44.55128      0
## 6  55394.63 10769.461 67.37055 151.6027 44.85791      0
## 7  62627.90  9730.261 64.54769 151.4220 45.01522      0
## 8  54936.56  8712.628 63.08035 153.9178 44.77898      0
## 9  52730.25  9260.990 63.41790 147.3275 44.78150      1
## 10 60525.27 10310.989 65.31023 154.1793 44.95833      0
```

(C)

See output table from imputed_vars

```
imputed_vars <- bestincome %>% select(age, female)
imputed_vars <- as.data.frame(describe(imputed_vars, na.rm = TRUE))
imputed_vars <- imputed_vars %>% select(mean, sd, min, max, n)
imputed_vars
```

```
##            mean        sd      min      max     n
## age    44.89069 0.2191325 43.97643 45.70361 10000
## female  0.46140 0.4985327  0.00000  1.00000 10000
```

(D)

See table printed below:

```
corr_matrix <- as_data_frame(cor(bestincome))
corr_matrix
```

```
## # A tibble: 6 x 6
##    lab_inc  cap_inc      hgt      wgt     age  female
##      <dbl>    <dbl>    <dbl>    <dbl>   <dbl>   <dbl>
## 1  1        0.00533  0.00279  0.00451  0.924  -0.167
## 2  0.00533  1        0.0216   0.00630  0.234  -0.0470
## 3  0.00279  0.0216   1        0.172   -0.0451 -0.135
## 4  0.00451  0.00630  0.172    1       -0.300  -0.777
## 5  0.924    0.234   -0.0451  -0.300    1       0.0724
## 6 -0.167   -0.0470  -0.135   -0.777    0.0724  1
```

Stationary and data drift (4 points)

```
print(incomeintel)
```

The equation: $salaryp4_i = b_0 + b_1 greqnt_i + e_i$

(A)

In plain words, the equation is for the salary is the predictor variable, and the GRE quantitative score is the predictor variable. b_0 is the Y-intercept. e_i is an error term. Using a linear regression model and inference from the graph, we compute the equation to be:

$salaryp4_i = 89541.29 - 25.76 * greqnt_i + e_i$

The estimated Y-intercept coefficient is $89,541.29$ and the estimated GRE score coefficient is -25.76. Their standard errors are 878.764 and 1.365 respectively. Both coefficients are significant at the 95% confidence level.

```
Y_hat <- mean(incomeintel$salary_p4)
Y_hat
```

```
## [1] 74173.29
```

```
incomeintel_lm <- lm(salary_p4 ~ gre_qnt, data = incomeintel)
incomeintel_lm
```

```
##
## Call:
## lm(formula = salary_p4 ~ gre_qnt, data = incomeintel)
##
## Coefficients:
## (Intercept)      gre_qnt
##    89541.29       -25.76
```

```
summary(incomeintel_lm)
```

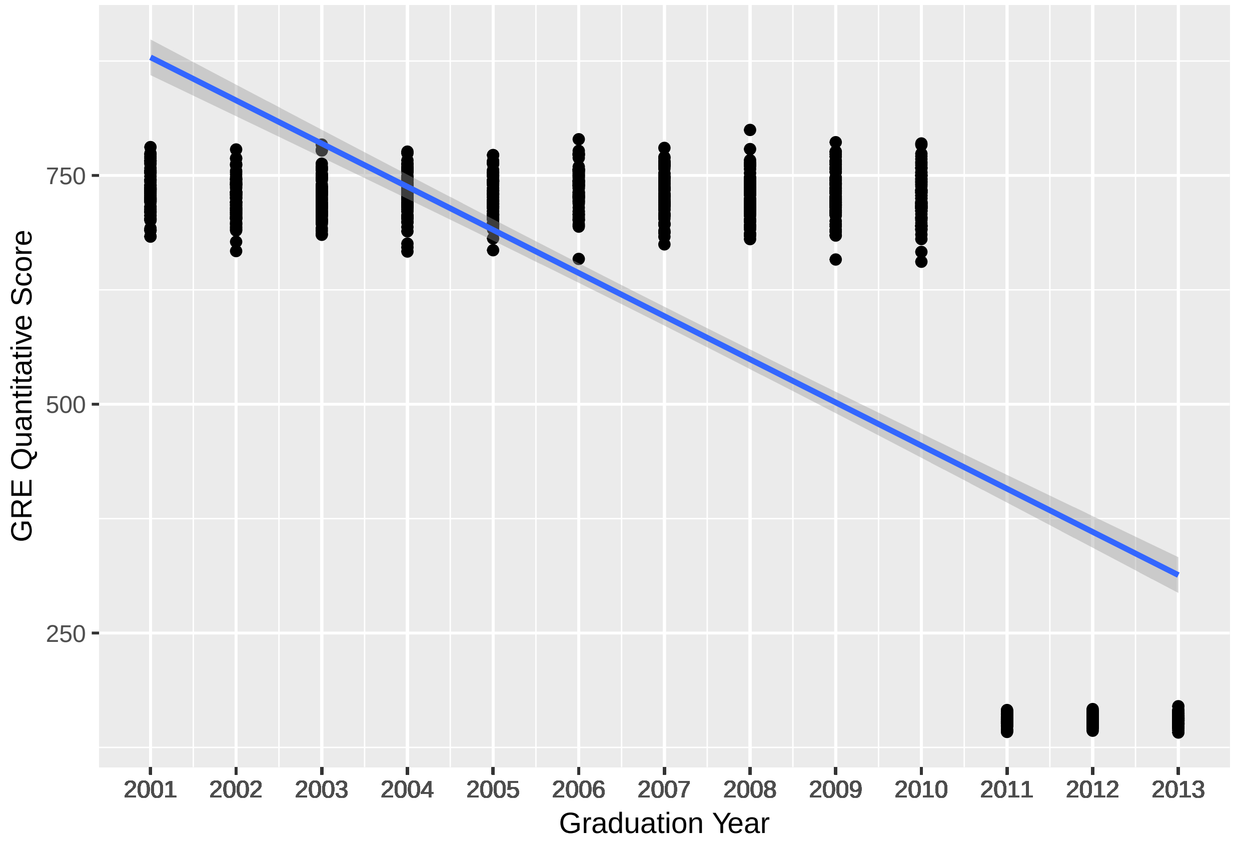
```
##
## Call:
## lm(formula = salary_p4 ~ gre_qnt, data = incomeintel)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -28761  -7049   -293   6549  37666
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 89541.293    878.764  101.89   <2e-16 ***
## gre_qnt       -25.763      1.365  -18.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## 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)

The GRE quantitative scoring scale changed in 2011, from (200-800) at 10-point increments to now (130-170) at 1-point increments. This change is represented on the plot shown below with the huge gap in scores from 2010 to 2011, and a steeply decreasing line of best fit.

As such the data need to be adjusted to better compare recent score data to pre-2011 scores. A simple google search shows how the ETS compares old to new scores. For example, for Quant new 166-170 = 800 old and new 151 = 640-650 old.

```
ggplot(data = incomeintel, aes(x = grad_year, y = gre_qnt)) +
  geom_point() +
  geom_smooth(method='lm') +
  scale_y_continuous(labels = scales::comma) +
  labs(x = "Graduation Year", y = "GRE Quantitative Score") +
  scale_x_continuous(labels = as.character(incomeintel$grad_year),
                     breaks = incomeintel$grad_year)
```



```
score_conv <- read_xlsx("new_scores.xlsx") #Reading in conversion table
```

```
## Warning in strptime(x, format, tz = tz): unknown timezone 'zone/tz/2018e.
## 1.0/zoneinfo/America/Chicago'
```

```
incomeintel <- incomeintel %>% #rounding scores from given data
  mutate("gre_qnt" <- ifelse(grad_year < 2011, round(gre_qnt, -1), round(gre_qnt,0))) %>%
  select(-(gre_qnt))

names(incomeintel)[3] <- "gre_qnt"
```

```
#joining on gre_qnt
income_est <- left_join(incomeintel, score_conv, by="gre_qnt")

income_est <- income_est %>%
  mutate("gre_qnt" <- ifelse(grad_year < 2011, current_scale, gre_qnt))

income_est <- income_est %>%
  select(-gre_qnt, -current_scale, -rank)
names(income_est)[3] <- "gre_qnt"
```

```
head(income_est)
```

```
##   grad_year salary_p4 gre_qnt
## 1      2001  67400.48     158
## 2      2001  67600.58     156
## 3      2001  58704.88     158
## 4      2001  64707.29     161
## 5      2001  51737.32     158
## 6      2001  64010.82     160
```

```
head(incomeintel)
```

```
##   grad_year salary_p4 gre_qnt
## 1      2001  67400.48     740
## 2      2001  67600.58     720
## 3      2001  58704.88     740
## 4      2001  64707.29     770
## 5      2001  51737.32     740
## 6      2001  64010.82     760
```

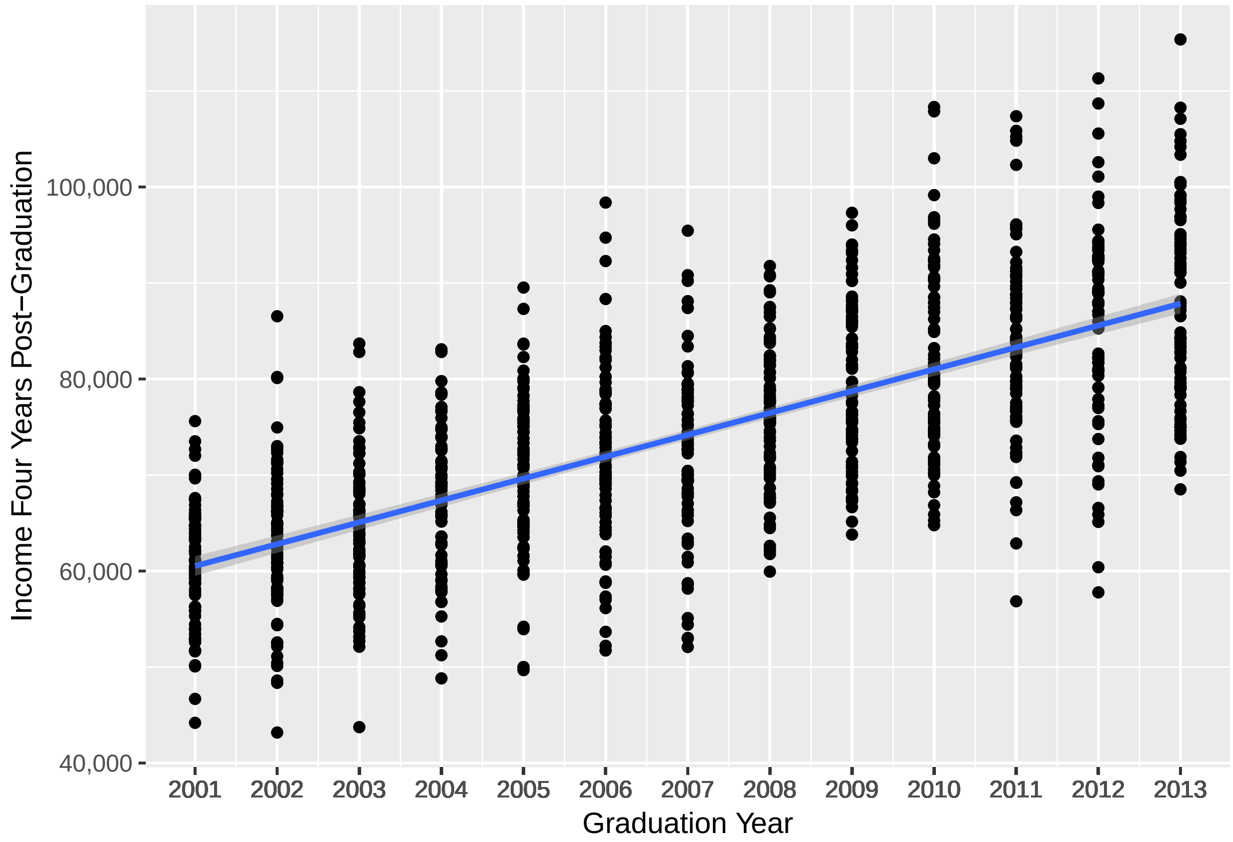
```
head(score_conv)
```

```
## # A tibble: 6 x 3
##   gre_qnt current_scale  rank
##     <dbl>         <dbl> <dbl>
## 1     800           166    91
## 2     790           164    87
## 3     780           163    84
## 4     770           161    78
## 5     760           160    76
## 6     750           159    73
```

```
#scores conversions are correct
```

(C)

```
ggplot(data = income_est, aes(x = grad_year, y = salary_p4)) +
  geom_point() +
  geom_smooth(method='lm') +
  scale_y_continuous(labels = scales::comma) +
  labs(x = "Graduation Year", y = "Income Four Years Post-Graduation") +
  scale_x_continuous(labels = as.character(incomeintel$grad_year),
                     breaks = incomeintel$grad_year)
```



```
#Used: https://github.com/UC-MACSS/persp-analysis_A18/issues/13

base_year <- 2011

#mean salary by each year
avg_inc_by_year <- income_est %>%
  group_by(grad_year) %>%
  summarize(mean_salary = mean(salary_p4))

#growth rate in salaries across all 13 years
avg_growth_rate <- avg_inc_by_year %>%
  mutate((mean_salary - (lag(mean_salary, default = first(mean_salary))))/(lag(mean_salary, default = fi
names(avg_growth_rate)[3] <- "sal_growth_rate"

income_est <- left_join(income_est, avg_growth_rate, by = "grad_year")
head(income_est)
```

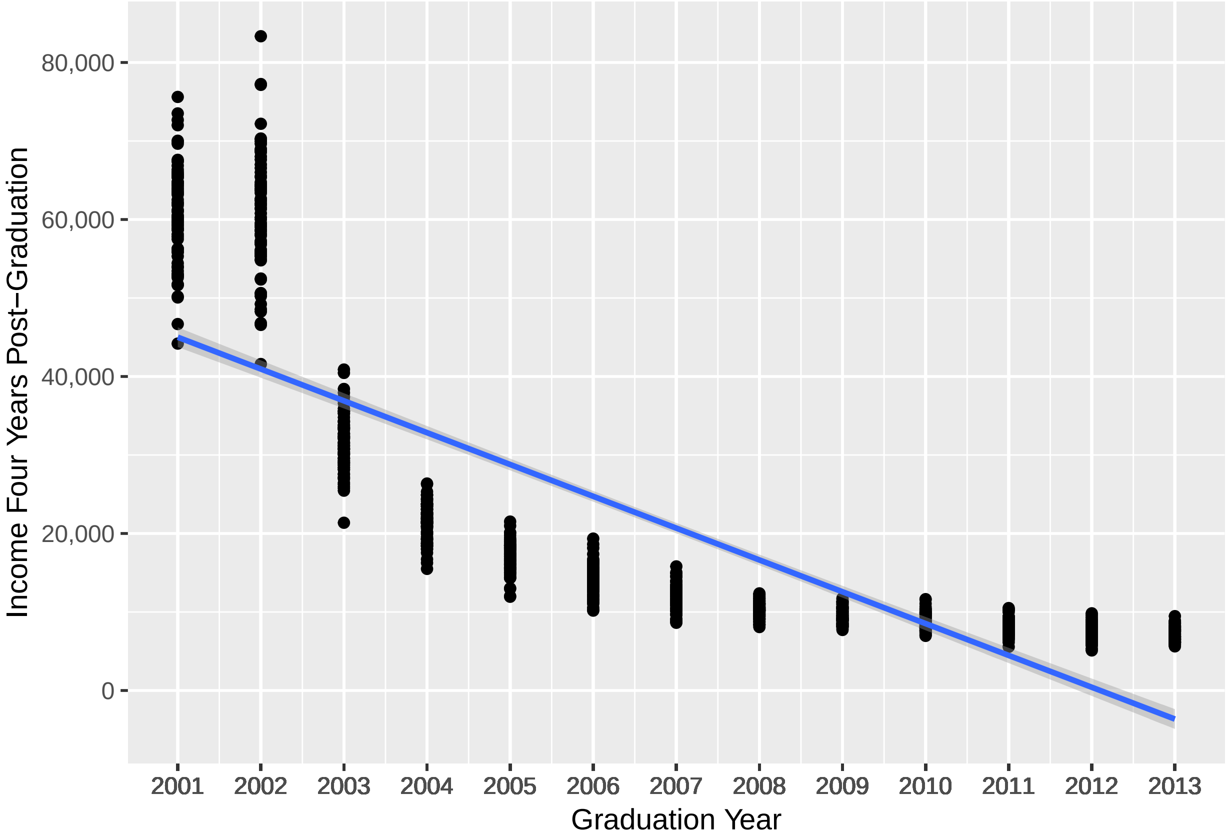
```
##   grad_year salary_p4 gre_qnt mean_salary sal_growth_rate
## 1      2001  67400.48     158    60710.71               0
## 2      2001  67600.58     156    60710.71               0
## 3      2001  58704.88     158    60710.71               0
## 4      2001  64707.29     161    60710.71               0
## 5      2001  51737.32     158    60710.71               0
## 6      2001  64010.82     160    60710.71               0
```

```
income_est <- income_est %>%
  mutate(new_salary = salary_p4 / ((1 + sal_growth_rate)*(grad_year - 2001)))
```

```r
income_estimate <- income_est %>%
  mutate("new_sal" = ifelse(new_salary == Inf, salary_p4, new_salary))
head(income_estimate)
```

```
##   grad_year salary_p4 gre_qnt mean_salary sal_growth_rate new_salary
## 1      2001  67400.48     158    60710.71               0        Inf
## 2      2001  67600.58     156    60710.71               0        Inf
## 3      2001  58704.88     158    60710.71               0        Inf
## 4      2001  64707.29     161    60710.71               0        Inf
## 5      2001  51737.32     158    60710.71               0        Inf
## 6      2001  64010.82     160    60710.71               0        Inf
##    new_sal
## 1 67400.48
## 2 67600.58
## 3 58704.88
## 4 64707.29
## 5 51737.32
## 6 64010.82
```

```r
ggplot(data = income_estimate, aes(x = grad_year, y = new_sal)) +
  geom_point() +
  geom_smooth(method='lm') +
  scale_y_continuous(labels = scales::comma) +
  labs(x = "Graduation Year", y = "Income Four Years Post-Graduation") +
  scale_x_continuous(labels = as.character(incomeintel$grad_year),
                     breaks = incomeintel$grad_year)
```



(D)

The new coefficients for the intercept and GRE quantitative score are -55284.7 and 485.5 respectively. standared errors are 28115.2 and 179.6. Essentially these salaries are adjusted according to the average growth rate for inflation on a year-to-year basis, so its showing that with each point-increase in the gre score, salary increases by only 485.50 whereas before, it decrased by 25.76. These are also significant at a 0.05 confidence level so we fail to reject the HO that higher intelligence (measured with GRE quant score) is associated with higher income.

```
income_lm <- lm(new_sal ~ gre_qnt, data = income_estimate)
income_lm
```

```
##
## Call:
## lm(formula = new_sal ~ gre_qnt, data = income_estimate)
##
## Coefficients:
## (Intercept)      gre_qnt
##    -55284.7        485.5
```

```
summary(income_lm)
```

```
##
## Call:
## lm(formula = new_sal ~ gre_qnt, data = income_estimate)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -20387 -11453  -8609   1925  63382
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -55284.7    28115.2  -1.966  0.04953 *
## gre_qnt        485.5      179.6   2.703  0.00698 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18500 on 998 degrees of freedom
## Multiple R-squared:  0.00727,    Adjusted R-squared:  0.006275
## F-statistic: 7.309 on 1 and 998 DF,  p-value: 0.006979
```

Assessment of Kossinets and Watts (2009) (3 points)

Questions

(a) State the research question of this paper. The research question is the fundamental question that the paper is trying to answer. The research question should be one sentence long and should end with a question mark. An example of a research question is, "What is the effect of an extra hour of storm advisory in a county on births nine months later in that county?"

(b) Describe the data that the authors used. How many data sources? How many observations (this question could have multiple dimensions)? What time period did the data span? Where can you find a description and definition of all the variables?

(c) Highlight a potential problem that the data cleaning process might introduce in a way that diminishes the authors' ability to answer the research question.

(d) In this paper, the underlying theoretical construct is "social relationships" and the data are email logs linked to other characteristics of the senders and receivers. Discuss one weakness of this match of data source and theoretical construct and describe how the authors address this weakness.

Answers

According to Kossinets and Watts (2009), homophily is described as the tendency of “like to associate with like,” or similar people associating through different social systems with one another. They identify two different types of homophily: $choice - homophily$ and $induced - homophily$, which are individual and structural views of the world. Neverthless, both of these homophilies are related by the fact that social environments are created by peoples’ choices. Kossinets and Watts use a university community to study the origins of homophily among students, faculty and staff. The research question they explore is ** (a) generally, what is the origin of homophily? Or in context, “What are the primary determinants (social, structural and individual factors), called”risk sets" of who forms a “tie” with who?" **

In order to fully develop the research question and analyze outcomes, the “Data and Methods” secion of their paper explain some of the answers to part **(b)**. The overall analysis used a dataset of a size 7,156,162 messages exchanged by 30,396 undergraduate and graduate students, faculty, and staff in a large U.S. university, who used their university e-mail accounts to both send and receive messages during one academic year or 270 days of observation. These were selected out of 34,574 users who were active throughout both semesters. For the specific research question, they sampled a total of about 100,000 cases (pairs of nodes that experienced a tie-transition) over a 210-day period.

The data set also codes for several individual attributes that may represent different dimensions of homophily (you can find a description and definition for these by seeing Appendix A). These include personal characteristics (gender, age, status, field, year, and state (U.S. native/foreigner)), organizational affiliations (primary department, school, campus, dormitory, academic field); course-related variables (courses taken, courses taught); and e-mail-related variables (days active, messages sent, messages received, in-degree, out-degree, not included (reciprocated degree). Table 1 on page 411 shows the status breakdown of respondents as well.

The dataset only includes information for individuals who both sent and received emails using their university addresses, for both semesters of the academic year. The sender ID, and the IDs of all recipients of the message, extracted from the mail server logs were anonymized and the actual university used was not specified. These were compiled from 3 different databases on: (1) the logs of e-mail interactions within the university over one academic year, (2) individual attributes (status, gender, age, department, number of years in the community, etc.), and (3) records of course registration, in which courses were recorded separately for each semester.

(c) The selection of only users active in both semesters and those who communicated only with university emails significantly narrows the scope of the analysis. When cleaning binomial variables that were coded differently for the same user, they selected the value using a set of heuristics– though, the narrowing of the dataset may have biases in it that are different from commonly used heuristics. They assumed that most recent modal values were the most likely to be correct, though a user/data collector may have edited a mistake they made in the first semester and updated it with the correct information in the second. In the case with flagging status, they applied two flags when one was missing, thinking that the student graduated and was employed when the student actually may not have graduated. These limitations generated by the cleaning process because those students who only communicated via these emails may be more homogenous, and overall at the university level homogenous more than the general population. A department may also require more email-communication but that may not have been reported in the details. The more specific issues just distort how social ties may be created based on different personal and academic characteristics.

In the same way that these cleaning processes have this effect, the limitation of these being electronic communications hinders a more complete analysis around social relationships and creating ties that speak to homophily. According to the authors, “Electronic communication, moreover, may differ in some systematic ways between formal and informal organizations and various communities, depending both on the demographics of their constituents and on the purpose of the organizations themselves.” Additionally privacy prohibitions prevent them from observing types of interactions that may have been visible with an observational research component or social experiment. They address this **(d)** by using the extra data about individual attributes and affiliation, and looking at when pairs become triads, building a larger social network. They assume that, *“. . . forward-looking individuals select into structural positions, such as classes, clubs, and even friendship circles, precisely in order to maximize their chances of meeting the people they want to meet.”* (pg. 435)

Nevertheless on a somewhat separate point, the lack of race and socioeconomic status is perhaps the most extreme limitation of this data especially since they mentioned multiple times that key interest in this study was to observe issues *"...such as segregation, inequality, and even the transmission of information between groups."* The history of segregation in the United States, both de facto and de jure, has been based on racial identity, with no observation of similar interests and group-membership at all– especially when thinking of emails that may have been shared between affinity group members like a Black Student Union (BSU) or how many schools have group chats for native Chinese students. I'd bet that group membership and racial identity would be strongly correlated. Or since greek life is very expensive to enter, controlling for socioeconomic status may give more complete results as well. Not using those two details as control variables in my opinion drastically decrease my confidence in the outcomes.