Darin Zlatarev #261081234

Professor Roman Galperin

ORGB 672 – Exercise 4

4 April 2023

Part 1: Loading and Preparation

Following up on the previous exercise, we now need to find out whether or not there is any correlation between the average processing time of each patent officer and any of their types of centralities. First, we had to run some of the code we had written for Exercise 3 in order to guess the race and gender of each officer as well as the length of time they have served in the USPTO.

In order to find out the processing time of each application, we had to create a new column called ‘decision_date’ which merges two existing columns (see *Fig. 1*). As there are only two outcomes when a decision is made, the two columns merged are the column of the date when the patent was granted, and the column showing the date when it was rejected. Thus, we can now calculate the processing time of each application by subtracting the date when a decision was made and the date when the application was filed; to do so we create another column called ‘app_proc_time’.

We then proceed by getting rid of all entries that are still in processing using the drop_na() function. We do so in order to smoothly compute our next variable called ‘examiner_PT’. This variable calculates the average processing time per officer, it is crucial to remove all entries currently processing as if an officer has even as little as one null value, their average processing time will not be able to be computed (see *Fig. 2*). Lastly, we export the new data frame into a freshly new CSV file for extra convenience.

```r
### Part 1: Calculate Processing Time

applications = applications %>% mutate(decision_date = pmax(patent_issue_date, abandon_date, na.rm = TRUE))
applications$decision_date

applications$app_proc_time = as.numeric(as.Date(as.character(applications$decision_date)) - as.Date(as.character(applications$filing_date)))
applications$app_proc_time
summary(applications$app_proc_time)
```

```r
applications = applications %>% drop_na(app_proc_time)

applications$examiner_PT = with(applications, ave(app_proc_time, examiner_id, FUN=mean))
applications$examiner_PT

write.csv(applications, "C:\\Users\\dobri\\OneDrive\\Desktop\\McGill Courses\\ORGB 672\\672_project_data\\APP2.csv", row.names=FALSE)
```

Fig. 1: R code for the new variables

	application_number	filing_date	examiner_name_last	examiner_name_first	examiner_name_middle	examiner_id	e	tc	gender	race	earliest_date	latest_date	tenure_days	app_proc_time	decision_date	examiner_PT
1	08284457	2000-01-26	HOWARD	JACQUELINE	V	96082		1700	female	white	2000-01-10	2016-04-01	5926	1119	2003-02-18	594.3036
2	08413193	2000-10-11	YILDIRIM	BEKIR	L	87678		1700	NA	white	2000-01-04	2016-09-09	6093	685	2002-08-27	752.4583
3	08531853	2000-05-17	HAMILTON	CYNTHIA	NA	63213		1700	female	white	2000-01-06	2017-05-20	6344	-1170	1997-03-04	927.8468
4	08637752	2001-07-20	MOSHER	MARY	NA	73788		1600	female	white	2000-01-04	2017-05-05	6331	1481	2005-08-09	1046.5262
5	08682726	2000-04-10	BARR	MICHAEL	E	77294		1700	male	white	2000-01-03	2017-05-05	6332	261	2000-12-27	795.1788
6	08687412	2000-04-28	GRAY	LINDA	LAMEY	68606		1700	female	white	2000-01-04	2017-05-19	6345	459	2001-07-31	921.3147
7	08765941	2000-06-23	FORD	VANESSA	L	97543		1600	female	white	NA	NA	NA	425	2001-08-22	1342.5204
8	08776818	2000-02-04	STRZELECKA	TERESA	E	98714		1600	female	white	2000-01-21	2017-05-22	6331	892	2002-07-15	1235.2117
9	08809677	2002-02-20	KIM	SUN	U	65530		1700	female	Asian	2000-01-03	2017-05-18	6345	1098	2005-02-22	1015.4815
10	08836939	2000-06-13	WOOD	ELIZABETH	D	77112		1700	female	white	2000-01-05	2017-05-22	6347	644	2002-03-19	988.2744
11	08901519	2000-09-26	DENT	ALANA	HARRIS	92931		1600	female	white	2000-01-03	2017-05-23	6350	294	2001-07-17	1312.9385
12	08913518	2004-04-06	AFTERGUT	JEFFRY	H	75406		1700	male	white	2000-01-10	2017-05-23	6343	693	2006-02-28	1100.0306
13	08930379	2002-04-08	KUMAR	SHAILENDRA	NA	95054		1600	NA	Asian	2000-01-07	2017-05-12	6335	631	2003-12-30	816.7913
14	08945309	2000-06-15	STARSIAK	JOHN	S	99360		1700	male	white	2000-01-04	2017-01-19	6225	1259	2003-11-26	1019.3056
15	08952426	2000-08-21	TRAN	SUSAN	T	73198		1600	female	Asian	2000-01-14	2017-05-19	6335	715	2002-08-06	1265.4411
16	08973360	2000-02-09	LI	QIAN	JANICE	76132		1600	male	Asian	2000-01-12	2017-05-22	6340	1009	2002-11-14	1294.4639
17	08974843	2000-01-11	PEESO	THOMAS	R	77284		2100	male	white	2000-01-03	2017-04-28	6325	1098	2003-01-13	1324.3761
18	08981219	2000-07-27	DAVIS	ROBERT	B	63176		1700	male	white	2000-01-18	2017-05-23	6335	1048	2003-06-10	941.7602

Fig. 2: Generated output showing the new variables

Part 2: Recurring Regressions

After another round of pre-processing, we can continue by regressing the average processing time per officer on their three types of centralities, namely: degree centrality, closeness centrality, and betweenness centrality (see *Fig. 3*). Our first multiple linear regression shows some interesting input, degree centrality is positively correlated with processing time with high statistical significance, meaning that officers with lots of connections are slower in processing applications, likely because they spend more time keeping up with their network; none of the other centralities had any statistically significant correlation (see *Fig. 4*). On average, when degree centrality increases by one connection, processing times increases by 0.8767 days or 21 hours on average holding all else constant (see *Fig. 11*). Thus, we only kept degree centrality and plotted a scatterplot of a simple linear regression of processing time on degree centrality (see *Fig. 5*). We did the same operation by regressing on tenure time to see if any similar correlations appear; we can indeed see a similar correlation level when comparing tenure time and degree centrality, implying that officers with more experience tend to, quite intuitively, also have more connections (see *Fig. 6*).

```
> nodes_ALL
# A tibble: 2,387 × 8
   name  gender race  tenure_days examiner_PT    BC    DC       CC
   <chr> <chr>  <chr>       <dbl>       <dbl> <dbl> <dbl>    <dbl>
 1 66266 female white        6119        825.  44      34 0.0455
 2 84356 male   white        6347       1158.   0       3 0.000439
 3 63519 male   white        6198        898. 163      40 0.0714
 4 98531 female white        6323       1021.   4.5     5 0.0833
 5 93865 male   Asian        4951       1197.   5.67   23 0.5
 6 92953 female Asian        6328       2394.   0      43 0.000500
 7 91818 female Asian        6289       1909.  32.2    28 0.0167
 8 61519 male   Asian        6334       1336.   0      14 0.000465
 9 72253 male   white        6349       1221.  16.5     8 0.5
10 67515 male   Asian        6327       1030.  13.9    16 0.25
# … with 2,377 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Fig. 3: All nodes in the dataset

```
> lm11 = lm(nodes_ALL$examiner_PT ~ nodes_ALL$DC + nodes_ALL$CC + nodes_ALL$BC)
> summary(lm11)

Call:
lm(formula = nodes_ALL$examiner_PT ~ nodes_ALL$DC + nodes_ALL$CC + 
    nodes_ALL$BC)

Residuals:
     Min       1Q   Median       3Q      Max 
-1000.26  -194.24   -15.27   198.36  1095.42 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.244e+03  1.239e+01 100.348  < 2e-16 ***
nodes_ALL$DC  1.271e+00  3.537e-01   3.592 0.000339 ***
nodes_ALL$CC  2.022e+01  2.255e+01   0.897 0.370008    
nodes_ALL$BC -9.606e-04  2.546e-03  -0.377 0.706014    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 293.3 on 1432 degrees of freedom
  (951 observations deleted due to missingness)
Multiple R-squared:  0.009625,	Adjusted R-squared:  0.00755 
F-statistic: 4.639 on 3 and 1432 DF,  p-value: 0.00311
```

Fig. 4: Multiple linear regression of processing time on all types of centrality studied

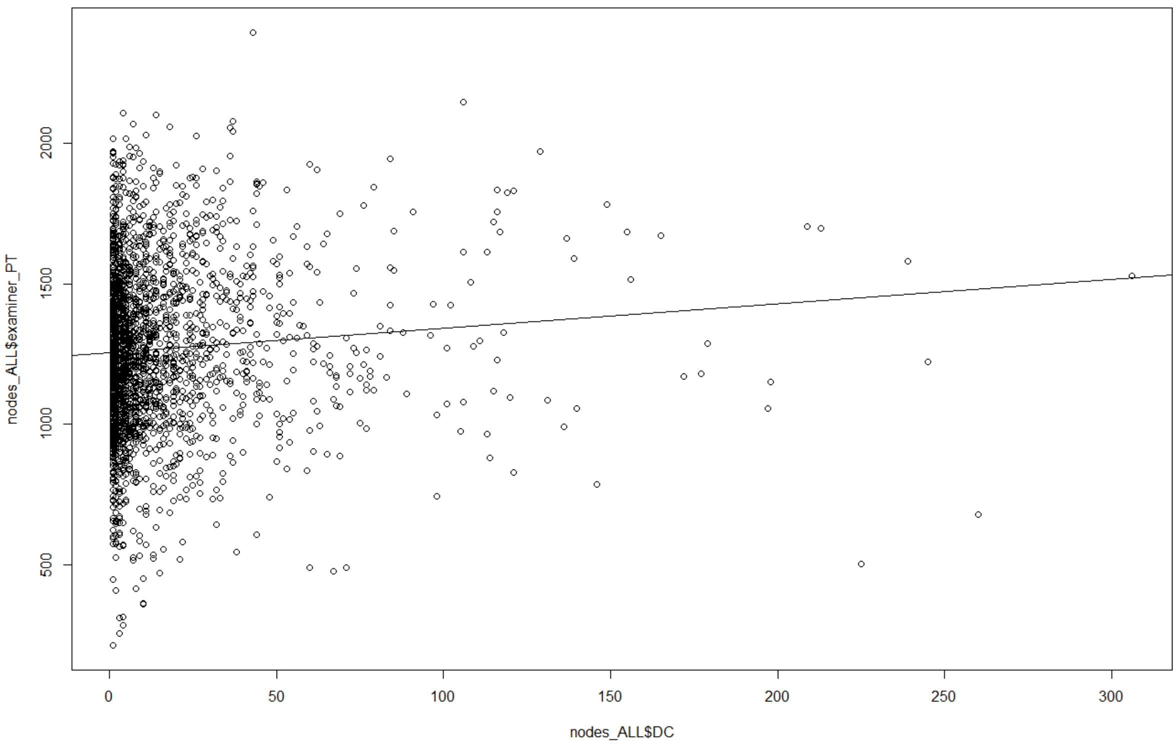

Fig. 5: Scatterplot of the simple liner regression on degree centrality

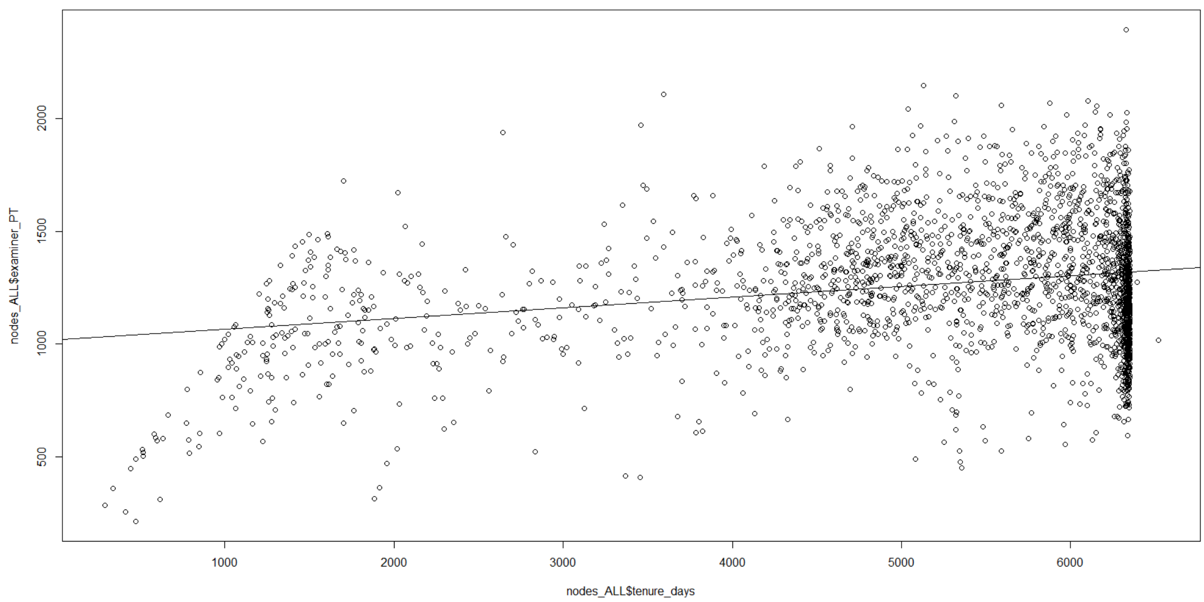

Fig. 6: Idem for tenure days

Part 3: Gender Comparisons

Next, we had to compare performance between both genders. An officer can either be male or female, so gender is a binary variable with values of 1 for male and 0 for female (see *Fig. 7*). After running a regression, we see that holding all else constant males process an application about 24 days slower than women do, which is only 2% slower than the dataset average of 1200 days. We then proceeded by plotting the regression with two trend lines on it: a red one for females and a blue one for males; data points are also colored accordingly (see *Fig. 8*). As expected, the blue line is plotted slightly higher, although the statistical significance is relatively low since we have a p-value of 0.09110 for the gender binary variable.

In addition, we can increase the accuracy of our regression by including an interaction term between gender and degree centrality, as the correlation between the two is not always linear. At higher levels of degree centrality, it is likely that the processing speed gender gap may decrease, or even get inverted. Putting in the interaction term increased the significance level of the gender variable since the p-value decreased to 0.02761; likewise, the interaction term has a p-value of 0.14390, which has relatively low statistical significance (see *Fig. 9*).

Interestingly enough, once plotted on a graph, we can see that the two lines do indeed cross each other. By solving for $(b_0 + b_2) + X * (b_1 + b_3) = (b_0) + X * (b_1)$ we can rearrange and simplify to find that $X = \frac{b_2}{-b_3}$ and after plugging in the coefficients, we can solve for $X = 42.36196$ and then plug in again for $Y = b_0 + X * b_1 = 1220.0889 + 1.4512X$ solving again for $Y = 1281.566$ therefore obtaining the following intersection point:

$$\text{POINT} = (42.36196; 1281.566)$$

We can plot the point on the graph as well (shown in green) and infer that once degree centrality reaches 43 (connections can only be integers) then females start processing applications slower than males on average and holding all other factors constant.

```
> lm55 = lm(nodes_ALL$examiner_PT ~ nodes_ALL$DC + nodes_ALL$gender)
> summary(lm55)

Call:
lm(formula = nodes_ALL$examiner_PT ~ nodes_ALL$DC + nodes_ALL$gender)

Residuals:
    Min      1Q  Median      3Q     Max 
-1041.7  -188.7   -17.2   186.0  1130.3 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         1229.6459    12.7531  96.419  < 2e-16 ***
nodes_ALL$DC           0.7862     0.2554   3.078  0.00211 ** 
nodes_ALL$gendermale  24.2857    14.3667   1.690  0.09110 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 290.3 on 2036 degrees of freedom
  (348 observations deleted due to missingness)
Multiple R-squared:  0.006106,	Adjusted R-squared:  0.005129 
F-statistic: 6.254 on 2 and 2036 DF,  p-value: 0.00196
```

Fig. 7: Multiple linear regression of processing time on degree centrality and gender

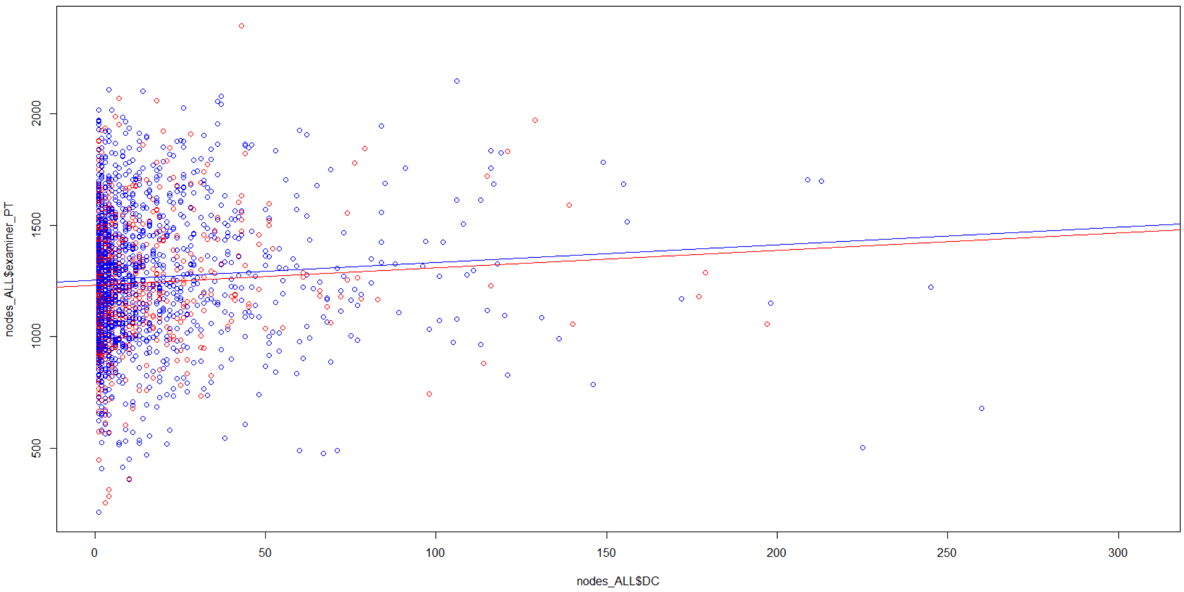

Fig. 8: Colour-coded scatterplot showing the trendlines of both genders

```
> lm66 = lm(nodes_ALL$examiner_PT ~ nodes_ALL$DC + nodes_ALL$gender + nodes_ALL$gender*nodes_ALL$DC)
> summary(lm66)

Call:
lm(formula = nodes_ALL$examiner_PT ~ nodes_ALL$DC + nodes_ALL$gender +
    nodes_ALL$gender * nodes_ALL$DC)

Residuals:
     Min       1Q   Median       3Q      Max
-1044.72  -190.36   -16.71   185.15  1111.29

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                       1220.0889    14.3277  85.156  < 2e-16 ***
nodes_ALL$DC                         1.4512     0.5217   2.782  0.00546 **
nodes_ALL$gendermale                37.0518    16.8088   2.204  0.02761 *
nodes_ALL$DC:nodes_ALL$gendermale   -0.8746     0.5983  -1.462  0.14390
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 290.2 on 2035 degrees of freedom
  (348 observations deleted due to missingness)
Multiple R-squared:  0.007149,  Adjusted R-squared:  0.005685
F-statistic: 4.884 on 3 and 2035 DF,  p-value: 0.002189
```

Fig. 9: Output of the regression with the interaction term added

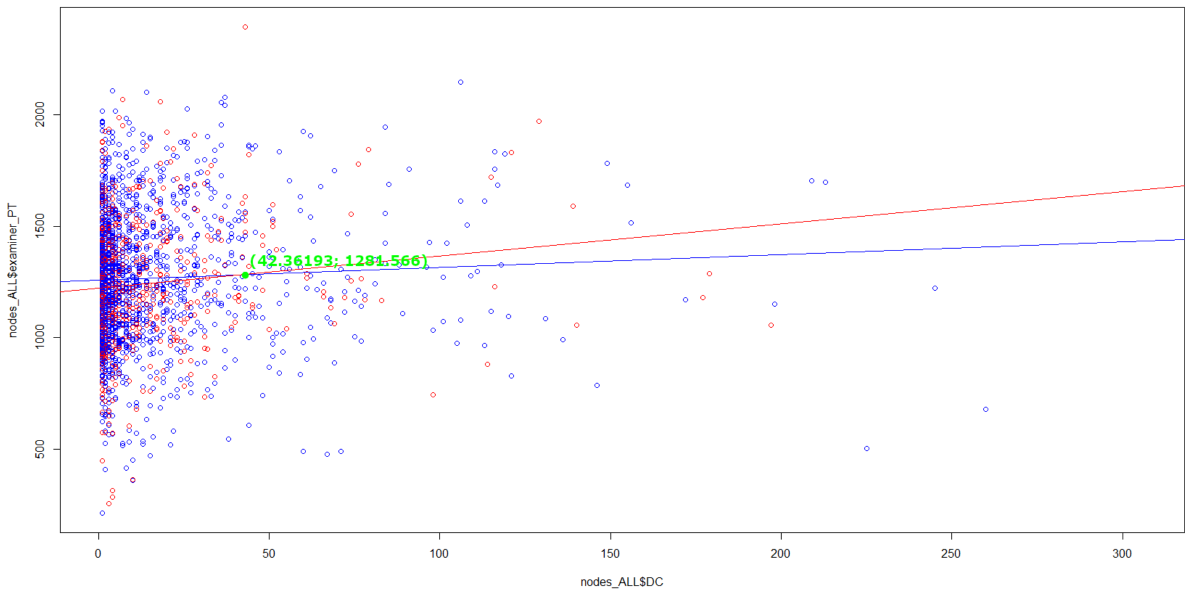

Fig. 10: Updated scatterplot showing crossing trendlines and their point of intersection

```
> lm22 = lm(nodes_ALL$examiner_PT ~ nodes_ALL$DC)
> summary(lm22)

Call:
lm(formula = nodes_ALL$examiner_PT ~ nodes_ALL$DC)

Residuals:
     Min       1Q   Median       3Q      Max
-1041.14  -188.40   -13.58   184.65  1102.82

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1253.2609     6.8092 184.055  < 2e-16 ***
nodes_ALL$DC    0.8767     0.2312   3.792 0.000153 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 287.3 on 2385 degrees of freedom
Multiple R-squared:  0.005992,  Adjusted R-squared:  0.005575
F-statistic: 14.38 on 1 and 2385 DF,  p-value: 0.0001533
```

Fig. 11: Output only when looking at degree centrality

Part 4: Implications for the USPTO & What About My Own Subset?

Regarding implications throughout the whole dataset, we can see that degree centrality plays quite a big role in officer efficiency. Idem for gender, albeit at a smaller significance level.

What I was really curious about was how the three unit subsets that I selected during last week's exercise differed from the whole dataset with all units put together. As a reminder, I selected units #161, #179, and #242. Last time I noticed that unit #161 was the most gender balanced with a 50/50 split but also the least racially diverse with over 75% Caucasian employees. Whereas, unit #242 was the least gender balanced with 4 to 1 male to female ratio (i.e. an 80/20 split) but with only 40% Caucasian employees. Hence, it was intuitive to infer that not all units are the same and that disparities in between are imminent, perhaps even intentionally designed like so.

Therefore, I decided to perform the exact same regression on my specific subset. By far the most striking difference was the fact that closeness centrality was now the most significant coefficient in the initial multiple linear regression (see *Fig. 12*). It had a coefficient of -63.7944 with a p-value of 0.0578, thus quite significant. Being negatively correlated with processing time, this meant that when closeness centrality was maxed out to 1.0 (as it is bounded by zero and one), mean processing time decreased by about more than 2 months holding everything else constant. Meaning that being in the centre of everything and being able to easily reach everyone made officers more efficient as having access to more human input was certainly a contributor to the shorter processing time. Putting in only closeness centrality, the coefficient increased to -66.50 with a more significant p-value of 0.0433, meaning that maxing out actually makes officers 3 months faster at processing holding all else constant (see *Fig. 13*).

Hence, moving forward, I disregarded the other centralities as they were not statistically significant and only kept closeness centrality. When adding in the gender variable, I noticed that significance decreased pretty noticeably as both p-values were now over 0.1 (see *Fig. 14*). Furthermore, the gender coefficient was slightly higher at 39 days, but less significant nonetheless; scatterplot results were also plotted (see *Fig. 15*).

Lastly, I had to perform one last iteration with the interaction term. The interaction term completely cancelled out the significance of the previously hyped over closeness centrality as its p-value was a 0.649 which is frankly laughable (see *Fig. 16*). None of the other coefficients were significant. I also *had* to plot the results, only that to find out that the trendlines do not cross within the graph bounds (see *Fig. 17*).

In conclusion, when looking at the entire dataset, we can observe that higher degree centrality is correlated with higher processing time since larger networks require more time to maintain and thus less time is spent on managing applications. One extra connection gained increases processing time by 21 hours holding all else constant. Moreover, males are about a month slower than females when looking at officers with a smaller number of connections; however, once the officer's network grows over 43 people, females become slower, at up to 140 days slower in certain cases. Nevertheless, some of the more specialized units such as #242, #179 and #161 exhibit special characteristics of their own. For example, closeness centrality was more impactful towards average processing time, suggesting that access to human input can speed things up, especially if it is in a unit full of experienced officers.

```
> lm1 = lm(nodes_APP$examiner_PT ~ nodes_APP$DC + nodes_APP$CC + nodes_APP$BC)
> summary(lm1)

Call:
lm(formula = nodes_APP$examiner_PT ~ nodes_APP$DC + nodes_APP$CC +
    nodes_APP$BC)

Residuals:
    Min      1Q  Median      3Q     Max
-753.60 -166.82   -0.12  144.50  890.89

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1220.4113    19.1590  63.699   <2e-16 ***
nodes_APP$DC    0.6654     0.6488   1.026   0.3057
nodes_APP$CC  -63.7944    33.5351  -1.902   0.0578 .
nodes_APP$BC   -0.3186     0.2654  -1.200   0.2306
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 245.2 on 437 degrees of freedom
  (256 observations deleted due to missingness)
Multiple R-squared:  0.01391,   Adjusted R-squared:  0.007137
F-statistic: 2.054 on 3 and 437 DF,  p-value: 0.1056
```

Fig. 12: Results of the subset regression, showing significant closeness centrality

```
> lm2 = lm(nodes_APP$examiner_PT ~ nodes_APP$CC)
> summary(lm2)

Call:
lm(formula = nodes_APP$examiner_PT ~ nodes_APP$CC)

Residuals:
   Min     1Q Median     3Q    Max
-752.0 -168.6   -4.9  144.5  885.5

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1227.03      15.98  76.785   <2e-16 ***
nodes_APP$CC   -66.50      32.81  -2.027   0.0433 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 245.2 on 439 degrees of freedom
  (256 observations deleted due to missingness)
Multiple R-squared:  0.00927,   Adjusted R-squared:  0.007013
F-statistic: 4.108 on 1 and 439 DF,  p-value: 0.0433
```

Fig. 13: Regression results using only closeness centrality

```
> lm5 = lm(nodes_APP$examiner_PT ~ nodes_APP$CC + nodes_APP$gender)
> summary(lm5)

Call:
lm(formula = nodes_APP$examiner_PT ~ nodes_APP$CC + nodes_APP$gender)

Residuals:
    Min      1Q  Median      3Q     Max
-760.00 -165.16    3.19  144.10  875.97

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           1194.61      25.28  47.255   <2e-16 ***
nodes_APP$CC           -53.53      34.53  -1.550    0.122
nodes_APP$gendermale    39.33      26.64   1.477    0.141
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 245.6 on 393 degrees of freedom
  (301 observations deleted due to missingness)
Multiple R-squared:  0.012,     Adjusted R-squared:  0.006977
F-statistic: 2.388 on 2 and 393 DF,  p-value: 0.09319
```

Fig. 14: Regression with only closeness centrality and gender (less significance)

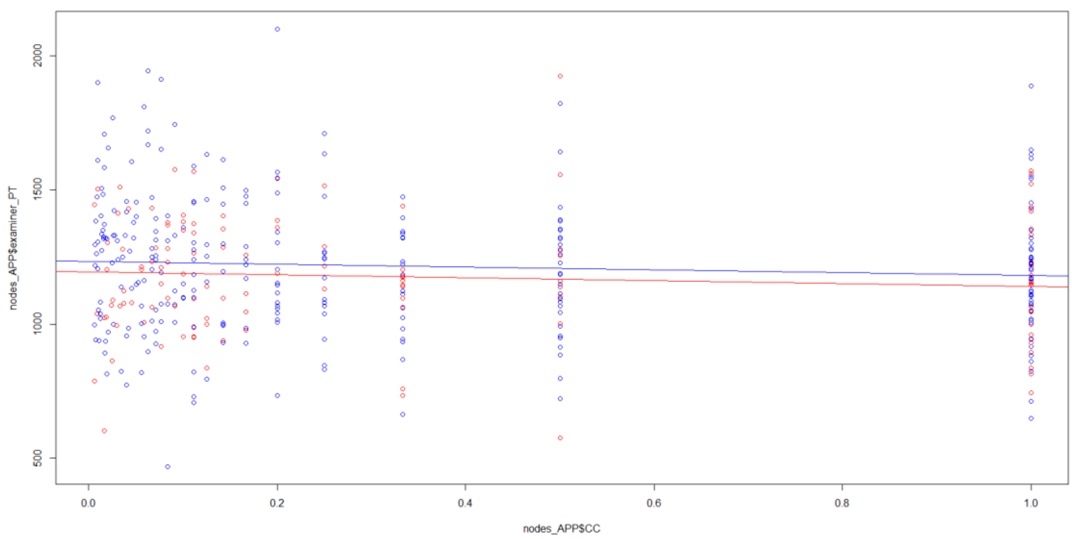

Fig. 15: Gender scatterplot results (without the interaction term)

```
> lm6 = lm(nodes_APP$examiner_PT ~ nodes_APP$CC + nodes_APP$gender + nodes_APP$gender*nodes_APP$CC)
> summary(lm6)

Call:
lm(formula = nodes_APP$examiner_PT ~ nodes_APP$CC + nodes_APP$gender +
    nodes_APP$gender * nodes_APP$CC)

Residuals:
   Min     1Q Median     3Q    Max
-763.1 -159.0    0.1  144.7  874.4

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        1185.28      30.85  38.423   <2e-16 ***
nodes_APP$CC                        -27.45      60.23  -0.456    0.649
nodes_APP$gendermale                 52.85      36.93   1.431    0.153
nodes_APP$CC:nodes_APP$gendermale   -38.89      73.54  -0.529    0.597
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 245.8 on 392 degrees of freedom
  (301 observations deleted due to missingness)
Multiple R-squared:  0.01271,	Adjusted R-squared:  0.005153
F-statistic: 1.682 on 3 and 392 DF,  p-value: 0.1703
```

Fig. 16: Regression output with the interaction term (even less significance)

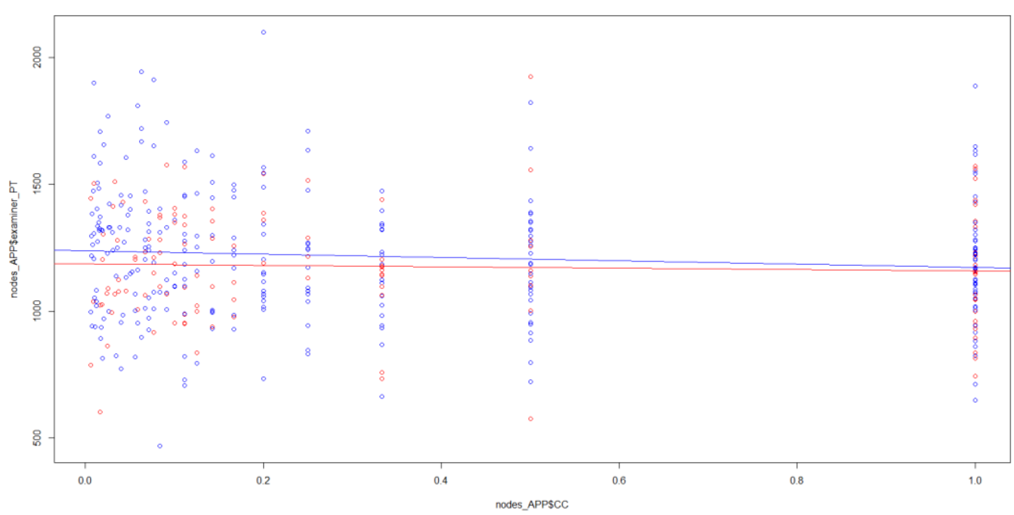

Fig. 17: Gender scatterplot results (with the interaction term)