Student ID: 224545
Final Data Analysis
Statistics 1
Prof. Mark Kayser

Task 1.

a) IV = committee
DV = industry

```
data:  task.1.a

             no trade       trade
no com    91.921841  8.078159
yes com   97.206245  2.793755

X-squared = 43.207, df = 1, p-value = 4.924e-11
```

b) The result is statistically significant because the p-value is below 0.05. Indeed, we have a X-squared of 43, much bigger than 3.84, which is the minimum value of the X-squared we need to assess the finding as statistically significant. Therefore, we can reject the null hypothesis that there is no relationship between being a member of the congress sit in the Natural Resources and/or the Climate Crisis committees, and trading in the Oil, Gas and Consumable Fuels; Automobiles; or Metals & Mining sectors.

c) Members who are in the committee are less likely to invest compared to the non members.

d) The difference is 5.284404%. **Being in the committee is associated with a decrease of around 5.28% of probability to invest** in Oil Gas and Consumable Fuels; Automobiles; or Metals & Mining sectors.

Task 2.

a) data: task.2.a
X-squared = 791.26, df = 1, p-value < 2.2e-16
There is a statistically significant relationship because the p-value is less than 0.05.

b) Republicans trade more in this sector. **Being in the Republican Party is associated with an increase of around 13.52% of probability to invest** in Oil Gas and Consumable Fuels; Automobiles; or Metals & Mining sectors.

Task 3.

a) Running the t-test, we can see that **democrats trade more that republicans**: the mean value of the total number of trades of the republicans (x) in the period is 59.69, while the democrats' (y) is 113.94.

b) Running the t-test, we see that the p-value is below 0.05; therefore the results are statistically significant. Indeed, the absolute value of the t-value is 2.02. Being it higher then 1.96, it is statistically significant. We can reject H0 saying that there are no differences between democrats and republicans regarding the trading behaviors.

Task 4.

Experience and acquaintances in the Congress can increase influence of members and provide privileged access to extensive information on the stock market of certain sectors. This could have an effect on the trading behavior of the congresspersons, leading to a possible conflict of interests between their private interest and the public service. Insider trading occurs when someone trades a security (such as a stock) while possessing material, non-public information about that security. The STOCK Act, which was signed into law in 2012, specifically prohibits members of Congress, as well as certain other government employees, from using non-public information for personal gain in the stock market. However, there are evidences of insider trading among the members of the congress[1]. There is evidence to think that more experienced members of Congress have more extensive access to non-public information, and could use that information to gain an unfair advantage in the stock market by doing insider trading. Furthermore, the relation between the tenure and the position in the stock market can be affected by the political and legislative power that the members have, as more activity can mean more influence in the Congress, and therefore more capacity to access non-public information on relevant topics[2].

This data analysis will test if, among the members of congress who trade in the stock market, those who have longer tenure and that are more politically and legislative active are more prone to have a better position in the stock market. Indeed, experience can be referred as the length of time a member has served in Congress. Members with longer tenures may have more familiarity with the legislative process and the inner workings of Congress, and may be more influential within the institution. As said, the relation between experience and trading behavior can be affected by the political influence that a member has within the Congress, as the more a member is politically and legislatively active, the more political power is attributed to her or him.

For this analysis, we will use the dataset *trades_aggregate.csv,* collecting the aggregate trading capacities by Members of Congress, and therefore, the analysis will test if we can safely generalize our results to all the Members of Congress that engage in trading of stocks in the relevant period. The independent variable used will be the expertise of the Members of the Congress, indicated by the number of years the Members have served, and operationalized by the variable *years_service*, conditional to the variable *bill_proposed*, which operationalize the legislative and political activity. Consequently, the main research question will be the following: *What is the extent of the impact of tenure in Congress on the position of the members in the stock market conditional on their political and legislative activity?*

H1: *An increase of years of service of the Congress-persons is associated with an increase in the position of the members in the stock market, conditional to the number of bill proposed.*
H0: *An increase of years of service of the Congress-persons is not associated with an increase in the position of the members in the stock market, conditional to the number of bill proposed.*

To be more precise, the model specification will include a set of control variables that could influence the relationship between experience and stock market, such as party membership, gender, the total member's activity in the stock market, and the geographical area. These control variable can be relevant for the relation between experience and the position of the members in the stock market. Particularly, the variables *area* is the result of a different operationalization of the variable *district. Area* report whether a representative has been elected in a district where the majority of population reside in urban or rural areas. It has been created by mutate the variable *district* into the categorical variable *area* that can assume the values of *urban* or *rural.* Such variables can influence the position in the stock market of members. In particular, I will include *sum_trades_absolutevalues* to operationalize the total members' activity in the stock market, that inevitably influences the stock market position, the dummy variable *party* to account of ideological differences that can influence the behavior, as well as the dummy variable *gender* and *area*, that respectively take into account the behavioral differences in trading preferences and aversion to risk that can possibly derive from gender and other differences in availability of information derived from the fact of being elected from a urban or rural area. Therefore, the regression model will be:

*net_trades_absolutevalues = a + b1year_service + b2bill_proposed + b3year_service*bill_proposed + b4sum_trades_absolutevalue + b5party + b6gender + b7area.*

[1] https://www.nytimes.com/interactive/2022/09/13/us/politics/congress-stock-trading-investigation.html

[2] Harvison, Thuong, Political Connections and Insider Trading (May 13, 2019). Available at SSRN: https://ssrn.com/abstract=3387495 or http://dx.doi.org/10.2139/ssrn.3387495

Before discussing the regression model, it is appropriate to show the descriptive statistics of the data we have.

Summary Statistics

Statistic	N	Mean	Median	St. Dev.	Min	Max	Pctl(25)	Pctl(75)
...1	153	77.0	77	44.3	1	153	39	115
congress_year	153	2,012.3	2,015	8.7	1,981	2,021	2,007	2,019
count	153	85.2	20	162.0	1	1,177	4	74
bill_proposed	153	5.2	5	2.1	0	10	4	6
year_service	153	9.7	7	8.7	1	41	3	15
net_trades_absolutevalue	153	1,116,763.0	192,289.7	3,371,568.0	457.0	27,210,848.0	39,372.7	520,216.1
sum_trades_absolutevalue	153	5,026,281.0	549,949.5	17,449,198.0	7,002.8	139,122,934.0	119,743.9	2,326,122.0
party_dummy	153	0.5	1	0.5	0	1	0	1
gender_dummy	153	0.8	1	0.4	0	1	1	1
area_dummy	153	0.6	1	0.5	0	1	0	1

Notice that, for dummy variables:

- party_dummy: value of 1 takes the value of "republican"
- gender_dummy: value of 1 takes the value of "male"
- area_dummy: value of 1 takes the value of "urban"

It is important to notice that the most important measure of central tendency for categorical variable is the mode, being republican (81) for *party*, men (116) for *gender*, and urban (89) for *area.*

From these summary statistics, it appears evident that some variables are not normally distributed: for example our dependent variable *net_trades_absolutevalues* is extremely right-skewed, as well as the variable *sum_trades_absolutevalue.* We can notice it because the median of both is much lower than the mean. The same is true for *year_service* and *count.* We can check by plotting the distribution of these variables using the histograms. I will take our dependent variable as an example, but the same reasoning applies to the other right-skewed variables. From the Figure 1 we can notice the skewness, while Figure 2 shows the distribution after having logged the variable.

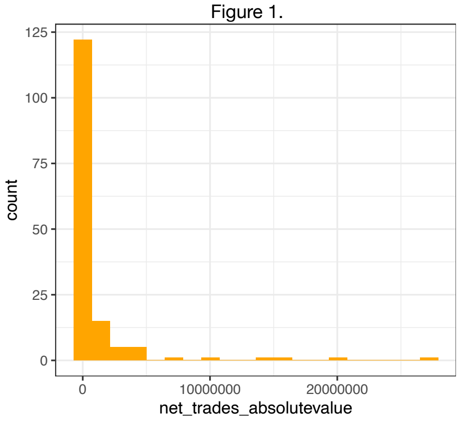


Figure 1.

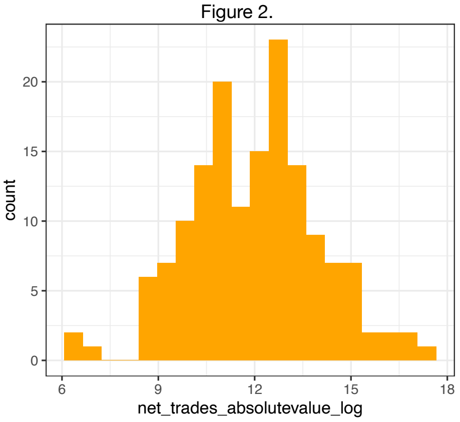


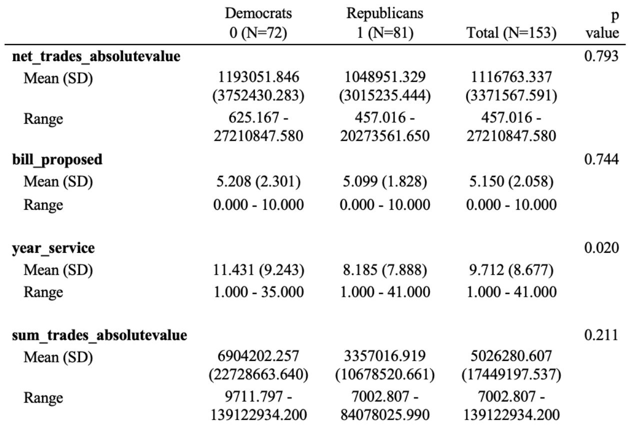
Figure 2.

It is necessary to log all the other right-skewed variables for the OLS assumptions to be valid and for the model to be BLUE.

Before continuing, it can be interesting to confront the values of the summary statistics among the dummy control variables.

Summary statistics in reference to *party*:

Figure 3.

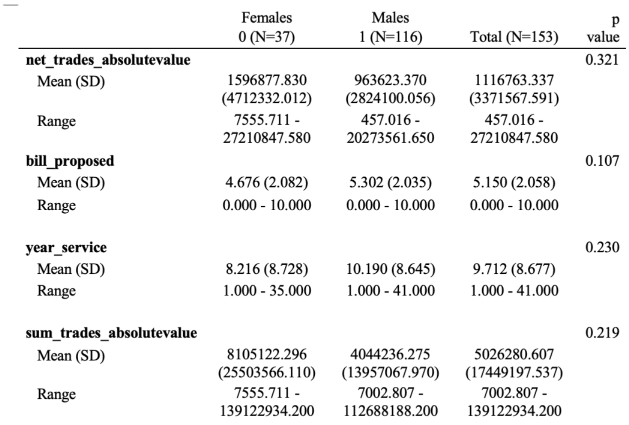


	Democrats 0 (N=72)	Republicans 1 (N=81)	Total (N=153)	p value
net_trades_absolutevalue				0.793
Mean (SD)	1193051.846 (3752430.283)	1048951.329 (3015235.444)	1116763.337 (3371567.591)	
Range	625.167 - 27210847.580	457.016 - 20273561.650	457.016 - 27210847.580	
bill_proposed				0.744
Mean (SD)	5.208 (2.301)	5.099 (1.828)	5.150 (2.058)	
Range	0.000 - 10.000	0.000 - 10.000	0.000 - 10.000	
year_service				0.020
Mean (SD)	11.431 (9.243)	8.185 (7.888)	9.712 (8.677)	
Range	1.000 - 35.000	1.000 - 41.000	1.000 - 41.000	
sum_trades_absolutevalue				0.211
Mean (SD)	6904202.257 (22728663.640)	3357016.919 (10678520.661)	5026280.607 (17449197.537)	
Range	9711.797 - 139122934.200	7002.807 - 84078025.990	7002.807 - 139122934.200	

We can notice that the mean of net trade is higher for democrats, the mean of the number of bill proposed is slightly higher, and the mean of the number of years of service is higher as well.

Summary statistics in reference to *gender*:

Figure 4.

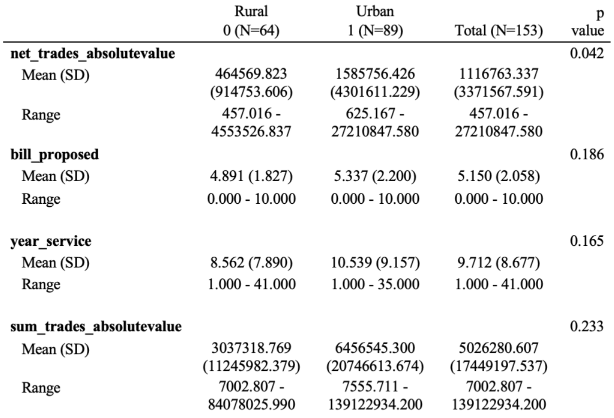


	Females 0 (N=37)	Males 1 (N=116)	Total (N=153)	p value
net_trades_absolutevalue				0.321
Mean (SD)	1596877.830 (4712332.012)	963623.370 (2824100.056)	1116763.337 (3371567.591)	
Range	7555.711 - 27210847.580	457.016 - 20273561.650	457.016 - 27210847.580	
bill_proposed				0.107
Mean (SD)	4.676 (2.082)	5.302 (2.035)	5.150 (2.058)	
Range	0.000 - 10.000	0.000 - 10.000	0.000 - 10.000	
year_service				0.230
Mean (SD)	8.216 (8.728)	10.190 (8.645)	9.712 (8.677)	
Range	1.000 - 35.000	1.000 - 41.000	1.000 - 41.000	
sum_trades_absolutevalue				0.219
Mean (SD)	8105122.296 (25503566.110)	4044236.275 (13957067.970)	5026280.607 (17449197.537)	
Range	7555.711 - 139122934.200	7002.807 - 112688188.200	7002.807 - 139122934.200	

When looking at the summary statistics between males and females, we notice that females have an higher *net_trade_ absolutevalue* compared to the males, but they have proposed less bills and have a shorter tenure.

Summary statistics in reference to *area:*

Figure 5.



	Rural 0 (N=64)	Urban 1 (N=89)	Total (N=153)	p value
net_trades_absolutevalue				0.042
Mean (SD)	464569.823 (914753.606)	1585756.426 (4301611.229)	1116763.337 (3371567.591)	
Range	457.016 - 4553526.837	625.167 - 27210847.580	457.016 - 27210847.580	
bill_proposed				0.186
Mean (SD)	4.891 (1.827)	5.337 (2.200)	5.150 (2.058)	
Range	0.000 - 10.000	0.000 - 10.000	0.000 - 10.000	
year_service				0.165
Mean (SD)	8.562 (7.890)	10.539 (9.157)	9.712 (8.677)	
Range	1.000 - 41.000	1.000 - 35.000	1.000 - 41.000	
sum_trades_absolutevalue				0.233
Mean (SD)	3037318.769 (11245982.379)	6456545.300 (20746613.674)	5026280.607 (17449197.537)	
Range	7002.807 - 84078025.990	7555.711 - 139122934.200	7002.807 - 139122934.200	

Looking at the area, we see that the volume of *net_trade* in the urban areas is higher in respect to the rural areas.
Moreover, representatives elected in the urban areas proposed more bills and have longer tenure.

We can also graphically show the relationship between the dummy variables and the value of *net_trade_absolutevalue* using box-plots:

Figure 6.

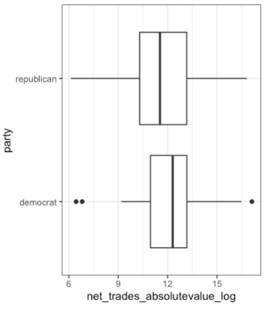


Figure 7.

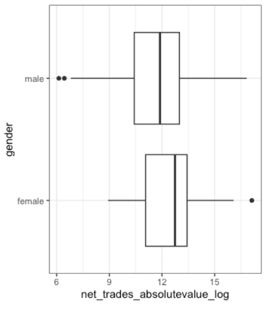
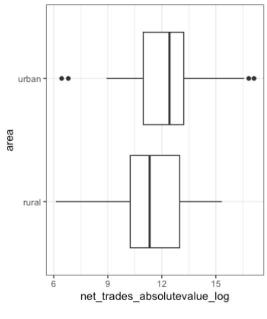


Figure 8.



To have a first look at the correlations between the variables we want to investigate, it is useful to visualize them graphically with a plot. Figure 9. shows the scatter plot between *year_service* and *net_trade_absolutevalue_log*, while Figure 10. indicates the correlation between *bill_proposed* and *net_trade_absolutevalue_log.*

Figure 9.

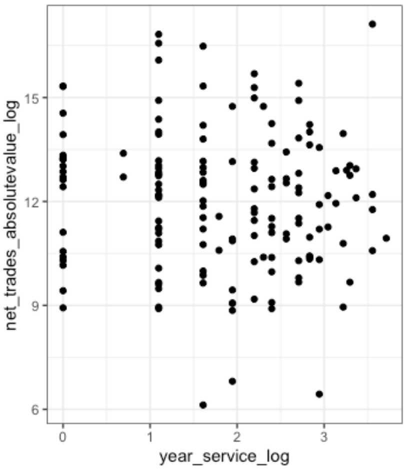
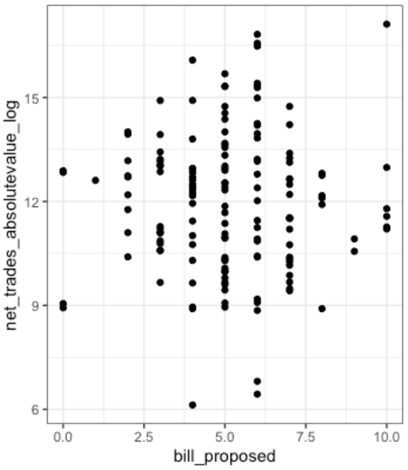


Figure 10.



It is already evident that there is a poor correlation between the variables. Indeed, the correlation expressed in Person's r between *year_service* and *net_trade_absolutevalue_log* is only of 0.043, while the correlation between *bill_proposed* and *net_trade_absolutevalue_log* is of 0.151.

To assess the association of an increase in year of service to an increase in the stock market position we need to run our regression. The first model is a naïve one: it only takes into account the years of service and the bill proposed

net_trades_absolutevalues_log = a + b1year_service_log + b2bill_proposed

This model shows that, everything else constant, a percentage unit increase in year service is associated with a percentage decrease of net trade of 0.172, and that a unit increase in bill proposed (one bill) is associated with a percentage increase of net trade of 0.030. However, these results are not statistically significant as the p-value will be higher than 0.05. Indeed, the t-values calculated as t = (b-H0)/SE are respectively of 0.654 and 0.37 (both much smaller than 1,96). Moreover, this model does not explain almost anything of our dependent variable, as R2 value is 0.008, therefore extremely low. The first model suffered of several omitted variable, but especially, it missed one important variable that cannot be omitted, namely the total value of trades executed, as the position is relative to the total. Therefore, the second model will include *sum_trades_absolutevalues_log* as a control variable.

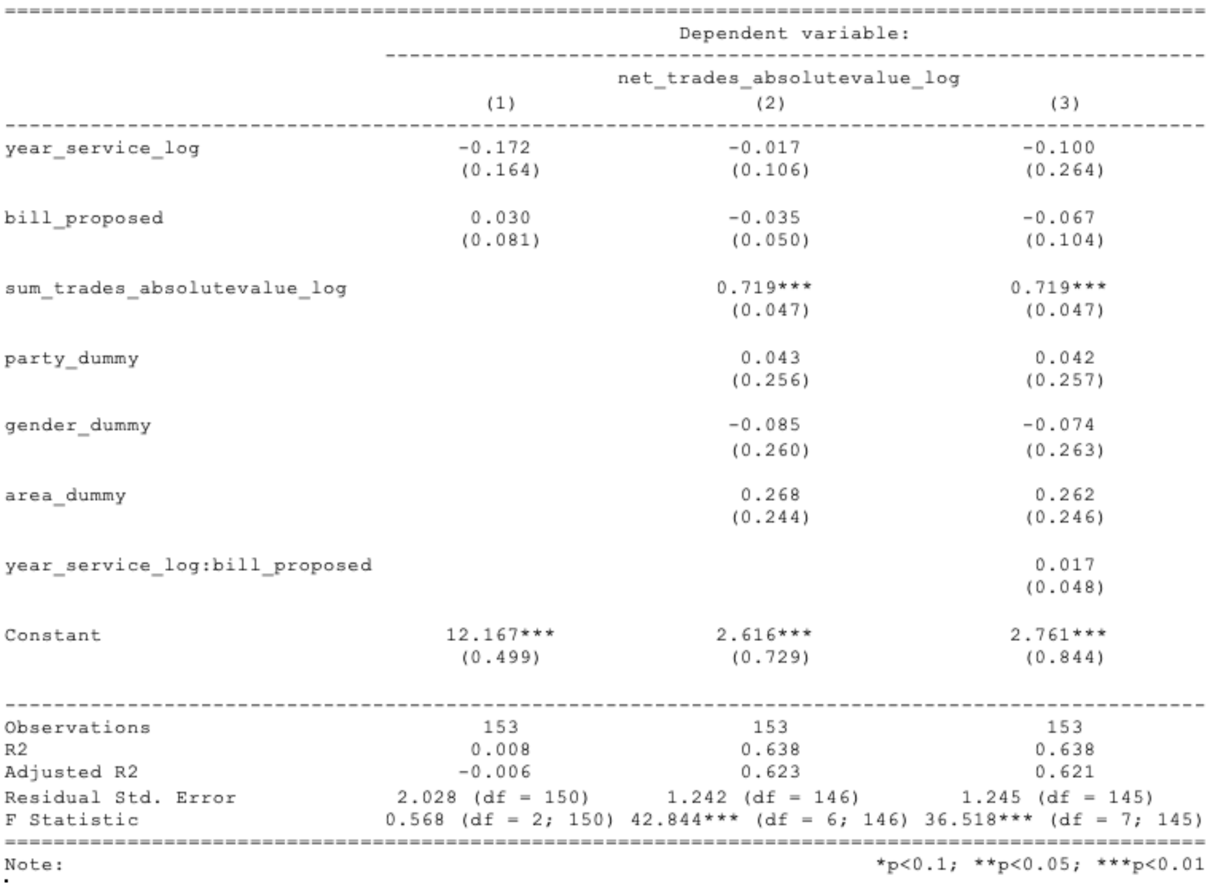
net_trades_absolutevalues_log = a + b1year_service_log + b2bill_proposed +
b3sum_trades_absolutevalues_log + b4party_dummy + b5gender_dummy + b6area_dummy

With this model we find a statistically significant association of one percentage unit increase of total value of trade (*sum_trades_absolutevalues*) with a percentage increase of 0.73 in net value. However, the coefficients for *year_service* and *bill_proposed* are still statistically not significant, and now *bill_proposed* became negative. Moreover, it is important to notice that by adding the new variables, the R2 extensively increased, being now equal to 0.634. Still, the only coefficient that is statistically significant, and who's coefficient is substantially significant is the variable *sum_trade_absolutevalue*, which seems to carry out most of the explanatory work. As you can see from the output (Figure 12), the only variable that could have influenced the R-squared is *sum_trade_absolutevalue* because its coefficient is big enough to comport a so drastic change in the R-squared, since the other coefficients are too small, beside being not statistically significant.

Finally, we can run the interaction we are testing for. Figure 11 compares the outputs of the three models.

net_trades_absolutevalues_log = a + b1year_service_log + b2bill_proposed +
b3sum_trades_absolutevalues_log + b4party_dummy + b5gender_dummy + b6area_dummy +
*b7year_service_log*bill_proposed*

Figure 11.



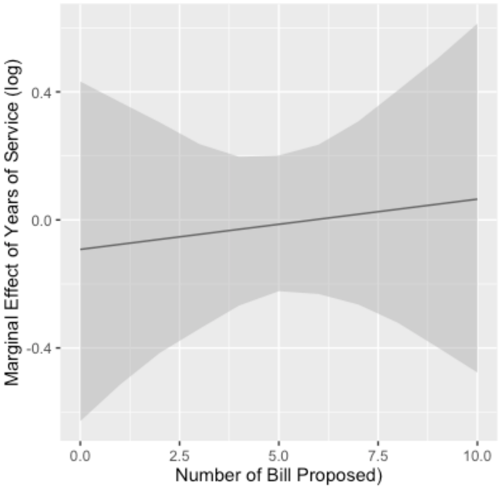
```
==================================================================================================
                                               Dependent variable:
                                  ----------------------------------------------------------------
                                           net_trades_absolutevalue_log
                                      (1)                  (2)                     (3)
--------------------------------------------------------------------------------------------------
year_service_log                    -0.172               -0.017                  -0.100
                                    (0.164)              (0.106)                 (0.264)

bill_proposed                        0.030               -0.035                  -0.067
                                    (0.081)              (0.050)                 (0.104)

sum_trades_absolutevalue_log                              0.719***                0.719***
                                                         (0.047)                 (0.047)

party_dummy                                               0.043                   0.042
                                                         (0.256)                 (0.257)

gender_dummy                                             -0.085                  -0.074
                                                         (0.260)                 (0.263)

area_dummy                                                0.268                   0.262
                                                         (0.244)                 (0.246)

year_service_log:bill_proposed                                                    0.017
                                                                                 (0.048)

Constant                            12.167***             2.616***                2.761***
                                    (0.499)              (0.729)                 (0.844)

--------------------------------------------------------------------------------------------------
Observations                          153                  153                     153
R2                                   0.008                0.638                   0.638
Adjusted R2                         -0.006                0.623                   0.621
Residual Std. Error           2.028 (df = 150)      1.242 (df = 146)        1.245 (df = 145)
F Statistic                 0.568 (df = 2; 150) 42.844*** (df = 6; 146) 36.518*** (df = 7; 145)
==================================================================================================
Note:                                                               *p<0.1; **p<0.05; ***p<0.01
```

Analyzing the outputs we find that, holding everything else constant, the interaction coefficient is b1+b7*number of bill proposed (we can z the number of bill proposed). dy/dx is indeed the marginal effect of tenure on net trade volume, that interacts with the number of bill proposed. We notice that the coefficient for the interaction is positive. Keeping fixed the years, we also notice that for every bill proposed, the net trade volume diminishes by -0.067. Keeping fixed the bill proposed instead, we see that at the one percentage unit increase in years, the net trade volume diminishes indistinctly from the number of bill proposed.

Now we consider the interaction: for every bill proposed, we see that there is an association of 0.017 increase in net trade. More precisely, if no bill are proposed, there is an association of -0.1 variation of net trade volume for a percentage unit increase in years (indeed, if z = 0, namely no bills are proposed, the relevant coefficient for the interaction is b1, namely -0.1), while for each bill proposed, the relevant coefficient will be -0.1 + 0.017*z, namely b1+b7z. As a consequence, we notice that increasing the number bill proposed, the relation between tenure and net trade volume becomes more and more positive. This relation is easily visualizable in the Figure 12 below. In the figure, the points in the line identify all the possible values of the marginal effect (the relevant coefficient to interpret the interaction) at the variation of the number of bill proposed. Still, this marginal effect plot has at least one flaw: the number of bill proposed here is indicated as a continuous variable, but in truth it should be interpreted as a categorial variable. Indeed, a representative cannot propose a decimal value of bill (e.g. 3.46 bills). Therefore, to make sense of this variable, we should only consider integers. We could have coded number of bills as a categorical variable that can assume only the value from 0 to 10. In this way we would have visualize the marginal effect in a slightly different way. Anyhow, the interpretation is the same, but coding the *bill_proposed* variable as a interval variable we have to be sure to interpret correctly the Figure 12, namely consider as valid only the integers values of *bill_proposed* comprised between 0 and 10 that we see in the plot, as the others have no substantive meaning. Moreover, we need to notice that this marginal effect is not statistically significant, as the lower bound of the confidence interval (the dark grey area) is always below the zero.

We can also notice this by the output, that clearly indicates that there is no statistical significance for the interaction term. We can check it using the t-value formula (t = b - H0 / SE = 0.017 / 0.048 = 0.35 < 1,96).

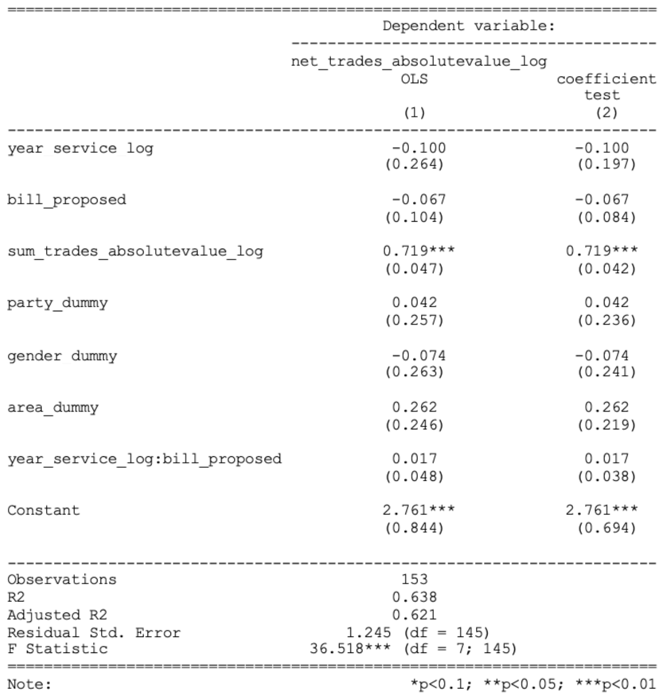
Figure 12.



As already mentioned, by comparing the three three models, we notice that the R-squared value for the last two models is much larger than the first. The R-squared of the second and the third models is very the same, meaning that they have the same explanatory power: they can explain the 63.8% of the change of net trade volumes. Still, looking at the Adjusted R-squared, we notice something unexpected, namely that this value is slightly larger for the model 2. This means that the interaction does not actually increase the explanatory value of the model, but indeed, slightly decrease it. To be sure, neither of the models are statistically significance, therefore the values in general are not really valuable. Anyhow, as mentioned before, it seems quite evident that the main explanatory driver is the variable *sum_trade_absolutevalue*, but this does not tells us much for our analysis.

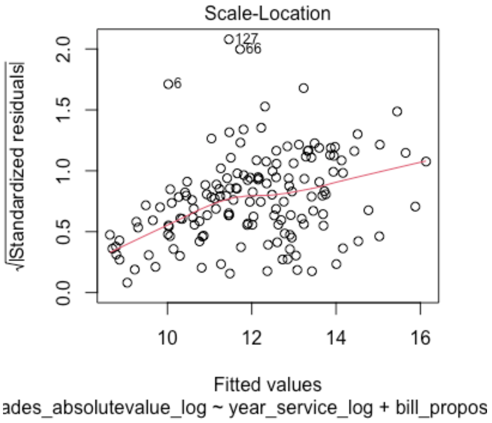
The last part of the analysis will focus on make sure that the main OLS assumptions are respected and detect possible problem in the formulation of the model 3. First of all we test for homoskedasticity, implying that the variance of errors cannot depend on the explanatory variables. The Breusch-Pagan test gives the result of 2.9893 with a p-value of 0.886, which fail to reject null hypothesis of heteroskedasticity, therefore concluding that our model is homoskedastic. Visually, we can plot the error and see if the red line is straight and horizontal with zero intercept. Since oddly it is not the case (Figure 14) even though the Breusch-Pagan test gave us a p-value over 0.05, it is better to use the robust standard error to evaluate our model. Indeed, 0.886 is more the 0.05, but it is still a quite low value. From Figure 13 we can see the comparison between the original Model 3 (1) and the Model 3 using RSE (2). We notice that no coefficient have changed, but obviously the standard error is reduced.

Figure 13.



```
==========================================================================
                                         Dependent variable:
                                  ----------------------------------------
                                  net_trades_absolutevalue_log
                                             OLS              coefficient
                                                                  test
                                             (1)                  (2)
--------------------------------------------------------------------------
year service log                           -0.100               -0.100
                                          (0.264)              (0.197)

bill_proposed                              -0.067               -0.067
                                          (0.104)              (0.084)

sum_trades_absolutevalue_log               0.719***             0.719***
                                          (0.047)              (0.042)

party_dummy                                 0.042                0.042
                                          (0.257)              (0.236)

gender dummy                               -0.074               -0.074
                                          (0.263)              (0.241)

area_dummy                                  0.262                0.262
                                          (0.246)              (0.219)

year_service_log:bill_proposed              0.017                0.017
                                          (0.048)              (0.038)

Constant                                   2.761***             2.761***
                                          (0.844)              (0.694)

--------------------------------------------------------------------------
Observations                                 153
R2                                          0.638
Adjusted R2                                 0.621
Residual Std. Error                   1.245 (df = 145)
F Statistic                        36.518*** (df = 7; 145)
==========================================================================
Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

Figure 14.



Checking for normality of residuals, we see that it is respected, as showed by Figure 15, as well as linearity (Figure 16).

Figure 15.

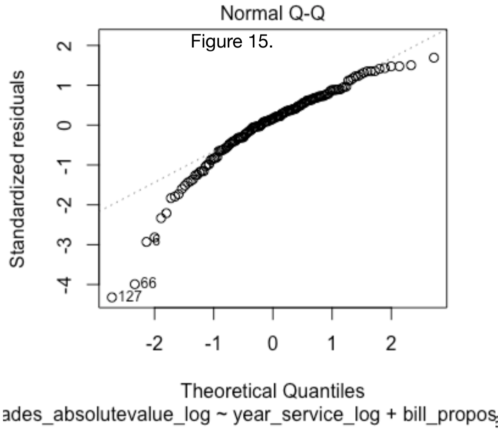
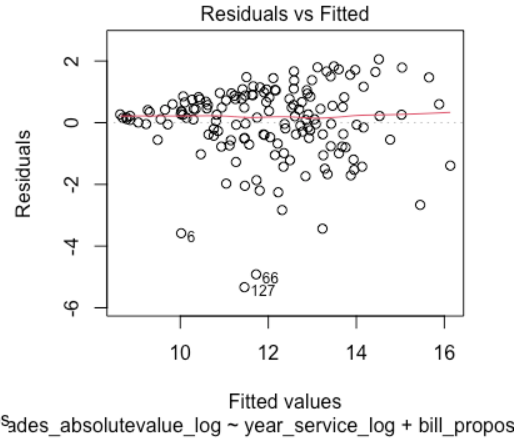


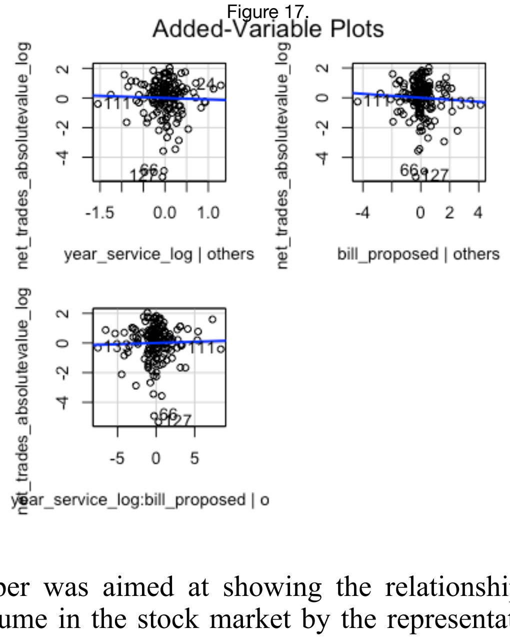
Figure 16.



Testing for multicollinearity by performing a VIF test, we notice high multicollinearity between the terms of the interaction, but no multicollinearity among the other terms. We decide to take into account of the interaction by running a VIF, type = “predictor”. In this way we obtain GVIF values that are all around 1, and never larger that 1.6. In this way, we are sure that multicollinearity is not a problem, as at worse, the √1.6 = 1.265*SE, not significantly inflating the relevant standard error. Moreover, if we consider GVIF^(1/(2*Df)), the highest value we encounter is 1.275, which gives an even smaller inflating value for the standard error.

We finally decide to check for outliers. From the Added-Variable plot of our main independent variables (Figure 17) we notice that 127 and 66 are outliers, but they have no high influence on our main independent variable (they have high discrepancy but low leverage, leading to a low influence, as they are not enough influent to tilt the line). We expect them not to have influence on the regression, so we decide not to run a fourth model.

Figure 17.



In conclusion, this paper was aimed at showing the relationship between the experience in the Congress and the net trade volume in the stock market by the representatives, conditioned by the political and legislative activity. After our analysis, evaluating three different models and after having checked for the OLS assumptions, we need to conclude that none of the three models have statistical significance, and therefore, the null hypothesis cannot be rejected, which equals to say that the alternative hypothesis cannot be accepted. Therefore, the the level of political experience expressed in years of service (tenure) have no influence on the net trade volume, conditional on the number of bill proposed.

Appendix.
R script.

```
setwd("/Users/giocopp/Library/Mobile  Documents/com~apple~CloudDocs/Desktop/Unive/Hertie
School/1st Semester/Statistics 1/Lab/R scripts and data")
rm(list=ls())

library(readr)
library(gmodels)
library(ggplot2)
library(stargazer)
library(car)
library(carData)
library(modelsummary)
library(corrplot)
library(dplyr)
library(AER)
library(corrplot)
library(lmtest)
library(survival)
library(zoo)
library(sandwich)

trades_aggregate <- read_csv("trades_aggregate.csv")
trades_individual <- read_csv("trades_individual.csv")

View(trades_individual)
View(trades_aggregate)

# Start with the individual-level dataset. Consider the variable committee select.
# This is coded as a dummy variable, where 1 denotes whether Members of
# Congress sit in the Natural Resources and/or the Climate Crisis committees
# or not (0). Then consider the variable industry select; this is a dummy variable
# which is coded as 1 if Members of Congress trade in the Oil, Gas and
# Consumable Fuels; Automobiles; or Metals & Mining sectors or not (0).

# 1. Based on these variables, cross-tabulate industry select vs. the dummy
# variable for committees. The CrossTable command from the gmodels
# package offers better formatted output than the table command but you
# needn’t show your cross-tab table.

# (a) Conduct a χ2-test and report your χ2 value.
# IV = committee (being part)
# DV = industry (trade)
task.1 <- table(trades_individual$commitee_select, trades_individual$industry_select)
sum(is.na(trades_individual$commitee_select))
sum(is.na(trades_individual$industry_select))

task.1
chisq.test(task.1)

# (b) Are the results statistically significant?
X-squared = 43.207, df = 1, p-value = 4.924e-11 # yes.

# (c/ d) Are members who sit on the Natural Resources and/or Climate Crisis
# committees more likely to trade stocks in the Oil, Gas and Consumable Fuels;
# Automobiles; or Metals and Mining sectors?
prop.table(task.1, 1)*100

2.793755 - 8.078159
```

```
-5.284404
# members who are in the committee are less likely to invest compared to the non members.

# 2. Create a dummy variable for party affiliation (Republican/Democrats).
# Conduct another χ2-test; this time between party affiliation and the industry select
variable.

trades_aggregate$party_dummy <- ifelse(trades_aggregate$party == 'republican', 1, 0)
trades_individual$party_dummy <- ifelse(trades_individual$party == 'republican', 1, 0)

# (a) Is there a statistically significant difference between Democrats and
# Republicans in terms of their participation in trades in stocks pertaining to the
# Oil, gas and consumable fuels, Automobiles or Metals & Mining sectors?
task.2 <- table(trades_individual$party_dummy, trades_individual$industry_select)
task.2
chisq.test(task.2.a)

prop.table(task.2.a, 1)
prop.table(task.2.a, 1)*100
0.16091003 - 0.02571916
0.1351909

# Now open the second dataset (“trades aggregate.csv”). In this dataset
# you can find the total value of trades conducted by Members of Congress,
# which is operationalized in two ways: 1. How often have members of
# Congress traded (count), 2. The absolute values of the net
# (purchases - sales) values of the total financial transactions conducted
# in the time period (net trades abs).

# 3. Examine the question of whether there are differences in trading behaviour
# between Democrats and Republicans. To do so, look at the count variable
# and conduct a two-sample t-test.

# (a) Report the substantive difference in trading count between the two
# parties. Which trades more?

View(trades_aggregate)
r_trade <- trades_aggregate[trades_aggregate$party=="republican",]
d_trade <- trades_aggregate[trades_aggregate$party=="democrat",]

t.test(r_trade$count, d_trade$count)
113.94444 - 59.69136 = 54.25308

# Is the difference statistically significant?
# Report and interpret the p-value for the difference.
t = -2.0205, df = 104.21, p-value = 0.04589 # yes.

# 4. Start your own analysis.
# DV: net trades absolute value (in trades_aggregate)
# -> absolute value of the sum of all purchases minus the sum of all sales by each member
of the House
# -> it captures the degree of change in each member’s position in the stock market,
# regardless of whether it is an increase or decrease in assets
# Goal: estimate the effect of an individual characteristic (of your own choosing)
# on this dependent variable and to persuade you readers of the accuracy and robustness
of this estimate.

summary(trades_aggregate)
str(trades_aggregate)
datasummary_correlation(trades_aggregate)
```

```
cor(trades_aggregate$net_trades_absolutevalue, trades_aggregate$bill_proposed)
cor(trades_aggregate$net_trades_absolutevalue, trades_aggregate$year_service)
cor(trades_aggregate$net_trades_absolutevalue, trades_aggregate$count)
cor(trades_aggregate$net_trades_absolutevalue, trades_aggregate$sum_trades_absolutevalue)

install.packages("dplyr")
library(dplyr)
aggregate(net_trades_absolutevalue ~ district, data = trades_aggregate, FUN = mean)

# create a new variable coding the districts as "rural" or "urban"
rural <- c("AL02", "AL04", "AL05","AR02", "AZ01", "CA03", "CO04", "CT02", "FL01", "FL02",
           "FL17", "GA08", "GA12", "GA14","IA01", "ID02", "IL16", "IL17", "IN03", "IN08",
"IN09",  "KY01",  "KY04",  "KY05",  "LA06",  "MI02",  "MI03",  "MI06",  "MN01",  "MO07","NC02",
"NC03",  "NC05",  "NC06",  "NC07",  "NC10",  "ME01",  "MS03",  "NC11",  "NE03",  "NY27",
"NY18","OH05",  "OK04",  "PA09",  "PA16",  "SC07",  "OH07",  "TN01",  "TN02",  "TN03",
"TN08","TN06",  "TN07",  "TX04",  "TX10",  "TX11",  "TX17",  "VA01","VA04",  "VA07",  "WA04",
"WI08", "WV01", "WV03")

urban <-c("AZ03", "CA06", "CA12", "CA16", "CA17", "CA19", "CA25", "CA27", "CA28",
          "CA30", "CA38", "CA47", "CA52", "CA53", "CO05", "CO07", "CT01", "FL04", "FL06",
          "FL12", "FL14", "FL15", "FL16", "FL18", "FL21", "FL23", "FL25", "FL27", "HI01",
          "IA03", "IL03", "IL08", "IL10", "IN05", "KS03", "KS04", "KY03", "KY04",
          "MA03", "MA04", "MA05", "MA06", "MA09", "MD06", "MD08", "MI12",
          "MN03", "MO02", "NJ05", "NJ06", "NC04", "NJ07", "NJ09", "NJ11", "NV03", “NY02",
           "NY03", "NY06", "NY08", "NY12", "NY25",c"NY26", "OH01",
          "OH14", "OH16", "OK01", "OK05", "OR03", "OR05", "PA03", "PA05",  "PA11",
          "RI02", "SC04",  "TN09", "TX02", "TX03", "TX15", "TX26", "TX32", "TX35",
          "UT01", "UT03",  "VA02", "VA03",  "VA08", "VA11", "WA01",  "WA08")

trades_aggregate1  <-  trades_aggregate  %>%  mutate(area  =  factor(case_when(district  %in%
urban ~ "urban",
                                                                    district %in% rural ~
"rural")))
View(trades_aggregate1) # new dataframe with urban/rural categorical variable

ggplot(data = trades_aggregate1, aes(x = net_trades_absolutevalue)) +
  geom_histogram(bins = 20, fill = "orange") +
  theme_bw() # right-skewed, therefore, log:
net_trades_absolutevalue_log <- log(trades_aggregate1$net_trades_absolutevalue)

ggplot(data = trades_aggregate1, aes(x = sum_trades_absolutevalue)) +
  geom_histogram(bins = 20, fill = "orange") +
  theme_bw() # right-skewed, therefore, log:
sum_trades_absolutevalue_log <- log(trades_aggregat1e$sum_trades_absolutevalue)

ggplot(data = trades_aggregate1, aes(x = year_service)) +
  geom_histogram(bins = 20, fill = "orange") +
  theme_bw() # right-skewed, therefore, log:
year_service_log <- log(trades_aggregate1$year_service)

ggplot(data = trades_aggregate, aes(x = count)) +
  geom_histogram(bins = 20, fill = "orange") +
  theme_bw() # right-skewed, therefore, log:
count_log <- log(trades_aggregate1$count)

ggplot(data = trades_aggregate, aes(x = bill_proposed)) +
  geom_histogram(bins = 20, fill = "orange") +
  theme_bw() # already normally distributed

ggplot(data = trades_aggregate1, aes(
  x = net_trades_absolutevalue_log,
  y = count_log,
))+
  geom_point() +
  theme_bw() # scatter plot to see the correlation between count and net trade

ggplot(data = trades_aggregate1, aes(x = party, y = net_trades_absolutevalue_log)) +
  geom_boxplot() +
  coord_flip() +
  theme_bw() # box plot to see the correlation between party and net trade

ggplot(data = trades_aggregate1, aes(x = gender, y = net_trades_absolutevalue_log)) +
  geom_boxplot() +
  coord_flip() +
  theme_bw() # box plot to see the correlation between gender and net trade

ggplot(data = trades_aggregate1, aes(x = area, y = net_trades_absolutevalue_log)) +
  geom_boxplot() +
  coord_flip() +
  theme_bw() # box plot to see the correlation between area and net trade


ggplot(data = trades_aggregate1, aes(
  x = bill_proposed,
  y = net_trades_absolutevalue_log,
)) +
  geom_point() +
  theme_bw() # scatter plot to see the correlation between bill proposed and net trades

cor(trades_aggregate1$year_service, trades_aggregate1$net_trades_absolutevalue)

ggplot(data = trades_aggregate1, aes(
  x = year_service_log,
  y = net_trades_absolutevalue_log,
)) +
  geom_point() +
  theme_bw() # scatter plot to see the correlation between years of service and net
trades

cor(trades_aggregate1$bill_proposed, trades_aggregate1$net_trades_absolutevalue)

trades_aggregate1$gender_dummy <- ifelse(trades_aggregate1$gender == 'male', 1, 0)
trades_aggregate1$party_dummy <- ifelse(trades_aggregate1$party == 'republican', 1, 0)
trades_aggregate1$area_dummy <- ifelse(trades_aggregate1$area == 'urban', 1, 0)

summary(trades_aggregate1[, c("net_trades_absolutevalue", "bill_proposed",
"year_service", "sum_trades_absolutevalue")])
trades_aggregate1_df <- data.frame(trades_aggregate1)
stargazer(trades_aggregate1_df, type = "text", out = "summary_stats.html",
          summary.stat = c("n","mean", "median", "sd", "min", "max", "p25", "p75"),
          title = "Summary Statistics",
          digits = 2)

library(arsenal)
tableFDA <- tableby(party_dummy ~ net_trades_absolutevalue + bill_proposed + year_service
+ sum_trades_absolutevalue,
                  data = trades_aggregate1)
write2word(tableFDA, "summary_stats_FDA.doc", title = "Summary Statistics FDA")
```

```
tableFDA2  <-  tableby(gender_dummy  ~  net_trades_absolutevalue  +  bill_proposed  +
year_service + sum_trades_absolutevalue,
                    data = trades_aggregate1)
write2word(tableFDA2, "summary_stats_FDA2.doc", title = "Summary Statistics FDA2")

tableFDA3 <- tableby(area_dummy ~ net_trades_absolutevalue + bill_proposed + year_service
+ sum_trades_absolutevalue,
                     data = trades_aggregate1)
write2word(tableFDA3, "summary_stats_FDA3.doc", title = "Summary Statistics FDA3")

# first model:
m1  <-  lm(net_trades_absolutevalue_log  ~    year_service_log  +  bill_proposed   ,  data  =
trades_aggregate1)
summary(m1)
stargazer(m1, type = "text")

# second model:
m2  <-  lm(net_trades_absolutevalue_log  ~  year_service_log  +  bill_proposed  +
sum_trades_absolutevalue_log , data = trades_aggregate1)
summary(m2)
stargazer(m2, type = "text")

# third model:
m3  <-  lm(net_trades_absolutevalue_log  ~    year_service_log  +  bill_proposed  +
sum_trades_absolutevalue_log
         + party_dummy + gender_dummy + area_dummy + year_service_log*bill_proposed, data
= trades_aggregate1)
summary(m3)
stargazer(m3, type = "text", out = "m3.doc")
vif(m3)

library(interplot)
interplot(m3, var1 = "bill_proposed", var2 = "year_service_log") +
  labs(x = "Years of Service (log)",
       y = "Marginal Effect of Bill Proposed")

bptest(m3) # homoskedasticity --> p-value = 0.886 --> okay
plot(m3, 3)
m3_robust <- coeftest(m3, vcov = vcovHC(m3, type = "HC1")) #use robust standard errors
stargazer(m3, m3_robust, type = "text", out = "m3robust.doc")

plot(m3, 2) # normality of residuals --> residuals on the diagonal line --> okay
plot(m3, 1) # linearity --> red line is straight and horizontal with zero intercept -->
okay


avPlots(m3, ~sum_trades_absolutevalue_log) # outliers: 127 and 66
avPlots(m3, ~year_service_log) # outliers: 127 and 66
avPlots(m3, ~bill_proposed) # outliers: 127 and 66
avPlots(m3, ~year_service_log*bill_proposed)
vif(m3) # multicollinearity --> high
vif(m3, type = "predictor") # multicollinearity ok

trades_aggregate1_out_df <- data.frame(trades_aggregate1[-c(127,66), ])

m3_out  <-  lm(net_trades_absolutevalue_log  ~    year_service_log  +  bill_proposed  +
sum_trades_absolutevalue_log
              + party_dummy + gender_dummy + area_dummy + year_service_log*bill_proposed,
data = trades_aggregate1_out_df)
summary(m3_out)
```

```
bptest(m3_out) # p-value = 0.1639 --> less than before but still okay
plot(m3_out, 3)
m3_out_robust <- coeftest(m3_out, vcov = vcovHC(m3_out, type = "HC1")) #use robust standard errors
stargazer(m3_out, m3, type = "text", out = "m3out.doc")

plot(m3_out, 2) # residuals on the diagonal line indicate a normal distribution --> ok

plot(m3_out, 1) # Red line is straight and horizontal with zero intercept if linear --> ok

stargazer(m1, m2, m3, m3_out, type = 'text', out = "models.doc")

stargazer(m1, m2, m3, type = 'text', out = "models2.doc")
```