Name: **Ramzi Abdel.**
Project: **Voting trends In Turkey**

The tidyverse is an opinionated collection of R packages designed for data science. All

Shiny makes it incredibly easy to build interactive web applications with R. Shiny has automatic

rmarkdown lets you insert R code into a markdown document. R then generates a final

Sparklyr is an R interface to Apache Spark, a fast and general engine for big data processing. This

ggplot 2 is an enhanced data visualization package for R. Create stunning multi-layered

- Research/Project Questions:
  - Who votes for the AKP party in Turkey?
  - What factors drive the voting tendency in Turkey?
- P.Sc Literature Hypotheses
  - People who are religious and conservative tend to vote for the AKP Party
  - People who have high economic optimism tend to vote for the AKP Party
  - Kurdish people do not tend to vote for the AKP Party
  - Women less vote for the AKP Party
  - People who are highly educated do not vote for the AKP Party
  - High social class people do not tend to vote for the AKP Party
  - Young and well educated people do not tend to vote for the AKP Party

Statistical Methodology:

- **binomial logistic regression**
  - The normal (*z*) distribution is a continuous distribution, which means that between any two data values we could (at least in theory) find another data value.
  - Binomial distribution is discrete, not continuous. In other words, it is NOT possible to find a data value between any two data values.
  - The key difference is that a binomial distribution is discrete, not continuous. In other words, it is NOT possible to find a data value between any two data values.
  - The requirements for a binomial distribution are
    1) The r.v. of interest is the count of successes in n trials
    2) The number of trials (or sample size), n, is fixed
    3) Trials are independent, with fixed value $p = P$(success on a trial)
    4) <u>There are only two possible outcomes on each trial, called "success" and "failure." (This is where the "bi" prefix in "binomial" comes from.</u>

R Methodology

- Stage 1: Importing, Renaming and Slicing Data
  - Libraries & Packages used:
    - library(readxl)
    - library(writexl)
    - library(boot)
    - library(mlogit)# require(mlogit) package
    - library(distr)# require(distr) package
    - library(stats)
    - library(pscl)# require(pscl) package
    - library(ROCR)# require (ROCR) package

- Importing:
  - WVS <- read_excel("C:/Users/ramsey/Desktop/Metro College/R/Project/WVS.xlsx")
- Browsing & Selecting Data based on Data code book
  - E.g. names(WVS)
  - #[166] "V145: How often do you attend religious services"
  - #[285] "V228: Which party would you vote for if there were a national election tomorrow"
  - #[308] "V239: Scale of incomes"
- Creating new data.frames
  - names(WVS) [c(285, 166, 99, 62, 60, 308, 309, 311, 318, 319, 329)]=c("Voting","Religiousity", "Ideology", "Eco_Satisfaction", "Marital", "Income", "Gender","Age","Ethnicity","Education", "Region")
  - mydata<-WVS[,c("Voting","Religiousity", "Ideology", "Eco_Satisfaction", "Marital", "Income", "Gender","Age","Ethnicity","Education", "Region")]
- dim(mydata)=1605X11

R Methodology

- Stage 2: Grouping, Leveling, plotting, & Relabeling Each Variable
  - E.g. #[60] "V57: Marital status"
  - Marital<-factor(mydata$Marital)
  - summary(Marital)
  - **levels(Marital)<-c("Divorced"=0, "Living together as married"=1, "Married"=1,"Separated"=0,"Single"=0, "Widowed"=0)**
  - Marital_Labeled <- factor(Marital, levels = c(0,1), labels = c("Unmarried", "Married"))#
  - summary(Marital_Labeled)
  - Marital<- as.numeric(Marital)
  - hist(Marital, freq = TRUE, labels = TRUE, nclass = 3, plot = TRUE)
  - plot(Marital_Labeled, main="Marital Status")
  - #Marital<- na.omit(Marital)#optional based on the methodology and research question

- Note: as the R levelling should follow the statistical binomial regression method: binary coding (0 1) was used with the base = 0 and target = 1
- It is recommended to download and use smbinning package
  - library(smbinning)
  - **Optimal Binning** categorizes a numeric characteristic into bins for ulterior usage in scoring modeling. This process, also known as ***supervised discretization***, utilizes Algorithm to categorize the numeric.

R Methodology

- Stage 3: Running Statistical Tests
  - E.g. Some Correlations Tests
  - cor_A_M<- subset(Categorized_Data, select = c("Age", "Marital"))
  - summary(cor_A_M)
  - cor(cor_A_M)
  - cor.test(Age, Marital, method = c("pearson", "kendall", "spearman"), exact = NULL, conf.level = 0.95, continuity = FALSE)

Correlation between Age and Marital Status	**Spearsman:** S = 401370000, p-value < 2.2e-16 (.000000000000000022) alternative hypothesis: true rho is not equal to 0 sample estimates: rho 0.4175342
	Pearson t = 18.428, df = 1603, p-value < 2.2e-16 alternative hypothesis: true correlation is not equal to 0 95 percent confidence interval: 0.3768976 0.4576825 sample estimates: cor 0.4181164
	Age Marital Age 1.0000000 0.4181164 Marital 0.4181164 1.0000000

R Methodology

- Although, not recommended, we can run a regression model through **lm()** to check **statistical significance** of the variables and their **correlations**.
- This might require **re-leveling** the variables several times into more or diffirent categories “discrtization” to reach better R(squared) and statistical significance
- This also requires to change the leveled variables into numeric, to run the regression model.
- Note: do not remove N.A’s when you create levels as this will affect the level of the variable and the lm() regression function will not run due to difference in length
- You can remove the NA’s at the lm() command via **na.omit**

  – E.g. levels(Marital)<-c("Divorced"=2, "Living together as married"=2, "Married"=1,"Separated"=2,"Single"=3, "Widowed"=2)

  – Marital_Labeled <- factor(Marital, levels = c(1,2,3), labels = c("Married", "X-Married", "Single"))#

  – **Marital<- as.numeric(Marital)**

  – #Marital<- na.omit(Marital)#

  – Regression:

  E.g. CAT_regression<-lm(Voting~Religiousity+Ideology+Eco_Satisfaction+Age+Income+Education+Gender+Ethnicity+Region+Marital, data = model1, na.omit)

lm() results

CAT_regression_Re<-
lm(Voting~**Religiousity_2nd**+Ideology+Eco_Satisfaction+Age+Income+Education+Gender+Ethnicity+Region+Marital, data = Categorized_Data)

```
Call:
lm(formula = Voting ~ Religiousity_2nd + Ideology + Eco_Satisfaction +
    Age + Income + Education + Gender + Ethnicity + Region +
    Marital, data = Categorized_Data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9889 -0.3373  0.1202  0.3365  0.9554

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.48935    0.15431   3.171  0.00156 **
Religiousity_2nd  0.16410    0.03120   5.260 1.73e-07 ***
Ideology          0.35100    0.02890  12.147  < 2e-16 ***
Eco_Satisfaction  0.03209    0.02281   1.407  0.15976
Age               0.04706    0.01671   2.816  0.00495 **
Income            0.03438    0.02250   1.528  0.12674
Education        -0.10606    0.02017  -5.258 1.75e-07 ***
Gender           -0.02610    0.03023  -0.864  0.38801
Ethnicity         0.09658    0.03253   2.969  0.00306 **
Region            0.05483    0.02004   2.737  0.00631 **
Marital          -0.05701    0.02746  -2.076  0.03812 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4396 on 1105 degrees of freedom
  (489 observations deleted due to missingness)
Multiple R-squared:  0.2336,    Adjusted R-squared:  0.2267
F-statistic: 33.68 on 10 and 1105 DF,  p-value: < 2.2e-16
```

Logit Model

- In the logit model, the response variable is log odds:
  - E.g. If Religiosity increases by 1, the log odds to vote for AKP party increases by 0.64

Logistic_regression_Labeled<-glm(Voting_Labeled~Religiousity_Labeled+Ideology_Labeled+Age_Labeled+Income_Labeled+Education_Labeled+Gender_Label+Ethnicity+Region_Labeled+Marital_Labeled, data = Binary_Logistic_Data, binomial) P.S: Coding all variable into 0 1 binary coding

```
Call:
glm(formula = Voting_Labeled ~ Religiousity_Labeled + Ideology_Labeled +
    Age_Labeled + Income_Labeled + Education_Labeled + Gender_Label +
    Ethnicity + Region_Labeled + Marital_Labeled, family = binomial,
    data = Binary_Logistic_Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0383  -0.8586   0.5530   0.9367   2.0810

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -1.41219    0.38505  -3.668 0.000245 ***
Religiousity_LabeledReligious     0.64390    0.15663   4.111 3.94e-05 ***
Ideology_LabeledRight             1.56106    0.14920  10.463  < 2e-16 ***
Age_Labeledy30+                  -0.60536    0.17792  -3.402 0.000668 ***
Income_LabeledMiddle-High         0.14898    0.21033   0.708 0.478747
Education_LabeledHigh-(College+) -0.79839    0.15300  -5.218 1.81e-07 ***
Gender_LabelM                    -0.02025    0.14854  -0.136 0.891558
Ethnicity1                        1.02193    0.33589   3.042 0.002347 **
Region_LabeledWestern            -0.61287    0.21256  -2.883 0.003936 **
Marital_LabeledMarried            0.44976    0.16445   2.735 0.006237 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1560.9  on 1126  degrees of freedom
Residual deviance: 1287.6  on 1117  degrees of freedom
  (478 observations deleted due to missingness)
AIC: 1307.6

Number of Fisher Scoring iterations: 4
```

Research Conclusion

- We can describe the voters of AKP party in Turkey as religious married grown-ups who are ethnically Turkish and ideologically rightists, with middle to high income and mostly living the eastern region of Turkey.
  - However, to what extent can use this description?
  - In other words, to what extent are we sure/accurate in our description?

McFadden R^2

- While no exact equivalent to the R^2 of linear regression exists, the McFadden R^2index is used intensively to assess the model fit.
- R package (pscl)
- pR2(Logistic_regression_Labeled)

llh	llhNull	G2	McFadden	r2ML	r2CU
-643.8244806	-870.6172318	453.5855025	0.2604965	0.3313347	0.4211747

glm model accuracy prediction

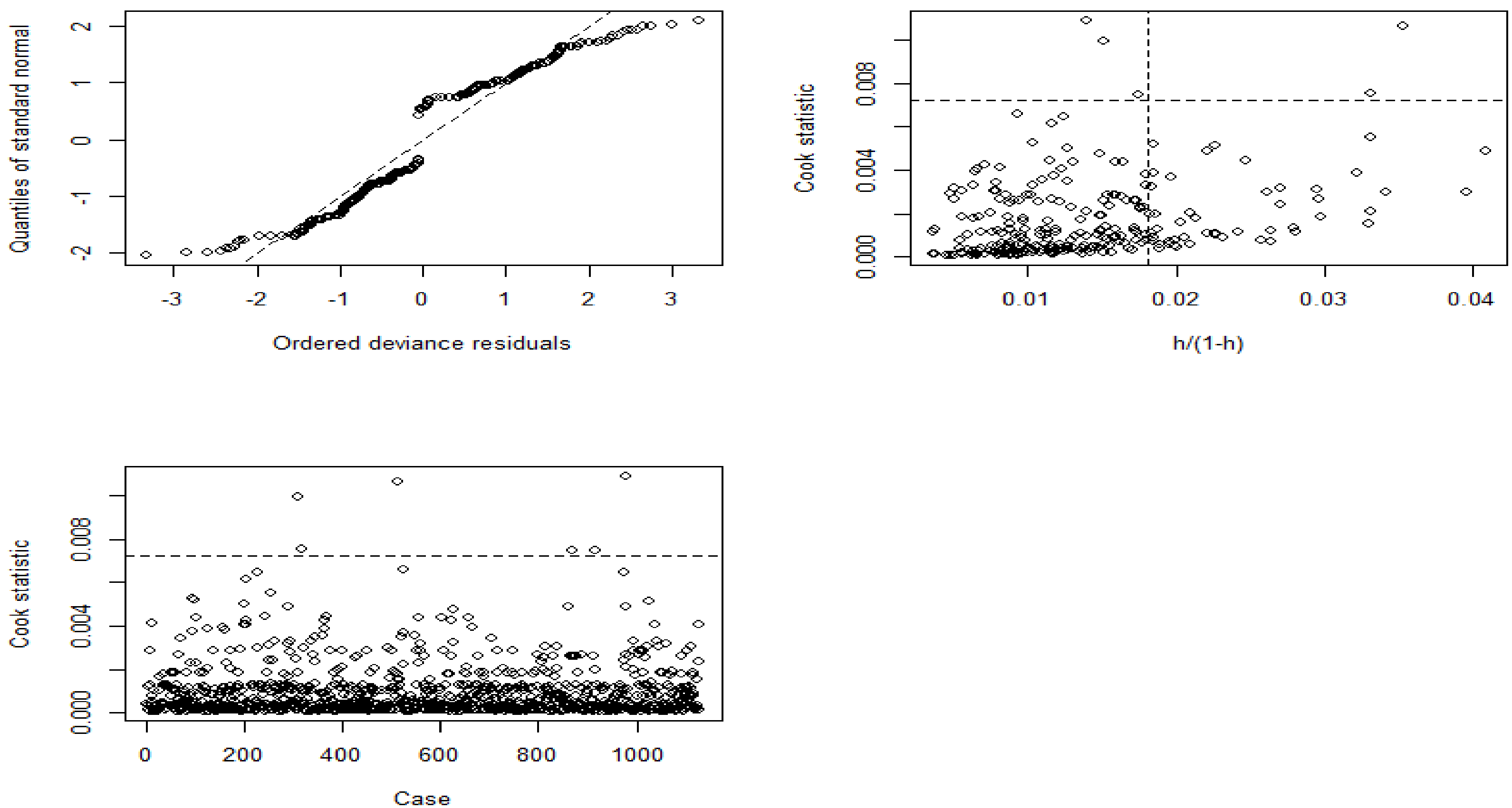
- R will output probabilities in the form of P(y=1|X). Our decision boundary will be 0.5. If P(y=1|X) > 0.5 then y = 1 otherwise y=0.

```
data.test<-Binary_Logistic_Data[,c("Voting","Religiousity", "Ideology", "Eco_Satisfaction", "Marita

library(ROCR)
p <- predict(Logistic_regression_Labeled, newdata=subset(data.test,select=c("Voting","Religiousity"
pr <- prediction(p, data.test)
summary(p)
fitted.results <- ifelse(p > 0.5,1,0)
misClasificError <- mean(fitted.results != data.test$Voting, na.rm=TRUE)
print(paste('Accuracy',1-misClasificError))
```

"Accuracy 0.707187222715173"

- **<u>Cook's distance</u>** can be used in several ways: to indicate influential data points that are particularly worth checking for validity; or **to indicate regions of the design space where it would be good to be able to obtain more data points.**
- R Package(boot)
- glm.diag.plots(Logistic_regression_Labeled, glmdiag = glm.diag(Logistic_regression_Labeled), subset = NULL, iden = FALSE, labels = NULL, ret = FALSE)



R Conclusion

- We can describe the voters of AKP party in Turkey as religious married grown-ups who are ethnically Turkish and ideologically rightists, with middle to high income and mostly living the eastern region of Turkey.

- The model presented in this paper has 26% explanatory power of association to the DP at 70% accuracy of predication.