EPPS6323 Knowledge Mining

Assignment 6

1. Run Lab_linearregression01.R in R
2. Review ISLR Chapters 4
3. Use the TEDS2016 dataset to run a logit (logistic regression) model using female as sole predictor. The dependent variable is the vote (1-0) for Tsai Ing-wen, the female candidate for the then opposition party Democratic Progressive Party (DPP). Access the data set using the following codes:

```
library(haven)
TEDS_2016 <-
read_stata("https://github.com/datageneration/home/blob/master/DataProgramming
/data/TEDS_2016.dta?raw=true")
```

   Hint: glm.vt=glm(votetsai~female, data=TEDS_2016,family=binomial)

   Are female voters more likely to vote for President Tsai? Why or why not?

   More diagnostics?

   Hint:
   summary(glm.vt)
   plot(glm.vt)

4. Add party ID variables (KMT, DPP) and other demographic variables (age, edu, income) to improve the model.

   What do you find? Which group of variables work better in explaining/predicting votetsai?

5. Try adding the following variables:

   Independence – Supporting Taiwan's Independence (vs. Unification with China)
   Econ_worse – Evaluations of economy (Negative)
   Govt_dont_care – Political Efficacy (Government does not care about people)
   Minnan_father – Descendent of local Taiwanese
   Mainland_father – Descendent of mainland China (migrated from mainland circa or after 1949)
   Taiwanese – Self-identified Taiwanese

6. (Optional) Run the model in Stata (available in EPPS labs)

   Hint: Use this command:

```
use
"https://github.com/datageneration/home/blob/master/DataProgramming/data/TEDS_20
16.dta?raw=true"
logit votetsai Independence Econ_worse Govt_dont_care Minnan_father
Mainland_father Taiwanese KMT DPP age edu female
```

a. Compare the results from R and Stata
b. Can you use mrobust to find the best predictor combination (by best consistency) as a prediction set?