Assignment 3

1. Lab03 in R
2. Review ISLR Chapters 3
3. Use the TEDS2016 dataset to run a multiple regression model. 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")
```

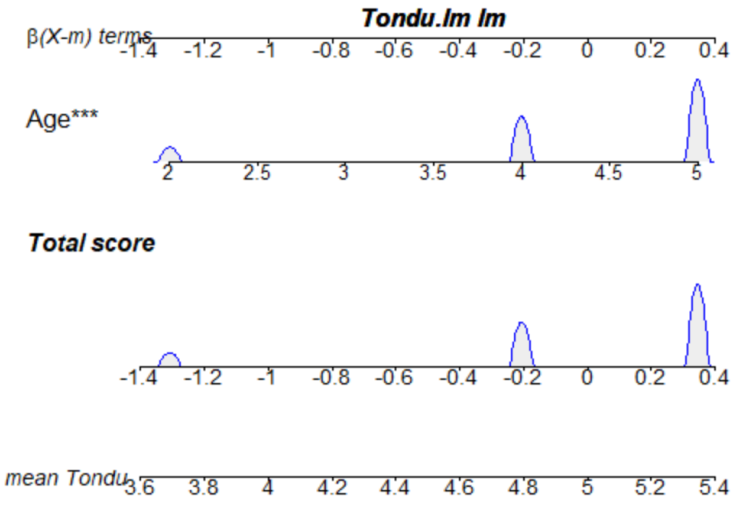
4. Select only relevant variables to create a subset of the dataset (`Tondu`, `female`, `DPP`, `age`, `income`, `edu`, `Taiwanese` and `Econ_worse`). Make sure the dependent variable `Tondu` is coded with right labels:

```
TEDS_2016$Tondu<-as.numeric(TEDS_2016$Tondu,labels=c("Unification now", "Status quo, unif. in future", "Status quo, decide later", "Status quo forever", "Status quo, indep. in future", "Independence now", "No response"))
```

```
> TEDS_2016$Tondu<-as.numeric(TEDS_2016$Tondu, labels=c("Unification now",
"Status quo unif in future", "Status quo decide later", "Status Quo forever",
"Status quo indep in future", "Independence now", "No response"))
> Tondu.lm=lm(Tondu ~ Age, Edu, income, data=TEDS_2016)
> summary(Tondu.lm)

Call:
lm(formula = Tondu ~ Age, data = TEDS_2016, subset = Edu, weights = income)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-6.6311 -6.6311  0.4625  8.5731  8.5731

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.58814    0.30028   8.619  < 2e-16 ***
Age          0.55126    0.06693   8.236 3.52e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.512 on 1688 degrees of freedom
Multiple R-squared:  0.03863, Adjusted R-squared:  0.03807
F-statistic: 67.84 on 1 and 1688 DF,  p-value: 3.518e-16
```

5. Run a regplot on the dependent variable using:
a. Age
b. Education
c. Income

```
> regplot(Tondu.lm)
Regression  Tondu.lm lm formula:
Tondu `~` Age
Replicate integer weights assumed
Note: non-integer weights have been floored
Distributions estimated with "nsamp=10000" random sub-sample of 11055
[1] "note: points tables not constructed unless points=TRUE "
```



6. What is the problem? Why? (hint: how many categories in the DV?)

There appears to be too many of sub-categories of the DV represented in the regression model that was used by regplot.

7. What can be done to improve prediction of the dependent variable?

Each sub-category of DV should be treated as a stand alone DV and then ran against IVs Age, Education, Income