**Assignment 4**

**Author:** Keerthy Raaj Shanmugam, 8779954

**Course:** PROG8430

**1. Reduce Dimensionality**

**1.1. Apply the Missing Value Filter to remove appropriate columns of data.**

id\_KS group\_KS hs.grad\_KS

Min. : 1 Length:2033 Length:2033

1st Qu.: 509 Class :character Class :character

Median :1017 Mode :character Mode :character

Mean :1017

3rd Qu.:1525

Max. :2033

nation\_KS gender\_KS age\_KS

Length:2033 Length:2033 Min. :18.00

Class :character Class :character 1st Qu.:29.00

Mode :character Mode :character Median :41.00

Mean :41.23

3rd Qu.:53.00

Max. :65.00

m.status\_KS political\_KS

Length:2033 Length:2033

Class :character Class :character

Mode :character Mode :character

n.child\_KS income\_KS food\_KS

Min. : 0.000 Min. : 0 Min. :0.0000

1st Qu.: 0.000 1st Qu.: 38000 1st Qu.:0.2500

Median : 1.000 Median : 76409 Median :0.5100

Mean : 1.474 Mean : 76876 Mean :0.5054

3rd Qu.: 2.000 3rd Qu.:114950 3rd Qu.:0.7600

Max. :201.000 Max. :166378 Max. :1.0000

housing\_KS other\_KS score\_KS

Min. :0.0000 Min. :0.0000 Min. :-3.09000

1st Qu.:0.0600 1st Qu.:0.0700 1st Qu.:-0.64000

Median :0.1800 Median :0.1800 Median : 0.03000

Mean :0.2426 Mean :0.2521 Mean : 0.02229

3rd Qu.:0.3700 3rd Qu.:0.3900 3rd Qu.: 0.65000

Max. :0.9800 Max. :0.9800 Max. : 3.12000

Pol\_KS time1 time2\_KS

Min. :-1.8092 Min. :0.0000 Min. :-3.7000

1st Qu.: 0.4094 1st Qu.:0.1200 1st Qu.:-1.0350

Median : 1.4511 Median :0.4500 Median : 0.3750

Mean : 1.4514 Mean :0.4646 Mean : 0.2465

3rd Qu.: 2.4475 3rd Qu.:0.8000 3rd Qu.: 1.5550

Max. : 5.2240 Max. :1.0000 Max. : 4.1600

NA's :1973

time3\_KS scr\_KS

Min. :-0.02767 Min. :-1.2985

1st Qu.: 0.12114 1st Qu.: 0.5461

Median : 0.45124 Median : 0.9660

Mean : 0.46486 Mean : 0.9635

3rd Qu.: 0.80573 3rd Qu.: 1.3761

Max. : 1.01798 Max. : 2.9733

From the above-given output, it’s very clear that **[time2\_KS - Time Taken on Section 1 (Standardized)]** variable have no value for **1973** records out of **2033** records.

**97.04%** of records have no value, so we can remove the variable **time2\_KS.**

Therefore, dimensionality is reduced by applying the **Missing Value Filter.**

**1.2. Apply the Low Variance Filter to remove appropriate columns of data.**

id\_KS group\_KS hs.grad\_KS

nbr.val 2033.0000000 NA NA

nbr.null 0.0000000 NA NA

nbr.na 0.0000000 NA NA

min 1.0000000 NA NA

max 2033.0000000 NA NA

range 2032.0000000 NA NA

sum 2067561.0000000 NA NA

median 1017.0000000 NA NA

mean 1017.0000000 NA NA

SE.mean 13.0192166 NA NA

CI.mean.0.95 25.5324038 NA NA

var 344593.5000000 NA NA

std.dev 587.0208685 NA NA

coef.var 0.5772083 NA NA

nation\_KS gender\_KS age\_KS

nbr.val NA NA 2033.0000000

nbr.null NA NA 0.0000000

nbr.na NA NA 0.0000000

min NA NA 18.0000000

max NA NA 65.0000000

range NA NA 47.0000000

sum NA NA 83822.0000000

median NA NA 41.0000000

mean NA NA 41.2306936

SE.mean NA NA 0.3053630

CI.mean.0.95 NA NA 0.5988572

var NA NA 189.5702779

std.dev NA NA 13.7684523

coef.var NA NA 0.3339370

m.status\_KS political\_KS n.child\_KS

nbr.val NA NA 2033.0000000

nbr.null NA NA 529.0000000

nbr.na NA NA 0.0000000

min NA NA 0.0000000

max NA NA 201.0000000

range NA NA 201.0000000

sum NA NA 2997.0000000

median NA NA 1.0000000

mean NA NA 1.4741761

SE.mean NA NA 0.1016570

CI.mean.0.95 NA NA 0.1993627

var NA NA 21.0092983

std.dev NA NA 4.5835901

coef.var NA NA 3.1092555

income\_KS food\_KS

nbr.val 2033.000000 2033.000000000

nbr.null 54.000000 14.000000000

nbr.na 0.000000 0.000000000

min 0.000000 0.000000000

max 166378.090362 1.000000000

range 166378.090362 1.000000000

sum 156288806.799335 1027.440000000

median 76408.892900 0.510000000

mean 76875.950221 0.505381210

SE.mean 1000.141989 0.006534969

CI.mean.0.95 1961.410584 0.012815938

var 2033577369.116162 0.086820930

std.dev 45095.203394 0.294653915

coef.var 0.586597 0.583032985

housing\_KS other\_KS

nbr.val 2033.000000000 2033.000000000

nbr.null 119.000000000 105.000000000

nbr.na 0.000000000 0.000000000

min 0.000000000 0.000000000

max 0.980000000 0.980000000

range 0.980000000 0.980000000

sum 493.120000000 512.440000000

median 0.180000000 0.180000000

mean 0.242557796 0.252060994

SE.mean 0.004939360 0.004985204

CI.mean.0.95 0.009686737 0.009776643

var 0.049599655 0.050524638

std.dev 0.222709801 0.224776863

coef.var 0.918172097 0.891755839

score\_KS Pol\_KS

nbr.val 2033.00000000 2033.00000000

nbr.null 8.00000000 0.00000000

nbr.na 0.00000000 0.00000000

min -3.09000000 -1.80915681

max 3.12000000 5.22399449

range 6.21000000 7.03315130

sum 45.31000000 2950.75395388

median 0.03000000 1.45109710

mean 0.02228726 1.45142841

SE.mean 0.02121719 0.03080753

CI.mean.0.95 0.04160971 0.06041763

var 0.91519393 1.92952782

std.dev 0.95665769 1.38907445

coef.var 42.92396991 0.95703959

time1\_KS time3\_KS

nbr.val 2033.000000000 2033.000000000

nbr.null 114.000000000 0.000000000

nbr.na 0.000000000 0.000000000

min 0.000000000 -0.027667705

max 1.000000000 1.017979555

range 1.000000000 1.045647260

sum 944.550000000 945.052027402

median 0.450000000 0.451235905

mean 0.464608952 0.464855891

SE.mean 0.007703489 0.007705237

CI.mean.0.95 0.015107559 0.015110988

var 0.120645824 0.120700584

std.dev 0.347341077 0.347419896

coef.var 0.747598761 0.747371179

scr\_KS

nbr.val 2033.00000000

nbr.null 0.00000000

nbr.na 0.00000000

min -1.29846354

max 2.97325759

range 4.27172114

sum 1958.79354465

median 0.96595730

mean 0.96349904

SE.mean 0.01346945

CI.mean.0.95 0.02641537

var 0.36883931

std.dev 0.60732142

coef.var 0.63032904

From the above-given output, it’s very clear that **all variables** have a **high Variance.**

There is no **coef.var** value that is very less or strange and so we cannot remove any variable from the data.

Therefore, no dimensionality is reduced by applying the **Low Variance Filter.**

**1.3. Apply the High Correlation Filter to remove appropriate columns of data.**

id\_KS age\_KS n.child\_KS

id\_KS 1.000000000 -0.010180449 0.021168960

age\_KS -0.010180449 1.000000000 -0.003923146

n.child\_KS 0.021168960 -0.003923146 1.000000000

income\_KS 0.007600309 0.017314689 0.022275343

food\_KS 0.026412515 0.004059447 0.004574709

housing\_KS -0.011693077 -0.004570835 -0.009248095

other\_KS -0.024547540 0.005268611 -0.008210990

score\_KS 0.003199035 -0.023853563 -0.001211450

Pol\_KS -0.026302761 0.065959760 0.006923419

time1\_KS 0.028396935 0.016235922 -0.015738751

time3\_KS 0.030541659 0.016229475 -0.016923946

scr\_KS -0.010172169 0.002487941 0.014408082

income\_KS food\_KS housing\_KS

id\_KS 0.007600309 0.026412515 -0.011693077

age\_KS 0.017314689 0.004059447 -0.004570835

n.child\_KS 0.022275343 0.004574709 -0.009248095

income\_KS 1.000000000 -0.010025015 -0.001051356

food\_KS -0.010025015 1.000000000 -0.668185310

housing\_KS -0.001051356 -0.668185310 1.000000000

other\_KS 0.004049181 -0.671058677 0.021199437

score\_KS 0.021756090 -0.064192597 0.043013985

Pol\_KS 0.015552803 -0.008514522 0.031397068

time1\_KS -0.038314249 -0.045094851 0.043938147

time3\_KS -0.038800346 -0.045360595 0.043763999

scr\_KS 0.473612069 -0.022315164 0.024842918

other\_KS score\_KS Pol\_KS

id\_KS -0.024547540 0.003199035 -0.026302761

age\_KS 0.005268611 -0.023853563 0.065959760

n.child\_KS -0.008210990 -0.001211450 0.006923419

income\_KS 0.004049181 0.021756090 0.015552803

food\_KS -0.671058677 -0.064192597 -0.008514522

housing\_KS 0.021199437 0.043013985 0.031397068

other\_KS 1.000000000 0.050330095 -0.033883739

score\_KS 0.050330095 1.000000000 0.134022818

Pol\_KS -0.033883739 0.134022818 1.000000000

time1\_KS 0.016709662 -0.001889094 -0.023651348

time3\_KS 0.016429551 -0.002298464 -0.022824323

scr\_KS -0.008323459 0.250161458 0.022302018

time1\_KS time3\_KS scr\_KS

id\_KS 0.028396935 0.030541659 -0.010172169

age\_KS 0.016235922 0.016229475 0.002487941

n.child\_KS -0.015738751 -0.016923946 0.014408082

income\_KS -0.038314249 -0.038800346 0.473612069

food\_KS -0.045094851 -0.045360595 -0.022315164

housing\_KS 0.043938147 0.043763999 0.024842918

other\_KS 0.016709662 0.016429551 -0.008323459

score\_KS -0.001889094 -0.002298464 0.250161458

Pol\_KS -0.023651348 -0.022824323 0.022302018

time1\_KS 1.000000000 0.998631881 -0.025751485

time3\_KS 0.998631881 1.000000000 -0.026037495

scr\_KS -0.025751485 -0.026037495 1.000000000

From the above-given output, it’s very clear that variables **time1\_KS** and **time3\_KS** are **highly Correlated (0.99 - strong linear relationship)**, so we can remove the variable **time3\_KS.**

Therefore, dimensionality is reduced by applying the **High Correlation Filter.**

**2. Data Transformation**

**As demonstrated in class, transform any variables that are required to conduct the regression analysis (e.g. categorical variables to dummies).**

**NOTE: This was/will be demonstrated in the lecture on Multivariate Regression. If you have not yet had this lecture, you can skip this step for now.**

'data.frame': 2033 obs. of 29 variables:

$ age\_KS : num 21 53 44 35 20 35 55 37 49 21 ...

$ n.child\_KS : num 0 1 0 1 0 2 3 3 3 2 ...

$ income\_KS : num 138147 21953 87000 66420 146000 ...

$ food\_KS : num 0.13 0.75 0.82 0.67 0.26 0.04 0.87 0.8 0.22 0.2 ...

$ housing\_KS : num 0.87 0.2 0.08 0.05 0.18 0.17 0.11 0.19 0.67 0.2 ...

$ other\_KS : num 0 0.05 0.1 0.28 0.56 0.79 0.02 0.01 0.11 0.6 ...

$ score\_KS : num -0.15 -0.71 -0.16 -0.19 0.73 0.46 -0.55 -1.55 2.08 0.75 ...

$ Pol\_KS : num 0.932 1.892 1.708 3.41 2.849 ...

$ time1\_KS : num 0.57 0.52 0.16 0.61 0.09 0.03 0.78 0.88 0.16 0.5 ...

$ scr\_KS : num 0.783 0.906 1.327 1.177 1.717 ...

$ group.Control\_KS : num 1 0 0 0 0 1 1 0 1 0 ...

$ group.Treat\_KS : num 0 1 1 1 1 0 0 1 0 1 ...

$ hs.grad.No\_KS : num 1 0 1 1 1 0 1 1 0 0 ...

$ hs.grad.Yes\_KS : num 0 1 0 0 0 1 0 0 1 1 ...

$ nation.Asia\_KS : num 0 0 0 1 1 0 0 0 1 0 ...

$ nation.Europe\_KS : num 0 0 0 0 0 0 1 1 0 0 ...

$ nation.NorthAmerica\_KS : num 1 1 1 0 0 1 0 0 0 1 ...

$ nation.Southern\_KS : num 0 0 0 0 0 0 0 0 0 0 ...

$ gender.Female\_KS : num 0 0 1 0 0 0 0 1 0 0 ...

$ gender.Male\_KS : num 0 0 0 1 1 0 0 0 0 0 ...

$ gender.Undis\_KS : num 1 1 0 0 0 1 1 0 1 1 ...

$ political.Conservative\_KS: num 0 0 0 0 0 0 1 0 1 0 ...

$ political.Liberal\_KS : num 0 0 1 0 1 0 0 1 0 1 ...

$ political.NewDemocrat\_KS : num 1 1 0 1 0 0 0 0 0 0 ...

$ political.Other\_KS : num 0 0 0 0 0 1 0 0 0 0 ...

$ m.status\_divorced\_KS : num 0 0 0 1 0 0 1 0 0 1 ...

$ m.status\_married\_KS : num 0 0 0 0 0 0 0 1 1 0 ...

$ m.status\_never\_KS : num 0 0 1 0 1 1 0 0 0 0 ...

$ m.status\_widowed\_KS : num 1 1 0 0 0 0 0 0 0 0 ...

To get the above-given output, the following **Data Transformations** are done,

1. Converted **integer** variables like **(age\_KS, n.child\_KS)** to a **numeric** datatype.
2. Converted **character** variables like **(group\_KS, hs.grad\_KS, nation\_KS, gender\_KS, political\_KS, m.status\_KS)** to **Dummy Variables.**
3. Removed unwanted **character** variables like **(group\_KS, hs.grad\_KS, nation\_KS, gender\_KS, political\_KS, m.status\_KS).**
4. Removed **integer** variable **(id\_KS), a primary key** that is of no use in **Regression Analysis.**
5. Renamed all the new **Dummy Variables (group.Control\_KS, group.Treat\_KS, hs.grad.No\_KS,**

**hs.grad.Yes\_KS, nation.Asia\_KS, nation.Europe\_KS, nation.NorthAmerica\_KS,**

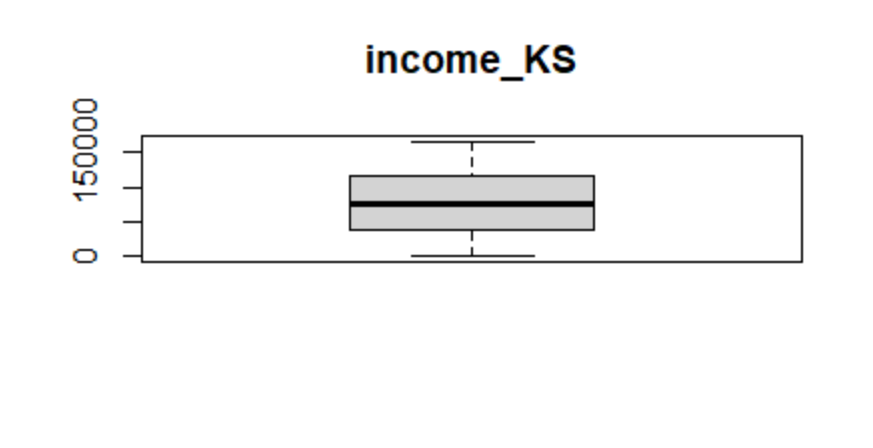
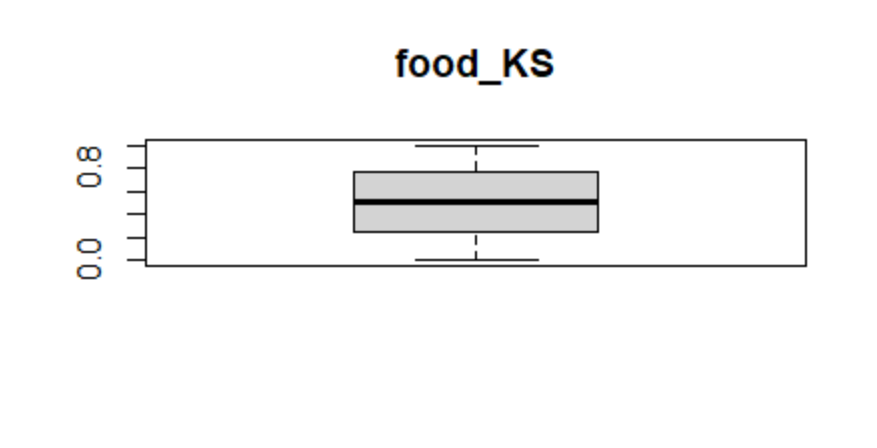
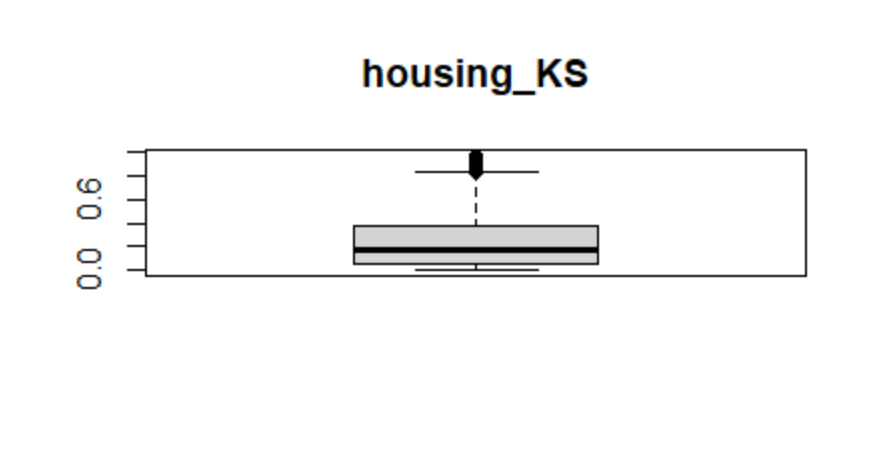
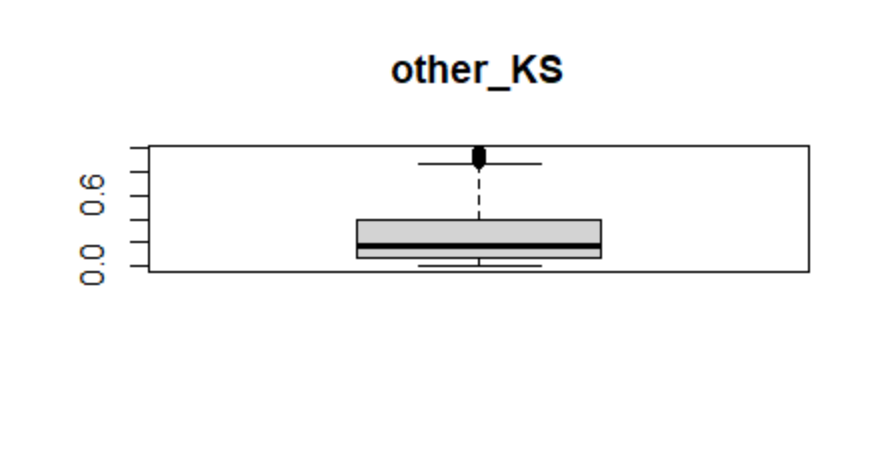
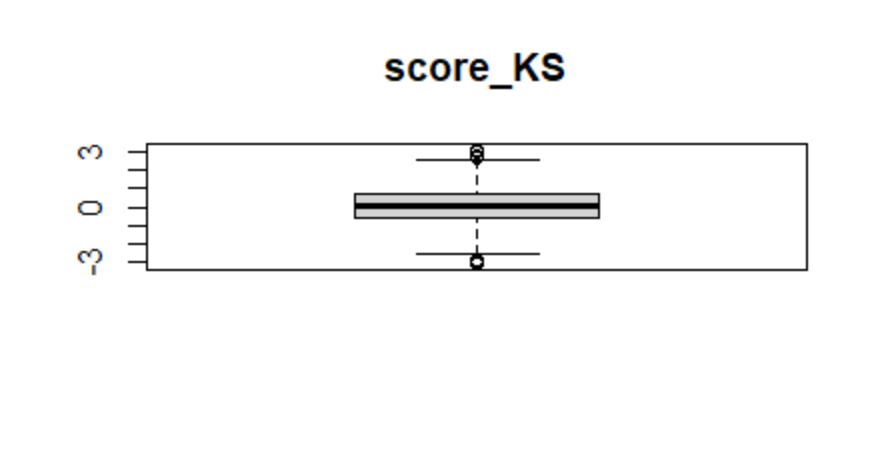
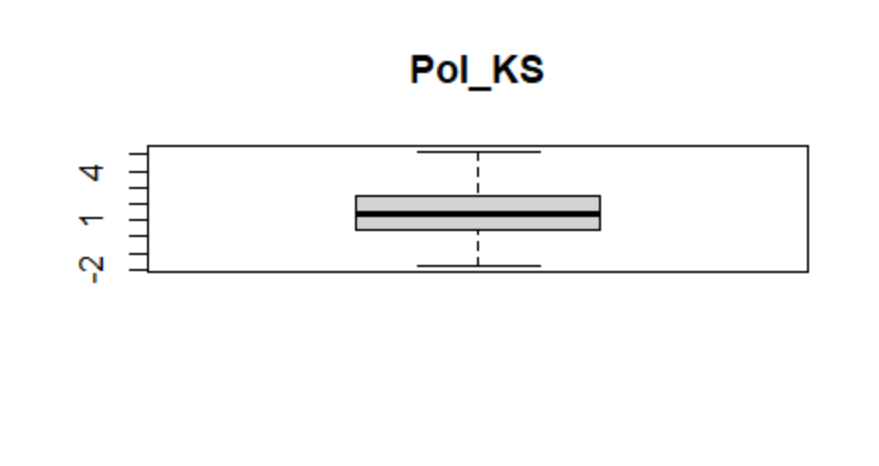
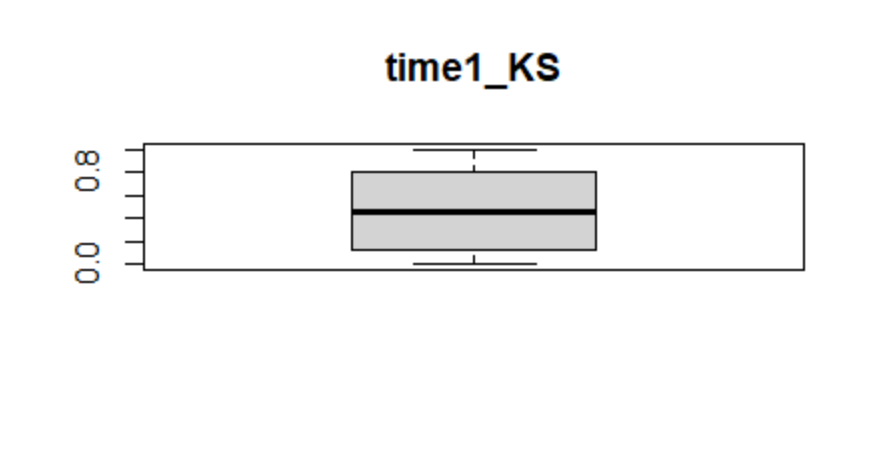
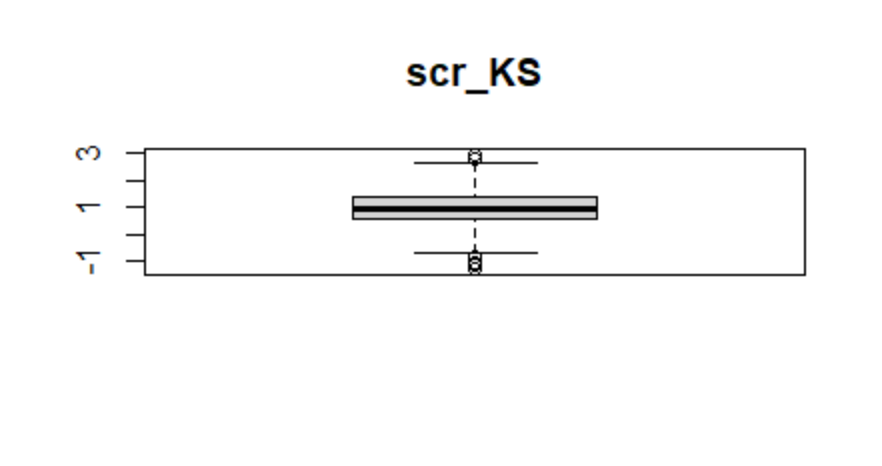
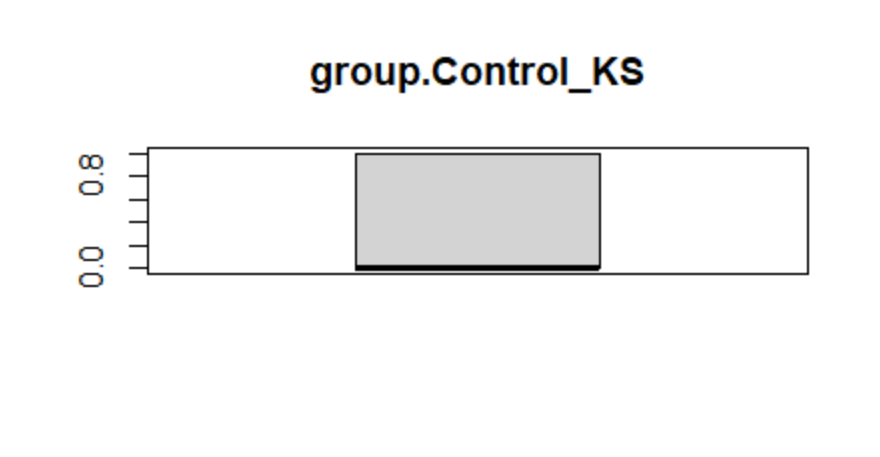
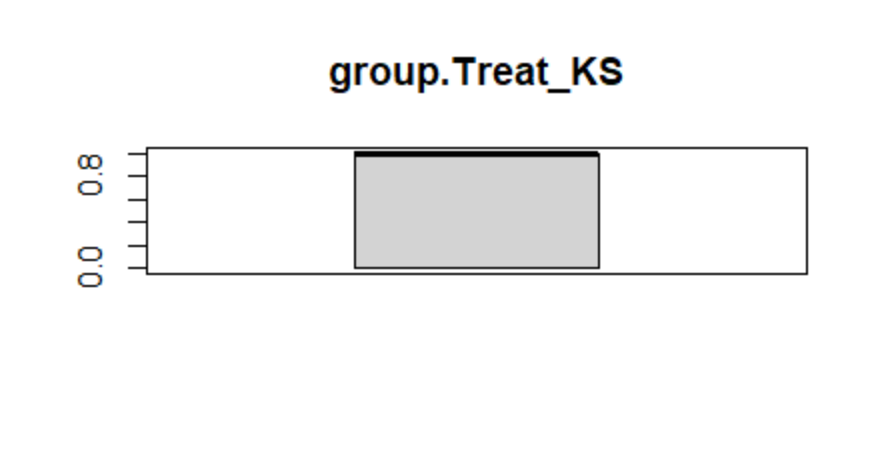
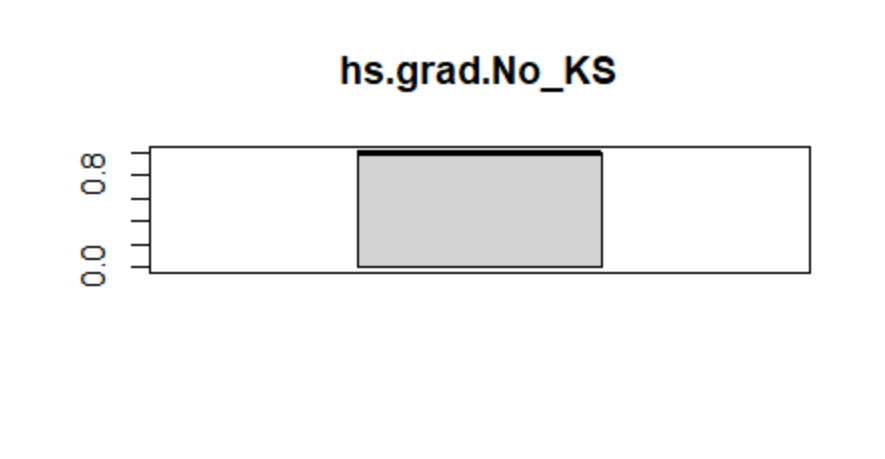
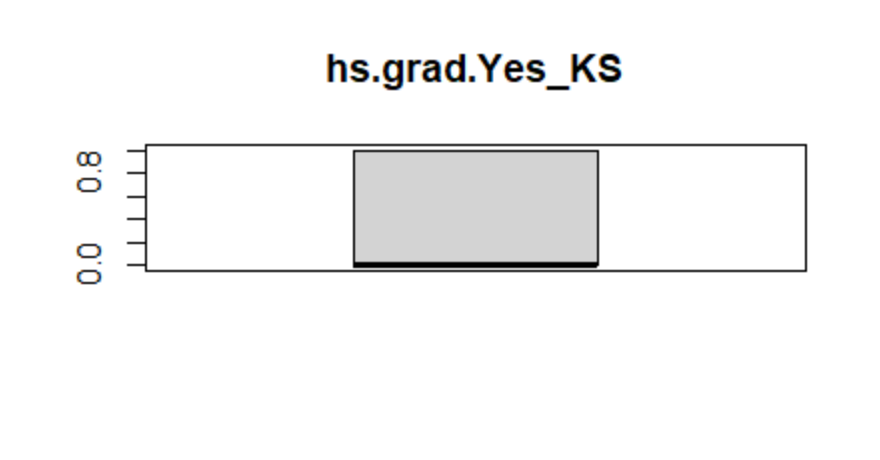
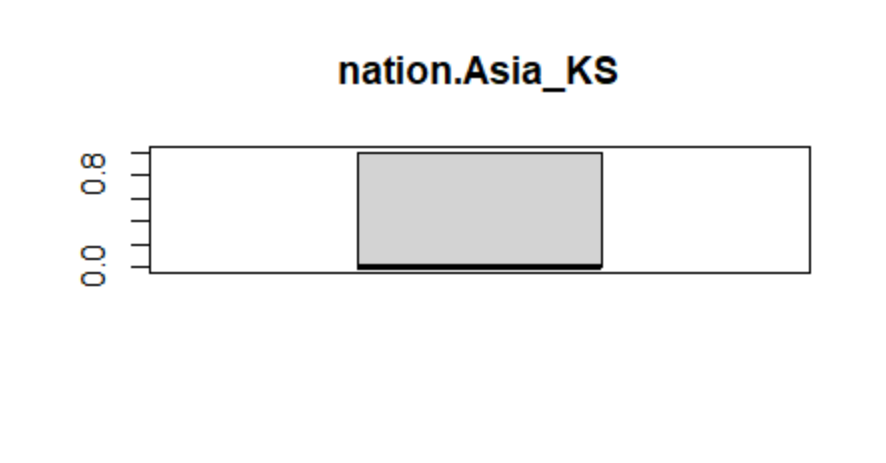
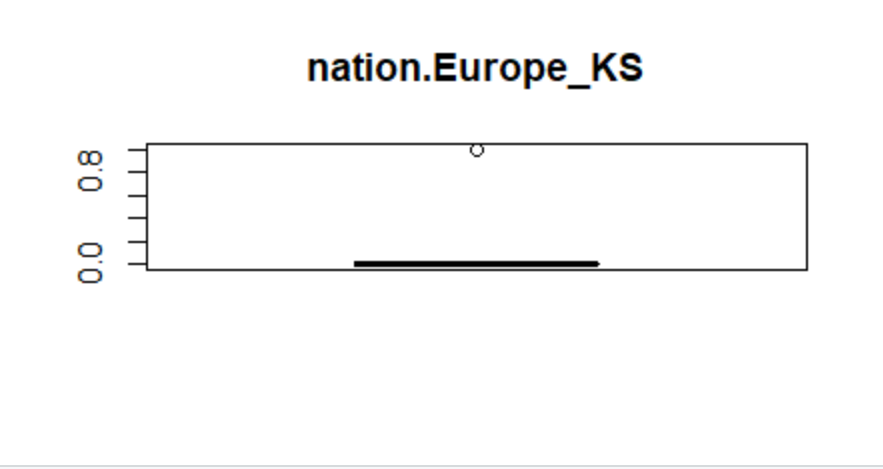
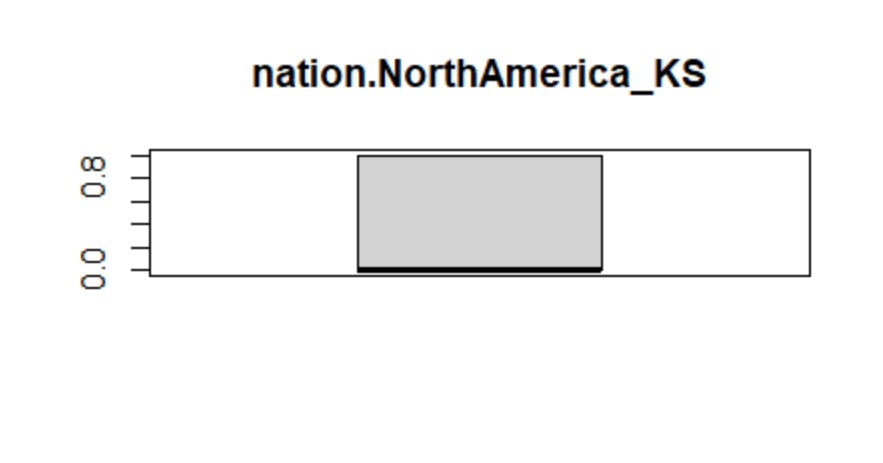
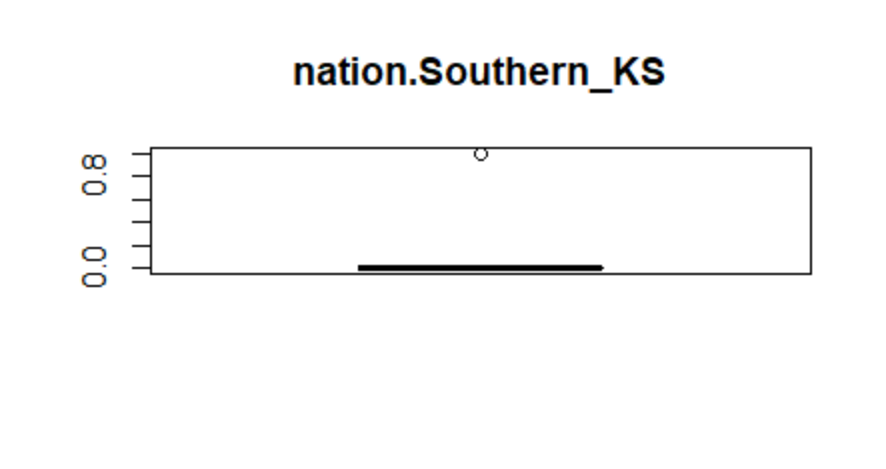
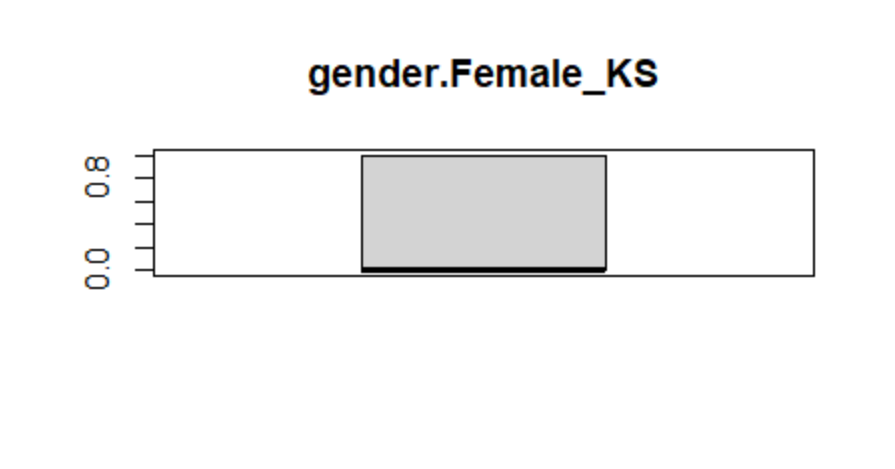
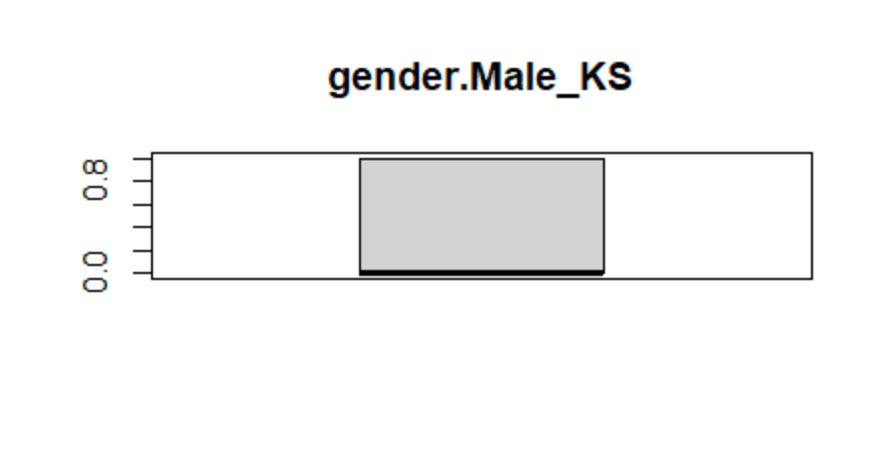
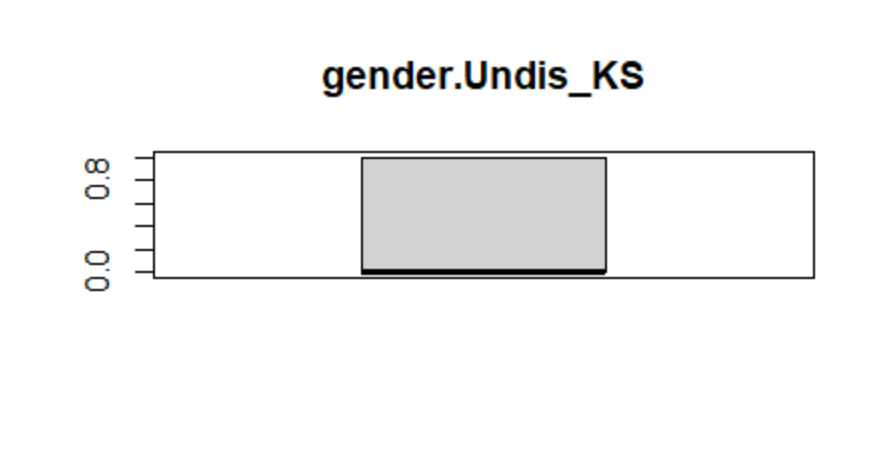
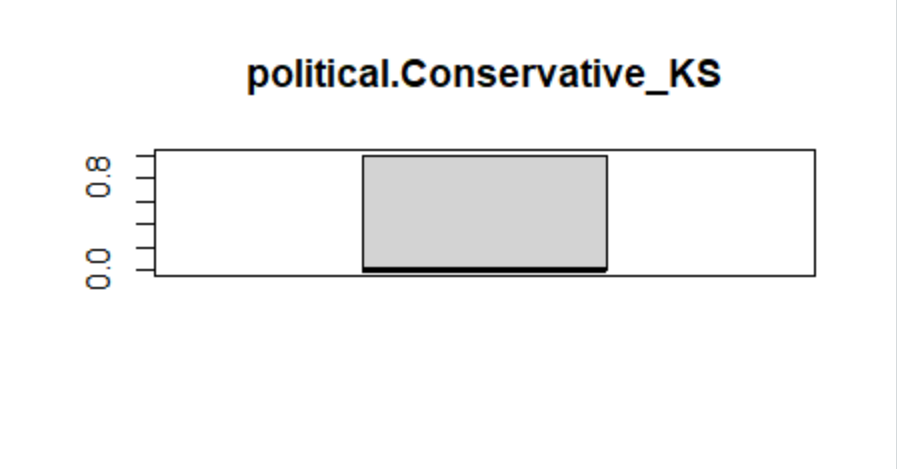
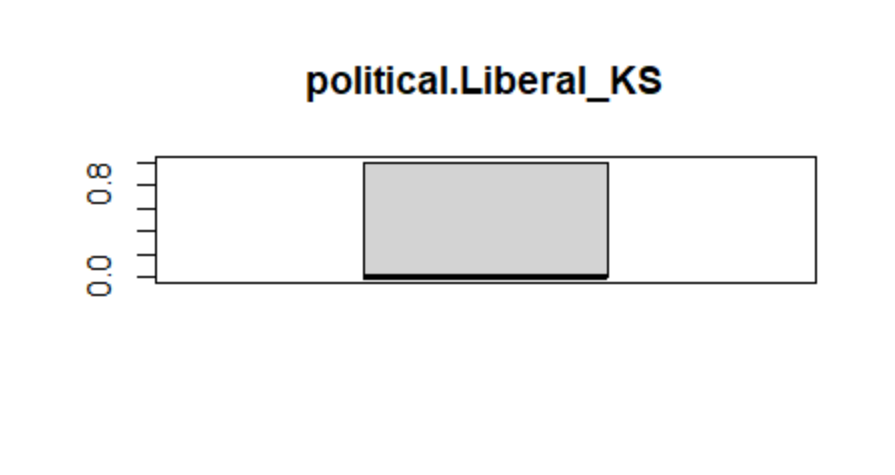
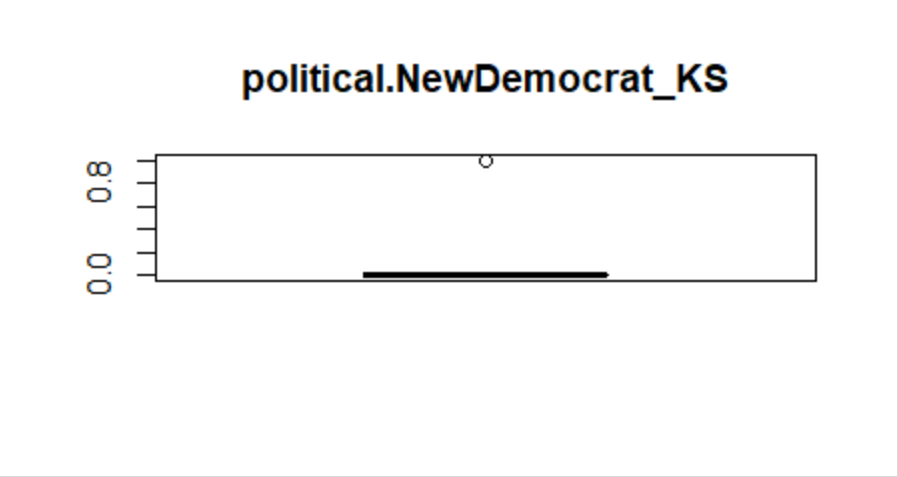
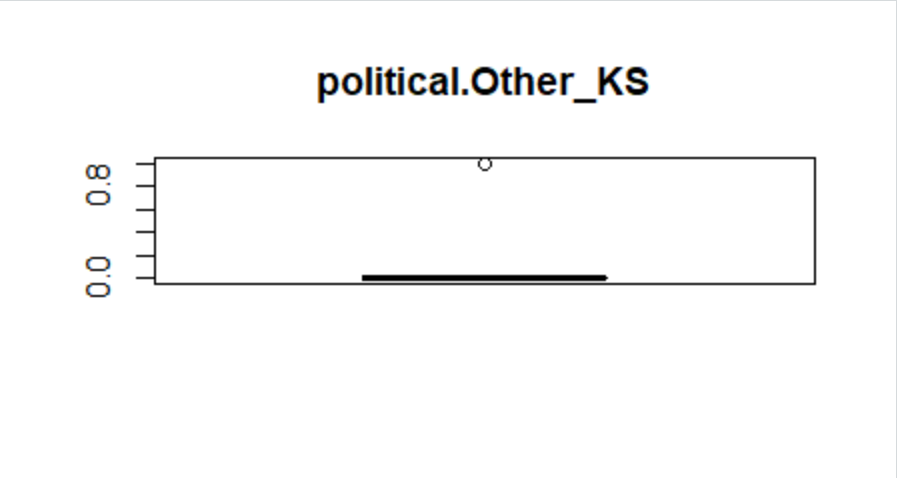
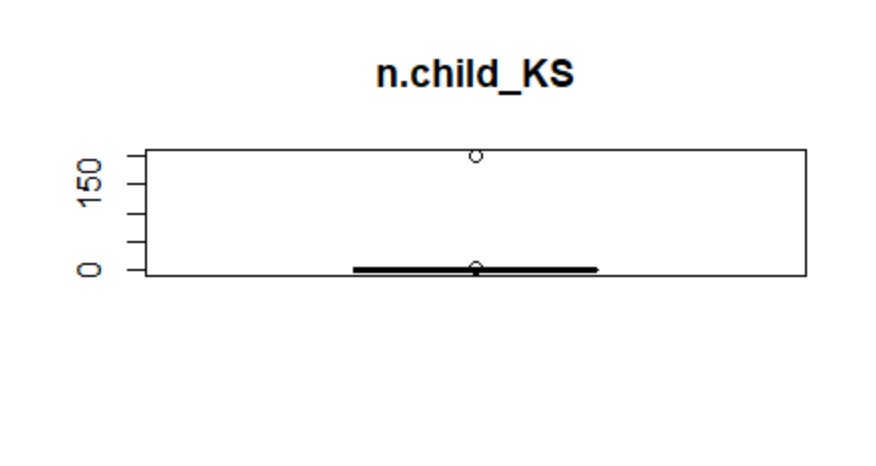
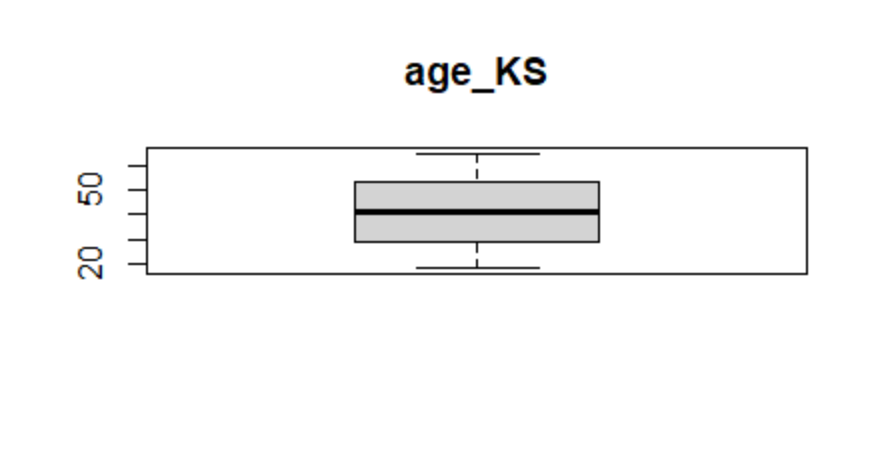
**nation.Southern\_KS, gender.Female\_KS, gender.Male\_KS, gender.Undis\_KS,**

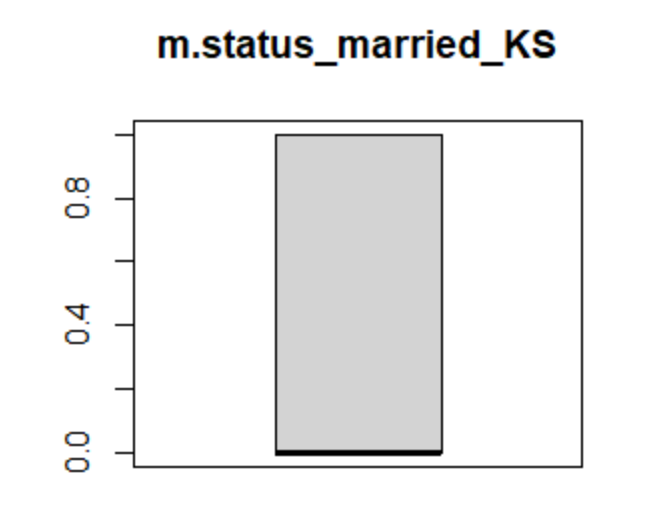
**political.Conservative\_KS, political.Liberal\_KS, political.NewDemocrat\_KS,**

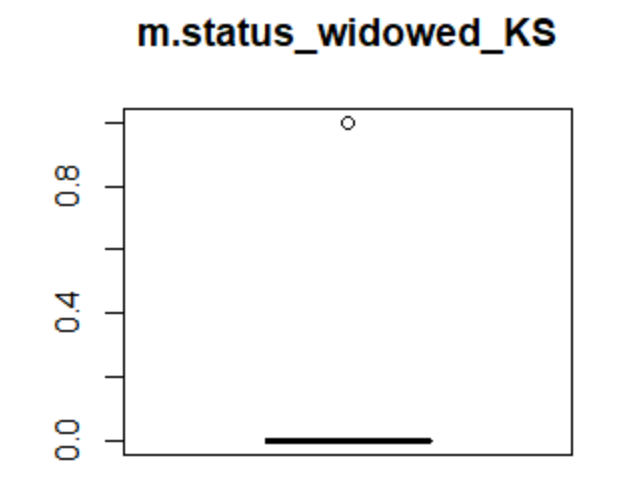
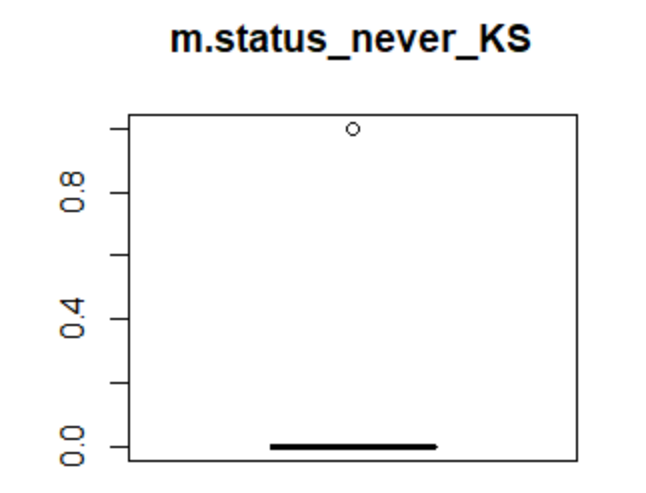
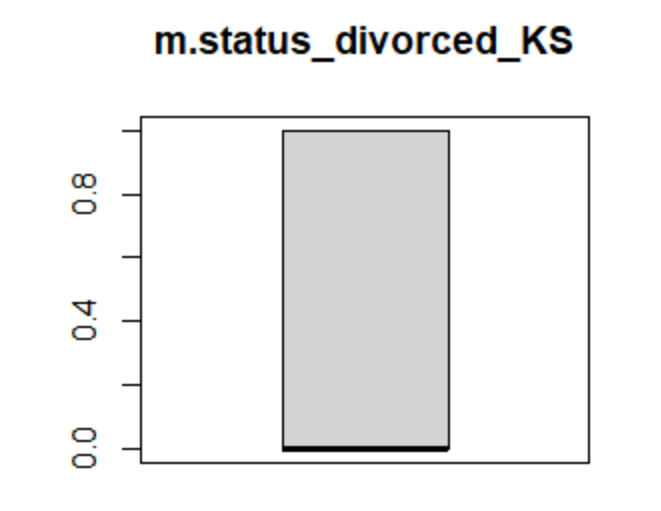
**political.Other\_KS,m.status\_divorced\_KS,m.status\_married\_KS,m.status\_never\_KS, m.status\_widowed\_KS)**

**3. Outliers**

**3.1. Create boxplots of all relevant variables (i.e. numeric, non-binary) to determine outliers.**

** **

****

****

From the above-given output, it’s very clear that the **Box Plot** of variables **n.child\_KS** has outliers in its observation because the **Maximum data point** in that variable is surely more than 1.5 times away from the **Median**.

**3.2. Comment on any outliers you see and deal with them appropriately.**

**Comment on outliers,**

age\_KS n.child\_KS income\_KS

Min. :18.00 Min. : 0.000 Min. : 0

1st Qu.:29.00 1st Qu.: 0.000 1st Qu.: 38000

Median :41.00 Median : 1.000 Median : 76409

Mean :41.23 Mean : 1.474 Mean : 76876

3rd Qu.:53.00 3rd Qu.: 2.000 3rd Qu.:114950

Max. :65.00 Max. :201.000 Max. :166378

food\_KS housing\_KS other\_KS

Min. :0.0000 Min. :0.0000 Min. :0.0000

1st Qu.:0.2500 1st Qu.:0.0600 1st Qu.:0.0700

Median :0.5100 Median :0.1800 Median :0.1800

Mean :0.5054 Mean :0.2426 Mean :0.2521

3rd Qu.:0.7600 3rd Qu.:0.3700 3rd Qu.:0.3900

Max. :1.0000 Max. :0.9800 Max. :0.9800

score\_KS Pol\_KS time1\_KS

Min. :-3.09000 Min. :-1.8092 Min. :0.0000

1st Qu.:-0.64000 1st Qu.: 0.4094 1st Qu.:0.1200

Median : 0.03000 Median : 1.4511 Median :0.4500

Mean : 0.02229 Mean : 1.4514 Mean :0.4646

3rd Qu.: 0.65000 3rd Qu.: 2.4475 3rd Qu.:0.8000

Max. : 3.12000 Max. : 5.2240 Max. :1.0000

scr\_KS group.Control\_KS group.Treat\_KS

Min. :-1.2985 Min. :0.0000 Min. :0.0000

1st Qu.: 0.5461 1st Qu.:0.0000 1st Qu.:0.0000

Median : 0.9660 Median :0.0000 Median :1.0000

Mean : 0.9635 Mean :0.4919 Mean :0.5081

3rd Qu.: 1.3761 3rd Qu.:1.0000 3rd Qu.:1.0000

Max. : 2.9733 Max. :1.0000 Max. :1.0000

hs.grad.No\_KS hs.grad.Yes\_KS nation.Asia\_KS nation.Europe\_KS

Min. :0.0000 Min. :0.0000 Min. :0.000 Min. :0.0000

1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:0.0000

Median :1.0000 Median :0.0000 Median :0.000 Median :0.0000

Mean :0.5283 Mean :0.4717 Mean :0.362 Mean :0.1545

3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.000 3rd Qu.:0.0000

Max. :1.0000 Max. :1.0000 Max. :1.000 Max. :1.0000

nation.NorthAmerica\_KS nation.Southern\_KS gender.Female\_KS

Min. :0.0000 Min. :0.0000 Min. :0.0000

1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000

Median :0.0000 Median :0.0000 Median :0.0000

Mean :0.3788 Mean :0.1048 Mean :0.3335

3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:1.0000

Max. :1.0000 Max. :1.0000 Max. :1.0000

gender.Male\_KS gender.Undis\_KS political.Conservative\_KS

Min. :0.0000 Min. :0.0000 Min. :0.0000

1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000

Median :0.0000 Median :0.0000 Median :0.0000

Mean :0.3384 Mean :0.3281 Mean :0.2941

3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000

Max. :1.0000 Max. :1.0000 Max. :1.0000

political.Liberal\_KS political.NewDemocrat\_KS political.Other\_KS

Min. :0.0000 Min. :0.0000 Min. :0.0000

1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000

Median :0.0000 Median :0.0000 Median :0.0000

Mean :0.3546 Mean :0.1938 Mean :0.1574

3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:0.0000

Max. :1.0000 Max. :1.0000 Max. :1.0000

m.status\_divorced\_KS m.status\_married\_KS m.status\_never\_KS

Min. :0.0000 Min. :0.000 Min. :0.0000

1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:0.0000

Median :0.0000 Median :0.000 Median :0.0000

Mean :0.2774 Mean :0.392 Mean :0.2391

3rd Qu.:1.0000 3rd Qu.:1.000 3rd Qu.:0.0000

Max. :1.0000 Max. :1.000 Max. :1.0000

m.status\_widowed\_KS

Min. :0.00000

1st Qu.:0.00000

Median :0.00000

Mean :0.09149

3rd Qu.:0.00000

Max. :1.00000

From the above-given output, it’s very clear that variables **n.child\_KS** has outliers in its observation because the **Maximum data point** in that variable is **201**, and it’s **100** times of **(IQR- 3rd Quartile)** away from the **Median**.

Where **Maximum data point = 201, Median = 1,** **IQR- 3rd Quartile = 2.**

**Dealing with Outliers appropriately,**

'data.frame': 2032 obs. of 29 variables:

$ age\_KS : num 21 53 44 35 20 35 55 37 49 21 ...

$ n.child\_KS : num 0 1 0 1 0 2 3 3 3 2 ...

$ income\_KS : num 138147 21953 87000 66420 146000 ...

$ food\_KS : num 0.13 0.75 0.82 0.67 0.26 0.04 0.87 0.8 0.22 0.2 ...

$ housing\_KS : num 0.87 0.2 0.08 0.05 0.18 0.17 0.11 0.19 0.67 0.2 ...

$ other\_KS : num 0 0.05 0.1 0.28 0.56 0.79 0.02 0.01 0.11 0.6 ...

$ score\_KS : num -0.15 -0.71 -0.16 -0.19 0.73 0.46 -0.55 -1.55 2.08 0.75 ...

$ Pol\_KS : num 0.932 1.892 1.708 3.41 2.849 ...

$ time1\_KS : num 0.57 0.52 0.16 0.61 0.09 0.03 0.78 0.88 0.16 0.5 ...

$ scr\_KS : num 0.783 0.906 1.327 1.177 1.717 ...

$ group.Control\_KS : num 1 0 0 0 0 1 1 0 1 0 ...

$ group.Treat\_KS : num 0 1 1 1 1 0 0 1 0 1 ...

$ hs.grad.No\_KS : num 1 0 1 1 1 0 1 1 0 0 ...

$ hs.grad.Yes\_KS : num 0 1 0 0 0 1 0 0 1 1 ...

$ nation.Asia\_KS : num 0 0 0 1 1 0 0 0 1 0 ...

$ nation.Europe\_KS : num 0 0 0 0 0 0 1 1 0 0 ...

$ nation.NorthAmerica\_KS : num 1 1 1 0 0 1 0 0 0 1 ...

$ nation.Southern\_KS : num 0 0 0 0 0 0 0 0 0 0 ...

$ gender.Female\_KS : num 0 0 1 0 0 0 0 1 0 0 ...

$ gender.Male\_KS : num 0 0 0 1 1 0 0 0 0 0 ...

$ gender.Undis\_KS : num 1 1 0 0 0 1 1 0 1 1 ...

$ political.Conservative\_KS: num 0 0 0 0 0 0 1 0 1 0 ...

$ political.Liberal\_KS : num 0 0 1 0 1 0 0 1 0 1 ...

$ political.NewDemocrat\_KS : num 1 1 0 1 0 0 0 0 0 0 ...

$ political.Other\_KS : num 0 0 0 0 0 1 0 0 0 0 ...

$ m.status\_divorced\_KS : num 0 0 0 1 0 0 1 0 0 1 ...

$ m.status\_married\_KS : num 0 0 0 0 0 0 0 1 1 0 ...

$ m.status\_never\_KS : num 0 0 1 0 1 1 0 0 0 0 ...

$ m.status\_widowed\_KS : num 1 1 0 0 0 0 0 0 0 0 ...

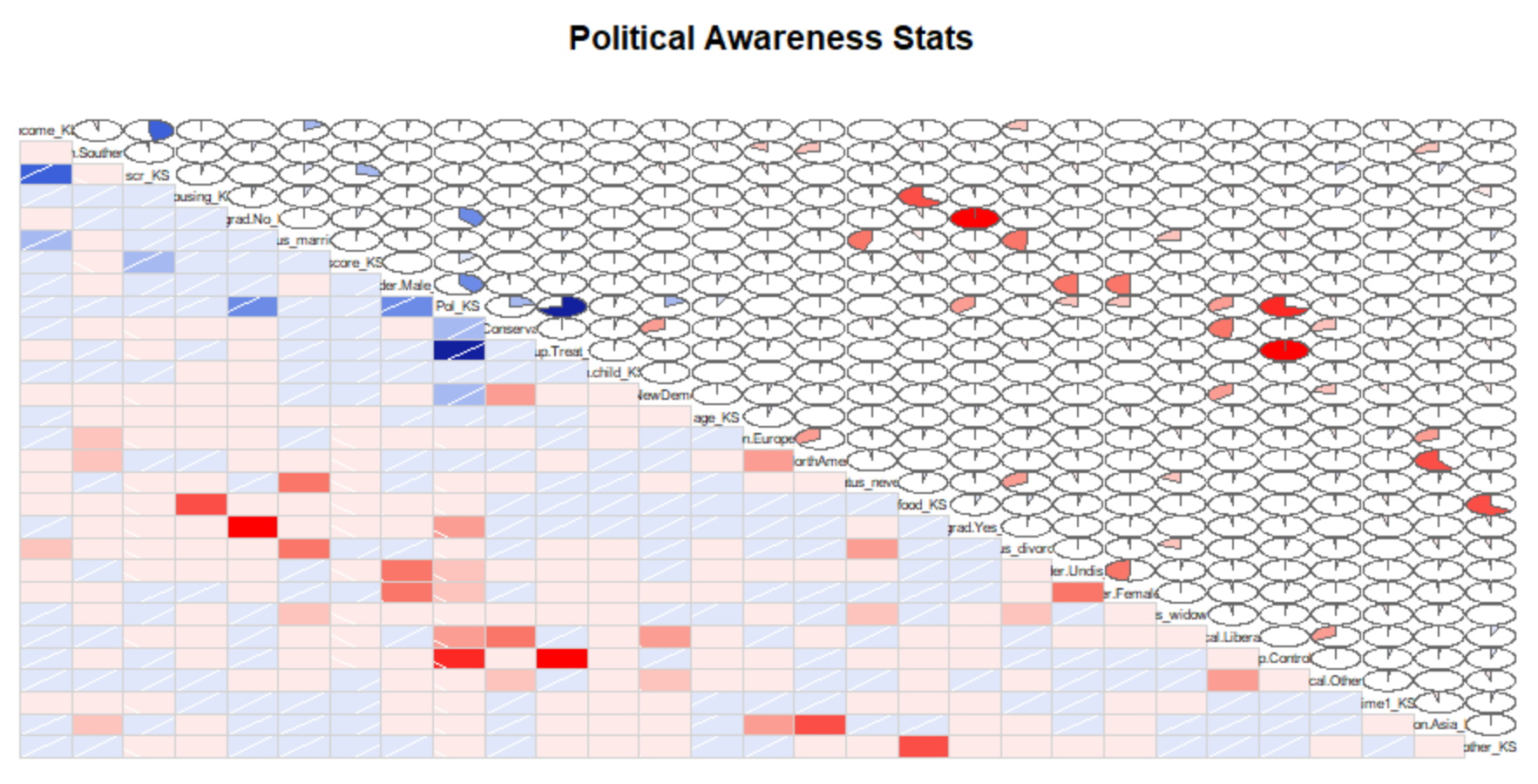
From the above-given output, it’s very clear that the observation with **n.child\_KS** **= 201** is **ignored** for further Regression. Now, the data consists of **2032 observations** instead of **2033 observations.**

The **exploration of data** is done and removed the only record which has an outlier as it will not affect the **size of the dataset**. Later, **methodology** **-** from where and how the data is taken should be noted. Finally, **precedence -** the rules that the company follows for dealing with outliers should also be followed.

**4. Exploratory Analysis**

**4.1. Correlations: Create both numeric and graphical correlations (as demonstrated in class) and comment on noteworthy correlations you observe. Are these surprising? Do they make sense?**

**Graphical correlations:**

****

**Numeric correlations:**

age\_KS n.child\_KS income\_KS food\_KS housing\_KS

age\_KS 1.00 0.00 0.02 0.00 0.00

n.child\_KS 0.00 1.00 0.02 0.01 -0.01

income\_KS 0.02 0.02 1.00 -0.01 0.00

food\_KS 0.00 0.01 -0.01 1.00 -0.67

housing\_KS 0.00 -0.01 0.00 -0.67 1.00

other\_KS 0.01 -0.01 0.00 -0.67 0.02

score\_KS -0.02 0.00 0.02 -0.06 0.04

Pol\_KS 0.07 0.01 0.01 -0.01 0.03

time1\_KS 0.02 -0.02 -0.04 -0.04 0.04

scr\_KS 0.00 0.01 0.47 -0.02 0.02

group.Control\_KS -0.02 -0.01 0.01 -0.01 -0.03

group.Treat\_KS 0.02 0.01 -0.01 0.01 0.03

hs.grad.No\_KS -0.03 -0.02 0.00 -0.05 0.03

hs.grad.Yes\_KS 0.03 0.02 0.00 0.05 -0.03

nation.Asia\_KS 0.01 -0.02 0.02 0.01 0.00

nation.Europe\_KS 0.03 0.00 0.01 0.02 -0.02

nation.NorthAmerica\_KS 0.00 0.02 -0.01 0.00 -0.01

nation.Southern\_KS -0.04 0.00 -0.03 -0.04 0.04

gender.Female\_KS -0.03 -0.01 0.00 0.01 0.00

gender.Male\_KS 0.04 0.02 0.02 -0.03 0.01

gender.Undis\_KS 0.00 -0.01 -0.02 0.02 -0.01

political.Conservative\_KS 0.03 0.02 0.00 0.00 -0.01

political.Liberal\_KS -0.01 0.00 0.01 -0.02 -0.02

political.NewDemocrat\_KS 0.00 -0.01 -0.02 0.03 -0.02

political.Other\_KS -0.01 -0.02 0.01 -0.01 0.07

m.status\_divorced\_KS -0.02 -0.01 -0.19 0.05 -0.03

m.status\_married\_KS 0.00 0.01 0.16 -0.08 0.04

m.status\_never\_KS 0.00 0.01 0.00 0.03 -0.02

m.status\_widowed\_KS 0.02 -0.02 0.03 0.00 0.01

other\_KS score\_KS Pol\_KS time1\_KS scr\_KS

age\_KS 0.01 -0.02 0.07 0.02 0.00

n.child\_KS -0.01 0.00 0.01 -0.02 0.01

income\_KS 0.00 0.02 0.01 -0.04 0.47

food\_KS -0.67 -0.06 -0.01 -0.04 -0.02

housing\_KS 0.02 0.04 0.03 0.04 0.02

other\_KS 1.00 0.05 -0.03 0.02 -0.01

score\_KS 0.05 1.00 0.13 0.00 0.25

Pol\_KS -0.03 0.13 1.00 -0.02 0.02

time1\_KS 0.02 0.00 -0.02 1.00 -0.03

scr\_KS -0.01 0.25 0.02 -0.03 1.00

group.Control\_KS 0.05 -0.04 -0.74 0.03 0.01

group.Treat\_KS -0.05 0.04 0.74 -0.03 -0.01

hs.grad.No\_KS 0.03 0.05 0.35 0.03 0.00

hs.grad.Yes\_KS -0.03 -0.05 -0.35 -0.03 0.00

nation.Asia\_KS 0.00 0.03 0.00 -0.04 0.04

nation.Europe\_KS 0.01 -0.03 -0.01 0.04 -0.05

nation.NorthAmerica\_KS -0.03 0.00 0.00 0.01 0.01

nation.Southern\_KS 0.03 -0.02 0.01 0.00 -0.02

gender.Female\_KS -0.01 0.02 -0.20 0.00 0.02

gender.Male\_KS 0.03 0.00 0.39 0.00 0.02

gender.Undis\_KS -0.02 -0.01 -0.19 0.00 -0.04

political.Conservative\_KS 0.00 0.00 0.23 0.03 -0.01

political.Liberal\_KS 0.06 -0.01 -0.32 0.02 0.00

political.NewDemocrat\_KS -0.04 0.00 0.17 -0.07 -0.05

political.Other\_KS -0.03 0.02 -0.05 0.01 0.07

m.status\_divorced\_KS -0.03 0.01 -0.03 0.00 -0.05

m.status\_married\_KS 0.04 0.02 0.03 0.02 0.07

m.status\_never\_KS -0.02 -0.02 0.00 -0.01 -0.02

m.status\_widowed\_KS 0.01 -0.01 -0.01 -0.03 -0.01

group.Control\_KS group.Treat\_KS hs.grad.No\_KS

age\_KS -0.02 0.02 -0.03

n.child\_KS -0.01 0.01 -0.02

income\_KS 0.01 -0.01 0.00

food\_KS -0.01 0.01 -0.05

housing\_KS -0.03 0.03 0.03

other\_KS 0.05 -0.05 0.03

score\_KS -0.04 0.04 0.05

Pol\_KS -0.74 0.74 0.35

time1\_KS 0.03 -0.03 0.03

scr\_KS 0.01 -0.01 0.00

group.Control\_KS 1.00 -1.00 0.01

group.Treat\_KS -1.00 1.00 -0.01

hs.grad.No\_KS 0.01 -0.01 1.00

hs.grad.Yes\_KS -0.01 0.01 -1.00

nation.Asia\_KS 0.01 -0.01 0.01

nation.Europe\_KS -0.01 0.01 -0.01

nation.NorthAmerica\_KS 0.01 -0.01 -0.02

nation.Southern\_KS -0.01 0.01 0.03

gender.Female\_KS 0.03 -0.03 -0.01

gender.Male\_KS -0.04 0.04 0.02

gender.Undis\_KS 0.00 0.00 -0.01

political.Conservative\_KS 0.00 0.00 -0.02

political.Liberal\_KS 0.00 0.00 0.02

political.NewDemocrat\_KS 0.01 -0.01 -0.01

political.Other\_KS -0.01 0.01 0.00

m.status\_divorced\_KS 0.04 -0.04 -0.02

m.status\_married\_KS -0.05 0.05 0.00

m.status\_never\_KS 0.00 0.00 0.01

m.status\_widowed\_KS 0.02 -0.02 0.00

hs.grad.Yes\_KS nation.Asia\_KS nation.Europe\_KS

age\_KS 0.03 0.01 0.03

n.child\_KS 0.02 -0.02 0.00

income\_KS 0.00 0.02 0.01

food\_KS 0.05 0.01 0.02

housing\_KS -0.03 0.00 -0.02

other\_KS -0.03 0.00 0.01

score\_KS -0.05 0.03 -0.03

Pol\_KS -0.35 0.00 -0.01

time1\_KS -0.03 -0.04 0.04

scr\_KS 0.00 0.04 -0.05

group.Control\_KS -0.01 0.01 -0.01

group.Treat\_KS 0.01 -0.01 0.01

hs.grad.No\_KS -1.00 0.01 -0.01

hs.grad.Yes\_KS 1.00 -0.01 0.01

nation.Asia\_KS -0.01 1.00 -0.32

nation.Europe\_KS 0.01 -0.32 1.00

nation.NorthAmerica\_KS 0.02 -0.59 -0.33

nation.Southern\_KS -0.03 -0.26 -0.15

gender.Female\_KS 0.01 -0.02 0.01

gender.Male\_KS -0.02 -0.01 0.00

gender.Undis\_KS 0.01 0.03 -0.01

political.Conservative\_KS 0.02 0.02 -0.02

political.Liberal\_KS -0.02 0.00 0.02

political.NewDemocrat\_KS 0.01 -0.02 0.03

political.Other\_KS 0.00 0.00 -0.04

m.status\_divorced\_KS 0.02 -0.05 0.04

m.status\_married\_KS 0.00 0.02 0.01

m.status\_never\_KS -0.01 0.02 -0.02

m.status\_widowed\_KS 0.00 0.02 -0.04

nation.NorthAmerica\_KS nation.Southern\_KS

age\_KS 0.00 -0.04

n.child\_KS 0.02 0.00

income\_KS -0.01 -0.03

food\_KS 0.00 -0.04

housing\_KS -0.01 0.04

other\_KS -0.03 0.03

score\_KS 0.00 -0.02

Pol\_KS 0.00 0.01

time1\_KS 0.01 0.00

scr\_KS 0.01 -0.02

group.Control\_KS 0.01 -0.01

group.Treat\_KS -0.01 0.01

hs.grad.No\_KS -0.02 0.03

hs.grad.Yes\_KS 0.02 -0.03

nation.Asia\_KS -0.59 -0.26

nation.Europe\_KS -0.33 -0.15

nation.NorthAmerica\_KS 1.00 -0.27

nation.Southern\_KS -0.27 1.00

gender.Female\_KS 0.01 0.00

gender.Male\_KS 0.01 0.00

gender.Undis\_KS -0.02 0.00

political.Conservative\_KS 0.01 -0.02

political.Liberal\_KS -0.04 0.02

political.NewDemocrat\_KS 0.02 -0.03

political.Other\_KS 0.01 0.02

m.status\_divorced\_KS 0.03 -0.01

m.status\_married\_KS -0.02 0.00

m.status\_never\_KS -0.01 0.02

m.status\_widowed\_KS 0.01 -0.01

gender.Female\_KS gender.Male\_KS gender.Undis\_KS

age\_KS -0.03 0.04 0.00

n.child\_KS -0.01 0.02 -0.01

income\_KS 0.00 0.02 -0.02

food\_KS 0.01 -0.03 0.02

housing\_KS 0.00 0.01 -0.01

other\_KS -0.01 0.03 -0.02

score\_KS 0.02 0.00 -0.01

Pol\_KS -0.20 0.39 -0.19

time1\_KS 0.00 0.00 0.00

scr\_KS 0.02 0.02 -0.04

group.Control\_KS 0.03 -0.04 0.00

group.Treat\_KS -0.03 0.04 0.00

hs.grad.No\_KS -0.01 0.02 -0.01

hs.grad.Yes\_KS 0.01 -0.02 0.01

nation.Asia\_KS -0.02 -0.01 0.03

nation.Europe\_KS 0.01 0.00 -0.01

nation.NorthAmerica\_KS 0.01 0.01 -0.02

nation.Southern\_KS 0.00 0.00 0.00

gender.Female\_KS 1.00 -0.51 -0.49

gender.Male\_KS -0.51 1.00 -0.50

gender.Undis\_KS -0.49 -0.50 1.00

political.Conservative\_KS 0.02 -0.01 -0.01

political.Liberal\_KS -0.02 0.03 -0.02

political.NewDemocrat\_KS 0.00 0.00 0.00

political.Other\_KS 0.00 -0.03 0.02

m.status\_divorced\_KS -0.01 0.02 -0.01

m.status\_married\_KS 0.01 -0.03 0.02

m.status\_never\_KS 0.01 0.03 -0.03

m.status\_widowed\_KS -0.01 -0.01 0.02

political.Conservative\_KS political.Liberal\_KS

age\_KS 0.03 -0.01

n.child\_KS 0.02 0.00

income\_KS 0.00 0.01

food\_KS 0.00 -0.02

housing\_KS -0.01 -0.02

other\_KS 0.00 0.06

score\_KS 0.00 -0.01

Pol\_KS 0.23 -0.32

time1\_KS 0.03 0.02

scr\_KS -0.01 0.00

group.Control\_KS 0.00 0.00

group.Treat\_KS 0.00 0.00

hs.grad.No\_KS -0.02 0.02

hs.grad.Yes\_KS 0.02 -0.02

nation.Asia\_KS 0.02 0.00

nation.Europe\_KS -0.02 0.02

nation.NorthAmerica\_KS 0.01 -0.04

nation.Southern\_KS -0.02 0.02

gender.Female\_KS 0.02 -0.02

gender.Male\_KS -0.01 0.03

gender.Undis\_KS -0.01 -0.02

political.Conservative\_KS 1.00 -0.48

political.Liberal\_KS -0.48 1.00

political.NewDemocrat\_KS -0.32 -0.36

political.Other\_KS -0.28 -0.32

m.status\_divorced\_KS 0.00 0.01

m.status\_married\_KS 0.03 -0.02

m.status\_never\_KS -0.03 0.03

m.status\_widowed\_KS -0.01 -0.02

political.NewDemocrat\_KS political.Other\_KS

age\_KS 0.00 -0.01

n.child\_KS -0.01 -0.02

income\_KS -0.02 0.01

food\_KS 0.03 -0.01

housing\_KS -0.02 0.07

other\_KS -0.04 -0.03

score\_KS 0.00 0.02

Pol\_KS 0.17 -0.05

time1\_KS -0.07 0.01

scr\_KS -0.05 0.07

group.Control\_KS 0.01 -0.01

group.Treat\_KS -0.01 0.01

hs.grad.No\_KS -0.01 0.00

hs.grad.Yes\_KS 0.01 0.00

nation.Asia\_KS -0.02 0.00

nation.Europe\_KS 0.03 -0.04

nation.NorthAmerica\_KS 0.02 0.01

nation.Southern\_KS -0.03 0.02

gender.Female\_KS 0.00 0.00

gender.Male\_KS 0.00 -0.03

gender.Undis\_KS 0.00 0.02

political.Conservative\_KS -0.32 -0.28

political.Liberal\_KS -0.36 -0.32

political.NewDemocrat\_KS 1.00 -0.21

political.Other\_KS -0.21 1.00

m.status\_divorced\_KS 0.00 -0.01

m.status\_married\_KS 0.00 -0.02

m.status\_never\_KS -0.02 0.03

m.status\_widowed\_KS 0.02 0.01

m.status\_divorced\_KS m.status\_married\_KS

age\_KS -0.02 0.00

n.child\_KS -0.01 0.01

income\_KS -0.19 0.16

food\_KS 0.05 -0.08

housing\_KS -0.03 0.04

other\_KS -0.03 0.04

score\_KS 0.01 0.02

Pol\_KS -0.03 0.03

time1\_KS 0.00 0.02

scr\_KS -0.05 0.07

group.Control\_KS 0.04 -0.05

group.Treat\_KS -0.04 0.05

hs.grad.No\_KS -0.02 0.00

hs.grad.Yes\_KS 0.02 0.00

nation.Asia\_KS -0.05 0.02

nation.Europe\_KS 0.04 0.01

nation.NorthAmerica\_KS 0.03 -0.02

nation.Southern\_KS -0.01 0.00

gender.Female\_KS -0.01 0.01

gender.Male\_KS 0.02 -0.03

gender.Undis\_KS -0.01 0.02

political.Conservative\_KS 0.00 0.03

political.Liberal\_KS 0.01 -0.02

political.NewDemocrat\_KS 0.00 0.00

political.Other\_KS -0.01 -0.02

m.status\_divorced\_KS 1.00 -0.50

m.status\_married\_KS -0.50 1.00

m.status\_never\_KS -0.35 -0.45

m.status\_widowed\_KS -0.20 -0.25

m.status\_never\_KS m.status\_widowed\_KS

age\_KS 0.00 0.02

n.child\_KS 0.01 -0.02

income\_KS 0.00 0.03

food\_KS 0.03 0.00

housing\_KS -0.02 0.01

other\_KS -0.02 0.01

score\_KS -0.02 -0.01

Pol\_KS 0.00 -0.01

time1\_KS -0.01 -0.03

scr\_KS -0.02 -0.01

group.Control\_KS 0.00 0.02

group.Treat\_KS 0.00 -0.02

hs.grad.No\_KS 0.01 0.00

hs.grad.Yes\_KS -0.01 0.00

nation.Asia\_KS 0.02 0.02

nation.Europe\_KS -0.02 -0.04

nation.NorthAmerica\_KS -0.01 0.01

nation.Southern\_KS 0.02 -0.01

gender.Female\_KS 0.01 -0.01

gender.Male\_KS 0.03 -0.01

gender.Undis\_KS -0.03 0.02

political.Conservative\_KS -0.03 -0.01

political.Liberal\_KS 0.03 -0.02

political.NewDemocrat\_KS -0.02 0.02

political.Other\_KS 0.03 0.01

m.status\_divorced\_KS -0.35 -0.20

m.status\_married\_KS -0.45 -0.25

m.status\_never\_KS 1.00 -0.18

m.status\_widowed\_KS -0.18 1.00

**Noteworthy correlations:**

From the above-given output, these are the noteworthy correlations,

**PERFECT CORRELATIONS (=1),**

1. group.Control\_KS, group.Treat\_KS **(PERFECT NEGATIVE CORRELATIONSHIP (-1))**
2. hs.grad.Yes\_KS, hs.grad.No\_KS **(PERFECT NEGATIVE CORRELATIONSHIP (-1))**

**STRONG LINEAR RELATIONSHIP (0.70 - 0.99),**

1. group.Control\_KS , pol\_KS **(STRONG NEGATIVE CORRELATIONSHIP (-0.74))**
2. group.Treat\_KS, pol\_KS **(STRONG POSITIVE CORRELATIONSHIP (0.74))**

**MODERATE LINEAR RELATIONSHIP (0.50 - 0.69),**

1. housing\_KS , food\_KS **(MODERATE NEGATIVE CORRELATIONSHIP (-0.67))**
2. other\_KS, food\_KS **(MODERATE NEGATIVE CORRELATIONSHIP (-0.67))**
3. nation.Asia\_KS, nation.NorthAmerica\_KS **(MODERATE NEGATIVE CORRELATIONSHIP (-0.59))**
4. gender.Female\_KS, gender.Male\_KS **(MODERATE NEGATIVE CORRELATIONSHIP (-0.51))**
5. gender.Male\_KS, gender.Undis\_KS **(MODERATE NEGATIVE CORRELATIONSHIP (-0.50))**
6. m.status\_divorced\_KS, m.status\_married\_KS **(MODERATE NEGATIVE CORRELATIONSHIP (-0.50))**

**Are these surprising? Do they make sense?**

From the above-given output, it’s very clear that the variables

1. **(group.Control\_KS, group.Treat\_KS)** have **PERFECT NEGATIVE CORRELATIONSHIP**. There is no surprise because different groups have different opinions and characteristics. And they do make sense.
2. **(hs.grad.Yes\_KS, hs.grad.No\_KS)** have **PERFECT NEGATIVE CORRELATIONSHIP**. There is no surprise because grad yes and no should be opposite. And they do make sense.
3. **(group.Control\_KS,pol\_KS)** have **STRONG NEGATIVE CORRELATIONSHIP,** and **(group.Treat\_KS, pol\_KS)** have **STRONG POSITIVE CORRELATIONSHIP.** There is no surprise because different groups have different opinions and characteristics. And they do make sense.
4. **(housing\_KS, food\_KS)** **, (other\_KS, food\_KS)** and **(nation.Asia\_KS, nation.NorthAmerica\_KS)** have **MODERATE NEGATIVE CORRELATIONSHIP.** This is surprising as any of these don’t make sense.
5. **(gender.Female\_KS, gender.Male\_KS**) and **(gender.Male\_KS, gender.Undis\_KS)** have **MODERATE NEGATIVE CORRELATIONSHIP.** There is no surprise because different genders have different opinions and characteristics. And they do make sense.
6. **(m.status\_divorced\_KS, m.status\_married\_KS)** have **MODERATE NEGATIVE CORRELATIONSHIP.** There is no surprise because different marital status has different opinions and characteristics. And they do make sense.

**5. Simple Linear Regression**

**5.1. Create a simple linear regression model using Pol as the dependent variable and score as the independent. Create a scatter plot of the two variables and overlay the regression line.**

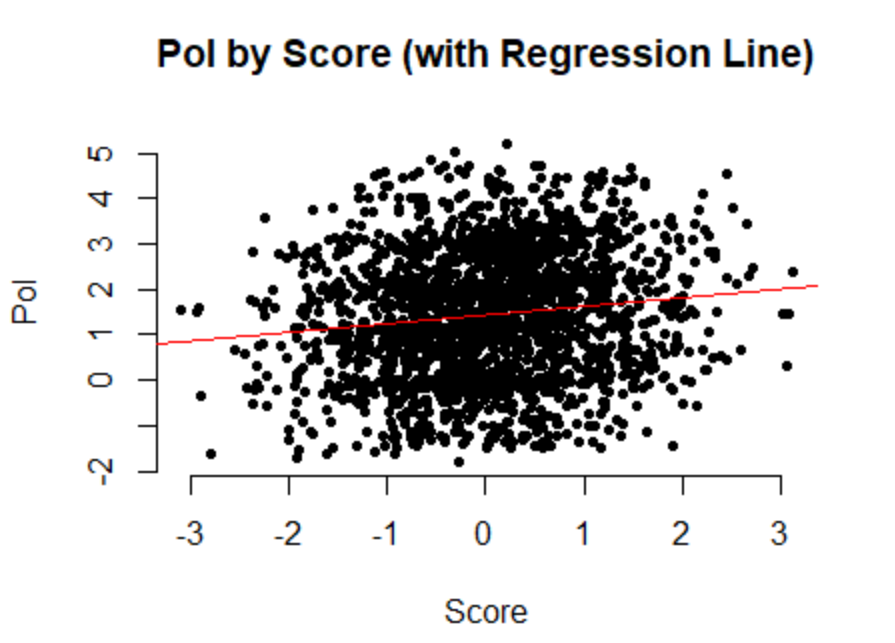
**simple linear regression model using Pol and score:**

Coefficients:

(Intercept) score\_KS

1.446 0.192

**scatter plot with a regression line using Pol and score:**



For **Measure of Political Involvement:**

**Pol\_KS = 0.192 \* score\_KS + 1.446**

For every additional **score\_KS(Score on Political Awareness Test),** there will be **0.192** additional **pol\_KS(Measure of Political Involvement).**

**5.2. Create a simple linear regression model using Pol as the dependent variable and scr as the independent. Create a scatter plot of the two variables and overlay the regression line.**

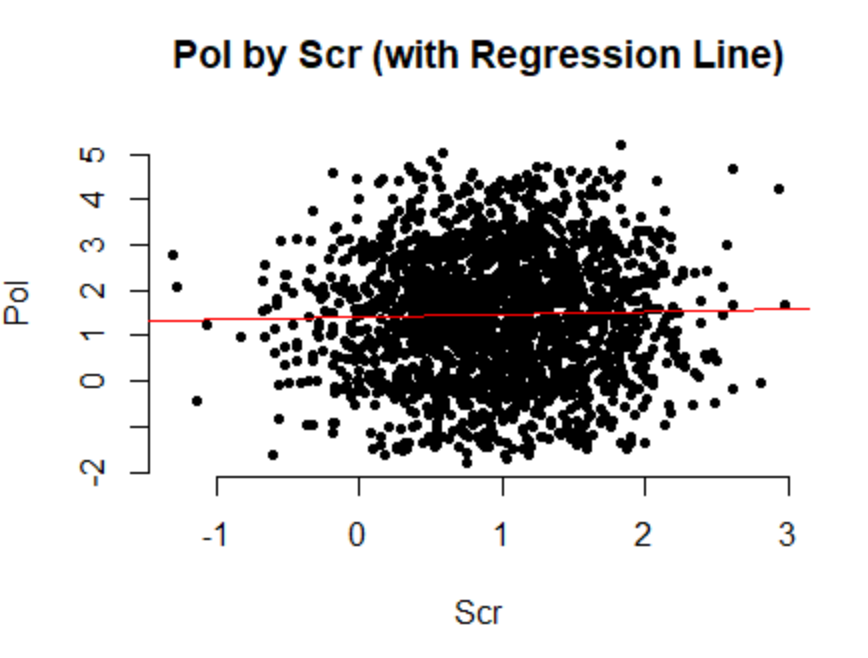
**simple linear regression model using Pol and scr:**

Coefficients:

(Intercept) scr\_KS

1.39707 0.05543

**scatter plot with a regression line using Pol and scr:**

****

For **Measure of Political Involvement:**

**Pol\_KS = 0.05543 \* scr\_KS + 1.39707**

For every additional **scr\_KS(Standardized Score Test**), there will be **0.05543** additional **pol\_KS(Measure of Political Involvement).**

**5.3. Compare the models. Which model is superior? Why?**

**Correlation (Pol\_KS,score\_KS):**

[1] 0.1323011

**Summary (Pol\_KS,score\_KS):**

Call:

lm(formula = Pol\_KS ~ score\_KS, data = pol\_eng)

Residuals:

Min 1Q Median 3Q Max

-3.2589 -1.0705 -0.0362 0.9586 3.7374

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.44623 0.03055 47.336 < 2e-16

score\_KS 0.19202 0.03193 6.014 0.00000000214

(Intercept) \*\*\*

score\_KS \*\*\*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.377 on 2030 degrees of freedom

Multiple R-squared: 0.0175, Adjusted R-squared: 0.01702

F-statistic: 36.17 on 1 and 2030 DF, p-value: 0.000000002144

**Correlation (Pol\_KS,scr\_KS):**

[1] 0.02424342

**Summary (Pol\_KS,scr\_KS):**

Call:

lm(formula = Pol\_KS ~ scr\_KS, data = pol\_eng)

Residuals:

Min 1Q Median 3Q Max

-3.2479 -1.0441 -0.0069 0.9951 3.7256

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.39707 0.05777 24.184 <2e-16 \*\*\*

scr\_KS 0.05543 0.05073 1.093 0.275

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.389 on 2030 degrees of freedom

Multiple R-squared: 0.0005877, Adjusted R-squared: 9.542e-05

F-statistic: 1.194 on 1 and 2030 DF, p-value: 0.2747

From the above-given output, it’s very clear that (**Pol\_KS,score\_KS)**  is superior because,

1. **F-stat, p-value,** and **Adjusted R-squared** value of (**Pol\_KS,score\_KS)** modelis better than **(Pol\_KS,scr\_KS)** model**.**
2. **Residuals** of both models are near to **zero (0).**
3. **T-value** of(**Pol\_KS,score\_KS)** model is less than **0.05** but **(Pol\_KS,scr\_KS)** model has greater than 0.05.
4. For both models **Coefficients Estimate** (positive or negative) which matches with correlation.
5. For both models **square root** of **Multiple R-squared** gives **Correlation.**

**Pol\_KS = 0.192 \* score\_KS + 1.446**

**6. Model Development - Multivariate**

**As demonstrated in class, create two models using two automatic variable selection techniques discussed in class (Full, Backward). For each model interpret and comment on the five main measures we discussed in class. (Your commentary should be yours, not simply copied from my example):**

**Full Model - Summary:**

Call:

lm(formula = Pol\_KS ~ age\_KS + n.child\_KS + income\_KS + food\_KS +

housing\_KS + other\_KS + score\_KS + time1\_KS + scr\_KS + group.Control\_KS +

group.Treat\_KS + hs.grad.No\_KS + hs.grad.Yes\_KS + nation.Asia\_KS +

nation.Europe\_KS + nation.NorthAmerica\_KS + nation.Southern\_KS +

gender.Female\_KS + gender.Male\_KS + gender.Undis\_KS + political.Conservative\_KS +

political.Liberal\_KS + political.NewDemocrat\_KS + political.Other\_KS +

m.status\_divorced\_KS + m.status\_married\_KS + m.status\_never\_KS +

m.status\_widowed\_KS, data = pol\_eng, na.action = na.omit)

Residuals:

Min 1Q Median 3Q Max

-0.97610 -0.16715 0.00044 0.17388 1.10184

Coefficients: (7 not defined because of singularities)

Estimate Std. Error t value

(Intercept) 1.0756138199 0.0443883021 24.232

age\_KS 0.0054479533 0.0004311633 12.635

n.child\_KS -0.0000754242 0.0049939461 -0.015

income\_KS 0.0000006816 0.0000001531 4.451

food\_KS 0.0214833757 0.0267784326 0.802

housing\_KS -0.0115702529 0.0353280237 -0.328

other\_KS NA NA NA

score\_KS 0.1242475460 0.0064869334 19.154

time1\_KS -0.0022685975 0.0171733765 -0.132

scr\_KS 0.0040142100 0.0116021060 0.346

group.Control\_KS -1.9626142731 0.0119031713 -164.882

group.Treat\_KS NA NA NA

hs.grad.No\_KS 1.0509461605 0.0119007218 88.309

hs.grad.Yes\_KS NA NA NA

nation.Asia\_KS -0.0057746985 0.0208157182 -0.277

nation.Europe\_KS -0.0168194004 0.0238206208 -0.706

nation.NorthAmerica\_KS -0.0046769342 0.0207140557 -0.226

nation.Southern\_KS NA NA NA

gender.Female\_KS -0.0066710767 0.0145788671 -0.458

gender.Male\_KS 1.1092926189 0.0145542800 76.218

gender.Undis\_KS NA NA NA

political.Conservative\_KS 0.6719386706 0.0185692055 36.186

political.Liberal\_KS -0.5224732337 0.0180323419 -28.974

political.NewDemocrat\_KS 0.6791428064 0.0202140808 33.598

political.Other\_KS NA NA NA

m.status\_divorced\_KS 0.0014214877 0.0227602124 0.062

m.status\_married\_KS -0.0321086248 0.0217943997 -1.473

m.status\_never\_KS -0.0133025581 0.0230462400 -0.577

m.status\_widowed\_KS NA NA NA

Pr(>|t|)

(Intercept) < 2e-16 \*\*\*

age\_KS < 2e-16 \*\*\*

n.child\_KS 0.988

income\_KS 0.00000902 \*\*\*

food\_KS 0.422

housing\_KS 0.743

other\_KS NA

score\_KS < 2e-16 \*\*\*

time1\_KS 0.895

scr\_KS 0.729

group.Control\_KS < 2e-16 \*\*\*

group.Treat\_KS NA

hs.grad.No\_KS < 2e-16 \*\*\*

hs.grad.Yes\_KS NA

nation.Asia\_KS 0.781

nation.Europe\_KS 0.480

nation.NorthAmerica\_KS 0.821

nation.Southern\_KS NA

gender.Female\_KS 0.647

gender.Male\_KS < 2e-16 \*\*\*

gender.Undis\_KS NA

political.Conservative\_KS < 2e-16 \*\*\*

political.Liberal\_KS < 2e-16 \*\*\*

political.NewDemocrat\_KS < 2e-16 \*\*\*

political.Other\_KS NA

m.status\_divorced\_KS 0.950

m.status\_married\_KS 0.141

m.status\_never\_KS 0.564

m.status\_widowed\_KS NA

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2666 on 2010 degrees of freedom

Multiple R-squared: 0.9635, Adjusted R-squared: 0.9631

F-statistic: 2528 on 21 and 2010 DF, p-value: < 2.2e-16

**Backward Model – Summary:**

Call:

lm(formula = Pol\_KS ~ age\_KS + income\_KS + score\_KS + group.Control\_KS +

hs.grad.No\_KS + gender.Male\_KS + political.Conservative\_KS +

political.Liberal\_KS + political.NewDemocrat\_KS + m.status\_married\_KS,

data = pol\_eng, na.action = na.omit)

Residuals:

Min 1Q Median 3Q Max

-0.95781 -0.16643 0.00095 0.17406 1.09697

Coefficients:

Estimate Std. Error t value

(Intercept) 1.0724518325 0.0269213039 39.837

age\_KS 0.0054361762 0.0004296336 12.653

income\_KS 0.0000006943 0.0000001329 5.224

score\_KS 0.1245458199 0.0061893800 20.123

group.Control\_KS -1.9623496036 0.0118447298 -165.673

hs.grad.No\_KS 1.0501177574 0.0118553966 88.577

gender.Male\_KS 1.1121736791 0.0125156932 88.862

political.Conservative\_KS 0.6721254106 0.0184570354 36.416

political.Liberal\_KS -0.5227747174 0.0178963540 -29.211

political.NewDemocrat\_KS 0.6793620797 0.0200409686 33.899

m.status\_married\_KS -0.0287945907 0.0122861492 -2.344

Pr(>|t|)

(Intercept) < 2e-16 \*\*\*

age\_KS < 2e-16 \*\*\*

income\_KS 0.000000193 \*\*\*

score\_KS < 2e-16 \*\*\*

group.Control\_KS < 2e-16 \*\*\*

hs.grad.No\_KS < 2e-16 \*\*\*

gender.Male\_KS < 2e-16 \*\*\*

political.Conservative\_KS < 2e-16 \*\*\*

political.Liberal\_KS < 2e-16 \*\*\*

political.NewDemocrat\_KS < 2e-16 \*\*\*

m.status\_married\_KS 0.0192 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2661 on 2021 degrees of freedom

Multiple R-squared: 0.9635, Adjusted R-squared: 0.9633

F-statistic: 5328 on 10 and 2021 DF, p-value: < 2.2e-16

**1. F-Stat**

**Full Model:**

**F-statistic: 2528 on 21 and 2010 DF, p-value: < 2.2e-16**

Here F-Stat is really good, and the p-value of the F-Stat is used for checking the **significance** of the model. Also, the p-value is **< 2.2e-16**, very less than 0.05 and it is good as well.

**Backward Model:**

**F-statistic: 5328 on 10 and 2021 DF, p-value: < 2.2e-16**

Here F-Stat is really good, and the p-value of the F-Stat is used for checking the **significance** of the model. Also, the p-value is **< 2.2e-16**, very less than 0.05 and it is good as well.

**2. R-Squared value**

**Full Model:**

**Adjusted R-squared: 0.9631**

Here **R-squared** is really good, and it is pretty close to **1**. Also, it checks for the **explanatory power** of the model.

**Backward Model:**

**Adjusted R-squared: 0.9633**

Here **R-squared** is really good, and it is pretty close to **1**. Also, it checks for the **explanatory power** of the model.

**3. Residuals**

**Full Model:**

**Residuals:**

**Min 1Q Median 3Q Max**

**-0.97610 -0.16715 0.00044 0.17388 1.10184**

Here we can see that all the residuals are almost equal to **0.** We need to look only into **1Q, Median, 3Q** but not the **Min** and **Max.** Also, it checks for the **Symmetry** of the model.

**Backward Model:**

**Residuals:**

**Min 1Q Median 3Q Max**

**-0.95781 -0.16643 0.00095 0.17406 1.09697**

Here we can see that all the residuals are almost equal to **0.** We need to look only into **1Q, Median, 3Q** but not the **Min** and **Max.** Also, it checks for the **Symmetry** of the model.

**4. Significant variables**

**Full Model:**

There are **29 variables** used in the Full model and out of which **10 variables** have t-values **less than 0.05** as it is really good**, 12 variables** have t-values **greater than 0.05** as it is really bad **and 7 variables** have t-value as **Not Available (NA).** Also,we need to check for the t-values of variables forthe **significance** of the model.

**Backward Model:**

There are **11 variables** used in the Backward model and out of which all the **11 variables** have t-values **less than 0.05** as it is really the best.Also,we need to check for the t-values of variables forthe **significance** of the model.

**5. Variable Co-Efficient**

**Full Model:**

There are **28 independent variables** used in the Full model and out of which **13 variables Coefficients Estimate** are matching (positive or negative) with the same variable’s correlation between **dependent variable (Pol\_KS)** and there are **8 variables (n.child\_KS,** **food\_KS,housing\_KS,nation.Asia\_KS,nation.NorthAmerica\_KS,m.status\_divorced\_KS,m.status\_married\_KS,m.status\_never\_KS)** are not matching, which results in **Multicollinearity.** Also, there are **7 variables** that haveCo-Efficient Estimate as **Not Available (NA).** Finally, it is used to check the **Sensibleness** of the model.

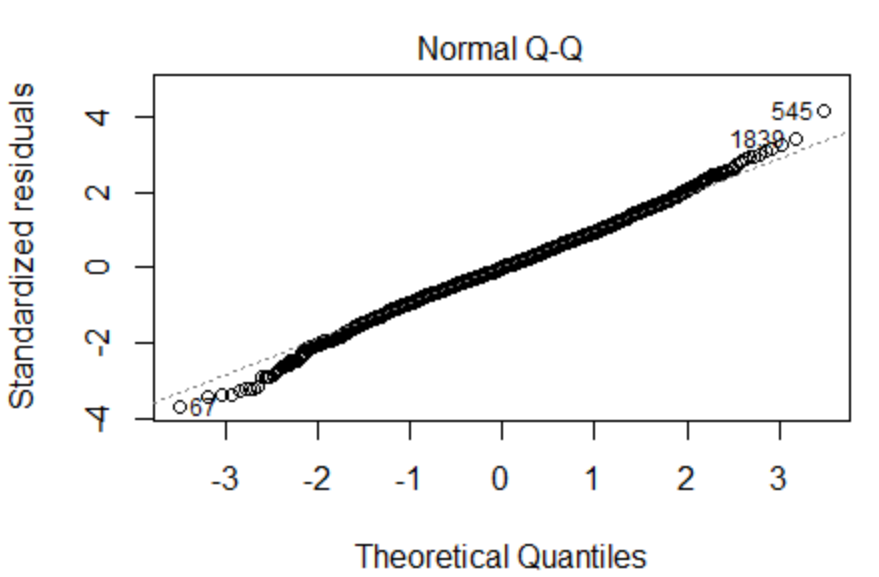
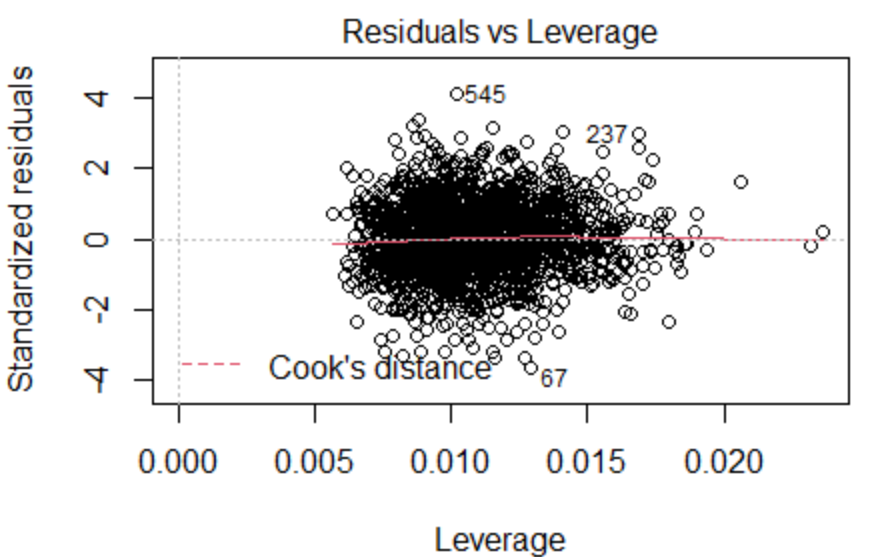
**Backward Model:**

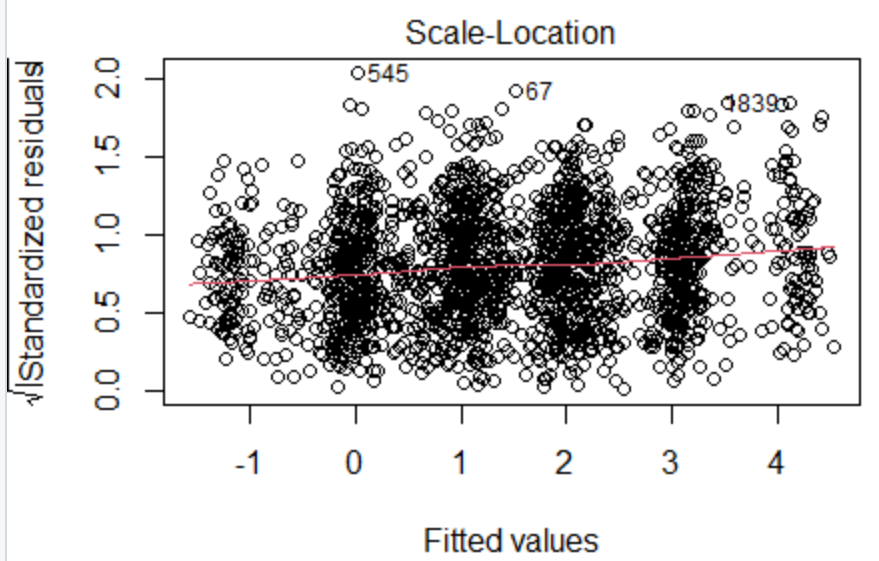
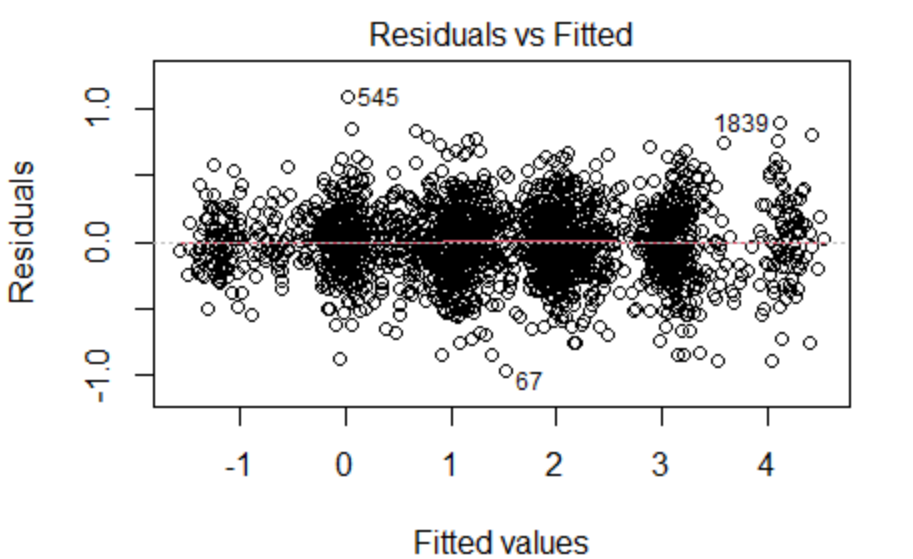
There are **10 independent variables** used in the backward model and out of which **9 variables** are matching with a correlation of **dependent Variable** (**Pol\_KS)** as it is really goodbut **1 variable (m.status\_married\_KS)** is not matching, which results in **Multicollinearity.** Finally, it is used to check the **Sensibleness** of the model.

**7. Model Evaluation – Verifying Assumptions – Multivariate**

**7.1. For both models (as discussed and demonstrated in class) evaluate the main assumptions of regression: Error terms mean of zero, constant variance and normally distributed.**

**Full Model:**

****

****

**Error terms mean of zero,**

From the above **Residuals and Fitted Chart,** we can tell that the Error terms mean of zero **(zero mean for the residuals)**. The red line is almost on the **fit line(y=0).** There are **no patterns** shown by the red line, and the residuals are scattered everywhere. Therefore, the assumptions are **true.**

**constant variance,**

From the **Scale-Location chart,** hereresiduals are standardized. The red line doesn’t go up and down. Also, residuals don’t increase or decrease with a uniform pattern. so, it has a **uniform variance (constant variance).**

**normally distributed,**

Shapiro-Wilk normality test

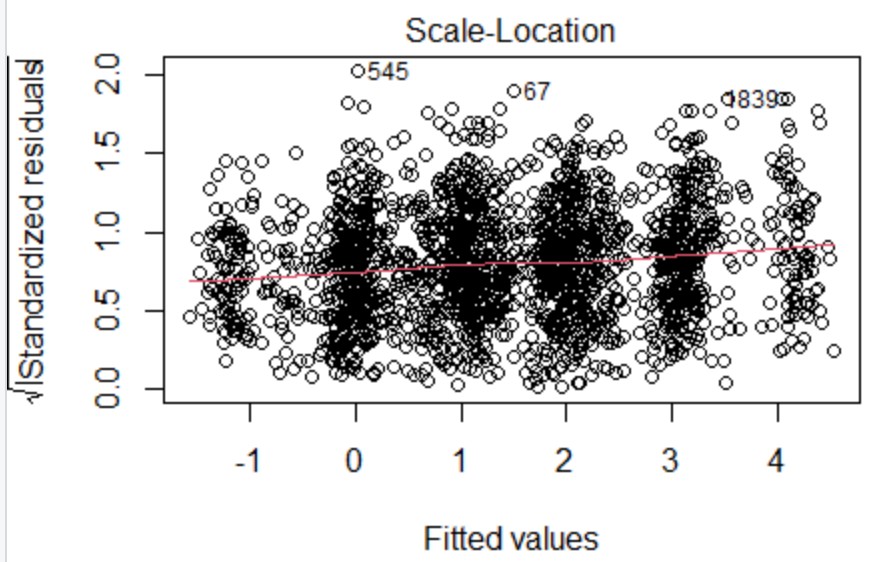
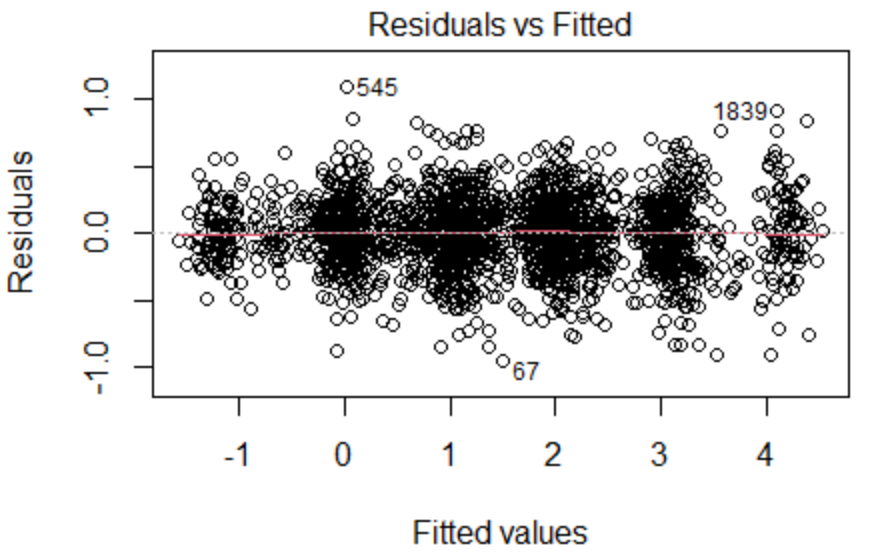
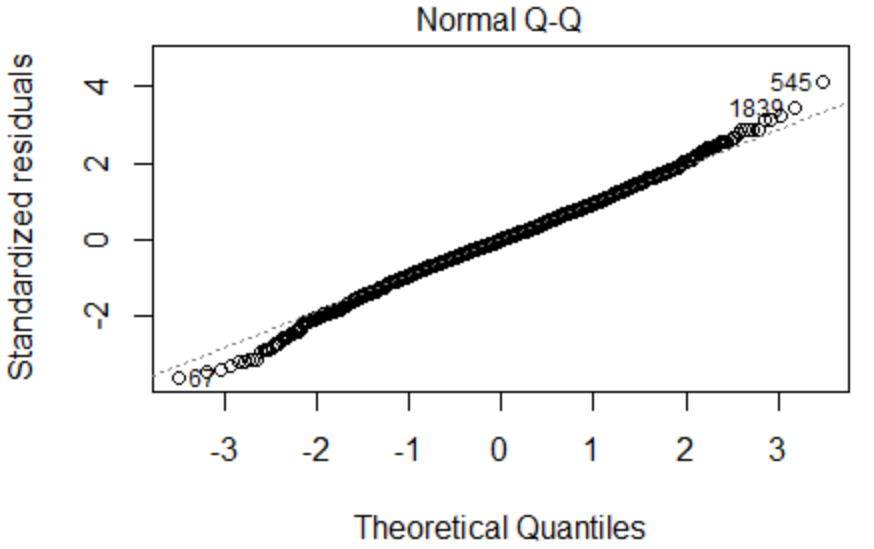
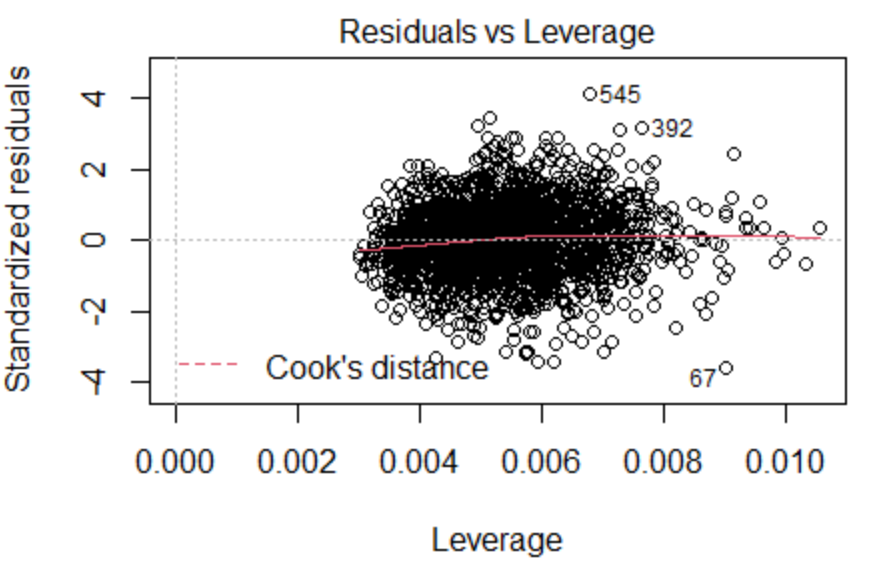
data: pol\_lm\_res

W = 0.99643, p-value = 0.0001037

From the **Shapiro-Wilk normality test,** it’s clear that the model is **not normally distributed** becausethe **p-value is <0.05.**

We can also similar normality in **Normal QQ-plot.**

**Backward Model:**

****

**Error terms mean of zero,**

From the above **Residuals and Fitted Chart,** we can tell that the Error terms mean of zero **(zero mean for the residuals)**. The red line is almost on the **fit line(y=0).** There are **no patterns** shown by the red line, and the residuals are scattered everywhere. Therefore, the assumptions are **true.**

**constant variance,**

From the **Scale-Location chart,** hereresiduals are standardized. The red line doesn’t go up and down. Also, residuals don’t increase or decrease with a uniform pattern. so, it has a **uniform variance (constant variance).**

**normally distributed,**

Shapiro-Wilk normality test

data: Bck\_pol\_lm\_res

W = 0.99661, p-value = 0.0001746

From the **Shapiro-Wilk normality test,** it’s clear that the model is **not normally distributed** becausethe **p-value is <0.05.**

We can also similar normality in **Normal QQ-plot.**

**Final Recommendation – Multivariate**

**1. Based on your preceding analysis, recommend which of the models should be used. NOTE – Even if none of the models meet all the assumptions of regression, choose the best of the two. In subsequent classes, we will learn how to deal with these issues.**

From the above two models, I will go with **the backward model** because,

1. It has a very good **F-statistic** while comparing the full model.
2. It has a good **p-value** as same as the full model.
3. It has good **residuals** which are near to 0 as same as the full model.
4. It has a good **t-value** <0.05 for all the variables but the full model has lots of variables which has a t-value > 0.05.
5. It has the same **Coefficients Estimate** (positive or negative) which matches with correlation for all variables except for one variable. But the full model has lots of mis matches.
6. It has an **error term mean of zero** as same as the full model.
7. It has **constant variance** as same as the full model.
8. It is not **normally distributed**. the full model is also not normally distributed.
9. According to **Cook’s Distance**, it has a **low influence** (<0.5) as same as the full model.

**Pol\_KS = (0.0054361762) \* age\_KS + (0.0000006943) \* income\_KS + (0.1245458199) \* score\_KS + (-1.9623496036) \* group.Control\_KS + (1.0501177574) \* hs.grad.No\_KS + (1.1121736791) \* gender.Male\_KS + (0.6721254106) \* political.Conservative\_KS + (-0.5227747174) \* political.Liberal\_KS + (0.6793620797) \* political.NewDemocrat\_KS + (-0.0287945907) \* m.status\_married\_KS**

**Professionalism, Clarity and Proper Citations**