Gurpinder Singh

STAT 108

12/2/2022

Load all the following library

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(nnet)
library(knitr)
library(broom)
```

Upload data

```
gss <- read_csv("data/gss2016.csv",
  na = c("", "Don't know", "No answer",
         "Not applicable"),
         guess_max = 2867) %>%
  select(natmass, age, sex, sei10, region, polviews) %>%
  drop_na()
```

```
## Rows: 2867 Columns: 935
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (810): wrkstat, marital, martype, childs, age, degree, sex, race, born, ...
## dbl (106): year, id_, hrs2, sphrs2, sibs, agekdbrn, educ, emailmin, emailhr,...
## lgl  (19): bigbang1, spwrkgvt, where6, away8, where8, away9, where9, mar10, ...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(gss)
```

```
## Rows: 2,590
## Columns: 6
## $ natmass  <chr> "Too little", "Too little", "Too much", "Too little", "About ~
## $ age      <chr> "47", "61", "43", "55", "53", "50", "23", "71", "86", "32", "~
## $ sex      <chr> "Male", "Male", "Female", "Female", "Female", "Male", "Female~
## $ sei10    <dbl> 87.9, 38.3, 21.8, 39.7, 44.6, 80.7, 20.1, 32.0, 13.2, 20.8, 2~
## $ region   <chr> "New england", "New england", "New england", "New england", "~
## $ polviews <chr> "Moderate", "Liberal", "Moderate", "Slightly liberal", "Sligh~
```

```
levels(as.factor(gss$natmass)  )
```

```
## [1] "About right" "Too little"  "Too much"
```

```
levels(as.factor(gss$polviews)  )
```

```
## [1] "Conservative"         "Extremely liberal"    "Extrmly conservative"
## [4] "Liberal"              "Moderate"             "Slghtly conservative"
## [7] "Slightly liberal"
```

```
length(unique(gss$natmass)  )
```

```
## [1] 3
```

```
length(unique(gss$polviews)  )
```

```
## [1] 7
```

Excersize 1

```
gss <- gss %>%
  mutate(natmass = relevel(as.factor(natmass), "About right"))
```

Excersize 2

```
gss <- gss %>%
  mutate(polviews = fct_relevel(as.factor(polviews), "Extremely liberal", "Liberal", "Slightly liberal"
levels(as.factor(gss$natmass)  )
```

```
## [1] "About right" "Too little"  "Too much"
```

```
levels(as.factor(gss$polviews)  )
```

```
## [1] "Extremely liberal"    "Liberal"              "Slightly liberal"
## [4] "Moderate"             "Slghtly conservative" "Conservative"
## [7] "Extrmly conservative"
```

```
length(unique(gss$natmass)  )
```

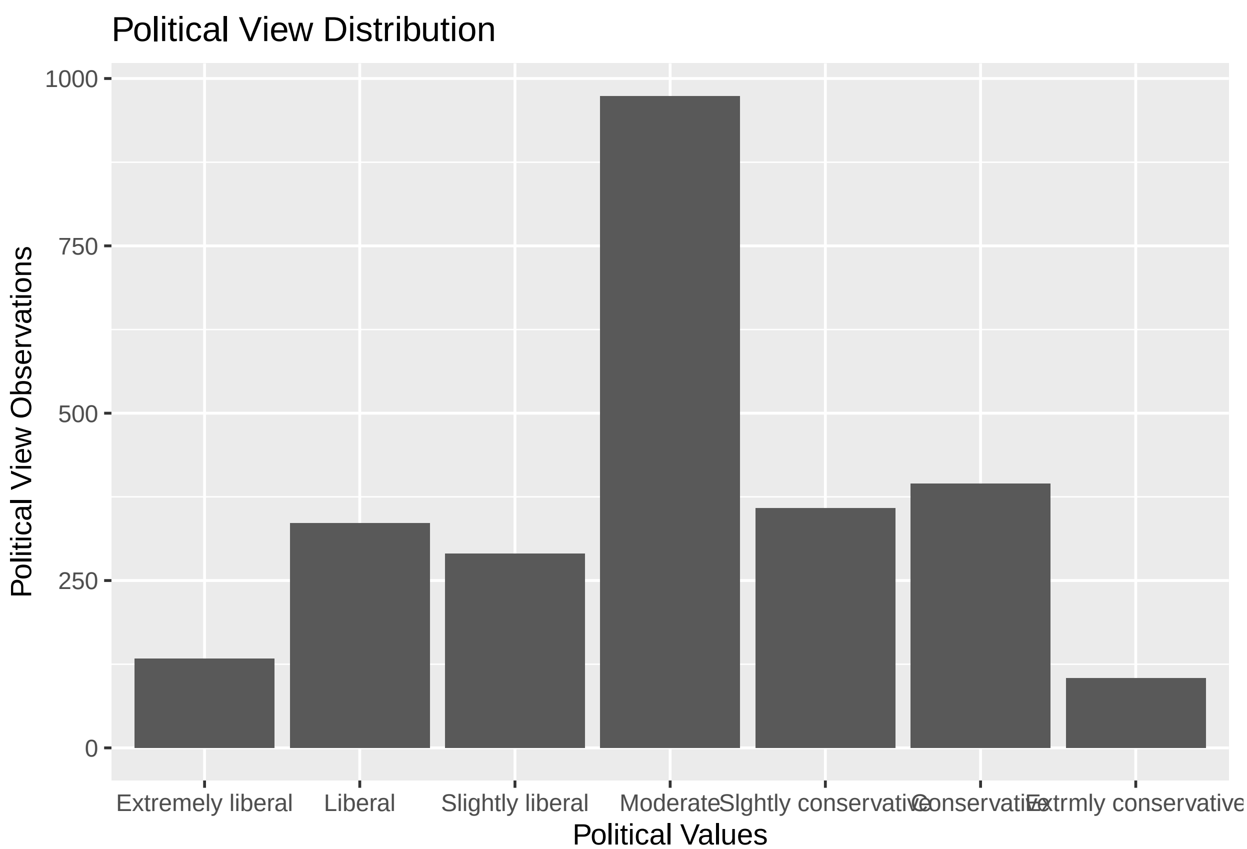
```
## [1] 3
```

```
length(unique(gss$polviews)  )
```

```
## [1] 7
```

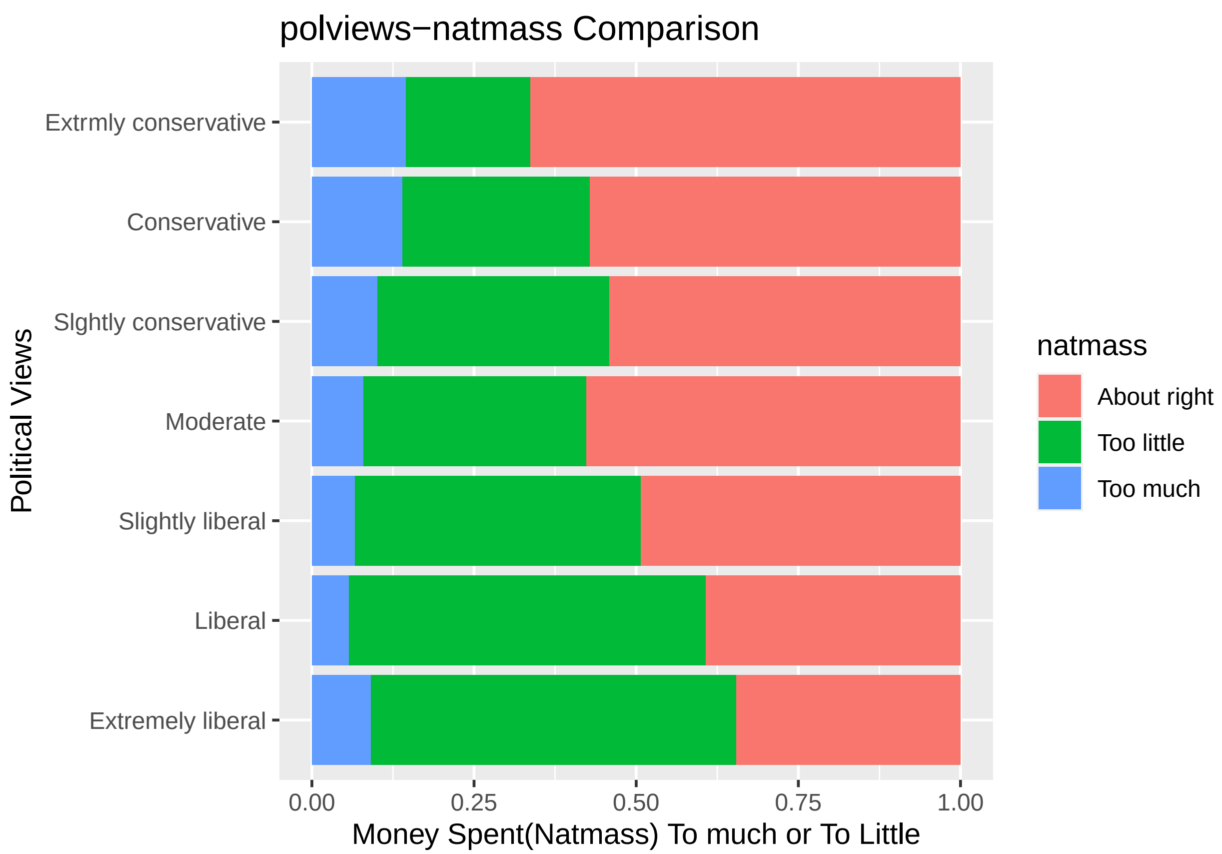
```
ggplot(data = gss, aes(x =polviews)) +
  geom_histogram(stat = "count") +
  labs(x ="Political Values",
       y = "Political View Observations",
       title = "Political View Distribution")
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



Exersize 3

```
ggplot(data = gss, aes(x = polviews, fill = natmass)) +
  geom_bar(position = "fill") +
  labs(x = "Political Views",y = "Money Spent(Natmass) To much or To Little", title = "polviews-natmass
  coord_flip()
```



Excersize 4

```
gss <- gss %>% mutate(age = case_when(
           age == "89 or older" ~ 89,
           TRUE ~ as.numeric(age)))
```

```
## Warning in eval_tidy(pair$rhs, env = default_env): NAs introduced by coercion
```

```
glimpse(gss$age)
```

```
##  num [1:2590] 47 61 43 55 53 50 23 71 86 32 ...
```

Excersize 5 Multinomial logistic regression is similar to binommial regression except it allows for more then two categorical variables for the response variable natmass

Exersize 6

```
names(gss)
```

```
## [1] "natmass"  "age"      "sex"      "sei10"    "region"   "polviews"
```

```
multi_model <- multinom(natmass ~ age + sex + sei10 + region,
             data=gss)
```

```
## # weights:  39 (24 variable)
## initial  value 2845.405828
## iter  10 value 2345.298055
## iter  20 value 2328.421434
## iter  30 value 2327.225660
## final  value 2327.223281
## converged
```

```
tidy(multi_model, conf.int = TRUE) %>% # output model
kable(digits = 3) # format model output
```

y.level	term	estimate	std.error	statistic	p.value	conf.low	conf.high
Too little	(Intercept)	-1.194	0.178	-6.692	0.000	-1.544	-0.844
Too little	age	0.004	0.002	1.607	0.108	-0.001	0.009
Too little	sexMale	0.196	0.085	2.300	0.021	0.029	0.364
Too little	sei10	0.009	0.002	5.282	0.000	0.006	0.013
Too little	regionE. sou. central	0.272	0.189	1.434	0.151	-0.099	0.643
Too little	regionMiddle atlantic	-0.030	0.164	-0.184	0.854	-0.352	0.292
Too little	regionMountain	0.183	0.177	1.034	0.301	-0.164	0.529
Too little	regionNew england	0.595	0.201	2.958	0.003	0.201	0.989
Too little	regionPacific	0.409	0.151	2.704	0.007	0.112	0.705
Too little	regionSouth atlantic	0.123	0.139	0.883	0.377	-0.150	0.396
Too little	regionW. nor. central	0.030	0.196	0.151	0.880	-0.355	0.414
Too little	regionW. sou. central	-0.086	0.169	-0.508	0.611	-0.417	0.245
Too much	(Intercept)	-2.413	0.298	-8.088	0.000	-2.998	-1.829
Too much	age	0.016	0.004	3.945	0.000	0.008	0.024
Too much	sexMale	0.553	0.145	3.808	0.000	0.269	0.838
Too much	sei10	-0.010	0.003	-3.018	0.003	-0.016	-0.003
Too much	regionE. sou. central	-0.285	0.350	-0.816	0.414	-0.970	0.400
Too much	regionMiddle atlantic	-0.162	0.278	-0.585	0.559	-0.707	0.382
Too much	regionMountain	-0.021	0.303	-0.070	0.944	-0.616	0.574
Too much	regionNew england	0.853	0.289	2.946	0.003	0.285	1.420
Too much	regionPacific	0.296	0.242	1.221	0.222	-0.179	0.771
Too much	regionSouth atlantic	-0.263	0.242	-1.086	0.278	-0.737	0.212
Too much	regionW. nor. central	0.138	0.302	0.457	0.647	-0.454	0.730
Too much	regionW. sou. central	-0.583	0.310	-1.878	0.060	-1.191	0.025

Exersize 7 To little intercept and two much is both negative which shows that with base lined values of age sexMale sei10 and region the value of Natmass will be negative which is closer to the about right value Exersize 8 Because the value of To little age is .004 we can state that as a persons age increases then people are more likeley to believe that spending on mass transportation is to little then about right. Exersize 9

```
newMultiModel <- multinom(natmass ~ age + sex + sei10 + region+polviews,
             data=gss)
```

```
## # weights:  57 (36 variable)
## initial  value 2845.405828
## iter  10 value 2308.054489
## iter  20 value 2277.361046
## iter  30 value 2276.038249
## iter  40 value 2275.922824
```

```
## final  value 2275.922640
## converged
```

```
tidy(newMultiModel, conf.int = TRUE) %>% # output model
kable(digits = 3) # format model output
```

y.level	term	estimate	std.error	statistic	p.value	conf.low	conf.high
Too little	(Intercept)	-0.415	0.258	-1.606	0.108	-0.921	0.092
Too little	age	0.006	0.003	2.448	0.014	0.001	0.011
Too little	sexMale	0.217	0.087	2.500	0.012	0.047	0.388
Too little	sei10	0.008	0.002	4.446	0.000	0.005	0.012
Too little	regionE. sou. central	0.334	0.192	1.736	0.083	-0.043	0.711
Too little	regionMiddle atlantic	-0.081	0.167	-0.487	0.627	-0.410	0.247
Too little	regionMountain	0.138	0.180	0.766	0.444	-0.215	0.490
Too little	regionNew england	0.466	0.205	2.270	0.023	0.064	0.868
Too little	regionPacific	0.364	0.154	2.364	0.018	0.062	0.665
Too little	regionSouth atlantic	0.132	0.142	0.930	0.353	-0.146	0.410
Too little	regionW. nor. central	0.031	0.199	0.153	0.878	-0.360	0.421
Too little	regionW. sou. central	-0.028	0.171	-0.161	0.872	-0.364	0.309
Too little	polviewsLiberal	-0.202	0.223	-0.906	0.365	-0.638	0.235
Too little	polviewsSlightly liberal	-0.597	0.227	-2.633	0.008	-1.041	-0.153
Too little	polviewsModerate	-0.969	0.203	-4.785	0.000	-1.367	-0.572
Too little	polviewsSlghtly conservative	-0.940	0.222	-4.226	0.000	-1.376	-0.504
Too little	polviewsConservative	-1.221	0.224	-5.456	0.000	-1.659	-0.782
Too little	polviewsExtrmly conservative	-1.696	0.320	-5.302	0.000	-2.323	-1.069
Too much	(Intercept)	-1.850	0.436	-4.246	0.000	-2.703	-0.996
Too much	age	0.014	0.004	3.480	0.001	0.006	0.022
Too much	sexMale	0.535	0.146	3.660	0.000	0.248	0.821
Too much	sei10	-0.010	0.003	-3.079	0.002	-0.016	-0.004
Too much	regionE. sou. central	-0.323	0.351	-0.922	0.357	-1.011	0.364
Too much	regionMiddle atlantic	-0.144	0.279	-0.514	0.607	-0.690	0.403
Too much	regionMountain	-0.025	0.305	-0.084	0.933	-0.623	0.572
Too much	regionNew england	0.879	0.292	3.007	0.003	0.306	1.451
Too much	regionPacific	0.340	0.244	1.396	0.163	-0.138	0.818
Too much	regionSouth atlantic	-0.274	0.243	-1.128	0.259	-0.750	0.202
Too much	regionW. nor. central	0.159	0.304	0.524	0.600	-0.436	0.755
Too much	regionW. sou. central	-0.602	0.311	-1.933	0.053	-1.212	0.008
Too much	polviewsLiberal	-0.631	0.411	-1.533	0.125	-1.437	0.175
Too much	polviewsSlightly liberal	-0.670	0.411	-1.630	0.103	-1.476	0.136
Too much	polviewsModerate	-0.680	0.351	-1.936	0.053	-1.368	0.008
Too much	polviewsSlghtly conservative	-0.401	0.377	-1.064	0.287	-1.140	0.337
Too much	polviewsConservative	-0.080	0.364	-0.219	0.826	-0.793	0.634
Too much	polviewsExtrmly conservative	-0.306	0.443	-0.692	0.489	-1.174	0.562

```
newMultiModel$AIC
```

```
## [1] 4623.845
```

```r
multi_model$AIC
```

```
## [1] 4702.447
```

```r
anova(newMultiModel, multi_model, test = "Chisq")
```

```
##                                    Model Resid. df Resid. Dev   Test    Df
## 1             age + sex + sei10 + region      5156   4654.447           NA
## 2 age + sex + sei10 + region + polviews      5144   4551.845 1 vs 2    12
##   LR stat.      Pr(Chi)
## 1       NA           NA
## 2 102.6013 2.220446e-16
```

Null hypothesis: Poltiical view is not a predictor Alternative:Political view is a predictor Using a chisquare test we see the p value is low and as well as see the AIC value is lower as well so we can there for conclude to reject the null hypothesis and say Model with political view is better. Exercise 11

```r
names(newMultiModel)
```

```
##  [1] "n"             "nunits"        "nconn"         "conn"
##  [5] "nsunits"       "decay"         "entropy"       "softmax"
##  [9] "censored"      "value"         "wts"           "convergence"
## [13] "fitted.values" "residuals"     "lev"           "call"
## [17] "terms"         "weights"       "deviance"      "rank"
## [21] "lab"           "coefnames"     "vcoefnames"    "contrasts"
## [25] "xlevels"       "edf"           "AIC"
```

```r
summary(newMultiModel$fitted.values)
```

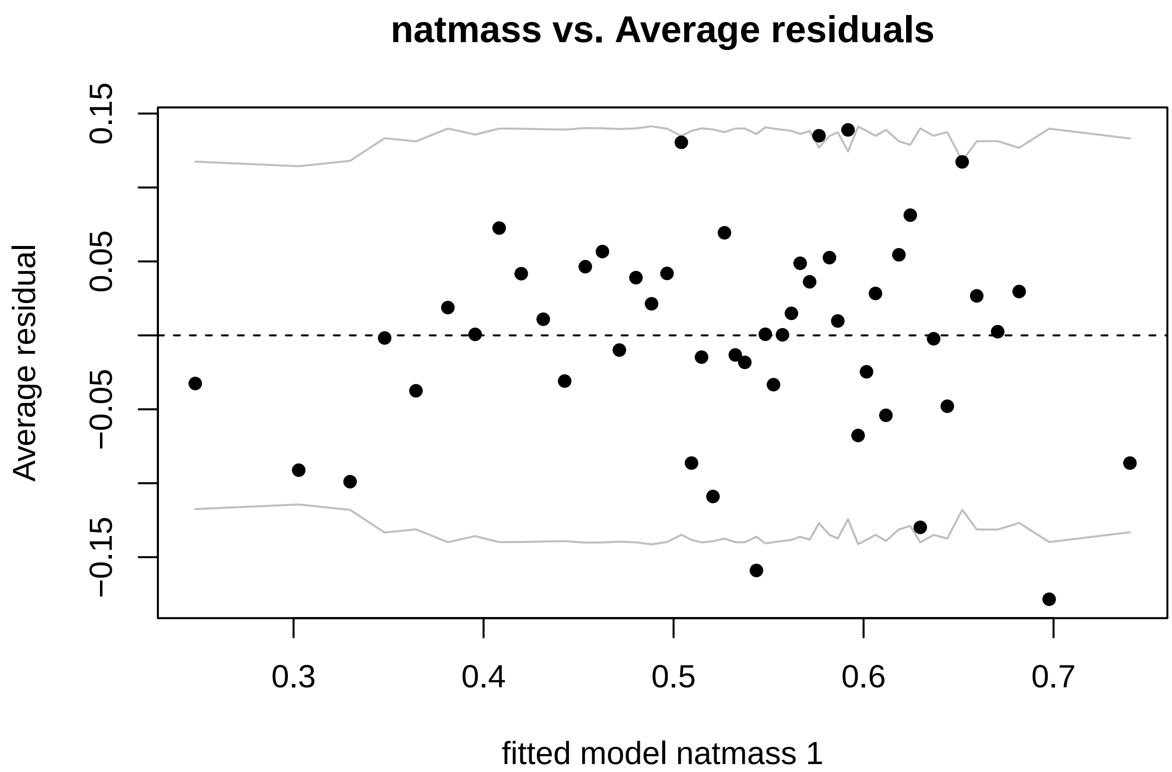
```
##   About right       Too little        Too much
##  Min.   :0.1710   Min.   :0.1234   Min.   :0.01392
##  1st Qu.:0.4627   1st Qu.:0.2980   1st Qu.:0.05318
##  Median :0.5458   Median :0.3593   Median :0.07730
##  Mean   :0.5297   Mean   :0.3803   Mean   :0.08996
##  3rd Qu.:0.6062   3rd Qu.:0.4490   3rd Qu.:0.11311
##  Max.   :0.8159   Max.   :0.7186   Max.   :0.41333
```

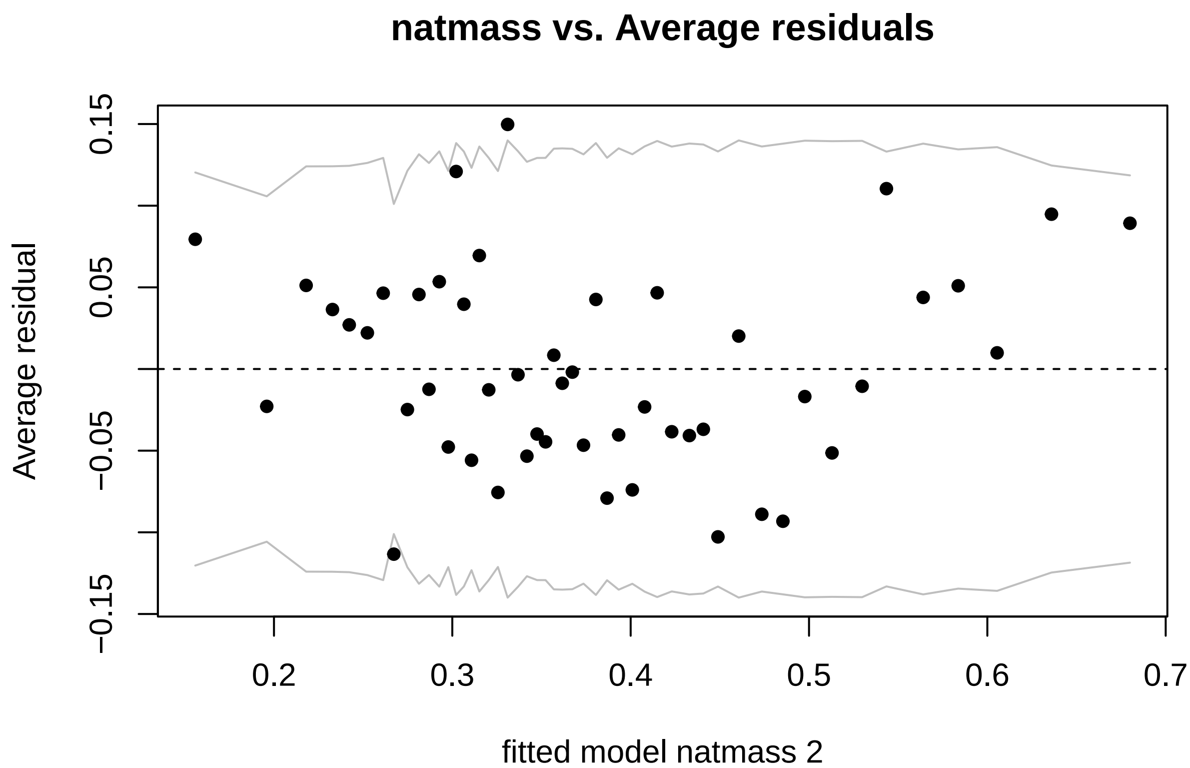
```r
summary(newMultiModel$residuals)
```

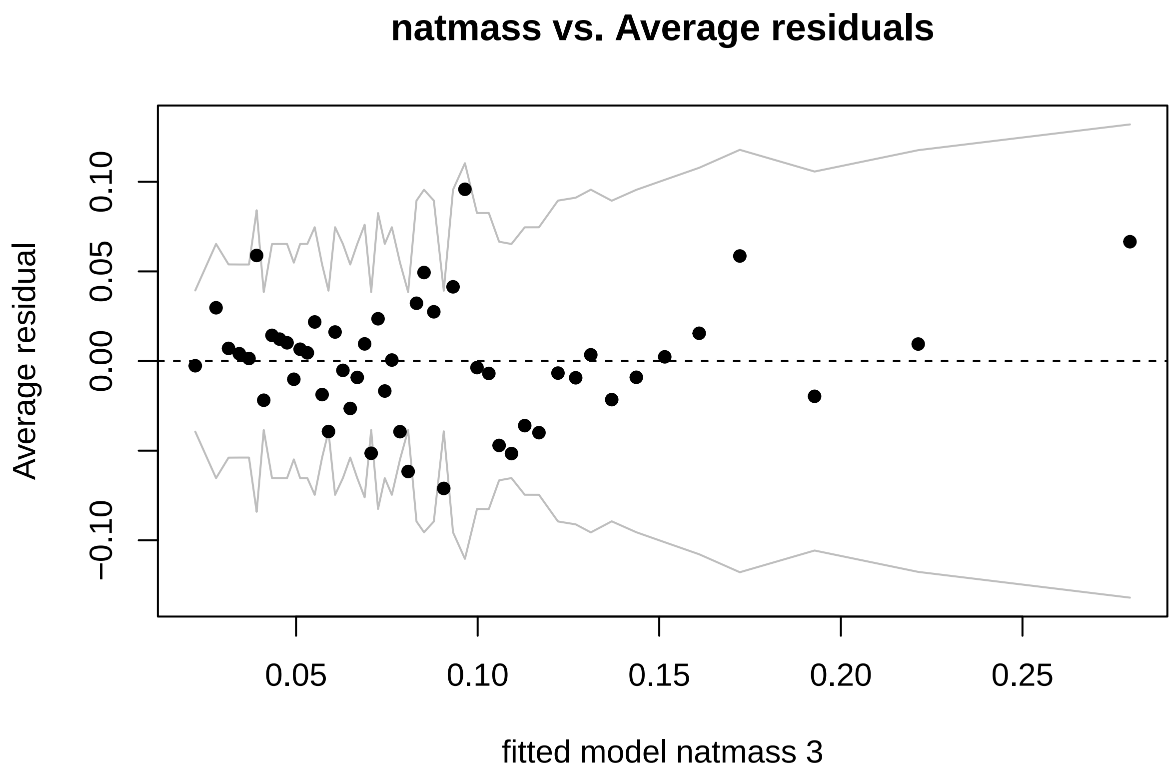
```
##   About right           Too little           Too much
##  Min.   :-0.7708944   Min.   :-0.7028361   Min.   :-0.3889465
##  1st Qu.:-0.5130073   1st Qu.:-0.3670163   1st Qu.:-0.1054340
##  Median : 0.3170380   Median :-0.2702892   Median :-0.0709375
##  Mean   :-0.0000022   Mean   : 0.0000017   Mean   : 0.0000005
##  3rd Qu.: 0.4408922   3rd Qu.: 0.5396924   3rd Qu.:-0.0451503
##  Max.   : 0.8289931   Max.   : 0.8689097   Max.   : 0.9812461
```

Exercie 12

```
for (i in 1:3){
x <-paste("fitted model natmass",as.character(i),sep=" ")
arm::binnedplot(newMultiModel$fitted.values[,i] ,newMultiModel$residuals[,i],
    xlab=x, ylab="Average residual",
    main="natmass vs. Average residuals", col.int="gray")
}
```




natmass vs. Average residuals
Average residual
0.15
0.05
−0.05
−0.15
0.2
0.3
0.4
0.5
0.6
0.7
fitted model natmass 2



Exercise 13:

```
for (i in 1:3){
print(paste("residual for",as.character(i),sep=" "))
print(mean(newMultiModel$residuals[,i]))
}
```

```
## [1] "residual for 1"
## [1] -2.238446e-06
## [1] "residual for 2"
## [1] 1.712403e-06
## [1] "residual for 3"
## [1] 5.260433e-07
```

Exercise 16: The model states that liberals will be more likeley to believe that there is to little spending in comparison to a conservative. Excersize 17:

```
gss %>%
  count(natmass, predict(newMultiModel, newdata = gss, type = "class"))
```

```
## # A tibble: 8 x 3
##   natmass     `predict(newMultiModel, newdata = gss, type = "class")`     n
##   <fct>       <fct>                                                   <int>
## 1 About right About right                                              1151
## 2 About right Too little                                                219
```

```
## 3 About right Too much                                                    2
## 4 Too little  About right                                                646
## 5 Too little  Too little                                                 339
## 6 Too much    About right                                                196
## 7 Too much    Too little                                                  36
## 8 Too much    Too much                                                     1
```

1,491 Correct values vs 2,590 observations 1,491/2,590= 0.5756756757