Homework 7

Jamila Sani and Kevin Linares

2024-11-04

```
library(faraway)
library(dplyr)
library(knitr)
library(tidyverse)
library(knitr)
data(teengamb)
summary(teengamb) |> kable()
```

	sex	status	income	verbal	gamble
	Min. :0.0000	Min. :18.00	Min. : 0.600	Min. : 1.00	Min. : 0.0
	1st Qu.:0.0000	1st Qu.:28.00	1st Qu.: 2.000	1st Qu.: 6.00	1st Qu.: 1.1
	Median :0.0000	Median :43.00	Median : 3.250	Median : 7.00	Median : 6.0
	Mean :0.4043	Mean :45.23	Mean : 4.642	Mean : 6.66	Mean : 19.3
	3rd Qu.:1.0000	3rd Qu.:61.50	3rd Qu.: 6.210	3rd Qu.: 8.00	3rd Qu.: 19.4
	Max. :1.0000	Max. :75.00	Max. :15.000	Max. :10.00	Max. :156.0

```
head(teengamb)
```

```
##   sex status income verbal gamble
## 1   1     51   2.00      8    0.0
## 2   1     28   2.50      8    0.0
## 3   1     37   2.00      6    0.0
## 4   1     28   7.00      4    7.3
## 5   1     65   2.00      8   19.6
## 6   1     61   3.47      6    0.1
```

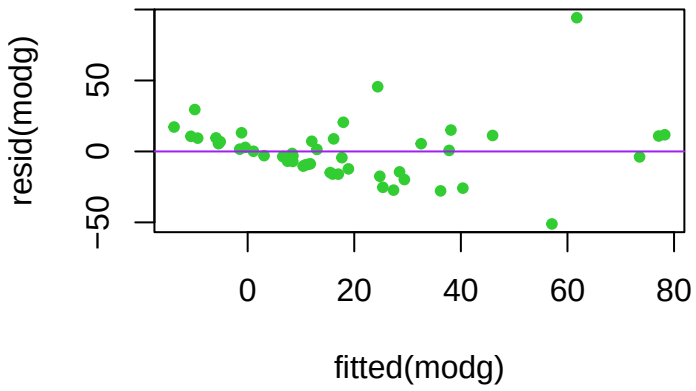
```
modg <- lm(gamble ~ sex + status + income + verbal, teengamb)
summary (modg)
```

```
##
## Call:
## lm(formula = gamble ~ sex + status + income + verbal, data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.55565   17.19680   1.312   0.1968
## sex         -22.11833    8.21111  -2.694   0.0101 *
## status        0.05223    0.28111   0.186   0.8535
## income        4.96198    1.02539   4.839 1.79e-05 ***
## verbal       -2.95949    2.17215  -1.362   0.1803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

1.A. Check the zero mean error assumption using residual plots. What do you conclude about whether the assumption is met?

- The zero mean error assumption is not violated as the points seem to be evenly distributed on either sides of the purple line.Also, the mean of the residuals is approximately zero (-1.6e-16), suggesting that the zero mean assumption is not violated.

```
plot(fitted(modg),resid(modg), pch=20, col="limegreen")+
abline(h=0,col="purple")
```




```
## integer(0)
```

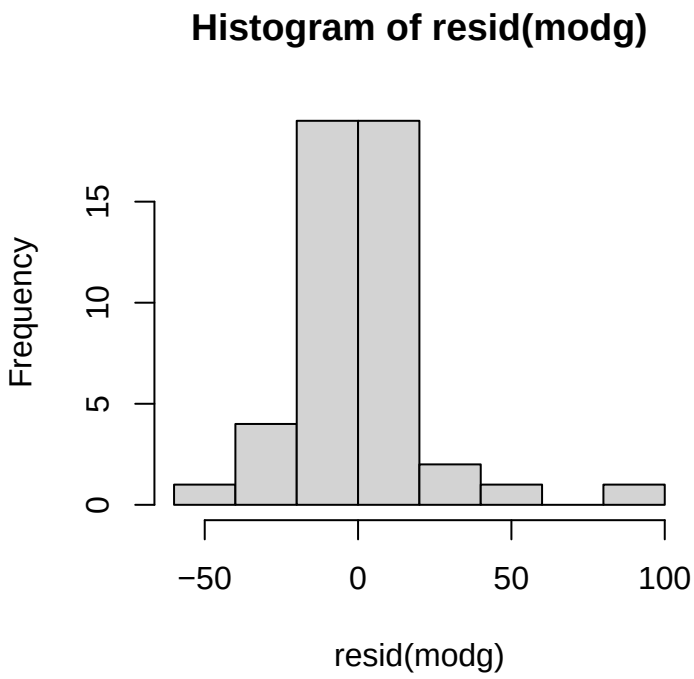
```
mean(resid(modg))
```

```
## [1] -1.556914e-16
```

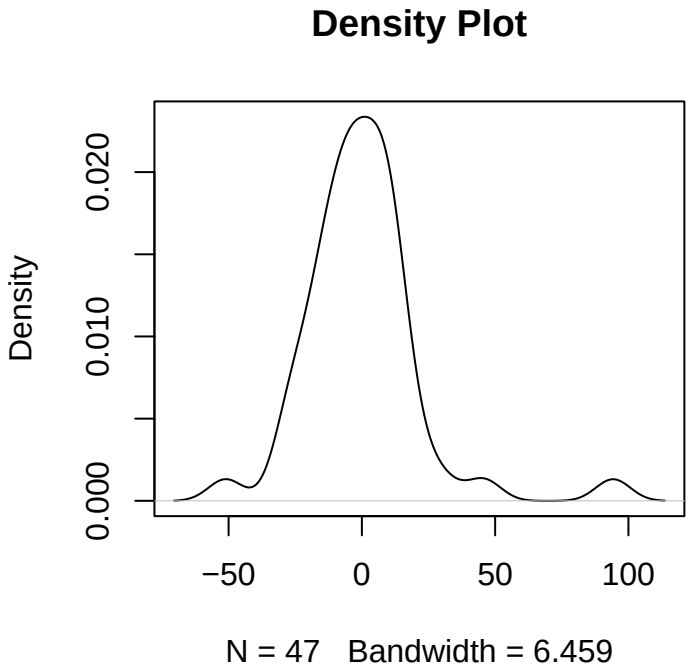
1.B. Check the constant error variance assumption using both residual plots and a formal statistical test. What do you conclude?

- The density plot shows evidence of heteroscadasticity. The resid vs. fitted plot from 1A above indicates the same as the points are more dense on the left side of the plot.
- Residuals seem to be related to fitted value from the low pvalues associated with the Beta1 (0.03) coefficient. We reject the null that true ratio of variances = 1 in favor of the alternative hypothesis => unequal variances, according to our F-test.

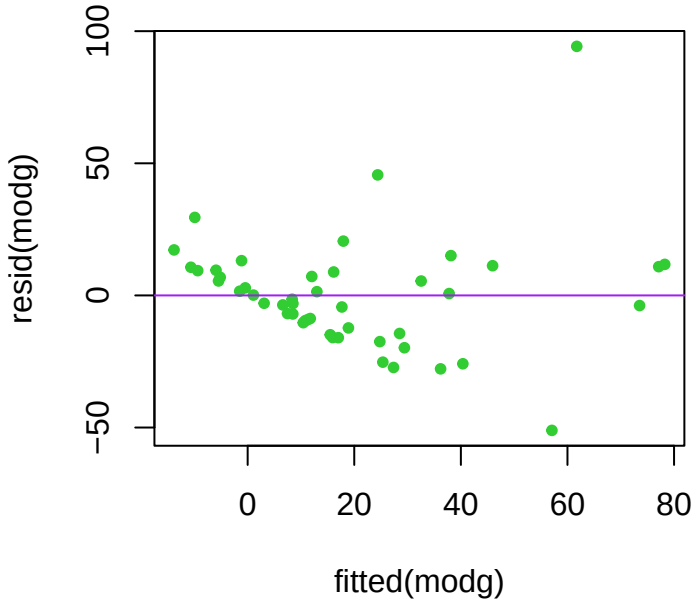
```
hist(resid(modg))
```




```
plot(density(resid(modg)),
     main="Density Plot")
```



```r
plot(fitted(modg),resid(modg), pch=20, col="limegreen")+
abline(h=0,col="purple")
```




```
## integer(0)
```

```r
summary(lm((residuals(modg)) ~ fitted(modg)))
```

```
##
## Call:
## lm(formula = (residuals(modg)) ~ fitted(modg))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.773e-16  4.203e+00       0        1
## fitted(modg)  0.000e+00  1.413e-01       0        1
##
## Residual standard error: 21.92 on 45 degrees of freedom
## Multiple R-squared:  2.813e-33,  Adjusted R-squared:  -0.02222
## F-statistic: 1.266e-31 on 1 and 45 DF,  p-value: 1
```

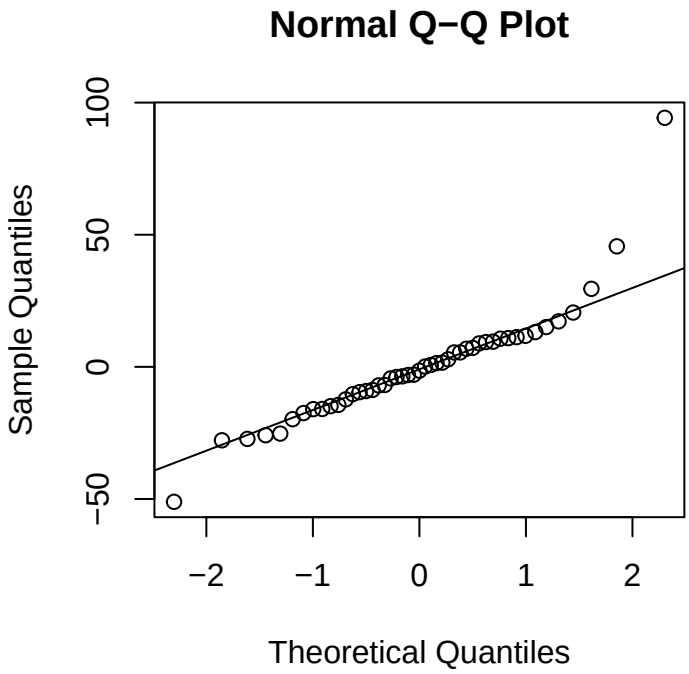
```r
var.test(resid(modg)[fitted(modg)<=30],
         resid(modg)[fitted(modg)>30])
```

```
##
##  F test to compare two variances
##
## data:  resid(modg)[fitted(modg) <= 30] and resid(modg)[fitted(modg) > 30]
## F = 0.16983, num df = 35, denom df = 10, p-value = 7.413e-05
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.05178625 0.41442186
## sample estimates:
## ratio of variances
##          0.1698276
```

1.C. Check the error normality assumption both graphically and statistically. What do you conclude? Which method do you think is preferable?

- Graphically: The data are not normally distributed. There are a few outliers on both ends of the qq plot.
- Statistically: The low pvalue means that we reject the null hypothesis that the residuals are normal.
- Both methods have their strengths; however, graphical method is preferable as they are more interpretable and easier to reveal structures that may not have been suspected; while statistical inferences could be harder to interpret and sometimes more sensitive e.g. Shapiro-Wilk normality test.

```r
qqnorm(resid(modg))
qqline(residuals(modg))
```




```
shapiro.test(resid(modg))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(modg)
## W = 0.86839, p-value = 8.16e-05
```

1.D. Check for observations with large leverage. Which observations have large leverage?

- Using the half-normal plots, observations 42 and 35 seem to have large leverages.

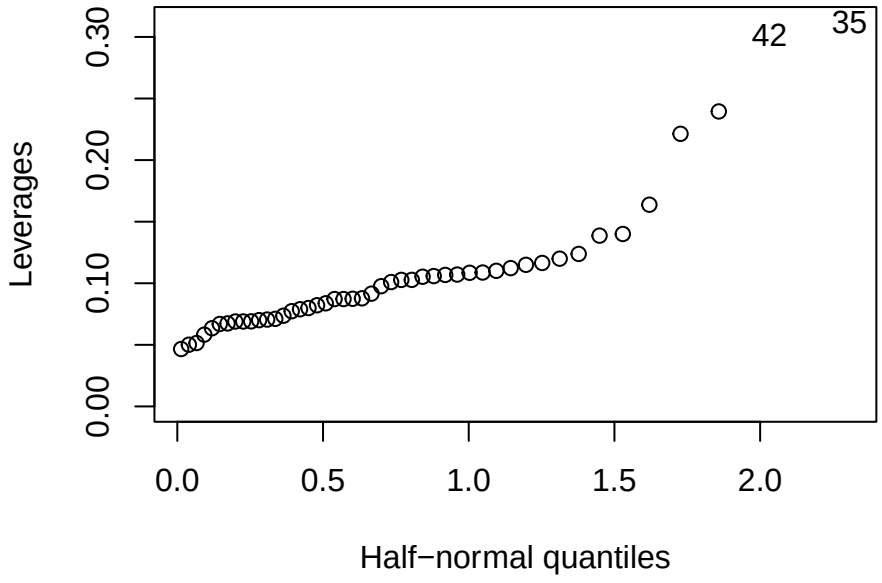
```
hatv <- hatvalues(modg)
head(hatv)
```

```
##          1          2          3          4          5          6
## 0.07988226 0.10851291 0.06347643 0.10273955 0.13866946 0.16378563
```

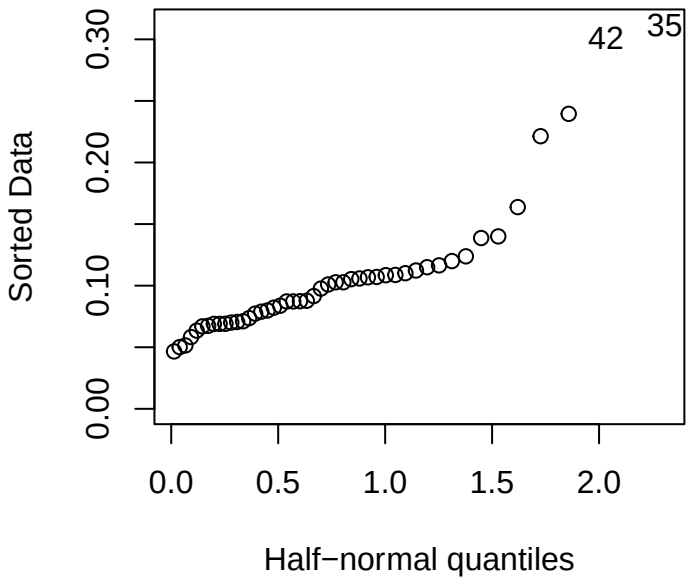
```r
sum(hatv) # sum of hat values should equal the number of parameters in model
```

```
## [1] 5
```

```r
highlev <- row.names (teengamb)
halfnorm (hatv, labs=highlev, ylab="Leverages")
```




```r
halfnorm(hatvalues(modg)) # half normal plot
```



1.E. Check for outliers. List any potential outliers.

- Observation 24 is a potential outlier. It is the largest of the studentized residuals (6.02). Also, it is greater than the Bonferroni critical value of -3.5 computed.

```
stud <- rstudent(modg)
stud[which.max(abs(stud))]
```

```
##       24
## 6.016116
```

```
abs(sort(stud)[1:47])
```

```
##          39          18          23          27          17          30
## 2.506089824 1.317398465 1.252647338 1.212679116 1.207094189 0.894246583
##           4          35          20          21          44           8
## 0.810686033 0.761255710 0.731951082 0.719525173 0.675872750 0.558091790
##          10          22          26          28           7          29
## 0.477911450 0.444490542 0.424594254 0.404491482 0.317374225 0.312262299
##          42          43          34           6          12          41
## 0.199979504 0.199861425 0.165636151 0.142158866 0.139821862 0.067239855
```

```
##          13          25          46          11          15          45
## 0.005399136 0.032148197 0.064230138 0.073183900 0.127853362 0.245946047
##           3           9          47          40          14           2
## 0.245991849 0.309727471 0.325550756 0.404973252 0.432273837 0.433163892
##           1          38          31          33          19          32
## 0.484870822 0.517035886 0.545192798 0.582020953 0.595780320 0.697965557
##          16          37           5          36          24
## 0.788830437 0.943967209 1.418582717 2.144825887 6.016116345
```

```
abs(sort(stud)[1:47])>abs(qt(0.05/(47),47-5))
```

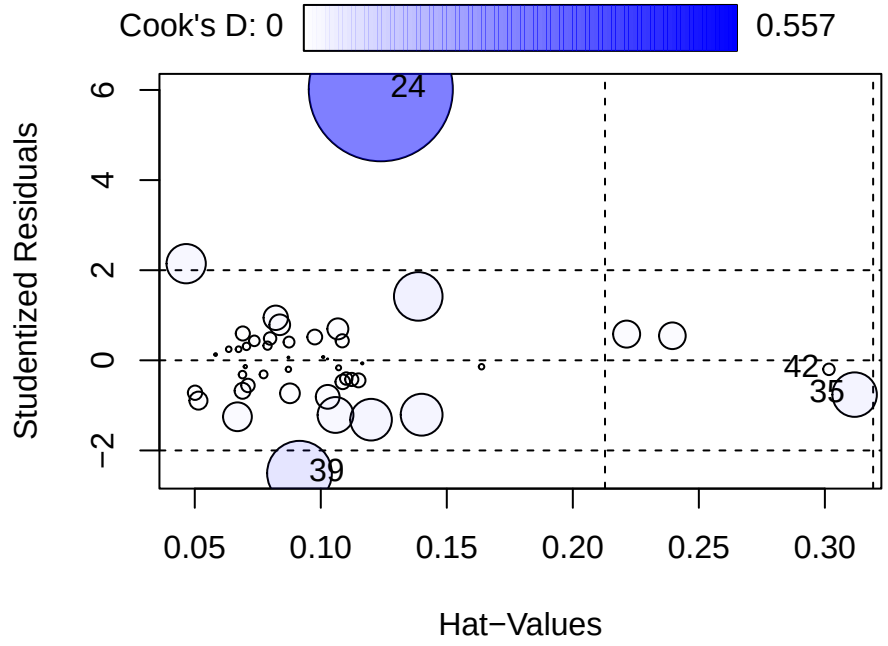
```
##    39    18    23    27    17    30     4    35    20    21    44     8    10
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    22    26    28     7    29    42    43    34     6    12    41    13    25
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    46    11    15    45     3     9    47    40    14     2     1    38    31
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    33    19    32    16    37     5    36    24
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
```

```
abs(stud)[24]
```

```
##       24
## 6.016116
```

Note. We can use the car package to detect outliers, influenctial observations, and leverage points to confirm our results below.

```
car::influencePlot(modg) #checks for influential points, outliers, and leverage.
```




```
##       StudRes        Hat      CookD
## 24  6.0161163 0.12380463 0.55650113
## 35 -0.7612557 0.31180294 0.05304304
## 39 -2.5060898 0.09155208 0.11244983
## 42 -0.1999795 0.30160877 0.00353499
```

1.F. Check for influential points. List any potential influential points.

- From the Cook's statistics, the three largest values identified as influential points are observations 24, 39, and 5. Observation 24 (corresponds to the potential outlier in 1E).

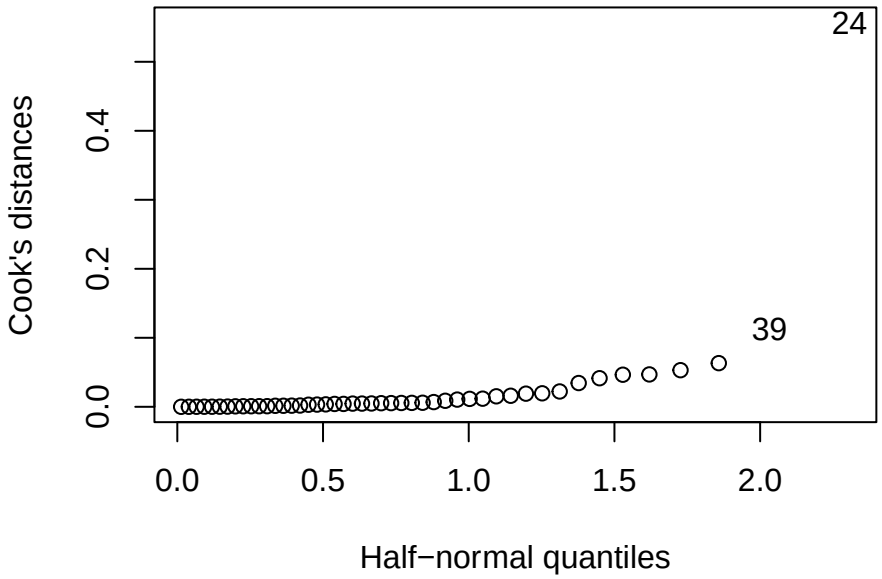
```
cookd<-cooks.distance(modg)
summary(cookd)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.0000007 0.0011908 0.0048478 0.0248308 0.0155806 0.5565011
```

```r
sort(cookd,dec=T)[1:10]
```

```
##          24          39           5          35          17          18          36
## 0.55650113 0.11244983 0.06327121 0.05304304 0.04693031 0.04649510 0.04141101
##          27          23          33
## 0.03442627 0.02221084 0.01956686
```

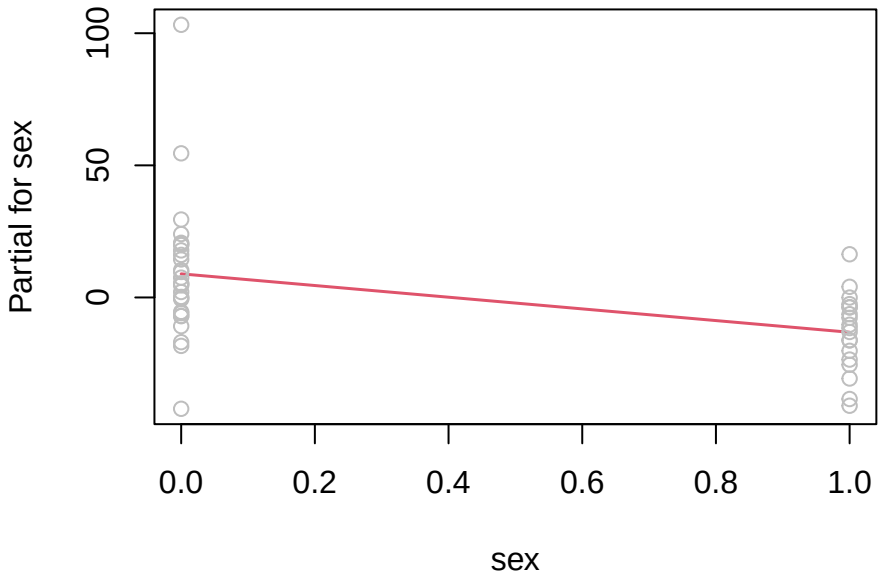
```r
halfnorm (cookd, 2, ylab="Cook's distances")
```



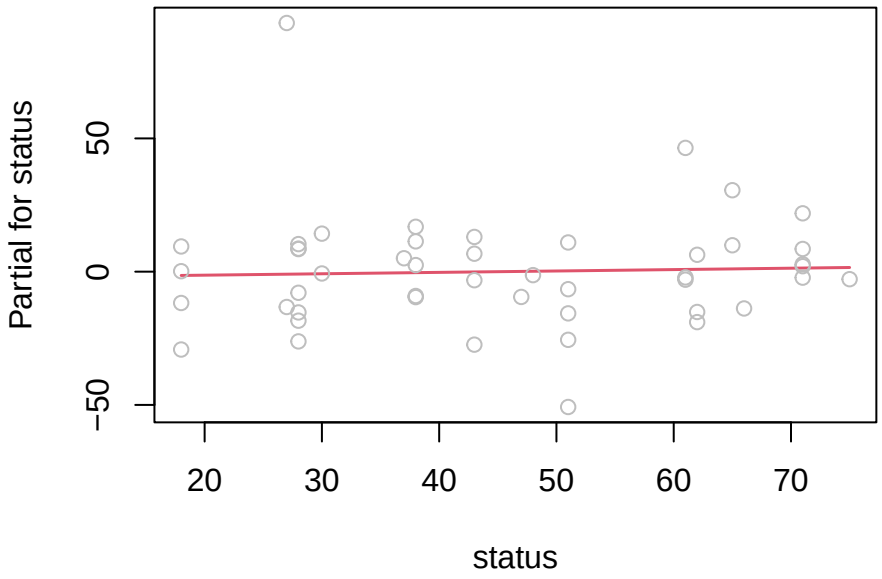
1.G. Check the structure of relationship between the predictors and the response.What do you observe?

- From the partial residual plots, all seem to be linear and there are no issues observed in the structure of the relationship between the predictors and the response. There is strong positive relationship between income and gambling; a negative linear relationship is observed in the plots for verbal and sex; while status shows

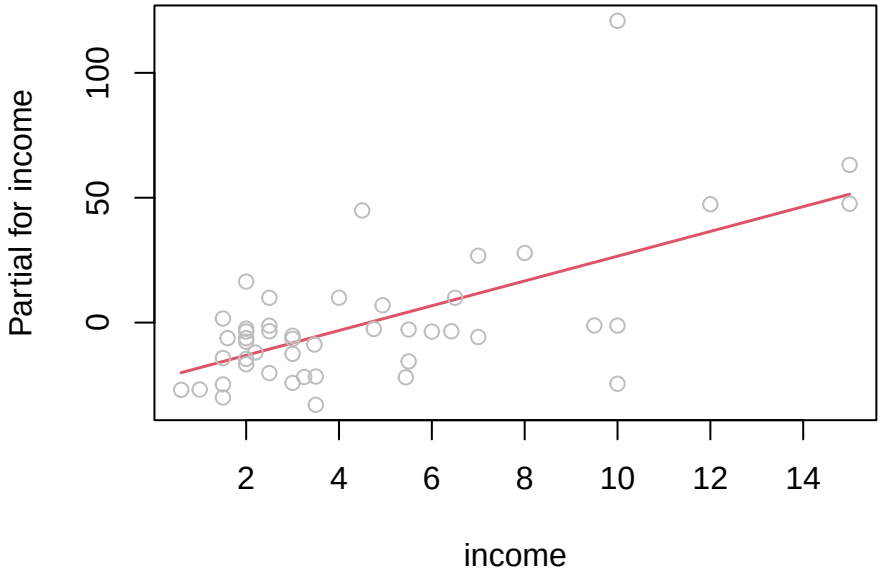
```r
termplot(modg, partial.resid = TRUE, terms=1)
```



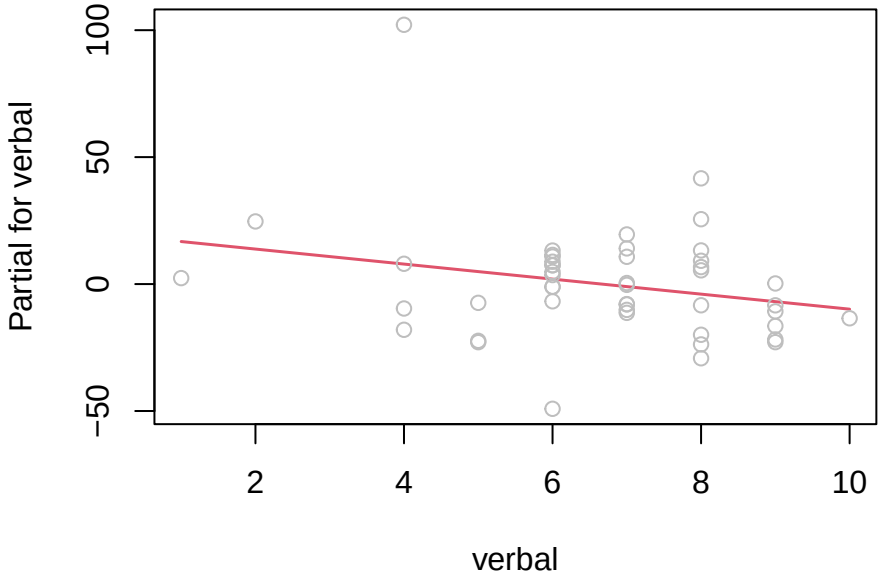

```
termplot(modg, partial.resid = TRUE, terms=2)
```



```r
termplot(modg, partial.resid = TRUE, terms=3)
```




```r
termplot(modg, partial.resid = TRUE, terms=4)
```


Partial for verbal
100
50
0
−50
2
4
6
8
10
verbal