HOMEWORK 7

AUTHOR
Kevin Linares and Jamila Sani

PUBLISHED
October 29, 2024

```
library(car)
library(gridExtra)
library(performance)
library(faraway)
library(knitr)
library(tidyverse)

options(scipen=999)
```

1.FARAWAY CHAPTER 4. EXERCISE 2. USING THE TEENGAMB DATASET FROM THE FARAWAY PACKAGE FIT A LINEAR REGRESSION MODEL WITH GAMBLE AS THE OUTCOME AND THE OTHER VARIABLES AS THE PREDICTORS.

- We model gamble, expenditure on gambling in pounds per year, as a function on sex (male, female), status score based on parents' occupation), income (in pounds per week), and verbal scores (in words out of 12 correctly defined).

```
data("teengamb")
```

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

```

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 0.0000179 ***
verbal       -2.95949    2.17215  -1.362    0.1803    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
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: 0.000001815
```

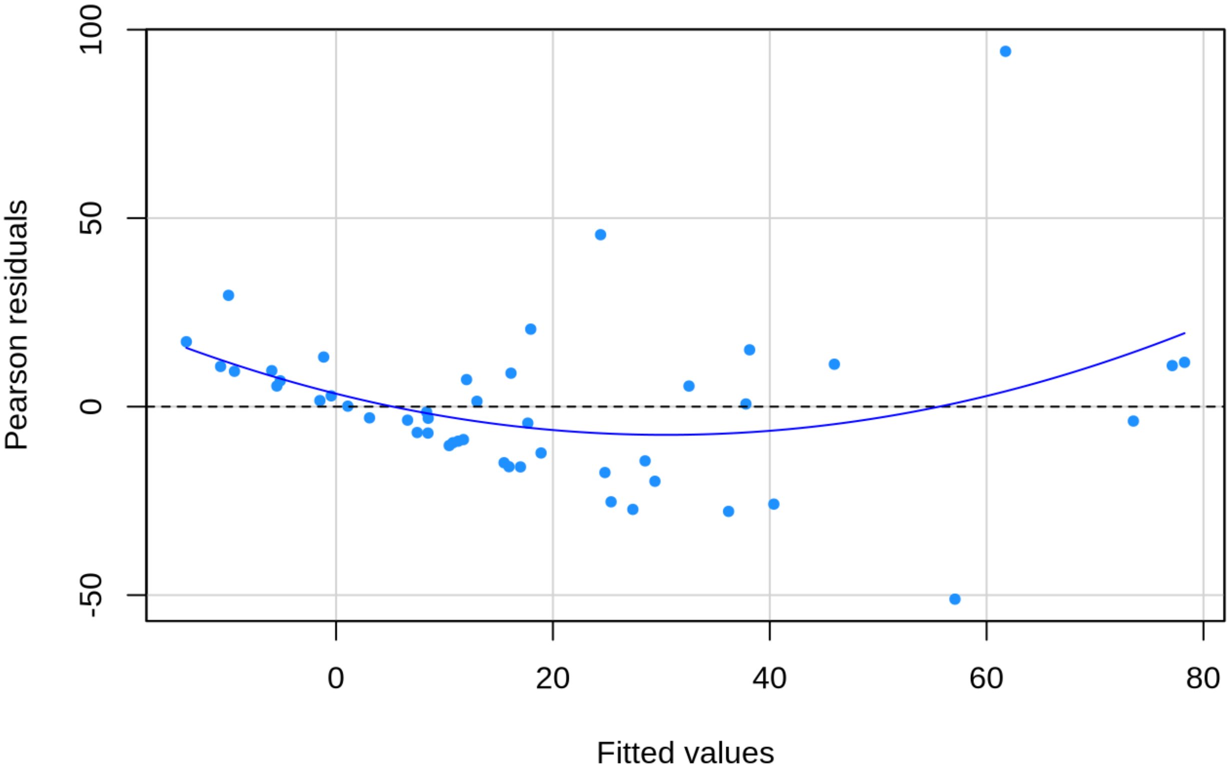
- Perform regression diagnostics on the model to answer the questions 1.A to 1.G. Display any plots that are relevant. Do not provide plots about which you have nothing to say. Suggest possible improvements or corrections to the model where appropriate.

- 1.A. Check the zero mean error assumption using residual plots. What do you conclude about whether the assumption is met?
  - We compute summary statistics of our model residuals and show that they are centered around 0. However, in examining the $\hat{\epsilon}$ against $\hat{y}$ we see that the reference line in green is not flat and horizontal. We see that for low and high fitted values the line is above 0, and for average fitted values the green reference line is below 0, suggesting that the residual errors don't always have a mean value of 0.

```r
summary(resid(mod))
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
-51.082 -11.320  -1.451   0.000   9.452  94.252 
```

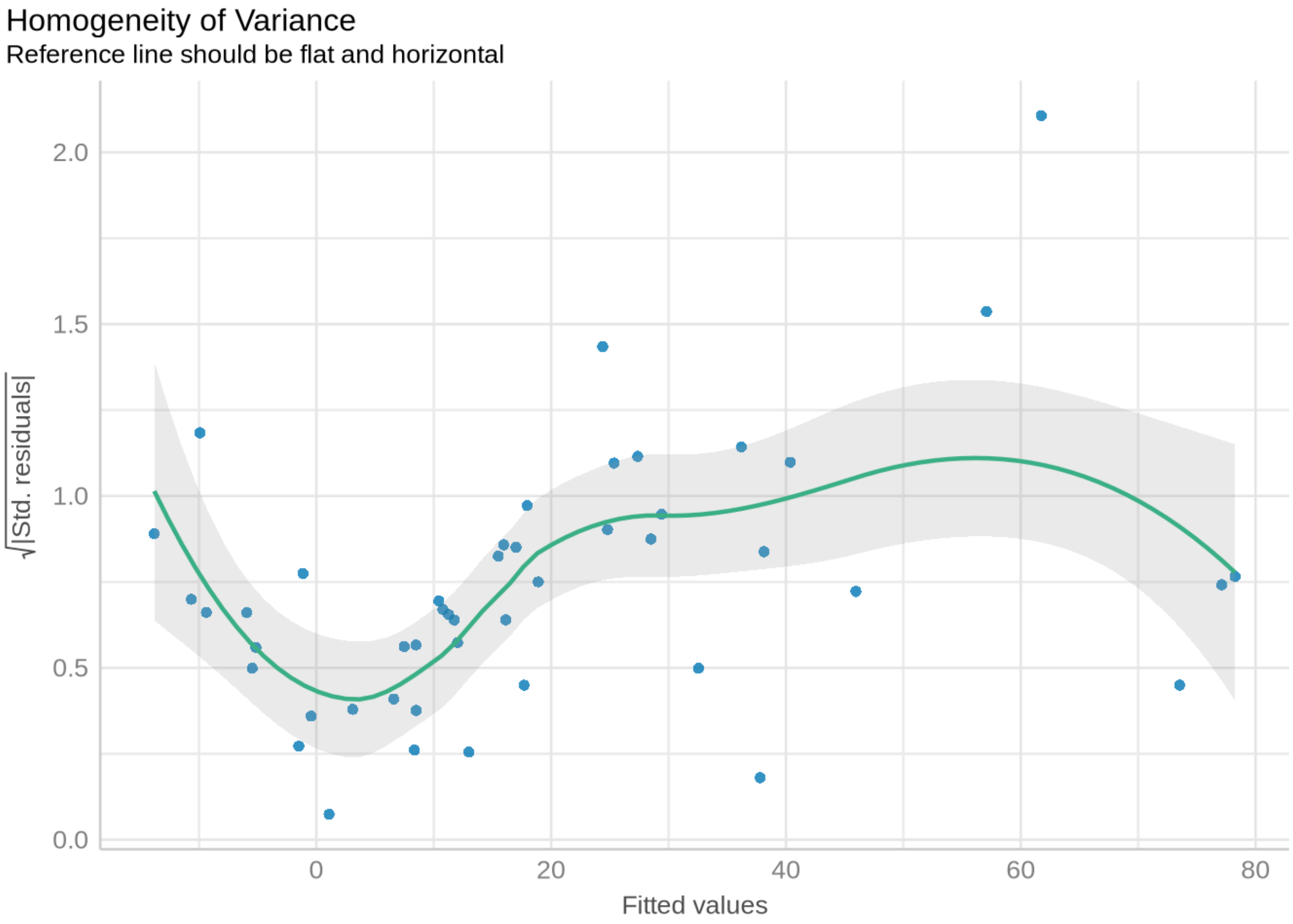
```r
residualPlots(mod, pch=20, col="dodgerblue", terms = ~ 1)
```



```
             Test stat Pr(>|Test stat|)
Tukey test    2.7242         0.006445 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- 1.B. Check the constant error variance assumption using both residual plots and a formal statistical test. What do you conclude?
    - We further examine the constant variance assumption by plotting $\sqrt{|\epsilon|}$ against $\hat{y}$ and it seems that we have non-constant variance based on the plot below. Numerically we take note of a weak correlation between the residuals and predictors, but when we model these values we see that for every one unit increase in fitted values residuals increase by about .03, suggesting that there is evidence for non-constant variance, or heteroscedasticity
        - Note. in a F-test for equal variances, for fitted values equal to or below 30 versus above 30 we see that the variances in the residuals are not equal based on that the confidence intervals (95% [.052, .414]) does not contain 0.

```
plot(check_heteroskedasticity(mod))
```



```
cor(fitted(mod), resid(mod))
```

```
[1] 0.000000000000000002404056
```

```r
summary(lm(sqrt(abs(residuals(mod))) ~ fitted(mod)))
```

```

Call:
lm(formula = sqrt(abs(residuals(mod))) ~ fitted(mod))

Residuals:
   Min     1Q Median     3Q    Max 
-3.055 -1.206 -0.072  0.733  5.176 

Coefficients:
            Estimate Std. Error t value         Pr(>|t|)    
(Intercept)  2.87838    0.31218   9.220 0.00000000000621 ***
fitted(mod)  0.02679    0.01050   2.552           0.0142 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.628 on 45 degrees of freedom
Multiple R-squared:  0.1265,	Adjusted R-squared:  0.107 
F-statistic: 6.514 on 1 and 45 DF,  p-value: 0.01417
```

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

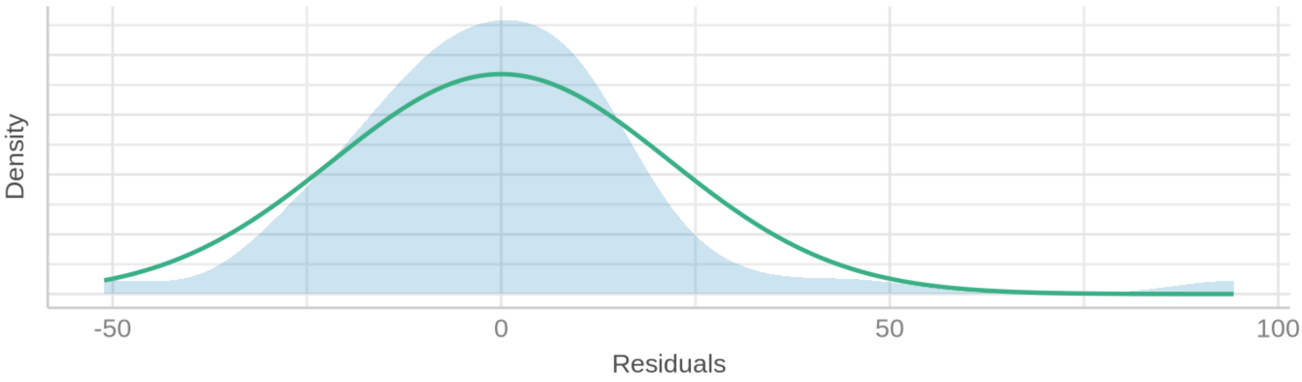
```

	F test to compare two variances

data:  resid(mod)[fitted(mod) <= 30] and +resid(mod)[fitted(mod) > 30]
F = 0.16983, num df = 35, denom df = 10, p-value = 0.00007413
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 
```

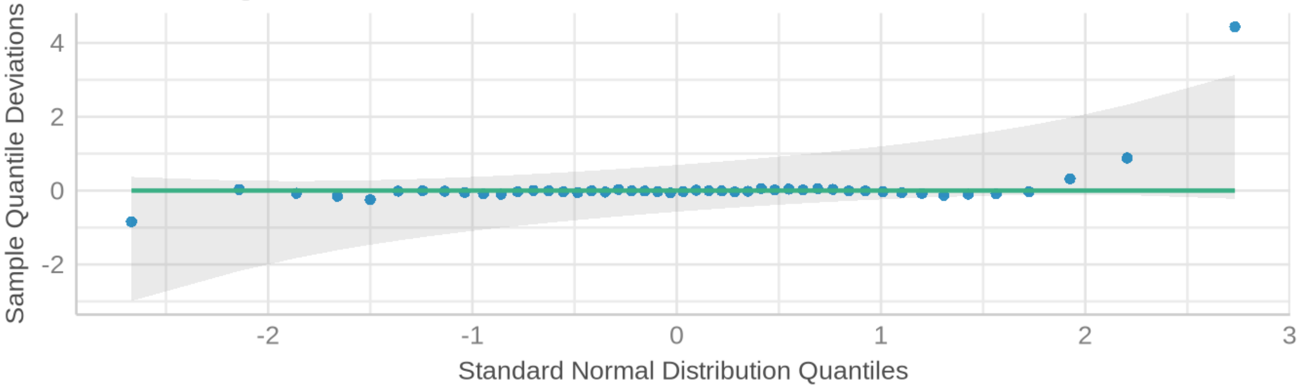
- 1.C. Check the error normality assumption both graphically and statistically. What do you conclude? Which method do you think is preferable?
    - Our visuals show that the distribution of residuals are not normally distributed. In the density plot we see a bump emerging at the right tail, while our qq-plot shows several observations outside of what is expected.
    - Additionally, we used the Shapiro-Wilk normality test and failed to reject the null hypothesis that the residuals are normally distributed in favor of the alternative hypothesis that they are probably not normal. We conclude that both graphically and statistically the assumption of error normality is violated and place more weight on the statistical test over our graphics.

```r
grid.arrange(
  plot(check_normality(mod), type="density"),
  plot(check_normality(mod), type="qq")
)
```

Normality of Residuals

Distribution should be close to the normal curve




Normality of Residuals

Dots should fall along the line

```
shapiro.test(resid(mod))
```

```

	Shapiro-Wilk normality test

data:  resid(mod)
W = 0.86839, p-value = 0.0000816
```

- 1.D. Check for observations with large leverage. Which observations have large leverage?
  - We assess the leverage H statistic using the `hatvalues()` function as well as plot these statistics using the `halfnorm()` function and find that for observations 31, 33, 35, and 42 we may want to further investigate how these observations influence the model estimates.
  - Recommendation: We recommend removing each one of these observations one by one and fitting the model, compare, residuals to determine how much the prediction line moves upon removing a high leveraged observation.

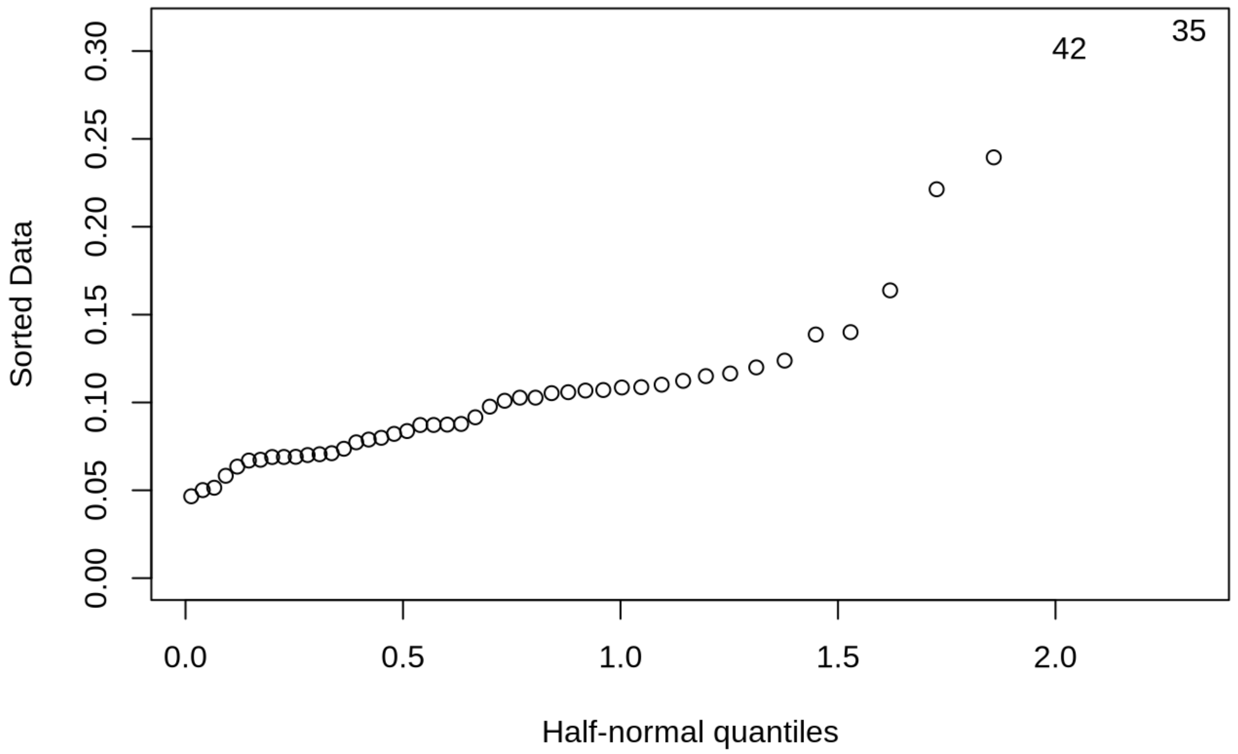
```
hat <- hatvalues(mod) |>
  as.data.frame() |>
  rowid_to_column() |>
  rename(hat=2) |>
  mutate(high_leverage = ifelse(hat  > 2*mean(hat), 1, 0))

hat |> summarise(mean(hat))
```

```
  mean(hat)
  mean(hat)
1  0.106383
```

```
hat |> filter(high_leverage==1) |> select(-high_leverage)
```

```
  rowid       hat
1    31 0.2395031
2    33 0.2213439
3    35 0.3118029
4    42 0.3016088
```

```
halfnorm(hatvalues(mod))
```

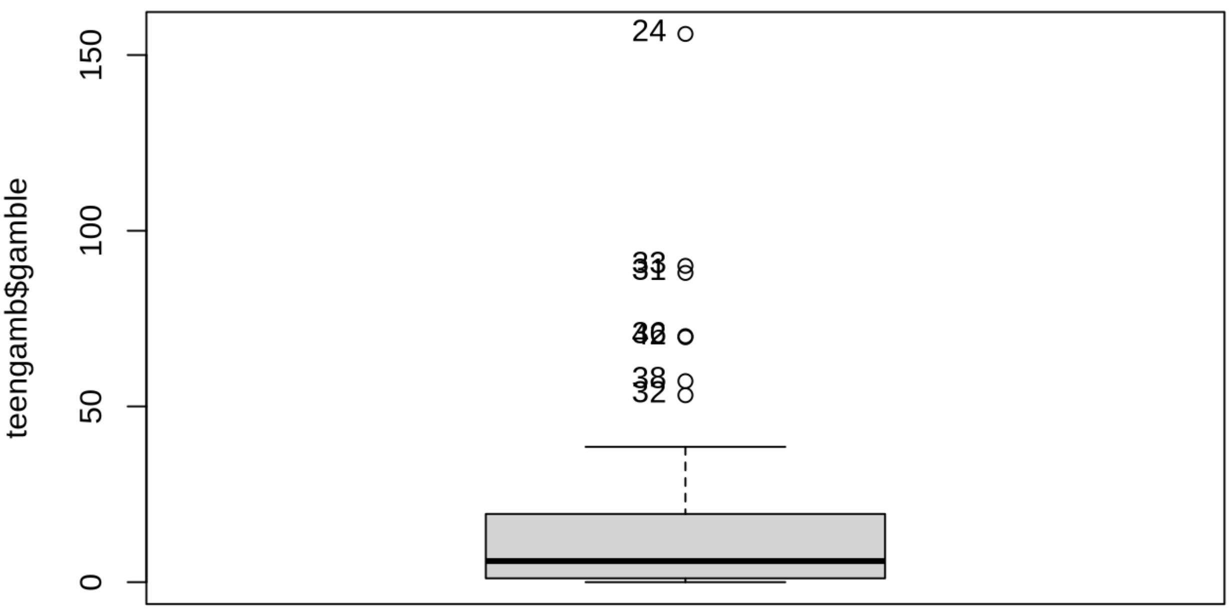


- 1.E. Check for outliers. List any potential outliers.
    - Using studentized residuals, any residual divided by its error if greater than 3, we determine that observation 24 is an outlier. We further examine a boxplot for raw gambling scores and take note that observation 24 is far away from the mean.
        - Note. Given that we test every observation, we make a Bonferroni correction test by adjusting the critical value.
    - Recommendation: We recommend removing observation 24 as it is a potential outlier and refitting the model.

```
r_it <- rstandard(mod) |>
  as.data.frame() |>
  rowid_to_column() |>
  rename(stud_resid = 2) |>
  mutate(outlier = ifelse(
    abs(stud_resid) > qt(1-.05/ n(),n()-2), 1, 0))

r_it |> filter(outlier == 1) |> select(-outlier)
```

```
  rowid stud_resid
1    24    4.43762
```

```
Boxplot(teengamb$gamble)
```



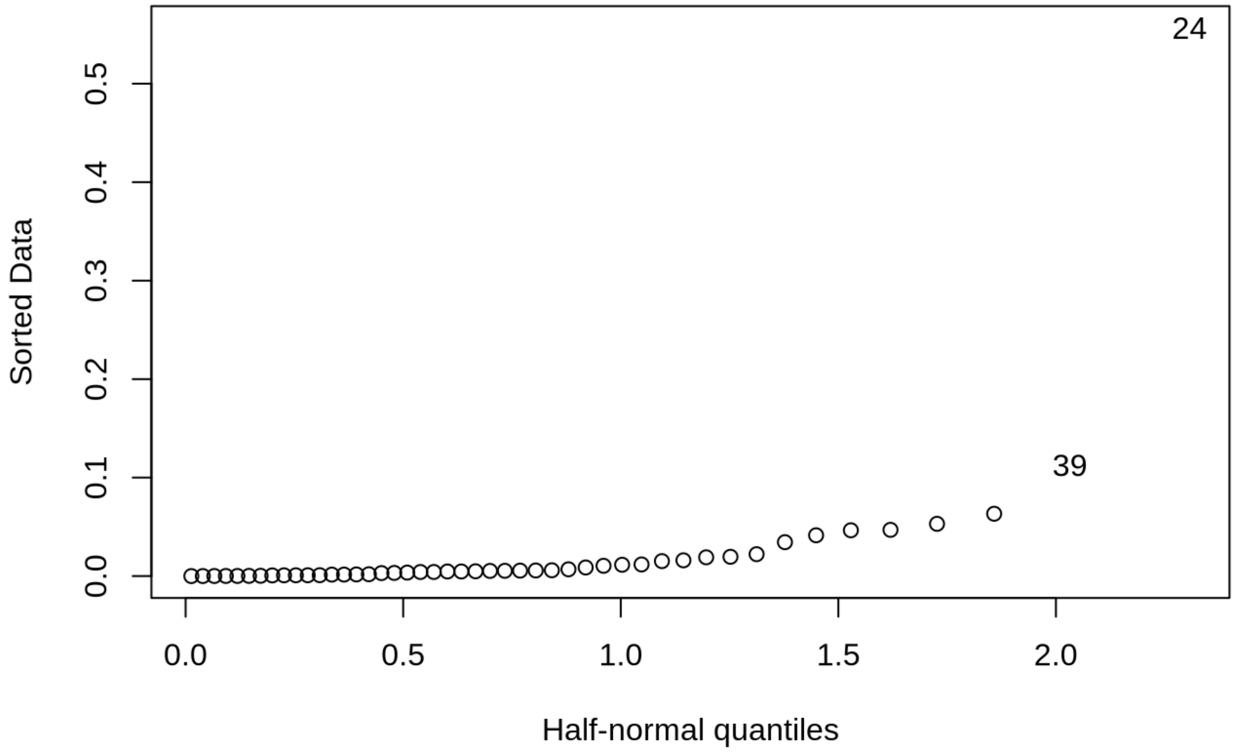
```
[1] 24 31 32 33 36 38 42
```

- 1.F. Check for influential points. List any potential influential points.
  - We use Cooks statistic to identify any influential observations and again find observation 24 to be problematic as seen in the halfnorm plot.
  - Recommendation: We recommend removing observation 24 as it is a potential outlier and influential value and refitting the model.

```
cd<-cooks.distance(mod)

summary(cd)
```

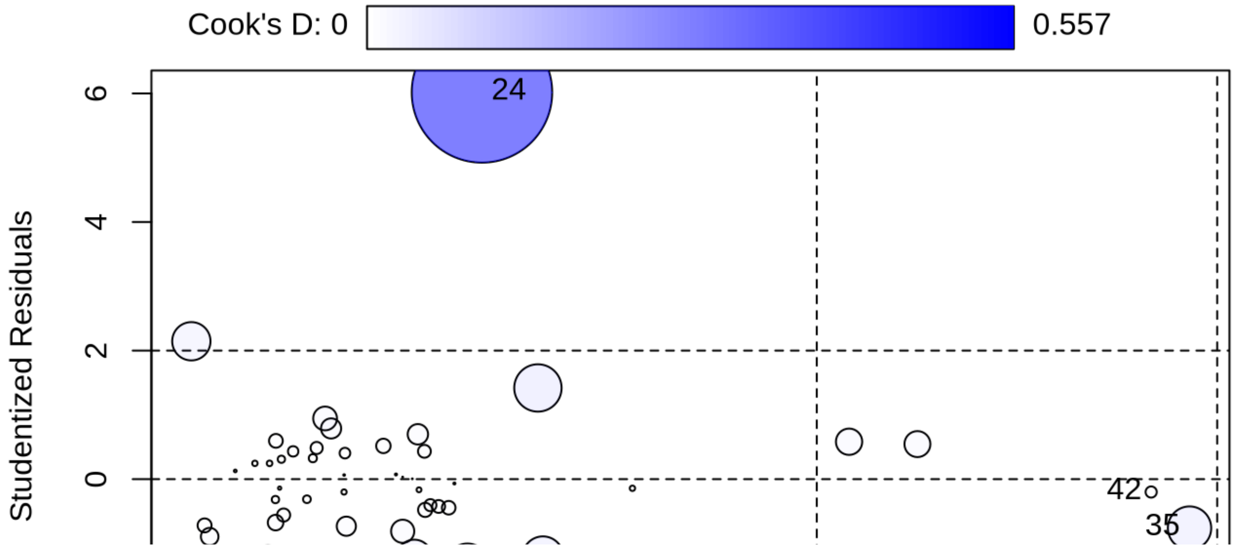
```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
0.0000007 0.0011908 0.0048478 0.0248308 0.0155806 0.5565011 
```

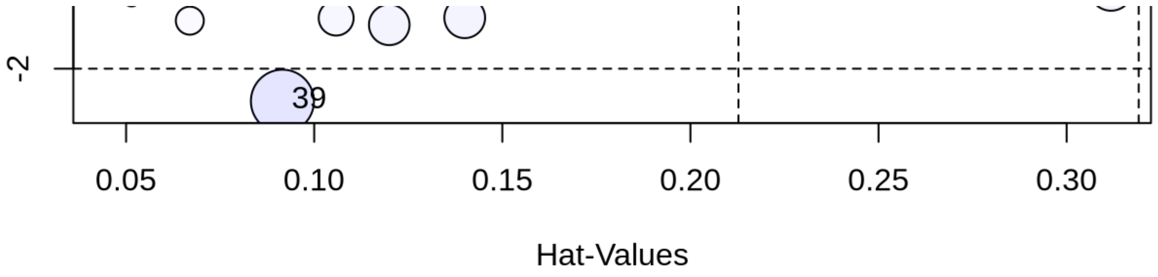
```
halfnorm(cd)
```



- We can also use the car package to see how well our assessment of high leverage, outliers, and influential observations we did. We find that we overlap in identifying 24, 42, and 35, but not 31 or 33 due to high hat values. In the car assessment we also see that 39 gets flagged; however, we did not identify this as a problem given a low studentized residual.
    - Note. in the plot below we take note that observation 24 has a high cook statistic as well as a high studentized residual despite having a low hat value.

```
influencePlot(mod)
```





```
      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.G. Check the structure of relationship between the predictors and the response. What do you observe?
    - We perform partial regression subsetting by sex. We observe that the model subset by males has an adjusted r-squared of .50 while for females it was .07, although there are almost twice as many males in the sample.
        - talk about the differences in coefficients and confidence intervals.

```
mod_male<-lm(gamble ~ status + income + verbal,
             subset(teengamb, sex == 0))

mod_female<-lm(gamble ~ status + income + verbal,
               subset(teengamb, sex == 1))

summary(mod_male); summary(mod_female)
```

```
Call:
lm(formula = gamble ~ status + income + verbal, data = subset(teengamb,
    sex == 0))

Residuals:
    Min      1Q  Median      3Q     Max
-56.654 -12.104  -2.061   7.729  83.903

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  27.6354    22.2192   1.244 0.225600
status       -0.1456     0.4181  -0.348 0.730748
income        6.0291     1.3288   4.537 0.000135 ***
verbal       -2.9748     3.0596  -0.972 0.340617
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.45 on 24 degrees of freedom
Multiple R-squared:  0.5536,    Adjusted R-squared:  0.4977
F-statistic: 9.919 on 3 and 24 DF,  p-value: 0.0001936


Call:
lm(formula = gamble ~ status + income + verbal, data = subset(teengamb,
    sex == 1))

Residuals:
    Min      1Q  Median      3Q     Max
-8.6972 -2.0567 -0.5836  2.6533 11.2536
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5.3778     7.1848  -0.749   0.4657
status        0.2073     0.1038   1.997   0.0643 .
income        0.6813     0.5177   1.316   0.2079
verbal       -0.1392     0.9259  -0.150   0.8825
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.974 on 15 degrees of freedom
Multiple R-squared:  0.2228,    Adjusted R-squared:  0.06738
F-statistic: 1.433 on 3 and 15 DF,  p-value: 0.2723
```

```
confint(mod_male); confint(mod_female)
```

```
                 2.5 %     97.5 %
(Intercept) -18.222842 73.4936478
status       -1.008459  0.7173273
income        3.286586  8.7715717
verbal       -9.289502  3.3399831

                   2.5 %    97.5 %
(Intercept) -20.69187319 9.9361899
status       -0.01396582 0.4286003
income       -0.42212178 1.7847306
verbal       -2.11265621 1.8341712
```