# Chapter 15: Putting It Altogether

Answers to Exercises

Brian Fogarty

## Contents



## Exercise 1

### Exercise 1.a

```
library(tidyverse)
library(readxl)

kenya <- read_xlsx("kenya_wvs.xlsx", .name_repair =
                     ~ str_sub(.x, start = 1, end = 7)) %>%
              select(starts_with("Q"))
```

```
kenya_subset <- kenya %>%
  rename(responsibility= `Q108: G`,
         ideology = `Q240: L`,
         age = `Q262: A`,
         sex = `Q260: S`,
         edu = `Q275R: `
         ) %>%
  mutate(across(everything(), ~replace(., .x < 0, NA))) %>%
  select(responsibility,ideology,age,sex,edu)

glimpse(kenya_subset)
```

```
Rows: 1,266
Columns: 5
```

```
$ responsibility <dbl> 10, 3, 1, 1, 2, 1, 1, 5, 10, 1, 10, 10, 2, 4, 8, 1, 1, ~
$ ideology       <dbl> 7, 7, 1, 5, NA, 5, 5, 6, NA, 3, 2, 1, 2, 7, 4, 3, 4, 3,~
$ age            <dbl> 63, 24, 26, 29, 37, 45, 21, 50, 26, 28, 30, 38, 40, 24,~
$ sex            <dbl> 1, 1, 1, 2, 1, 1, 1, 2, 1, 1, 2, 2, 1, 1, 2, 1, 1, 2, 2~
$ edu            <dbl> 2, 3, 3, 3, 2, 2, 2, 3, 1, 3, 1, 2, 1, 2, 3, 3, 1, 2, 2~
```

## Exercise 1.b

First, we state a generic null hypothesis for the outcome variable `responsibility`, and then state hypotheses for our predictors.

$H_{01}$: There is no relationship between `predictor`[1] and views on government versus individual responsibility.

Now, let's write out hypotheses for our five predictors ($H_1$ through $H_4$):

$H_1$: As political ideology increases, respondents are expected to have a higher belief in individual responsibility.

$H_2$: As age increases, respondents are expected to have a lower belief in individual responsibility.

$H_3$: Women respondents are expected to have a lower belief in individual responsibility than men respondents.

$H_4$: As level of education increases, respondents are expected to have a higher belief in individual responsibility.

## Exercise 1.c

```
summary(model.1 <- lm(responsibility ~ ideology + age + sex +
                        edu, data = kenya_subset))
```

```
Call:
lm(formula = responsibility ~ ideology + age + sex + edu, data = kenya_subset)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5379 -3.0027 -0.3571  2.3416  6.4185

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.811645   0.570473   6.682 3.69e-11 ***
ideology     0.073302   0.035304   2.076  0.03809 *
age         -0.015084   0.009599  -1.571  0.11637
sex          0.072549   0.191603   0.379  0.70502
edu          0.408421   0.128171   3.187  0.00148 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.201 on 1139 degrees of freedom
  (122 observations deleted due to missingness)
Multiple R-squared:  0.01551,   Adjusted R-squared:  0.01206
F-statistic: 4.487 on 4 and 1139 DF,  p-value: 0.001337
```

We see $R^2 = 0.01551$, which we interpret as *our model explains 1.55% of the variance in views on government versus individual responsibility.* We see that the $p$-value for the $F$-test is below 0.05 and thus our overall

[1] Substitute in the predictor name.

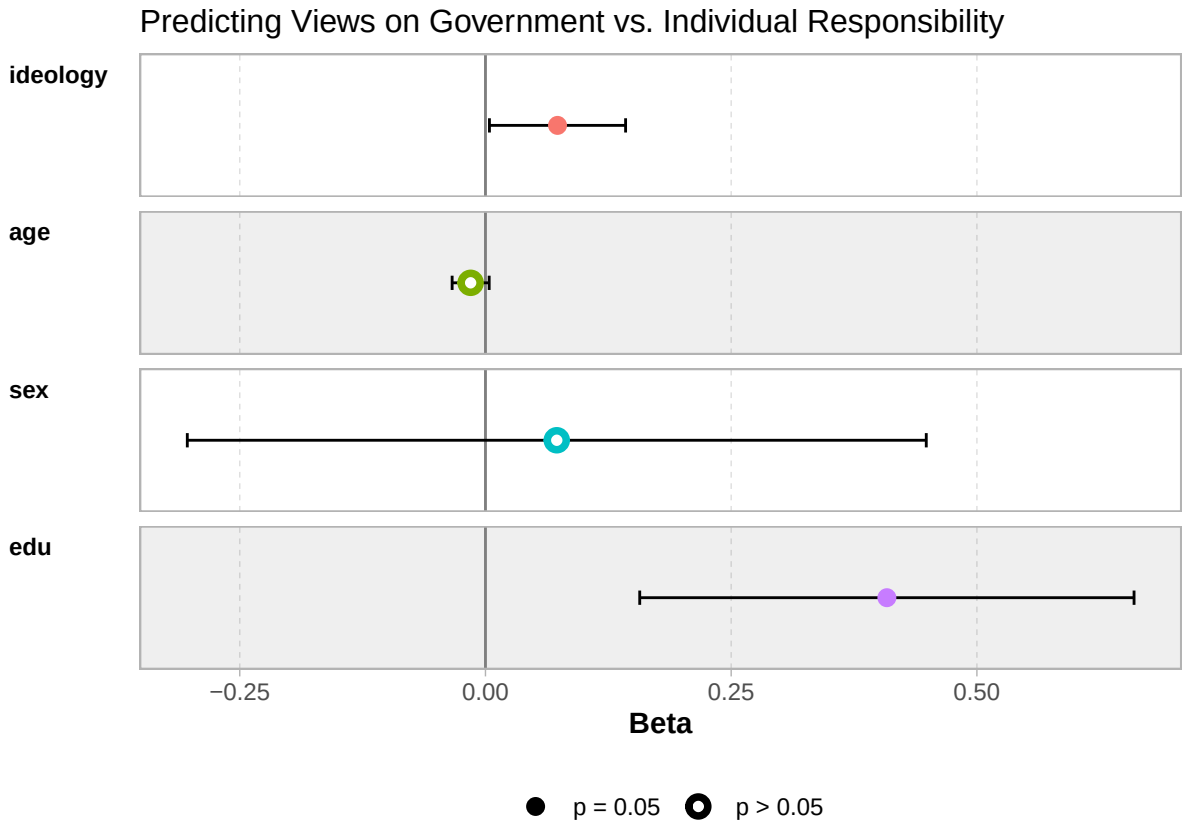
model is statistically significant. Again, this means that our model is better than a model where all the predictors equal 0.

Two predictors - `ideology` and `edu` - have a positive statistically significant effect on `responsibility`. (We conduct interpretations and discuss the predictors in Exercise 3.)

### Exercise 1.d

The predictor names are fairly clear, so we will not include new labels in the plot.

```
library(GGally)
ggcoef_model(model.1,
             show_p_values = FALSE,
             signif_stars = FALSE) +
  labs(title = "Predicting Views on Government vs. Individual Responsibility") +
  theme(
    plot.title = element_text(size = 12)
  )
```



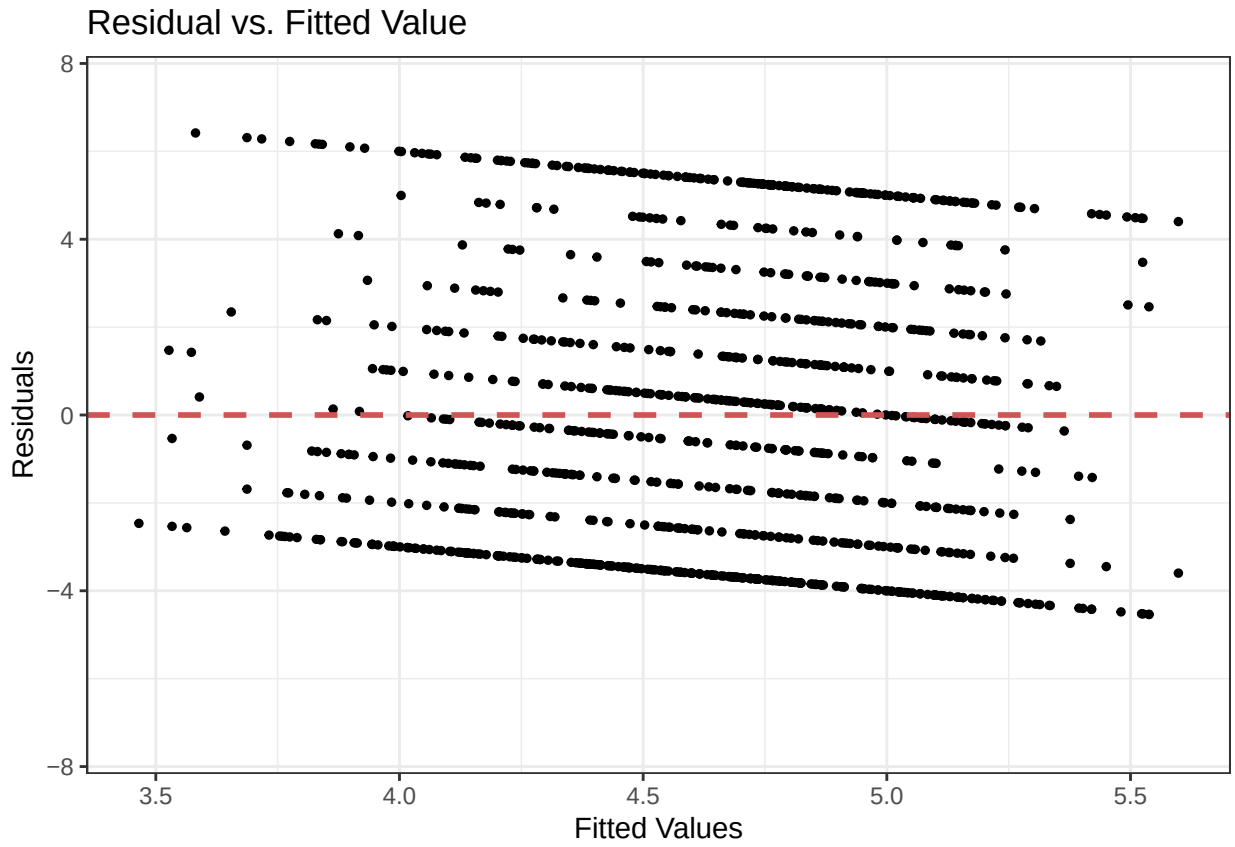

## Exercise 2

### Exercise 2.a - Functional Form

The first test is to plot the residuals and fitted values of our model.

```
library(lindia)

gg_resfitted(model.1) +
  theme_bw()
```




It is a little tricky to figure out the local means here and thus it is difficult to know if we violate functional form.

Next, we'll use the `resettest()` function from the `lmtest` package for the Ramsey RESET test.

```
library(lmtest)

resettest(model.1, power = 2:3, type = "fitted")
```

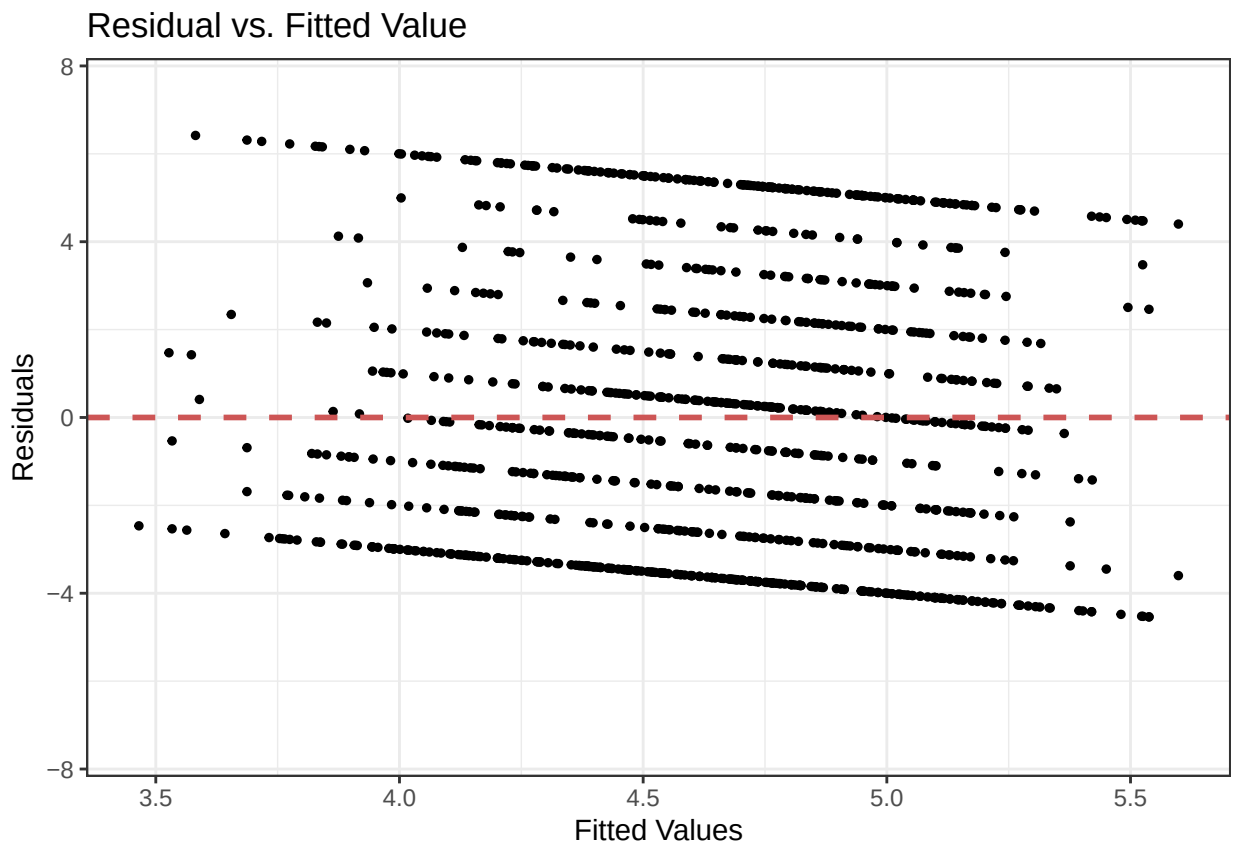
```
    RESET test

data:  model.1
RESET = 0.91203, df1 = 2, df2 = 1137, p-value = 0.402
```

We see that $p > 0.05$, we do not reject the null, and conclude that we do not violate the assumption of correct functional form.

## Exercise 2.b - Heteroscedasticity

```
gg_resfitted(model.1) +
  theme_bw()
```




There is a pattern to the residuals, slanting downwards, but given our non-continuous outcome variable is not entirely clear whether we have heteroscedasticity.

```
bptest(model.1, studentize = FALSE)
```

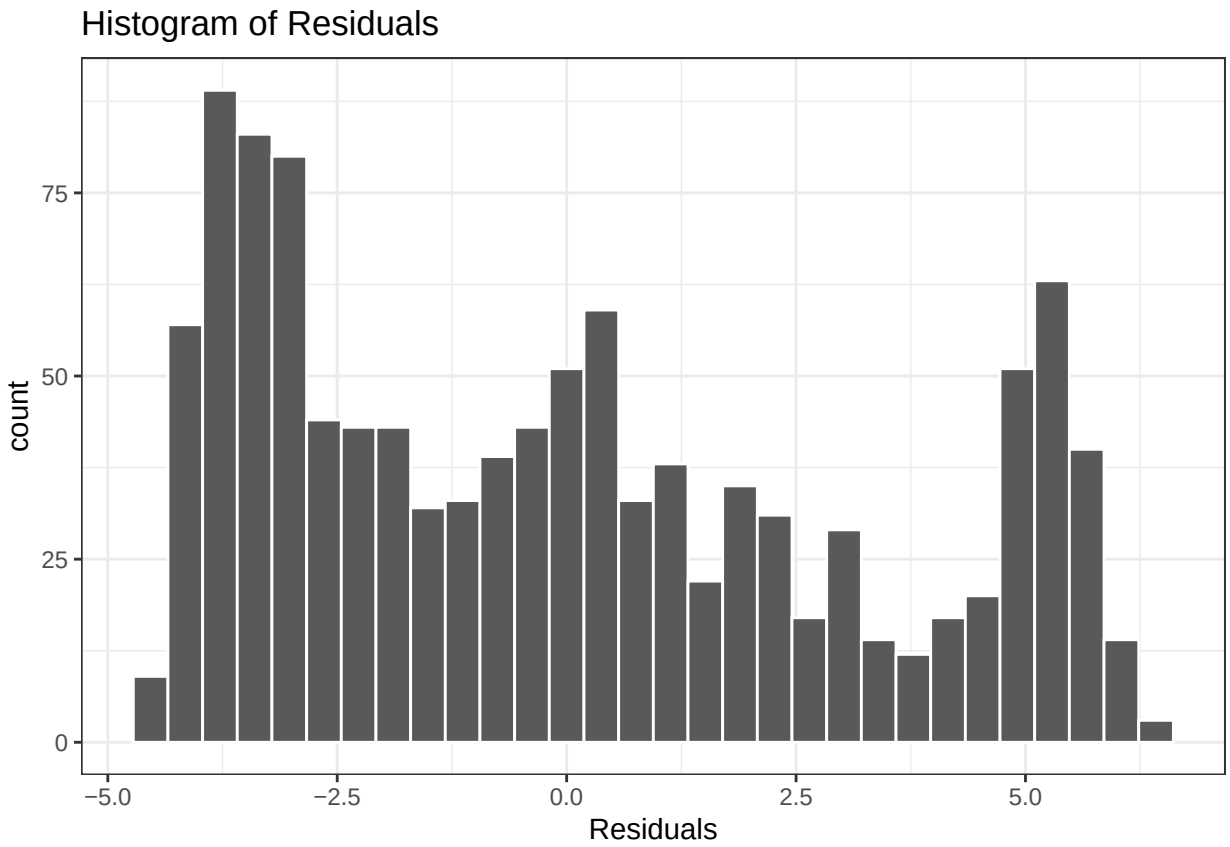
```
    Breusch-Pagan test

data:  model.1
BP = 3.5213, df = 4, p-value = 0.4746
```

We see that $p > 0.05$, we do not reject the null (of homoscedasticity), and conclude that we do not have heteroscedasticity.

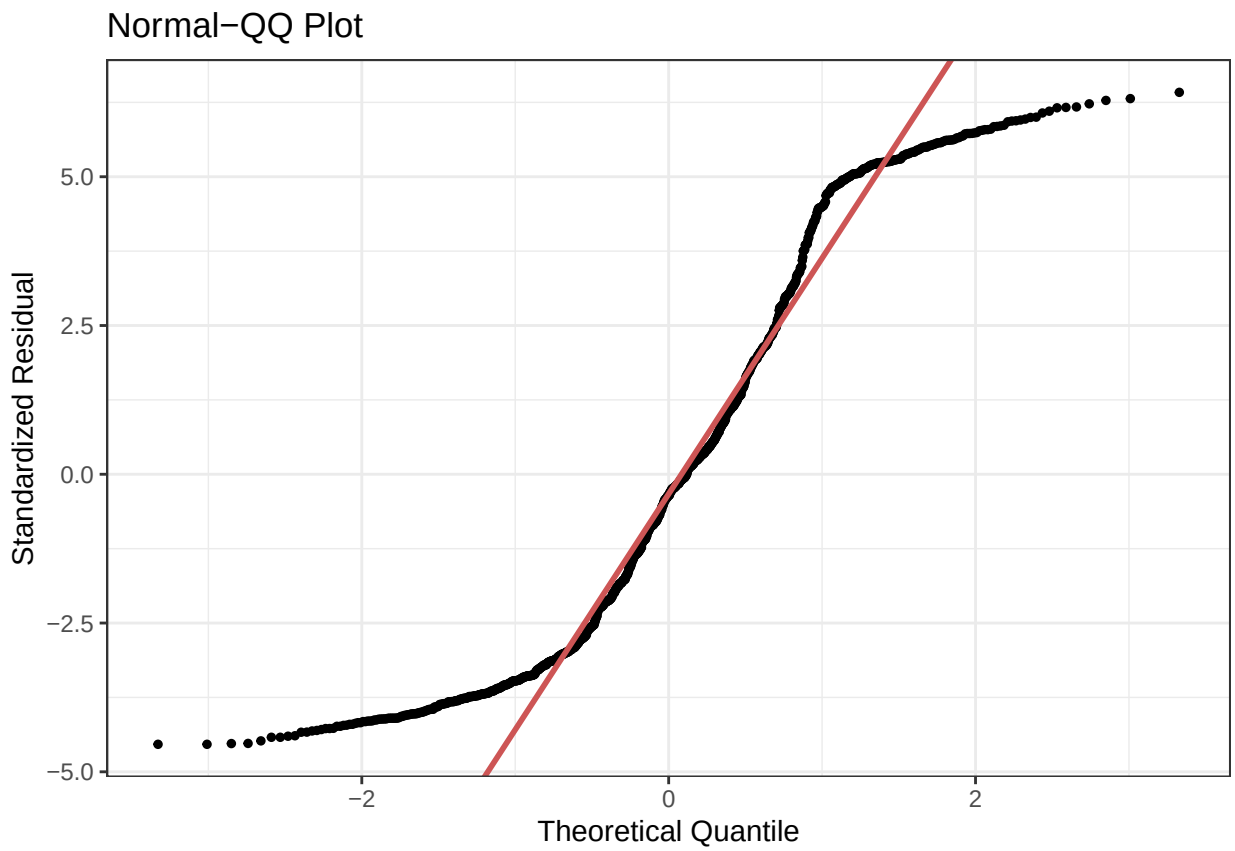
## Exercise 2.c - Normality

```
gg_reshist(model.1) +
  theme_bw()
```




This histogram shows the residuals are not normally distribution.

```
gg_qqplot(model.1) +
  theme_bw()
```




The Q-Q plot also suggests problems with normality.

```
library(nortest)
ad.test(model.1$residuals)
```

```
    Anderson-Darling normality test

data:  model.1$residuals
A = 27.329, p-value < 2.2e-16
```

We see that $p \leq 0.05$, we reject the null, and thus we cannot assume our residuals are normally distributed.

Although we can use a Box-Cox transformation (using the `powerTransform()` function), the nature of our outcome variable will make any possible transformation non-sensical. So, we will pass on finding a solution.

## Exercise 2.d - Multicollinearity

```
library(car)
vif(model.1)
```

```
ideology      age      sex      edu
1.005564 1.022506 1.023440 1.033289
```

None of the VIF values are near 10 and thus we don't have multicollinearity.

### Exercise 2.e - Outliers, Leverage, and Influential Data Points

We have 4 predictors and 1,144 observations ($n$) in `model.1`.[2] We calculate our leverage cut-point as:

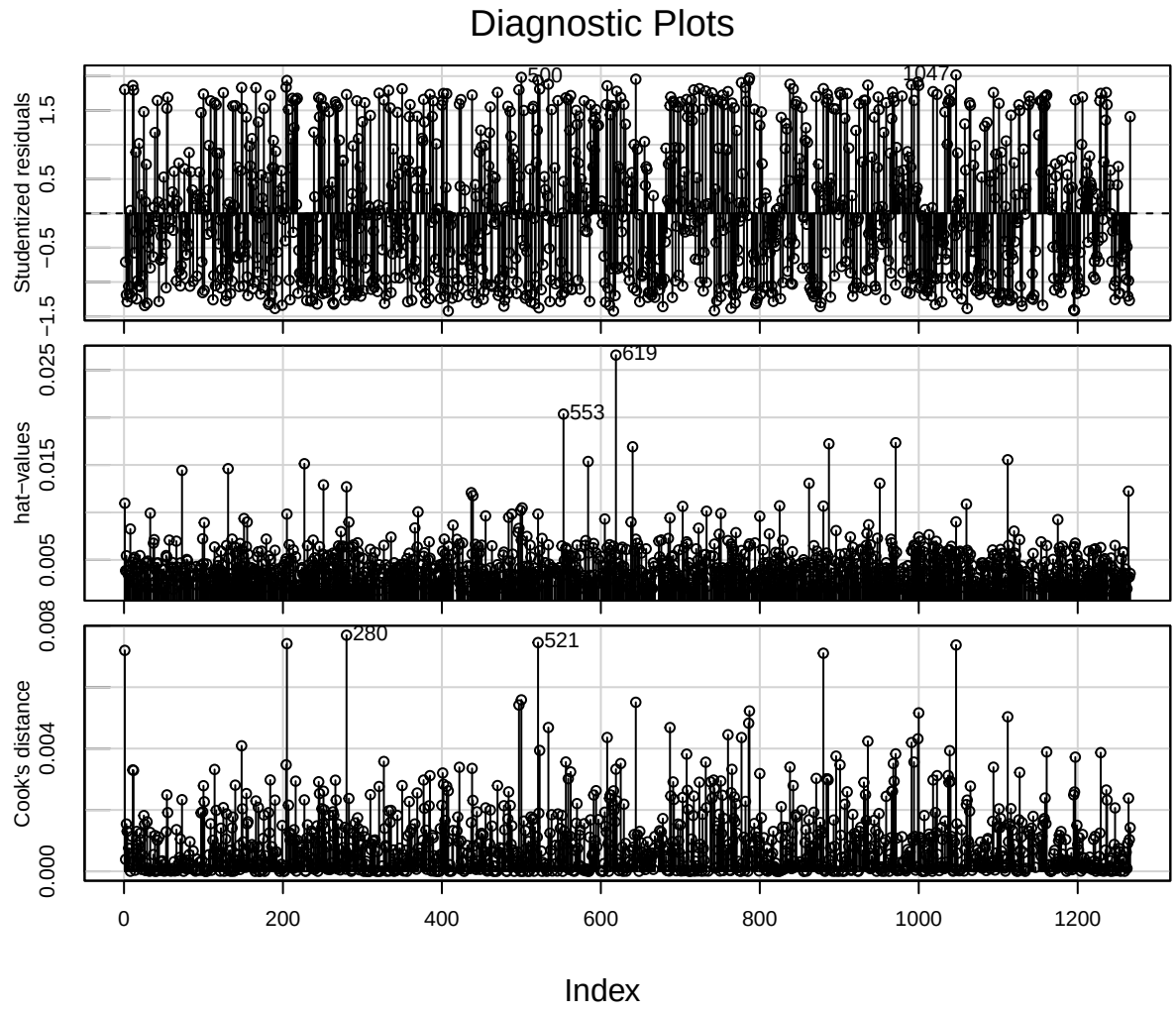
```
(2*(4+1))/1144
```

```
[1] 0.008741259
```

Thus, any data point that has a hat-value $\geq 0.0087$ is considered to have high leverage.

Since we have a relatively large number of observations, we'll consider any point that has a Cook's distance greater than 1 to be influential.

```
influenceIndexPlot(model.1,
                   vars = c("Studentized","hat","Cook"))
```




We see a lot of outliers (and large outliers), points with leverage, but none of the observations are influential.

[2] The data has 1,266 observations, but 122 are removed in `model.1` due to missingness.

## Exercise 3

Let's look again at `model.1`'s results.

```
summary(model.1)
```

```
Call:
lm(formula = responsibility ~ ideology + age + sex + edu, data = kenya_subset)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5379 -3.0027 -0.3571  2.3416  6.4185

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.811645   0.570473   6.682 3.69e-11 ***
ideology     0.073302   0.035304   2.076  0.03809 *
age         -0.015084   0.009599  -1.571  0.11637
sex          0.072549   0.191603   0.379  0.70502
edu          0.408421   0.128171   3.187  0.00148 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.201 on 1139 degrees of freedom
  (122 observations deleted due to missingness)
Multiple R-squared:  0.01551,   Adjusted R-squared:  0.01206
F-statistic: 4.487 on 4 and 1139 DF,  p-value: 0.001337
```

Two predictors - `ideology` and `edu` - have a statistically significant effect on `responsibility`. The coefficient interpretation for `ideology` is *for a one-unit increase in ideology, respondents are expected to increase their belief in individual responsibility by 0.073 units.* The phrasing of the outcome variable is a bit awkward, but the direction of the effect is as expected - belief in individual responsibility is one of the core tenets in conservatism. Specifically, respondents who are more conservative are expected to have a higher belief in individual responsibility as opposed to government responsibility.

The coefficient interpretation for `edu` is *for a one-unit increase in education, respondents are expected to increase their belief in individual responsibility by 0.408 units.* Therefore, respondents with a higher level of education are expected to have a higher belief in individual responsibility as opposed to government responsibility. (Without knowing more about Kenya, I'm somewhat hesitant to speculate about this result. If this was a "real" research project, I would the examine the academic literature and talk with experts on Kenyan society.)