Chapter 12 - OLS Assumptions and Diagnostic Testing

Exercises

Brian Fogarty

15 September 2018

Contents

EXERCISE I

Using the 2012 Smoking and Drug Use Amongst English Pupils Dataset (`2012smokedrugs.dta`), perform diagnostics on the second cigarette consumption multiple linear regression model from the Chapter 11 exercises (Question 3). To remind you, the outcome variable was `cigs7`, but recoded to remove all 0s, and the predictor variables were `free`, `schyear`, and `sex`.

1. **Functional Form Diagnostics:**
    (a) Create a plot to check for a functional form violation.
    (b) Perform a Ramsey RESET test.
    (c) If you violate the functional form assumption, try to find a solution using the techniques covered in this chapter.
2. **Heteroscedasticity Diagnostics:**
    (a) Create a plot to test for heteroscedasticity.
    (b) Perform a Breusch-Pagan test.
    (c) If heteroscedasticity is present re-run the regression using robust standard errors. Are any predictors that were statistically significant now not significant?
3. **Normality Diagnostics:**
    (a) Create a histogram of the residuals to test for non-normality.
    (b) Create a Q-Q plot to test for non-normality.
    (c) Perform a Shapiro-Wilk Normality test.
    (d) If you find non-normality in the residuals, try to find a solution using the techniques covered in this chapter.
4. **Multicollinearity Diagnostics:**
    (a) Perform a correlation between predictors to assess whether any have high correlations (over .8).
    (b) Perform a Variance Inflation Factor test.
5. **Outliers, Leverage, and Influential Data Points Diagnostics:**
    (a) Calculate the cut-point for assessing data points with high leverage.
    (b) Use the `influenceIndexPlot()` function to test for outliers, leverage, and influential data points.

ANSWERS FOR EXERCISE 1

Read-in 2012 Smoking and Drug Use Amongst English Pupils, and re-run regression.

```
setwd("C:/QSSD/Exercises/Chapter 12 - Exercises")
getwd()
```

```
[1] "C:/QSSD/Exercises/Chapter 12 - Exercises"
```

```
library(foreign)
drugs <- read.dta("2012smokedrugs.dta", convert.factors=FALSE)

library(car)
```

```
Loading required package: carData
```

```
drugs$cigs7a <- recode(drugs$cigs7, "0=NA")
table(drugs$cigs7a)
```

```

 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18
54 28 22 28 12 16 16  4  7  8  7  6  5  8  1  5  6  5
19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
```

```
  3   2   3   3   2   2   4   4   2   5   4   2   6   1   1   3   9   3 
 37  38  39  40  41  42  43  45  46  47  49  50  51  52  53  54  55  56 
  3   2   1   4   2   6   2   3   4   1   3   3   2   4   4   1   1   2 
 60  62  63  66  67  69  70  71  73  74  75  76  77  79  80  83  84  85 
  7   2   1   2   2   3  10   1   1   1   3   1   1   1   8   1   1   1 
 86  89  90  92  95 100 104 105 110 140 
  2   1   4   1   1   2   1   3   3   2 
```

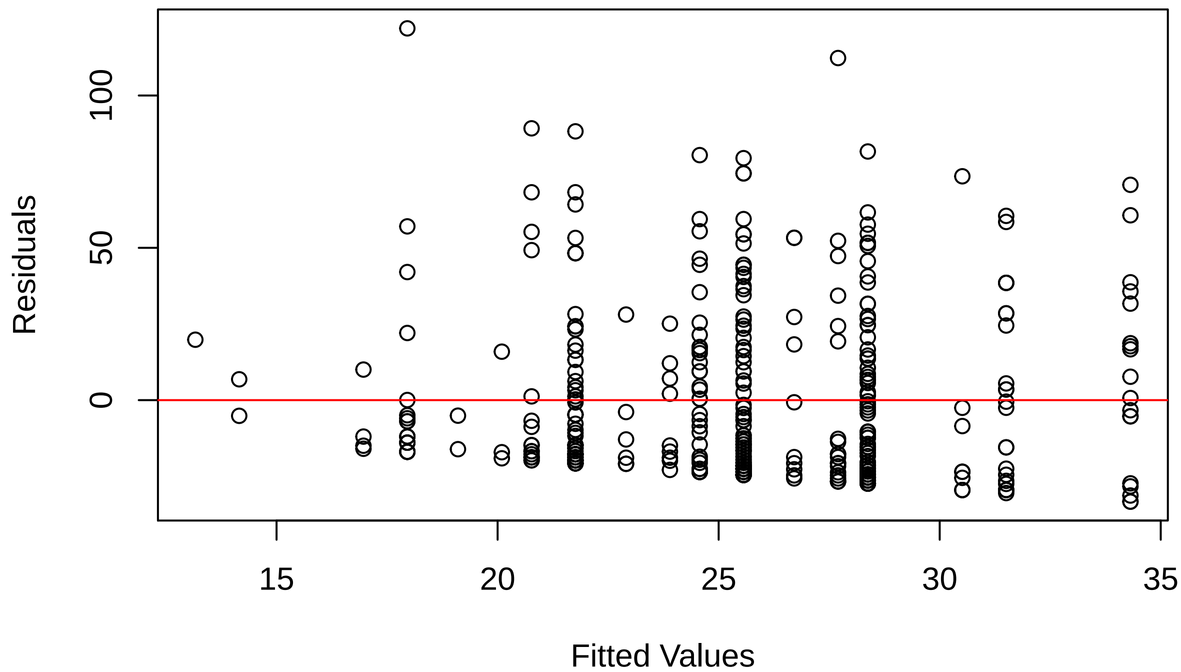
```r
model.1 <- lm(cigs7a ~ free + schyear + sex, data=drugs)
summary(model.1)
```

```
Call:
lm(formula = cigs7a ~ free + schyear + sex, data = drugs)

Residuals:
   Min     1Q Median     3Q    Max 
-33.31 -20.80 -12.47  15.99 122.04 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)    6.549      7.452   0.879   0.3800  
free           5.940      3.365   1.765   0.0783 .
schyear        3.803      1.625   2.340   0.0197 *
sex            2.811      2.855   0.984   0.3255  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.7 on 404 degrees of freedom
  (7181 observations deleted due to missingness)
Multiple R-squared:  0.01964,	Adjusted R-squared:  0.01236 
F-statistic: 2.698 on 3 and 404 DF,  p-value: 0.04555
```

Note: the general problem with regression diagnostics here is that 2 predictors are nominal and one is ordinal.

Question 1.1.a

```r
x11()
plot(y=model.1$residuals,x=model.1$fitted.values, xlab="Fitted Values", ylab="Residuals")
abline(h=0, col="red")
```



It is not clear whether there is a local mean of 0.

Question 1.1.b

```r
library(lmtest)
```

```
Loading required package: zoo


Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

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

```

    RESET test

data:  model.1
RESET = 0.76745, df1 = 2, df2 = 402, p-value = 0.4649
```

The p-value is above .05, thus we do not violate the functional form assumption.

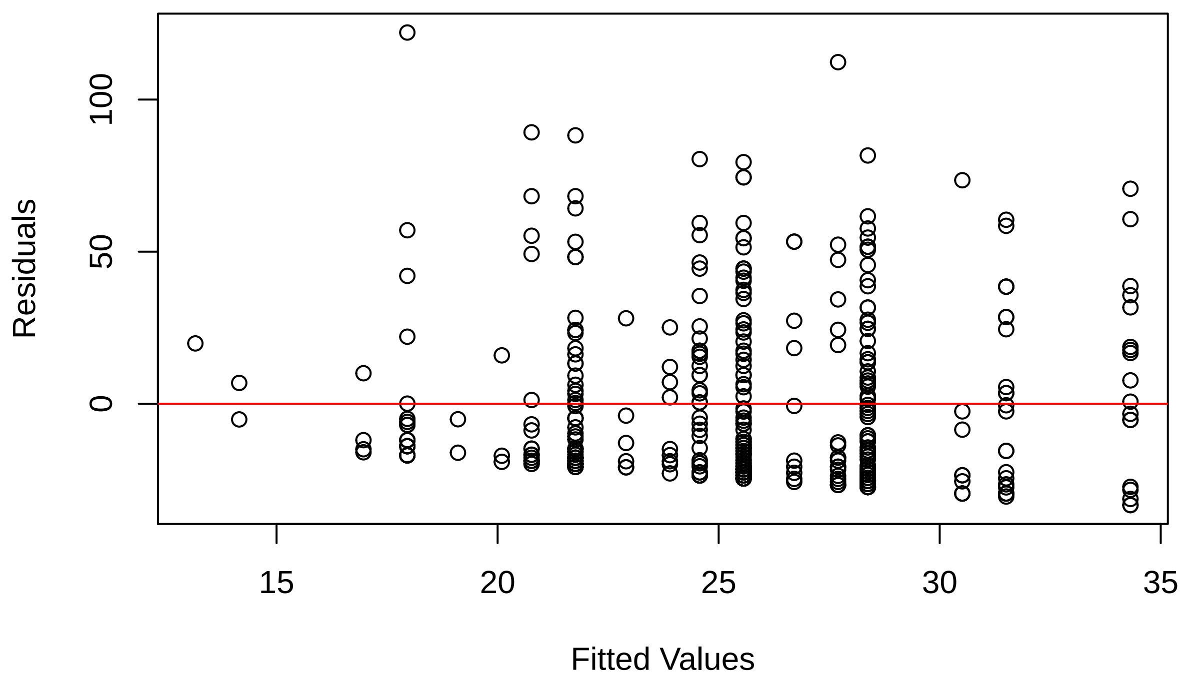
Question 1.1.c

Since we do not have an incorrect functional form, we do not need to attempt any corrections.

Question 1.2.a

This is just a repeat plot from Question 1.1.a.

```r
x11()
plot(y=model.1$residuals,x=model.1$fitted.values, xlab="Fitted Values", ylab="Residuals")
abline(h=0, col="red")
```



There is somewhat of a fan pattern, possibly indicating heteroscedasticity.

Question 1.2.b

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

```
	Breusch-Pagan test

data:  model.1
BP = 1.1206, df = 3, p-value = 0.7721
```

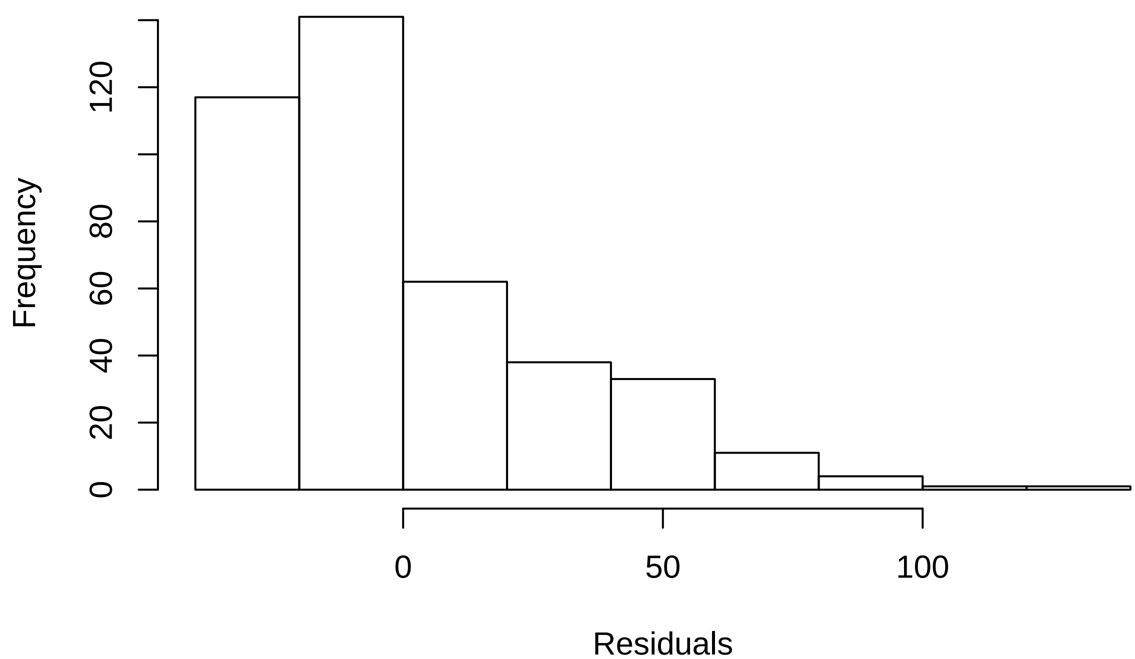
Since the p-value is above .05, we do not have heteroscedasticity.

Question 1.2.c

Since we did not have heteroscedasticity, we do not need to attempt any corrections.

Question 1.3.a

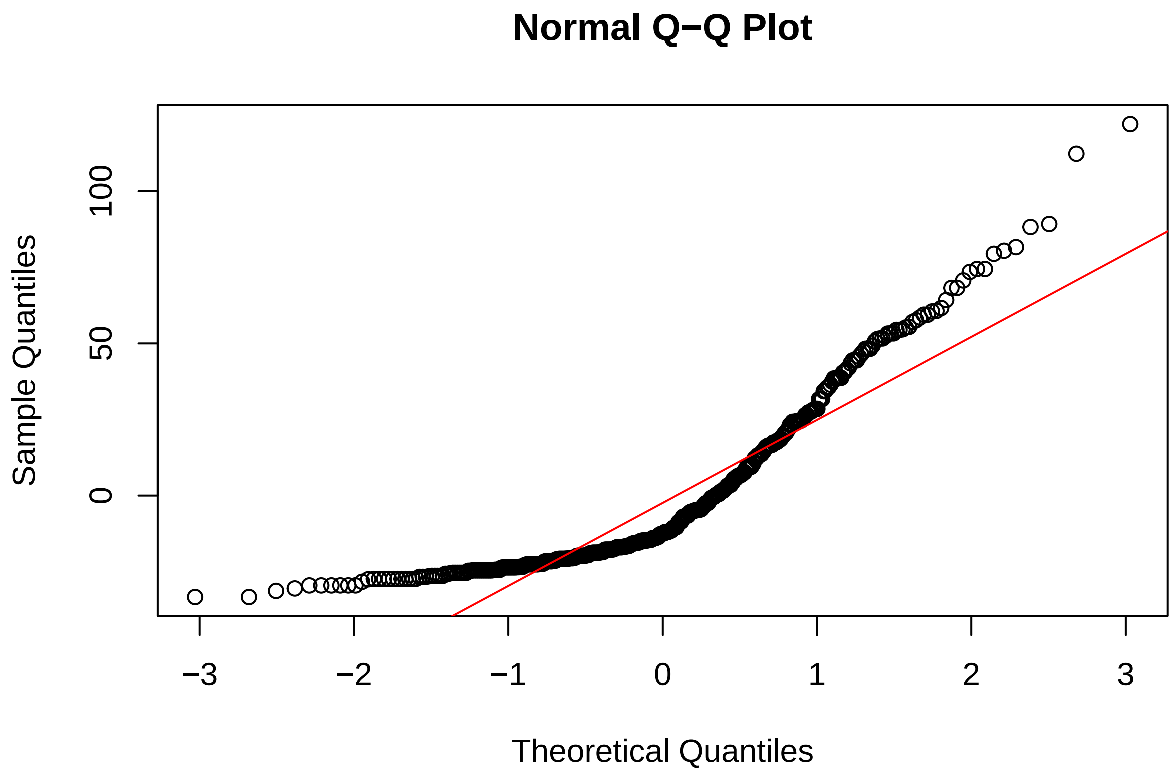
```
x11()
hist(model.1$residuals,xlab="Residuals",main="")
```



Definitely not normally distributed.

Question 1.3.b

```
x11()
qqnorm(model.1$residuals)
qqline(model.1$residuals,col="red")
```



Definitely not normally distributed.

Question 1.3.c

```r
shapiro.test(model.1$residuals)
```

```

	Shapiro-Wilk normality test

data:  model.1$residuals
W = 0.8451, p-value < 2.2e-16
```

Since the p-value is below .05, we violate the normality assumption.

Question 1.3.d

```r
drugs$cigs7b <- drugs$cigs7a + 1
summary(powerTransform(drugs$cigs7b))
```

```
bcPower Transformation to Normality
             Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
drugs$cigs7b    0.0063           0      -0.0827       0.0952

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
```

```
                              LRT df    pval
LR test, lambda = (0) 0.01904637  1 0.89023

Likelihood ratio test that no transformation is needed
                            LRT df       pval
LR test, lambda = (1) 441.0524  1 < 2.22e-16
```

The LR test says that we should transform the outcome variable and the suggested transformation is to raise it to .0063.

```r
model.1a <- lm(I(cigs7a^.0063) ~ free + schyear + sex, data=drugs)
summary(model.1a)
```

```

Call:
lm(formula = I(cigs7a^0.0063) ~ free + schyear + sex, data = drugs)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0179023 -0.0072318  0.0005963  0.0084505  0.0188354 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.0088510  0.0024222 416.494   <2e-16 ***
free        0.0016630  0.0010939   1.520   0.1292    
schyear     0.0013119  0.0005281   2.484   0.0134 *  
sex         0.0008289  0.0009281   0.893   0.3723    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.009328 on 404 degrees of freedom
  (7181 observations deleted due to missingness)
Multiple R-squared:  0.01931,   Adjusted R-squared:  0.01203 
F-statistic: 2.652 on 3 and 404 DF,  p-value: 0.0484
```

As we discussed in the chapter, transforming the outcome variable in a non-intuitive way makes it difficult to interpret the coefficients. Therefore, we may be better off leaving the outcome variable in its original form.

Question 1.4.a

```r
data <- data.frame(drugs$free, drugs$schyear, drugs$sex)
head(data)
```

```
  drugs.free drugs.schyear drugs.sex
1          0             3         0
2          1             1         0
3          1             1         0
4          1             2         0
5          0             1         0
6          0             4         0
```

```r
cor(data, use="pairwise.complete.obs")
```

```
               drugs.free drugs.schyear  drugs.sex
drugs.free     1.00000000   -0.02439321 0.01775831
```

```
drugs.schyear -0.02439321    1.00000000 0.01614925
drugs.sex      0.01775831    0.01614925 1.00000000
```

There are no high correlations.

Question 1.4.b

```r
vif(model.1)
```

```
    free  schyear      sex
1.038136 1.039147 1.001092
```

No multicollinearity.

Question 1.5.a

```r
(2*(3+1))/408
```

```
[1] 0.01960784
```

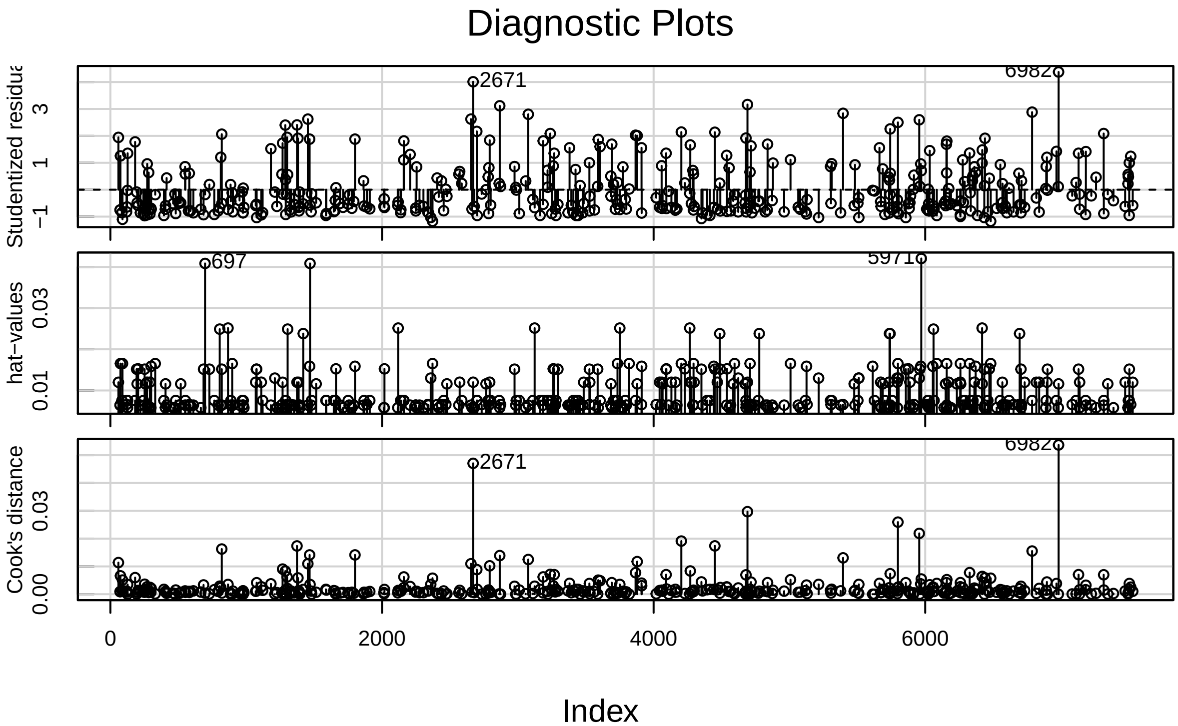
Cut-point for high leverage is .020.

Question 1.5.b

```r
x11()
influenceIndexPlot(model.1,
                   vars=c("Studentized","hat","Cook"),id.n=5)
```

```
Warning in plot.window(...): "id.n" is not a graphical parameter

Warning in plot.xy(xy, type, ...): "id.n" is not a graphical parameter

Warning in axis(side = side, at = at, labels = labels, ...): "id.n" is not
a graphical parameter

Warning in axis(side = side, at = at, labels = labels, ...): "id.n" is not
a graphical parameter

Warning in box(...): "id.n" is not a graphical parameter

Warning in title(...): "id.n" is not a graphical parameter

Warning in plot.xy(xy.coords(x, y), type = type, ...): "id.n" is not a
graphical parameter

Warning in plot.xy(xy.coords(x, y), type = type, ...): "id.n" is not a
graphical parameter

Warning in plot.window(...): "id.n" is not a graphical parameter

Warning in plot.xy(xy, type, ...): "id.n" is not a graphical parameter

Warning in axis(side = side, at = at, labels = labels, ...): "id.n" is not
a graphical parameter

Warning in axis(side = side, at = at, labels = labels, ...): "id.n" is not
```

```
a graphical parameter

Warning in box(...): "id.n" is not a graphical parameter

Warning in title(...): "id.n" is not a graphical parameter

Warning in plot.xy(xy.coords(x, y), type = type, ...): "id.n" is not a
graphical parameter

Warning in plot.xy(xy.coords(x, y), type = type, ...): "id.n" is not a
graphical parameter

Warning in plot.window(...): "id.n" is not a graphical parameter

Warning in plot.xy(xy, type, ...): "id.n" is not a graphical parameter

Warning in axis(side = side, at = at, labels = labels, ...): "id.n" is not
a graphical parameter

Warning in axis(side = side, at = at, labels = labels, ...): "id.n" is not
a graphical parameter

Warning in box(...): "id.n" is not a graphical parameter

Warning in title(...): "id.n" is not a graphical parameter

Warning in plot.xy(xy.coords(x, y), type = type, ...): "id.n" is not a
graphical parameter

Warning in plot.xy(xy.coords(x, y), type = type, ...): "id.n" is not a
graphical parameter
```



There are some outliers, points with high leverage, but no influential data points. Therefore, we do not need to make any corrections.

EXERCISE II

Using the 2011 England Health Survey dataset (`2011 England Health.dta`), perform diagnostics on the multiple linear regression model from Exercise II of Chapter 11 . To remind you, the outcome variable `bmival` and the predictor variables `employed`, `cigs`, `alcohol`, and `fruitveg`.

1. **Functional Form Diagnostics:**
   (a) Create a plot to check for a functional form violation.
   (b) Perform a Ramsey RESET test.
   (c) If you violate the functional form assumption, try to find a solution using the techniques covered in this chapter.
2. **Heteroscedasticity Diagnostics:**
   (a) Create a plot to test for heteroscedasticity.
   (b) Perform a Breusch-Pagan test.
   (c) If heteroscedasticity is present re-run the regression using robust standard errors. Are any predictors that were statistically significant now not significant?
3. **Normality Diagnostics:**
   (a) Create a histogram of the residuals to test for non-normality.

(b) Create a Q-Q plot to test for non-normality.

(c) Perform an Anderson-Darling Normality test.

(d) If you find non-normality in the residuals, try to find a solution using the techniques covered in this chapter.

4. **Multicollinearity Diagnostics:**

   (a) Perform a correlation between predictors to assess whether any have high correlations (over .8).

   (b) Perform a Variance Inflation Factor test.

5. **Outliers, Leverage, and Influential Data Points Diagnostics:**

   (a) Calculate the cut-point for assessing data points with high leverage.

   (b) Use the `influenceIndexPlot()` function to test for outliers, leverage, and influential data points.

ANSWERS FOR EXERCISE II

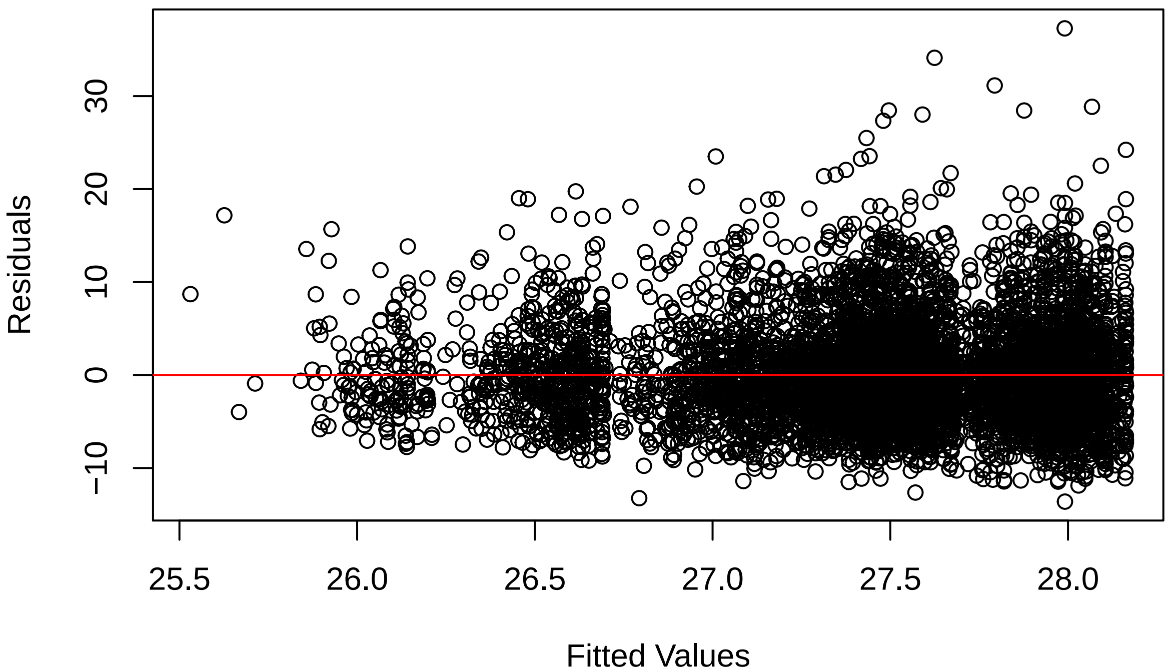
Read-in 2011 England Health Survey and re-run the regression.

```
health <- read.dta("2011 England Health.dta", convert.factors=FALSE)

summary(model.1 <- lm(bmival ~ employed + cigs + alcohol + fruitveg, data=health))
```

```

Call:
lm(formula = bmival ~ employed + cigs + alcohol + fruitveg, data = health)

Residuals:
    Min      1Q  Median      3Q     Max 
-13.611  -3.635  -0.771   2.699  37.286 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 28.653393   0.185322 154.614  < 2e-16 ***
employed    -0.492066   0.129601  -3.797 0.000148 ***
cigs        -0.490416   0.080824  -6.068 1.37e-09 ***
alcohol     -0.001092   0.003174  -0.344 0.730945    
fruitveg    -0.057178   0.025014  -2.286 0.022292 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.32 on 6865 degrees of freedom
  (3747 observations deleted due to missingness)
Multiple R-squared:  0.007778,	Adjusted R-squared:  0.007199 
F-statistic: 13.45 on 4 and 6865 DF,  p-value: 6.353e-11
```

Question 2.1.a

```
x11()
plot(y=model.1$residuals,x=model.1$fitted.values, xlab="Fitted Values", ylab="Residuals")
abline(h=0, col="red")
```



It is not clear whether or not the local means are 0.

Question 2.1.b

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

```
    RESET test

data:  model.1
RESET = 2.0877, df1 = 2, df2 = 6863, p-value = 0.1241
```

The *p*-value is above .05, thus we do not violate function form.

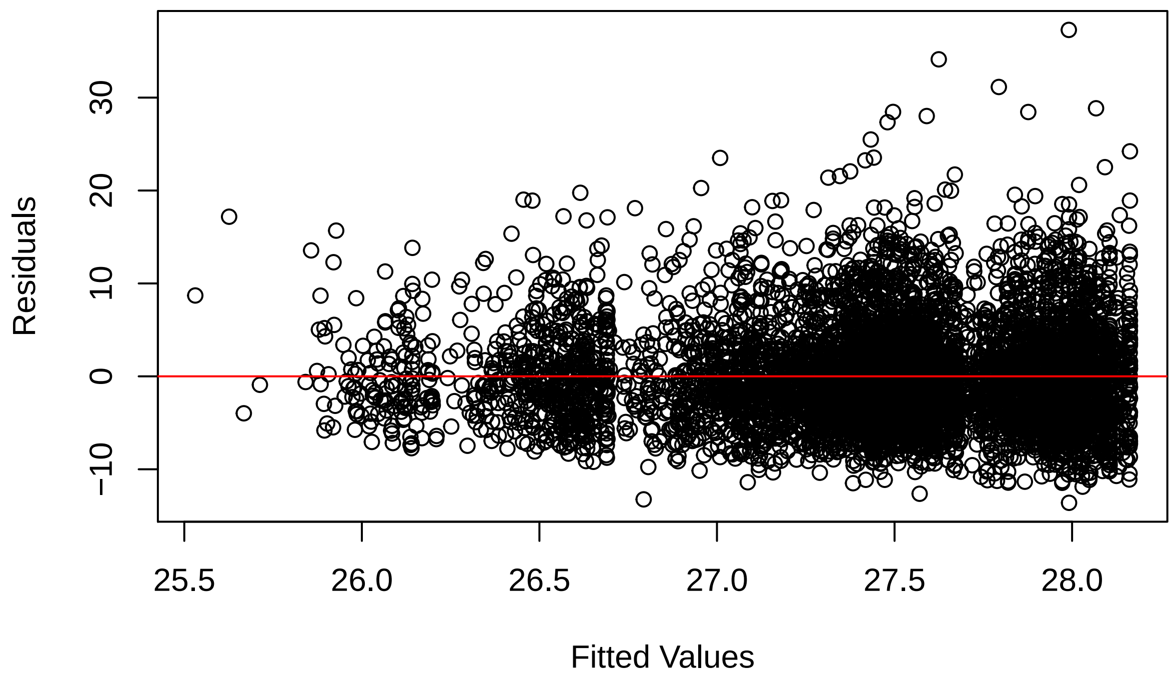
Question 2.1.c

Since we do not violate the functional form assumption, we do not need to attempt any corrections.

Question 2.2.a

This is just a repeat plot from Question 2.1.a.

```
x11()
plot(y=model.1$residuals,x=model.1$fitted.values, xlab="Fitted Values", ylab="Residuals")
abline(h=0, col="red")
```



There is somewhat of a fan pattern, possibly indicating heteroscedasticity.

Question 2.2.b

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

```
    Breusch-Pagan test

data:  model.1
BP = 68.175, df = 4, p-value = 5.509e-14
```

Since the p-value is below .05, we do have heteroscedasticity.

Question 2.2.c

```
library(sandwich)
coeftest(model.1, vcov = vcovHC)
```

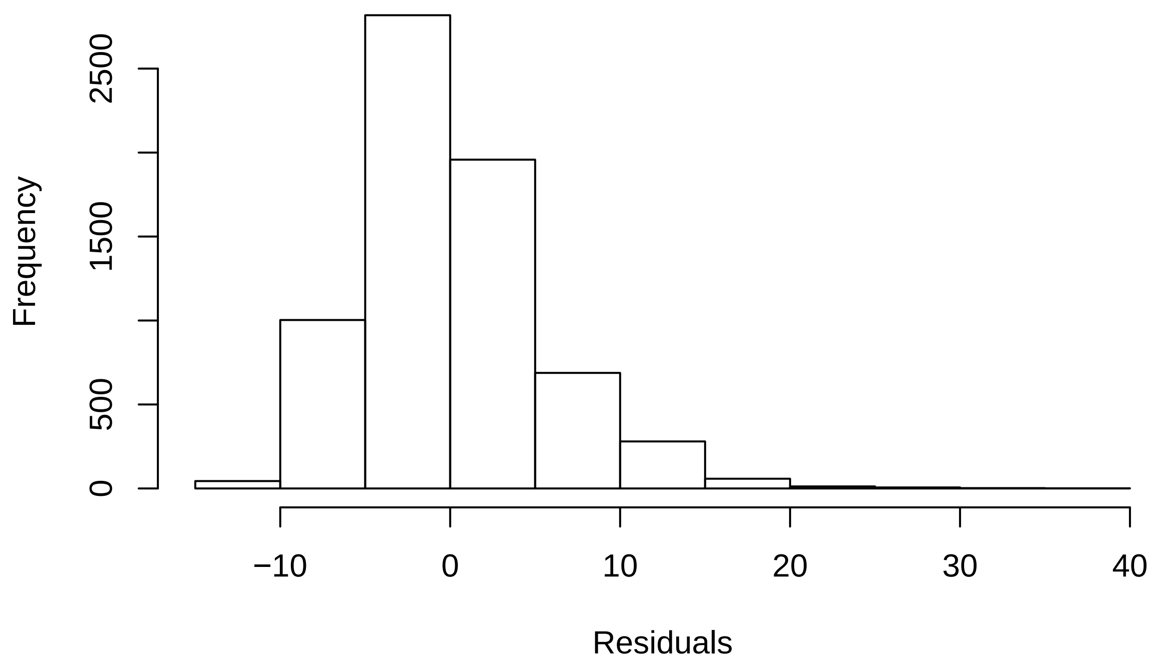
```
t test of coefficients:
```

```
              Estimate Std. Error  t value  Pr(>|t|)
(Intercept) 28.6533935  0.1920234 149.2183 < 2.2e-16 ***
employed    -0.4920658  0.1309879  -3.7566 0.0001737 ***
cigs        -0.4904161  0.0814090  -6.0241 1.788e-09 ***
alcohol     -0.0010916  0.0029950  -0.3645 0.7155083
fruitveg    -0.0571784  0.0243334  -2.3498 0.0188119 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The same predictors are statistically significant when using robust standard errors.

Question 2.3.a

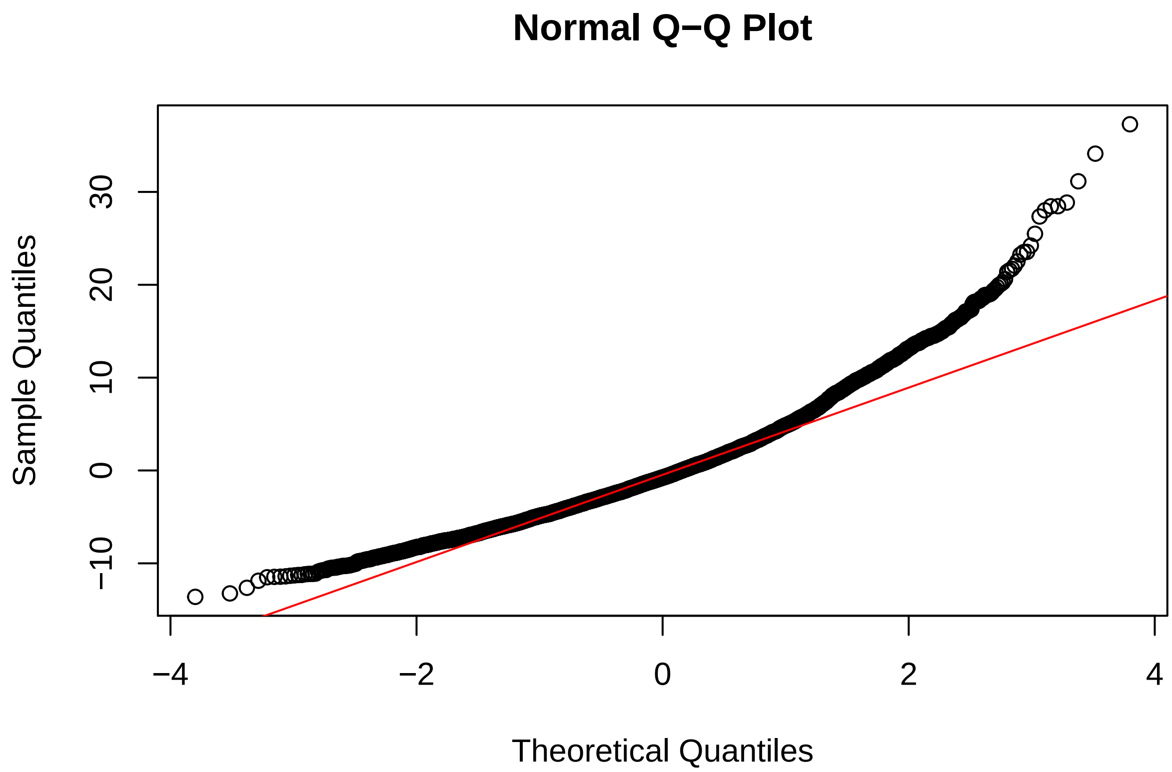
```r
x11()
hist(model.1$residuals,xlab="Residuals",main="")
```



Definitely not normally distributed.

Question 2.3.b

```r
x11()
qqnorm(model.1$residuals)
qqline(model.1$residuals,col="red")
```



Definitely not normally distributed.

Question 2.3.c

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

```

    Anderson-Darling normality test

data:  model.1$residuals
A = 70.542, p-value < 2.2e-16
```

Since the p-value is below .05, we violate the normality assumption.

Question 2.3.d

```
summary(powerTransform(health$bmival))
```

```
bcPower Transformation to Normality
              Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
health$bmival    0.3109        0.33       0.2392       0.3827

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
```

```
                            LRT df       pval
LR test, lambda = (0) 71.79366  1 < 2.22e-16

Likelihood ratio test that no transformation is needed
                            LRT df       pval
LR test, lambda = (1) 357.9705  1 < 2.22e-16
```

The LR test says that we should transform the outcome variable and the suggested transformation is to raise it to .3109.

```r
summary(model.1a <- lm(I(bmival^.3109) ~ employed + cigs + alcohol + fruitveg, data=health))
```

```
Call:
lm(formula = I(bmival^0.3109) ~ employed + cigs + alcohol + fruitveg,
    data = health)

Residuals:
     Min       1Q   Median       3Q      Max
-0.51901 -0.11034 -0.01412  0.09357  0.86057

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.826e+00  5.663e-03 499.108  < 2e-16 ***
employed    -1.380e-02  3.960e-03  -3.484 0.000496 ***
cigs        -1.602e-02  2.470e-03  -6.487 9.37e-11 ***
alcohol      9.825e-06  9.700e-05   0.101 0.919322
fruitveg    -1.610e-03  7.644e-04  -2.106 0.035261 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1626 on 6865 degrees of freedom
  (3747 observations deleted due to missingness)
Multiple R-squared:  0.007999,  Adjusted R-squared:  0.007421
F-statistic: 13.84 on 4 and 6865 DF,  p-value: 3.033e-11
```

As we discussed in the chapter, transforming the outcome variable in a non-intuitive way makes it difficult to interpret the coefficients. Therefore, we may be better off leaving the outcome variable in its original form.

Question 2.4.a

```r
data <- data.frame(health$employed, health$cigs, health$alcohol, health$fruitveg)
head(data)
```

```
  health.employed health.cigs health.alcohol health.fruitveg
1               0           1          0.058        4.000000
2               1           1          4.991        6.500000
3               0           1         49.029        1.000000
4               0           1          0.000        2.000000
5               1           1         30.230       10.333333
6               1           1         13.558        5.333333
```

```r
cor(data, use="pairwise.complete.obs")
```

```
                health.employed  health.cigs health.alcohol
```

```
health.employed     1.000000000  0.003196519     0.06582223
health.cigs         0.003196519  1.000000000     0.14631419
health.alcohol      0.065822235  0.146314190     1.00000000
health.fruitveg     0.037735103 -0.204998901    -0.06840007
                health.fruitveg
health.employed      0.03773510
health.cigs         -0.20499890
health.alcohol      -0.06840007
health.fruitveg      1.00000000
```

There are no high correlations.

Question 2.4.b

```
vif(model.1)
```

```
employed     cigs  alcohol fruitveg
1.004022 1.065813 1.026130 1.048789
```

No multicollinearity.

Question 2.5.a

```
(2*(4+1))/6870
```

```
[1] 0.001455604
```

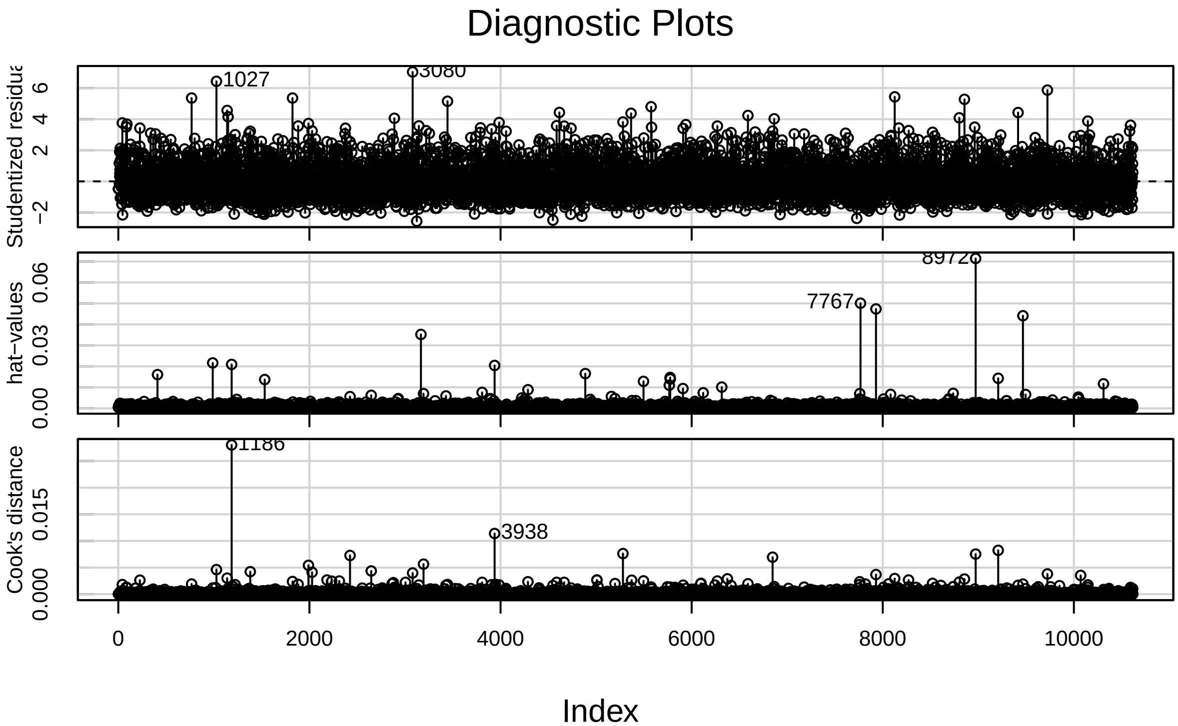
Cut-point for high leverage is .0015.

Question 2.5.b

```
x11()
influenceIndexPlot(model.1,
                   vars=c("Studentized","hat","Cook"),id.n=5)
```

```
Warning in plot.window(...): "id.n" is not a graphical parameter

Warning in plot.xy(xy, type, ...): "id.n" is not a graphical parameter

Warning in axis(side = side, at = at, labels = labels, ...): "id.n" is not
a graphical parameter

Warning in axis(side = side, at = at, labels = labels, ...): "id.n" is not
a graphical parameter

Warning in box(...): "id.n" is not a graphical parameter

Warning in title(...): "id.n" is not a graphical parameter

Warning in plot.xy(xy.coords(x, y), type = type, ...): "id.n" is not a
graphical parameter

Warning in plot.xy(xy.coords(x, y), type = type, ...): "id.n" is not a
graphical parameter
```

```
Warning in plot.window(...): "id.n" is not a graphical parameter

Warning in plot.xy(xy, type, ...): "id.n" is not a graphical parameter

Warning in axis(side = side, at = at, labels = labels, ...): "id.n" is not
a graphical parameter

Warning in axis(side = side, at = at, labels = labels, ...): "id.n" is not
a graphical parameter

Warning in box(...): "id.n" is not a graphical parameter

Warning in title(...): "id.n" is not a graphical parameter

Warning in plot.xy(xy.coords(x, y), type = type, ...): "id.n" is not a
graphical parameter

Warning in plot.xy(xy.coords(x, y), type = type, ...): "id.n" is not a
graphical parameter

Warning in plot.window(...): "id.n" is not a graphical parameter

Warning in plot.xy(xy, type, ...): "id.n" is not a graphical parameter

Warning in axis(side = side, at = at, labels = labels, ...): "id.n" is not
a graphical parameter

Warning in axis(side = side, at = at, labels = labels, ...): "id.n" is not
a graphical parameter

Warning in box(...): "id.n" is not a graphical parameter

Warning in title(...): "id.n" is not a graphical parameter

Warning in plot.xy(xy.coords(x, y), type = type, ...): "id.n" is not a
graphical parameter

Warning in plot.xy(xy.coords(x, y), type = type, ...): "id.n" is not a
graphical parameter
```



There are some outliers, a few points with high leverage, but no influential data points. Therefore, we do not need to make any corrections.