HW5 Math158

Joshua Jansen-Montoya

2022-10-06

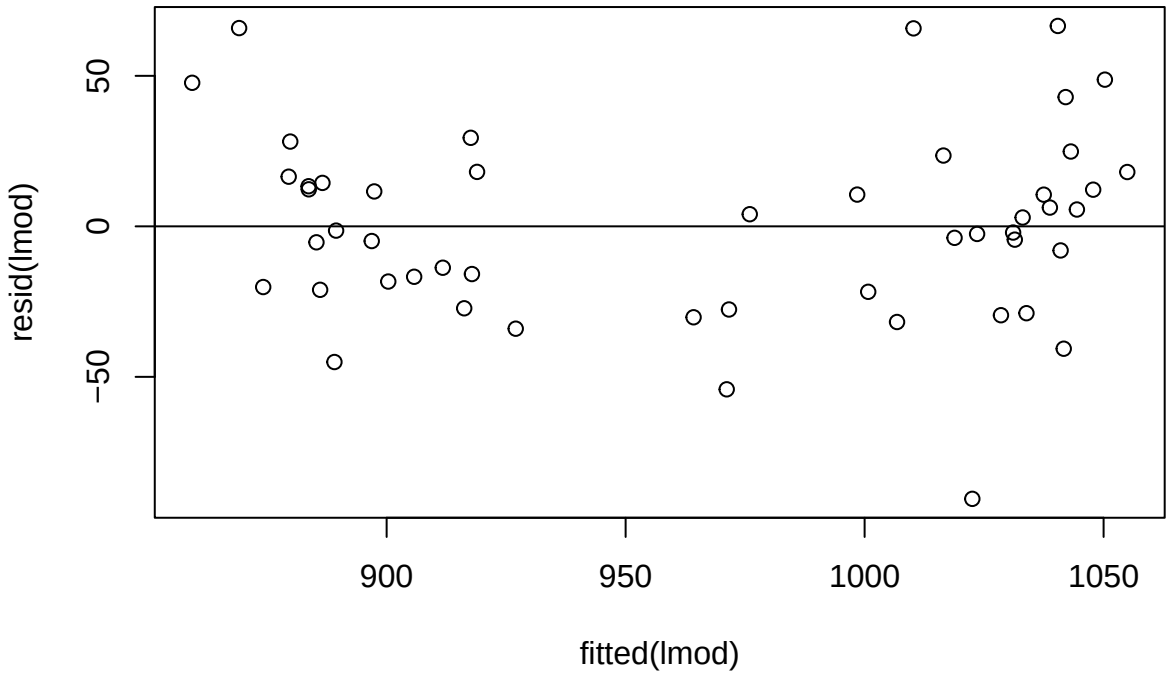
Problem 6.1

Using the sat dataset, fit a model with the total SAT score as the response and expend, salary, ratio and takers as predictors. Perform regression diagnostics on this model to answer the following questions. Display any plots that are relevant. Do not provide any plots about which you have nothing to say. Suggest possible improvements or corrections to the model where appropriate. 1. Check the constant variance assumption for the errors. 2. Check the normality assumption. 3. Check for large leverage points. 4. Check for outliers. 5. Check for influential points. 6. Check the structure of the relationship between the predictors and the response.

Answer 6.1

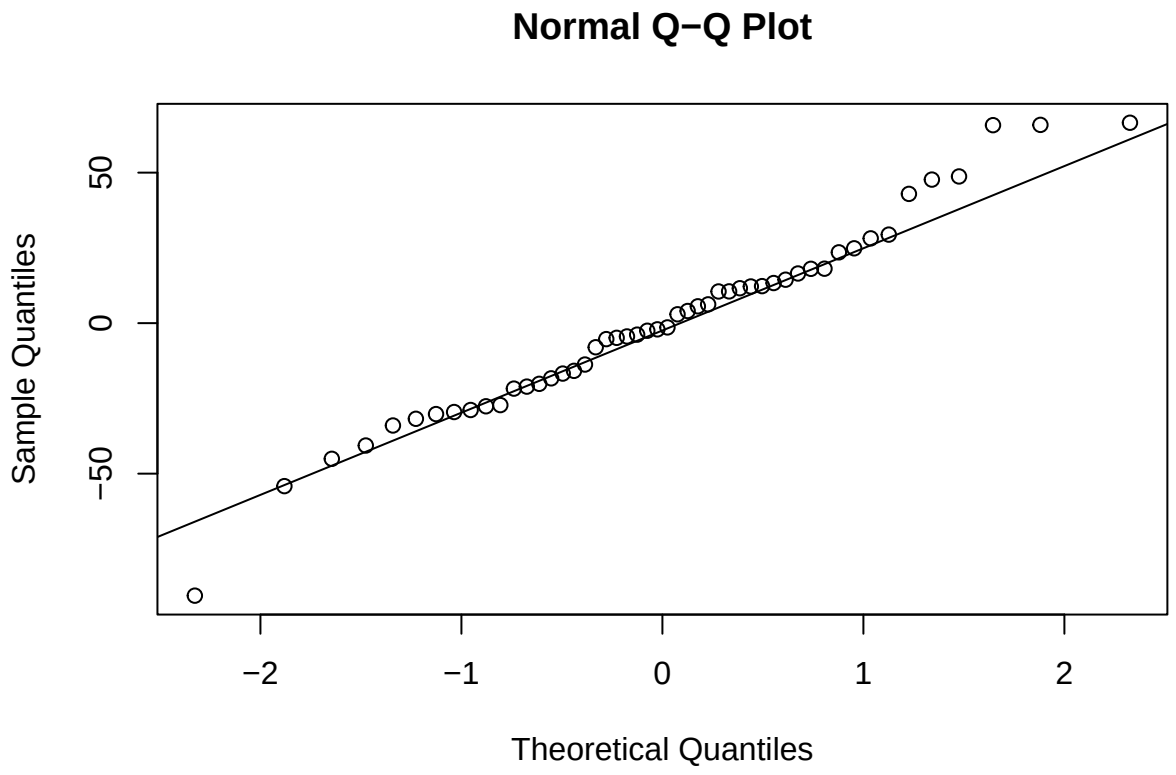
1. We can check the constant variance assumption for the errors with the following R-code,

```
library(faraway)
?sat
lmod <- lm(total ~ expend+salary+ ratio+ takers, data = sat)
plot(fitted(lmod), resid(lmod))
abline(h =0)
```



From which, we can see that there appears to be almost constant variance within our model and thus our assumption should be good. 2. We can check the normality assumption as follows,

```
qqnorm(resid(lmod))
qqline(resid(lmod))
```



Using this plot, we can see that our assumption of normality is indeed fufilled. 3. We can check for large leverage points as follows,

```
hatvalues(lmod) > 2*mean(hatvalues(lmod))
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##          FALSE          FALSE          FALSE          FALSE           TRUE
##       Colorado    Connecticut       Delaware        Florida        Georgia
##          FALSE           TRUE          FALSE          FALSE          FALSE
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##          FALSE          FALSE          FALSE          FALSE          FALSE
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##          FALSE          FALSE          FALSE          FALSE          FALSE
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##          FALSE          FALSE          FALSE          FALSE          FALSE
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##          FALSE          FALSE          FALSE          FALSE           TRUE
##     New Mexico       New York North Carolina   North Dakota           Ohio
##          FALSE          FALSE          FALSE          FALSE          FALSE
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##          FALSE          FALSE          FALSE          FALSE          FALSE
##   South Dakota      Tennessee          Texas           Utah        Vermont
##          FALSE          FALSE          FALSE           TRUE          FALSE
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##          FALSE          FALSE          FALSE          FALSE          FALSE
```

From this result, we can see that we have outliers for California, Connecticut, New Jersey, and Utah. 4. We can check for outliers as follows,

```
  rstandard(lmod)[abs(rstandard(lmod))>2]
```

```
## New Hampshire  North Dakota          Utah West Virginia
##      2.103095      2.123567      2.390264     -2.858505
```

Which we can see that the points corresponding to New Hampshire, North Dakota, Utah, and West Virginia are all high leverage points. 5. Now, we can check for influential points as follows,

```
  cooks.distance(lmod)[29] > 4/length(cooks.distance(lmod))
```

```
## New Hampshire
##         FALSE
```

```
  cooks.distance(lmod)[34] > 4/length(cooks.distance(lmod))
```

```
## North Dakota
##        FALSE
```

```
  cooks.distance(lmod)[44] > 4/length(cooks.distance(lmod))
```

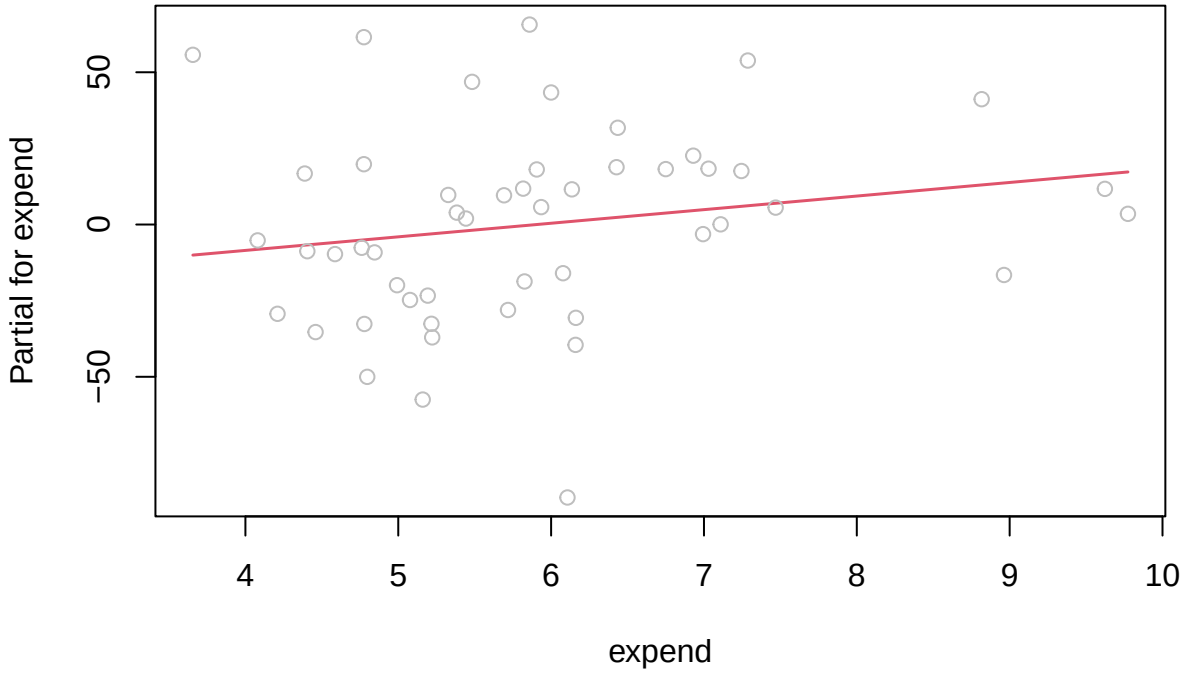
```
## Utah
## TRUE
```

```
  cooks.distance(lmod)[48] > 4/length(cooks.distance(lmod))
```

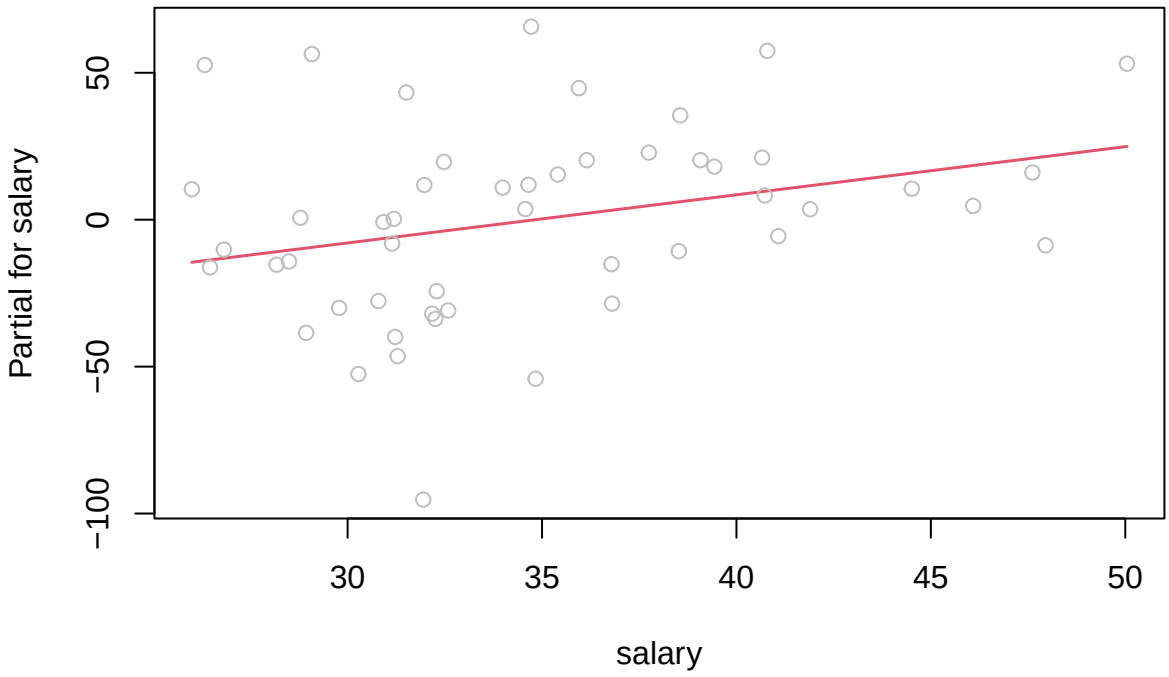
```
## West Virginia
##          TRUE
```

From which we can see that Utah and West Virginia are both influential points. 6. Checking for the structure between the response and predictor variable, we can see that,

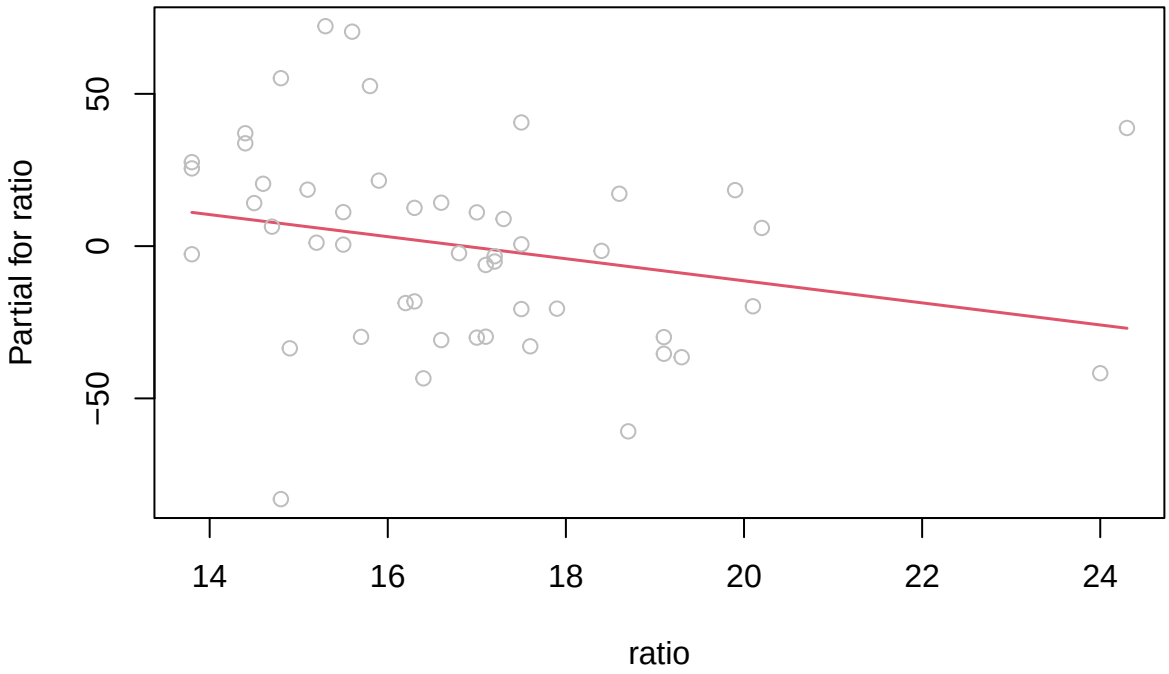
```
  termplot(lmod,partial.resid=TRUE, terms=1)
```



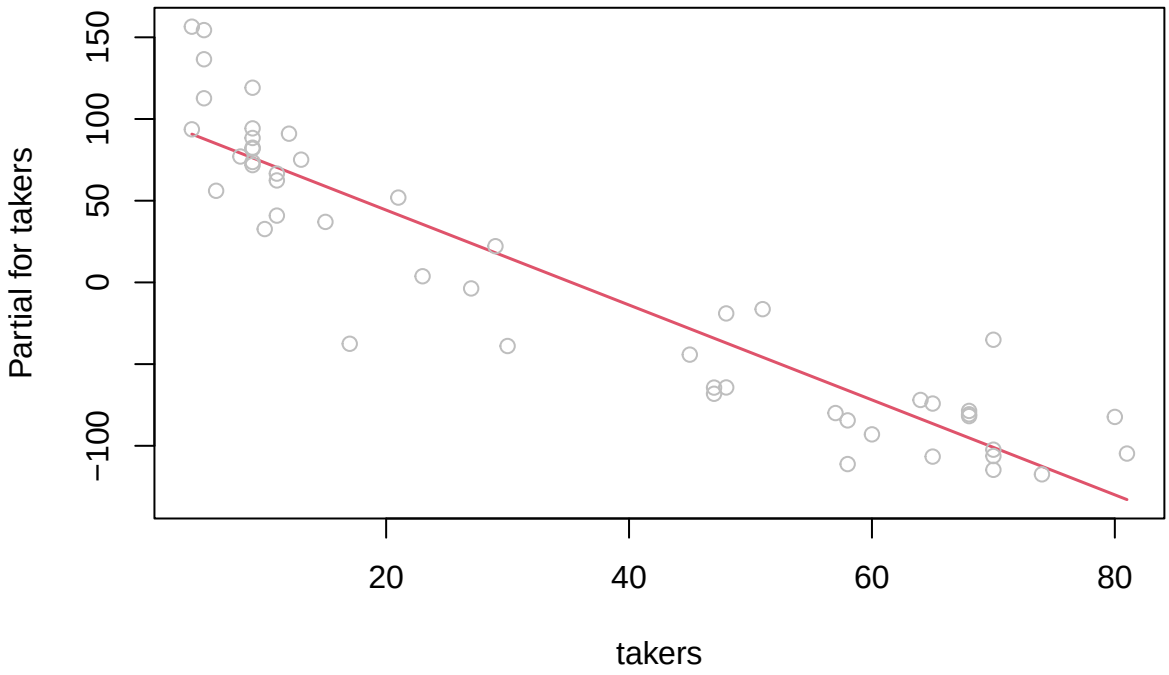

```
  termplot(lmod,partial.resid=TRUE, terms=2)
```




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




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



Looking at the output of our termplot, we can see that there seem to be slight positive linear relationships between expend and salary and their partials and a weak negative linear relationship between ratio and its partials, but a stronger negative relationship between takers and its partials.

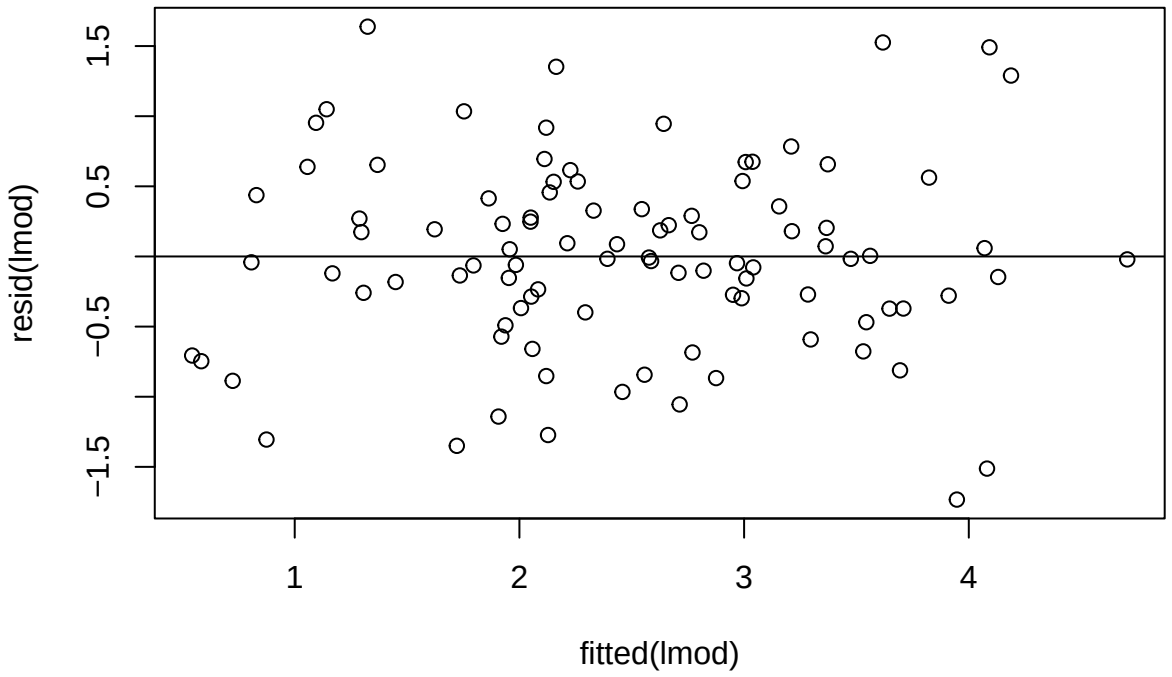
Problem 6.3

For the prostate data, fit a model with lpsa as the response and the other variables as predictors. Answer the questions posed in the first question.

Answer 6.3

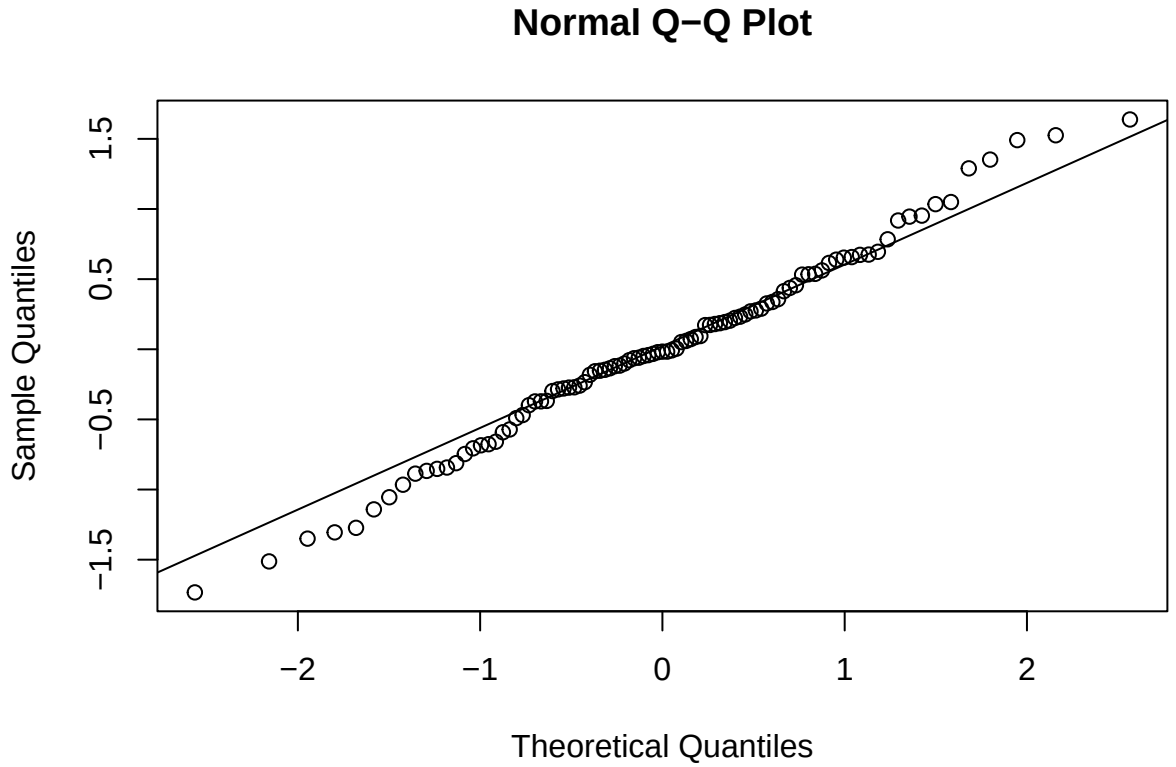
1. We can check the constant variance assumption for the errors with the following R-code,

```
library(faraway)
?prostate
lmod <- lm(lpsa ~ lcavol+lweight+age+lbph+svi+lcp+gleason+pgg45, data = prostate)
plot(fitted(lmod), resid(lmod))
abline(h =0)
```



From which, we can see that there appears to be almost constant variance within our model and thus our assumption should be good. 2. We can check the normality assumption as follows,

```
qqnorm(resid(lmod))
qqline(resid(lmod))
```



Using this plot, we can see that our assumption of normality is indeed fufilled. 3. We can check for large leverage points as follows,

```
  hatvalues(lmod) > 2*mean(hatvalues(lmod))
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12    13
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    14    15    16    17    18    19    20    21    22    23    24    25    26
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    27    28    29    30    31    32    33    34    35    36    37    38    39
## FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
##    40    41    42    43    44    45    46    47    48    49    50    51    52
## FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    53    54    55    56    57    58    59    60    61    62    63    64    65
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    66    67    68    69    70    71    72    73    74    75    76    77    78
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE
##    79    80    81    82    83    84    85    86    87    88    89    90    91
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    92    93    94    95    96    97
##  TRUE FALSE FALSE FALSE FALSE FALSE
```

From this result, we can see that we have outliers for entries 32, 37, 41, 74, and 92. 4. We can check for outliers as follows,

```
  rstandard(lmod)[abs(rstandard(lmod))>2]
```

```
##        39        47        69        95        97
## -2.534124 -2.316280  2.477016  2.323964  2.239719
```

Which we can see that the points corresponding to entries 39, 47, 69, 95, and 97. 5. Now, we can check for influential points as follows,

```
  cooks.distance(lmod)[39] > 4/length(cooks.distance(lmod))
```

```
##   39
## TRUE
```

```
  cooks.distance(lmod)[47] > 4/length(cooks.distance(lmod))
```

```
##   47
## TRUE
```

```
  cooks.distance(lmod)[69] > 4/length(cooks.distance(lmod))
```

```
##   69
## TRUE
```

```
  cooks.distance(lmod)[95] > 4/length(cooks.distance(lmod))
```

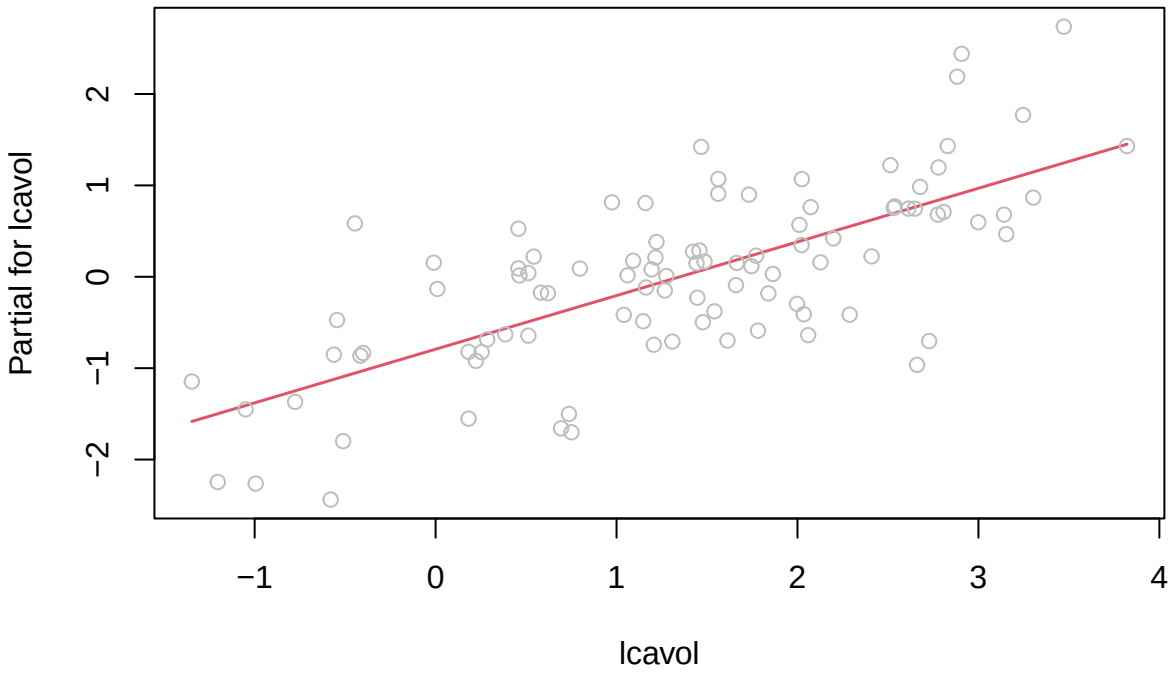
```
##   95
## TRUE
```

```
  cooks.distance(lmod)[97] > 4/length(cooks.distance(lmod))
```

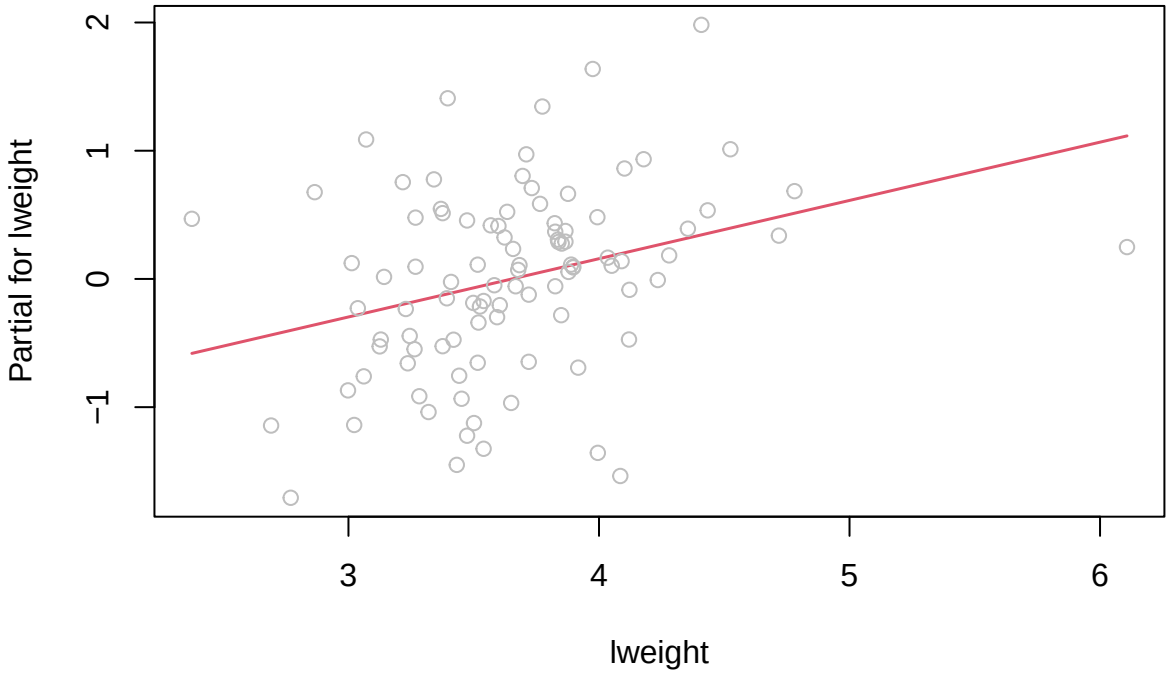
```
##   97
## TRUE
```

From which we can see that all of these points are influential points. 6. Checking for the structure between the response and predictor variable, we can see that,

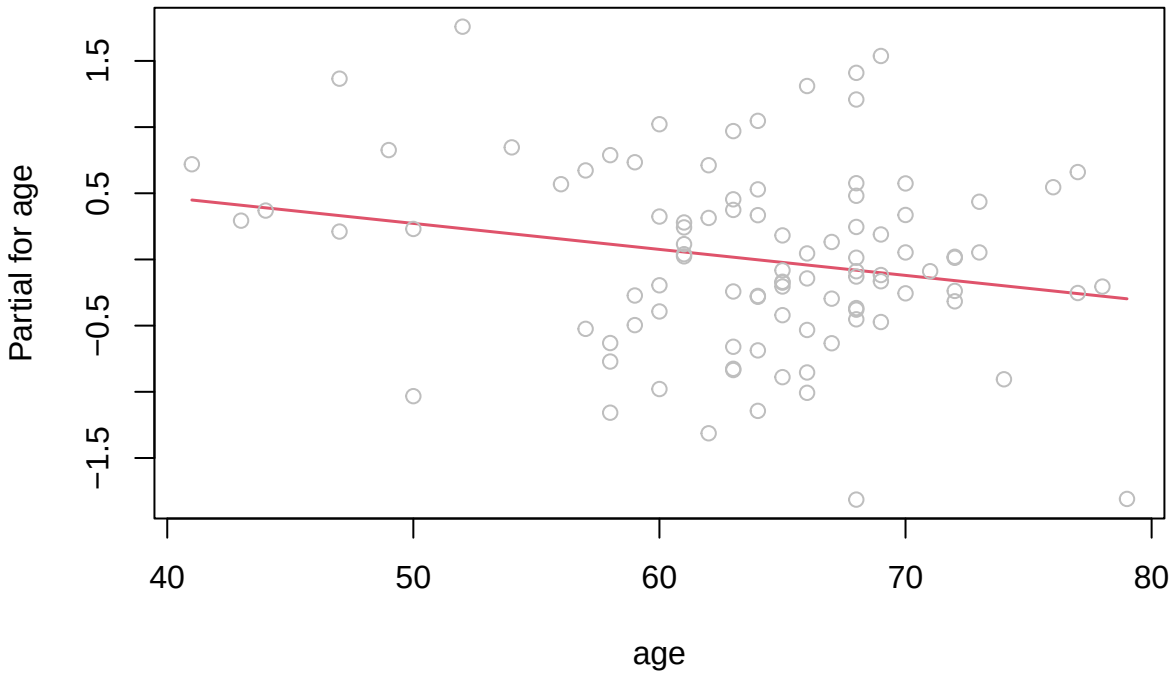
```
  termplot(lmod,partial.resid=TRUE, terms=1)
```



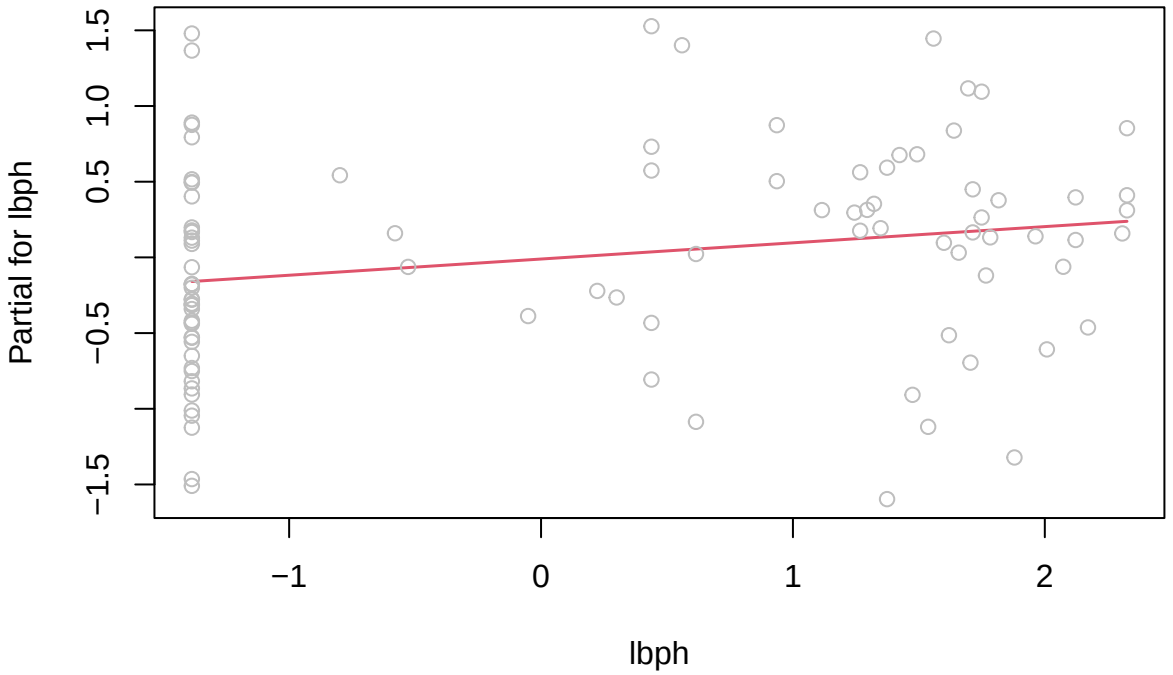

```
termplot(lmod,partial.resid=TRUE, terms=2)
```



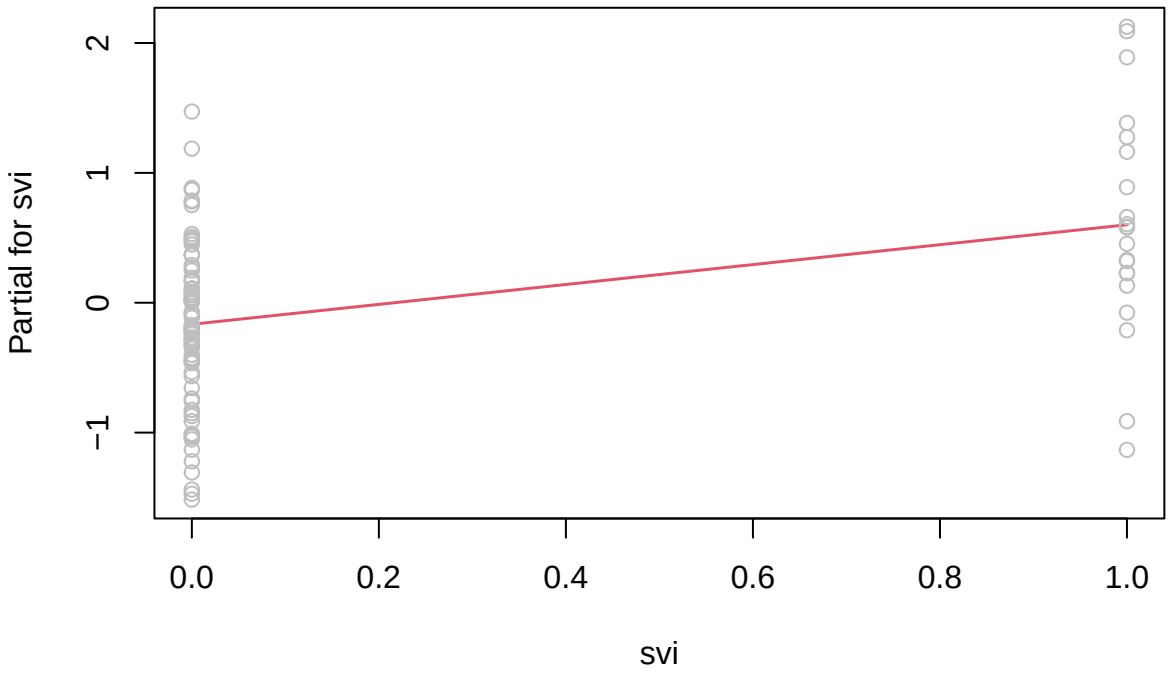

```
termplot(lmod,partial.resid=TRUE, terms=3)
```



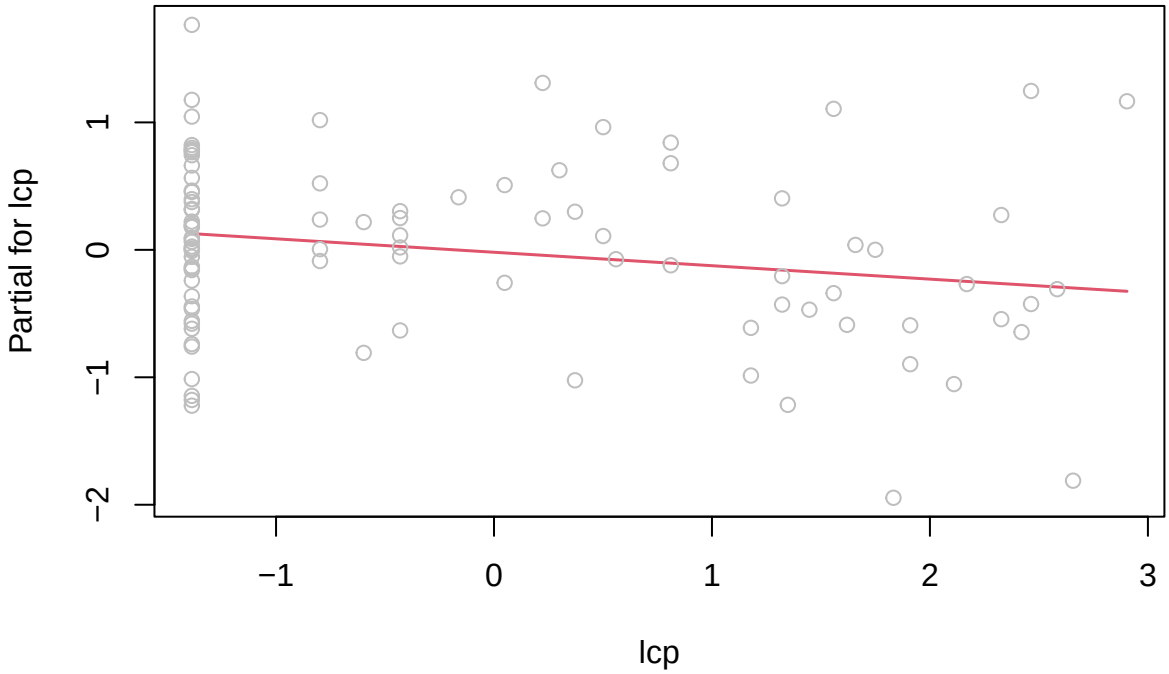

```
termplot(lmod,partial.resid=TRUE, terms=4)
```



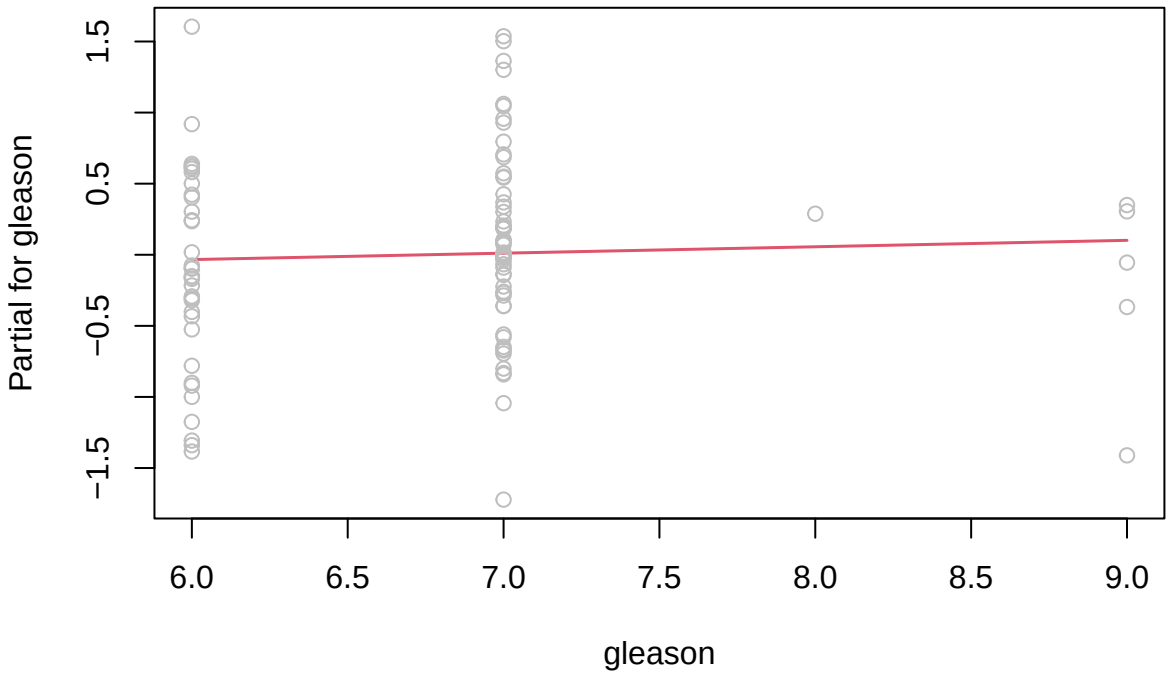

```
termplot(lmod,partial.resid=TRUE, terms=5)
```



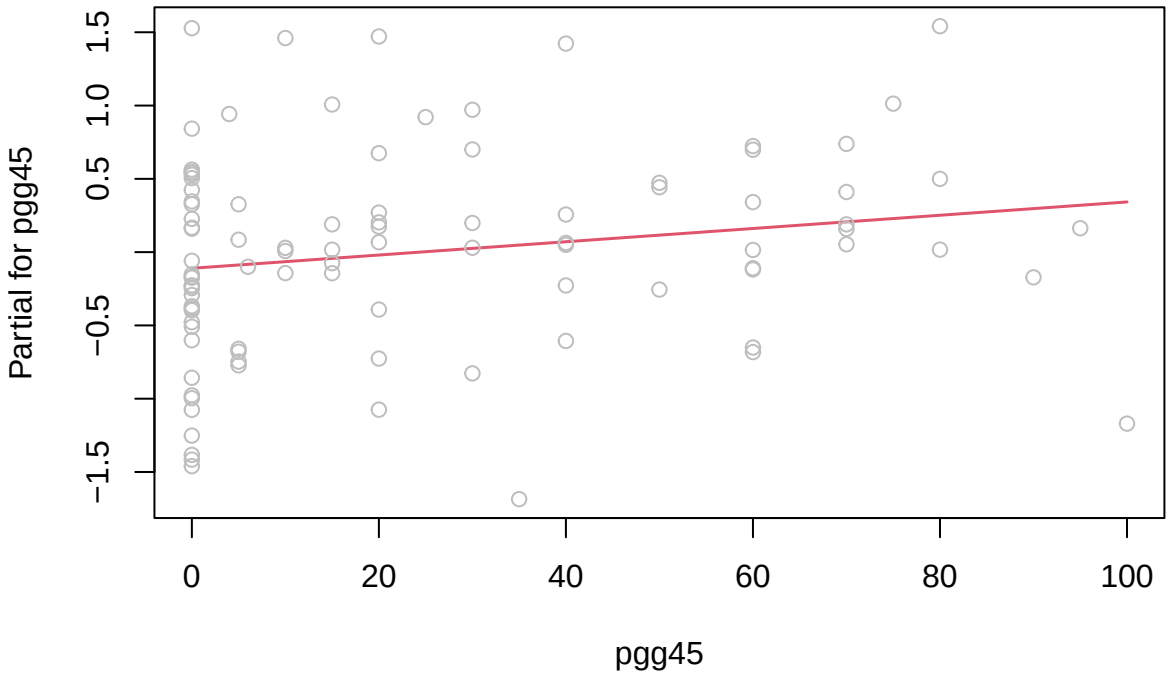

```
termplot(lmod,partial.resid=TRUE, terms=6)
```




```
termplot(lmod,partial.resid=TRUE, terms=7)
```




```
termplot(lmod,partial.resid=TRUE, terms=8)
```



Looking at our termplots, we can see that there seems to be very little linear relationship between lcp, gleason, lbph, svi and pgg45 and their respective partials, as well as no discernable structure between the partials of lweight and age and their respective variables, but that there seems to be a positive linear relationship between lcavol and the partials of lcavol.

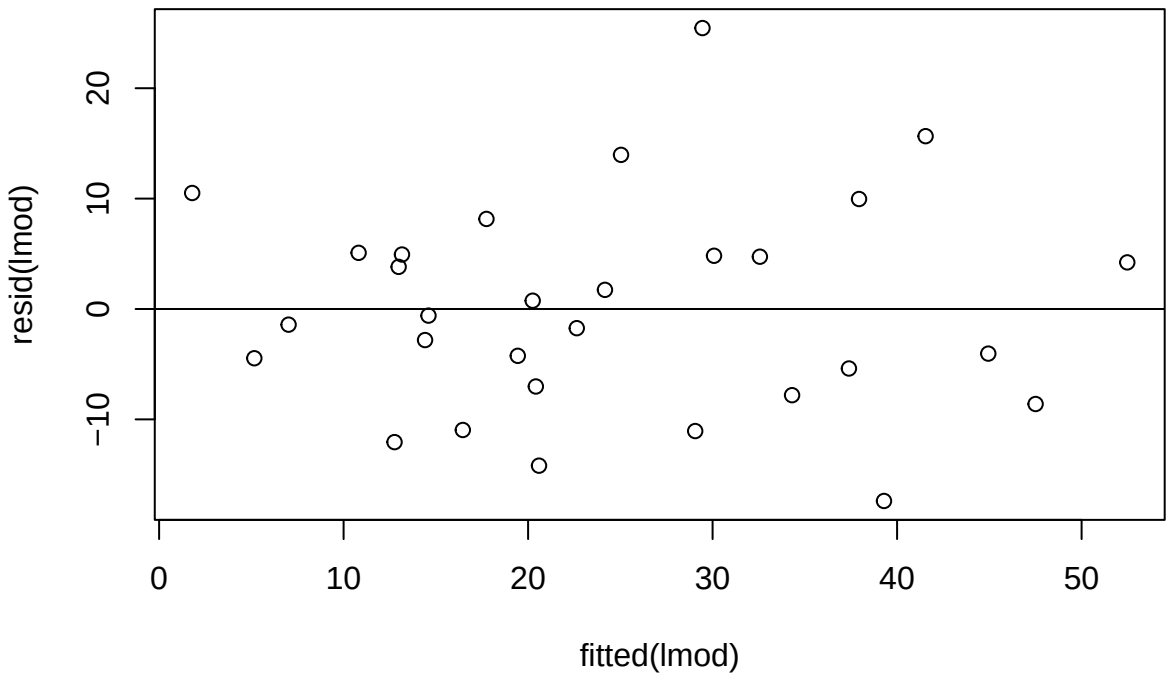
Problem 6.5

Using the cheddar data, fit a model with taste as the response and the other three variables as predictors. Answer the questions posed in the first question.

Answer 6.5

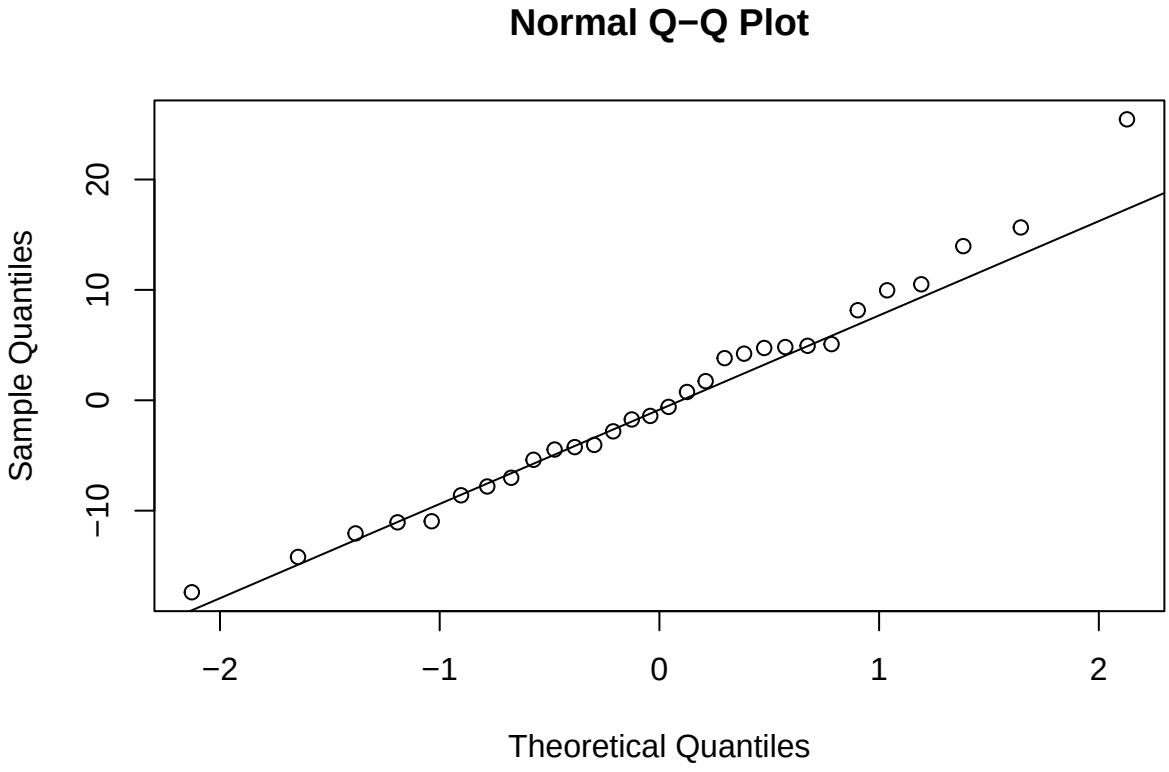
1. We can check the constant variance assumption for the errors with the following R-code,

```
library(faraway)
lmod <- lm(taste ~ Acetic+H2S+Lactic, data = cheddar)
plot(fitted(lmod), resid(lmod))
abline(h =0)
```



From which, we can see that there appears to be almost constant variance within our model and thus our assumption should be good. 2. We can check the normality assumption as follows,

```
qqnorm(resid(lmod))
qqline(resid(lmod))
```



Using this plot, we can see that our assumption of normality is indeed fufilled. 3. We can check for large leverage points as follows,

```
  hatvalues(lmod) > 2*mean(hatvalues(lmod))
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12    13
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    14    15    16    17    18    19    20    21    22    23    24    25    26
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    27    28    29    30
## FALSE FALSE FALSE FALSE
```

From this result, we can see that we do not have any high leverage points. 4. We can check for outliers as follows,

```
  rstandard(lmod)[abs(rstandard(lmod))>2]
```

```
##       15
## 2.633351
```

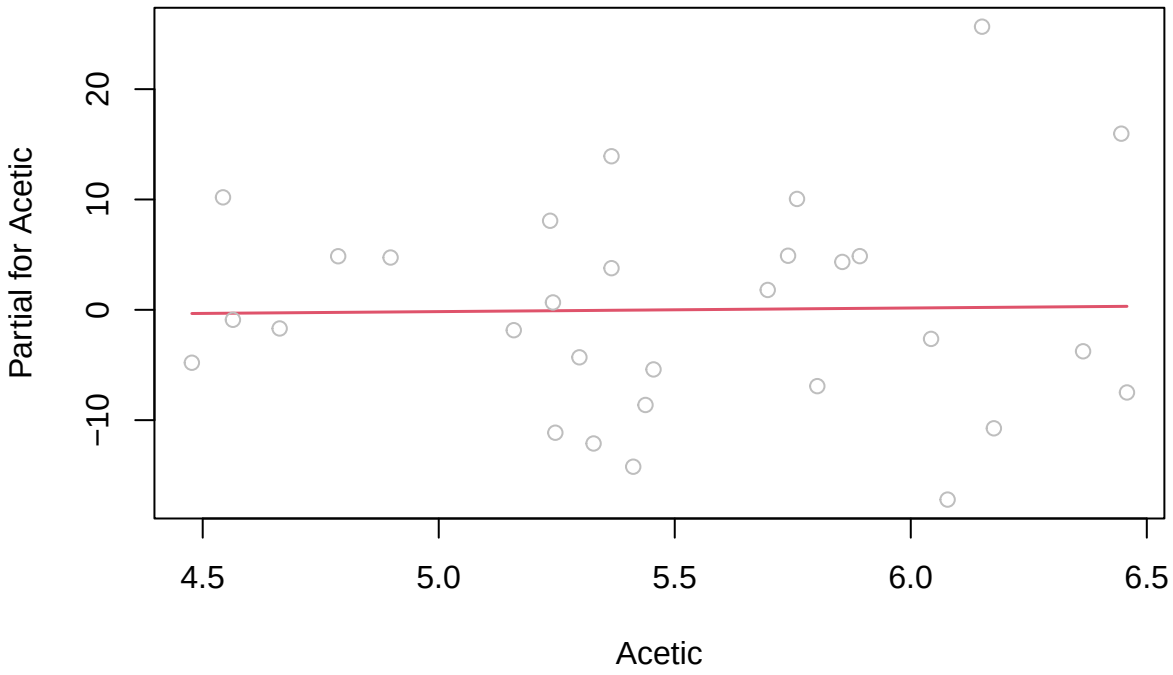
Which we can see that the points corresponding to entry 15 is an outlier. 5. Now, we can check for influential points as follows,

```
  cooks.distance(lmod)[15] > 4/length(cooks.distance(lmod))
```

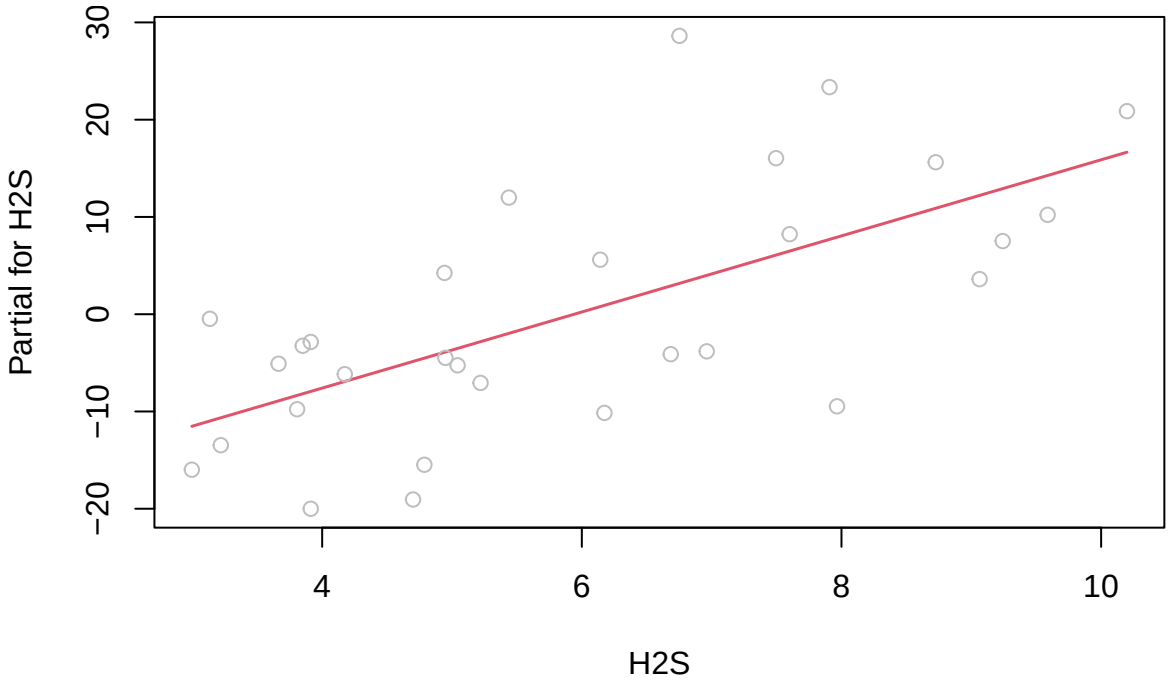
```
##   15
## TRUE
```

From which we can see that entry 15 is an influential points. 6. Checking for the structure between the response and predictor variable, we can see that,

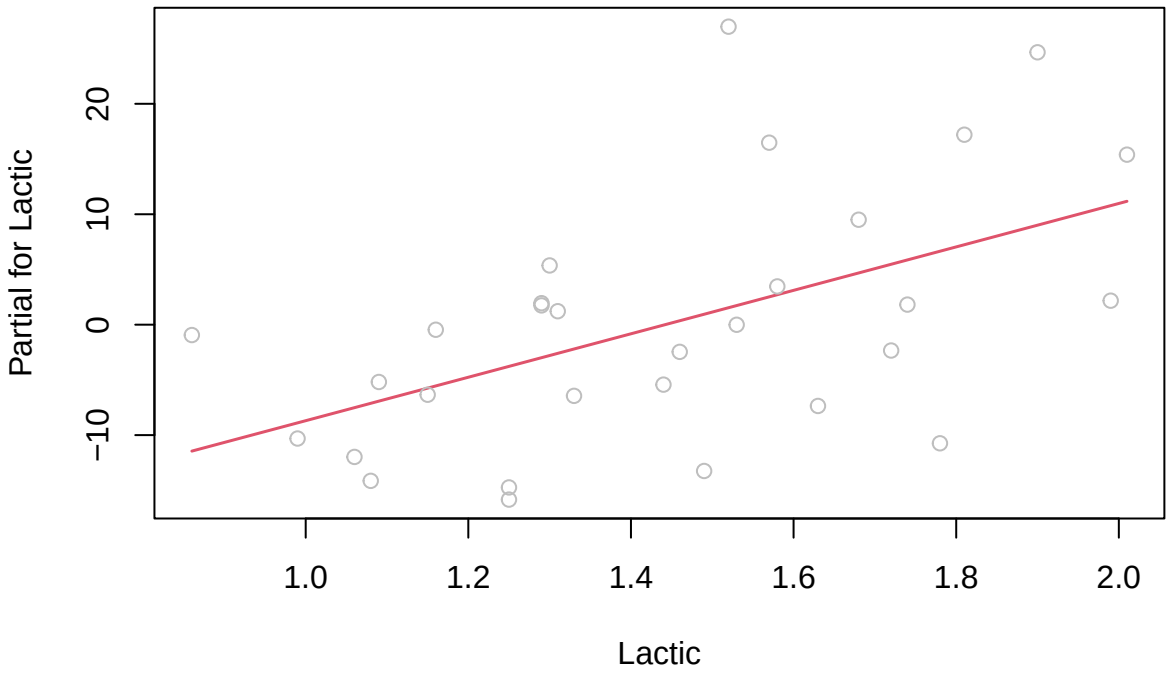
```
  termplot(lmod,partial.resid=TRUE, terms=1)
```




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




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



Looking at our plots, we can see that there does not seem to be any relationship between the partials of Acetic acid but that there does seem to be some positive linear relationship for the partials for H2S and Lactic acid, but nothing indicating any underlying structure.

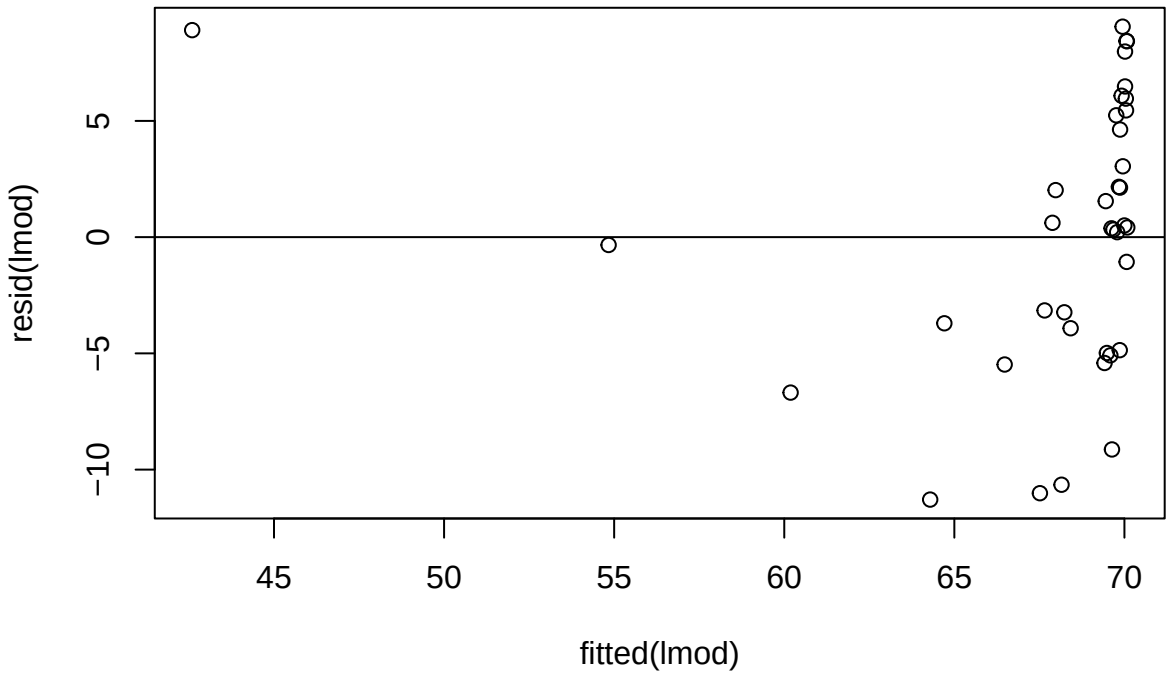
Problem 6.7

Using the tvdoctor data, fit a model with life as the response and the other two variables as predictors. Answer the questions posed in the first question.

Answer 6.7

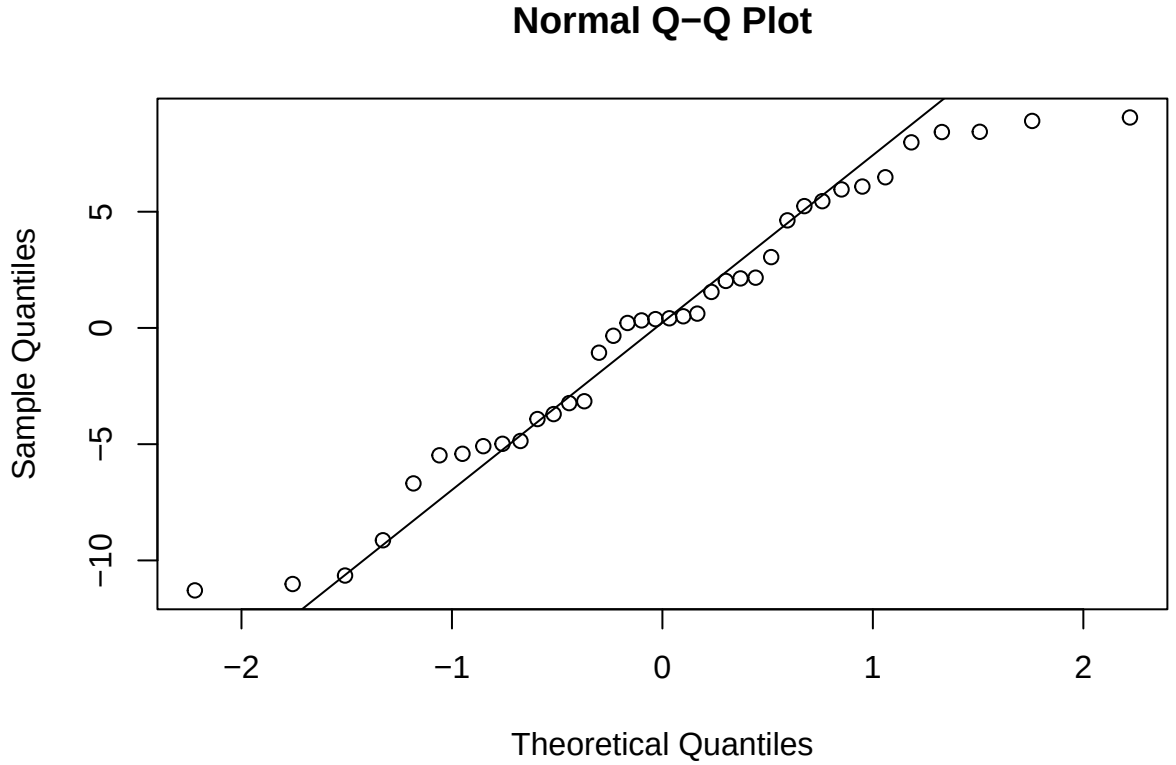
1. We can check the constant variance assumption for the errors with the following R-code,

```
  library(faraway)
?tvdoctor
  lmod <- lm(life ~ tv + doctor, data = tvdoctor)
  plot(fitted(lmod), resid(lmod))
  abline(h =0)
```



From which, we can see that there appears to be almost constant variance within our model and thus our assumption should be good. 2. We can check the normality assumption as follows,

```
qqnorm(resid(lmod))
qqline(resid(lmod))
```



Using this plot, we can see that our distribution is not exactly normal, but that it follows the trends fairly well. 3. We can check for large leverage points as follows,

```
  hatvalues(lmod) > 2*mean(hatvalues(lmod))
```

```
##     Argentina    Bangladesh        Brazil        Canada         China
##         FALSE          TRUE         FALSE         FALSE         FALSE
##      Colombia         Egypt      Ethiopia        France       Germany
##         FALSE         FALSE          TRUE         FALSE         FALSE
##         India     Indonesia          Iran         Italy         Japan
##         FALSE         FALSE         FALSE         FALSE         FALSE
##         Kenya    KoreaNorth    KoreaSouth        Mexico       Morocco
##         FALSE         FALSE         FALSE         FALSE         FALSE
##       Myanmar      Pakistan          Peru   Philippines        Poland
##          TRUE         FALSE         FALSE         FALSE         FALSE
##       Romania        Russia   SouthAfrica         Spain         Sudan
##         FALSE         FALSE         FALSE         FALSE         FALSE
##        Taiwan      Thailand        Turkey       Ukraine UnitedKingdom
##         FALSE         FALSE         FALSE         FALSE         FALSE
##  UnitedStates     Venezuela       Vietnam
##         FALSE         FALSE         FALSE
```

From this result, we can see that our high leverage points are Bangladesh, Ethiopia, and Myanmar. 4. We can check for outliers as follows,

```
  rstandard(lmod)[abs(rstandard(lmod))>2]
```

```
##  Ethiopia     Sudan
##  3.518939 -2.042465
```

Which we can see that the points corresponding to entries Ethiopia and Sudan are both outliers. 5. Now, we can check for influential points as follows,

```
  cooks.distance(lmod)[8] > 4/length(cooks.distance(lmod))
```

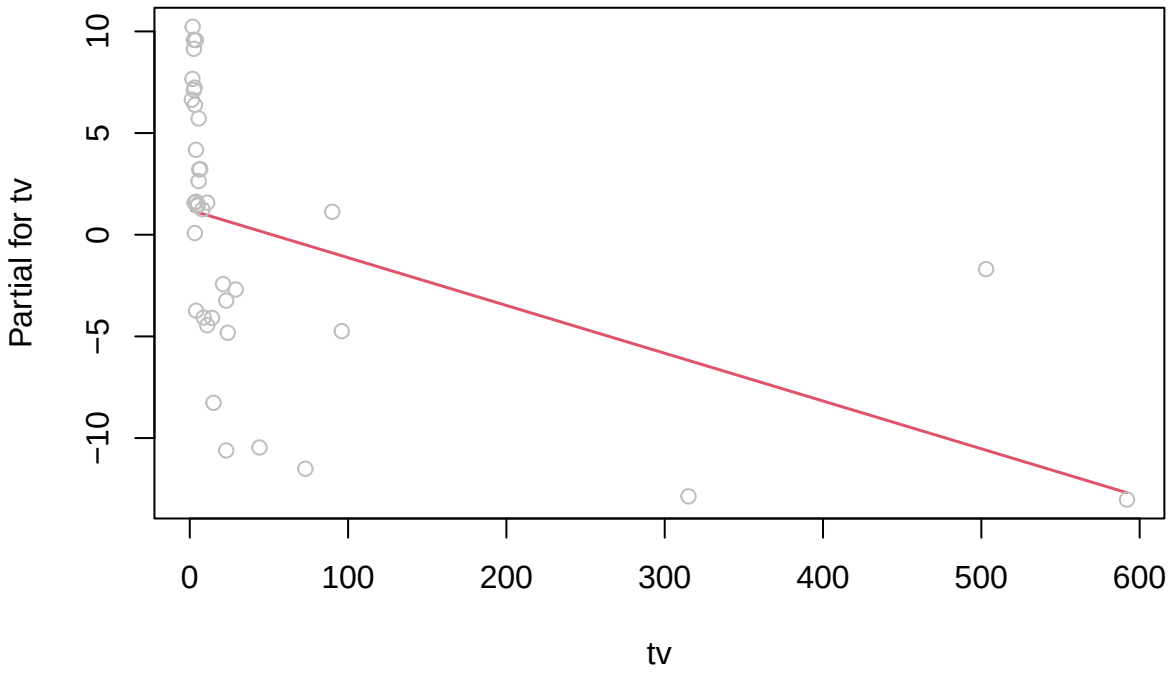
```
## Ethiopia
##     TRUE
```

```
  cooks.distance(lmod)[30] > 4/length(cooks.distance(lmod))
```

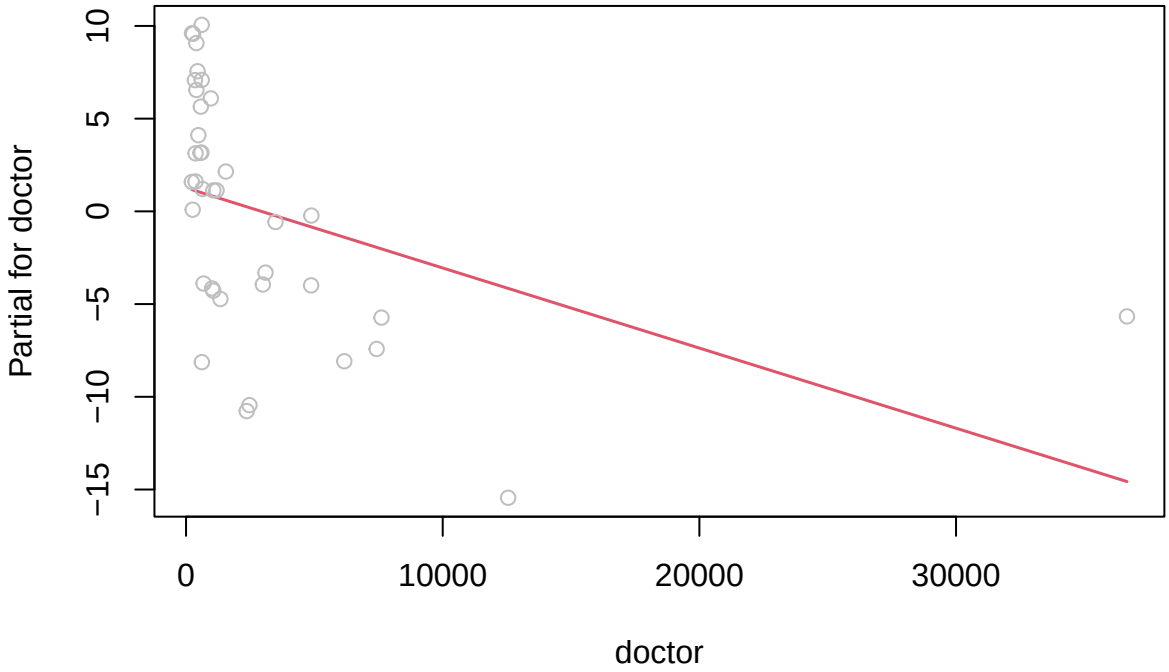
```
## Sudan
##  TRUE
```

From which we can see that both Ethiopia and Sudan are influential points. 6. Checking for the structure between the response and predictor variable, we can see that,

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




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



Looking at our plots, we can see that there does not seem to be any discernable relationship between our two variables and the partials that we obtain.