Week 1 Homework 621

Sam Reeves

1/31/2022

Exercises

1.1

The dataset teengamb concerns a study of teenage gambling in Britain. Make a numerical and graphical summary of the data, commenting on any features that you find interesting. Limit the output you present to a quantity that a busy reader would find sufficient to get a basic understanding of the data.

```r
data(teengamb, package = 'faraway')

(teengamb$sex <- factor(teengamb$sex))
```

```
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [39] 0 0 0 0 0 0 0 0 0
## Levels: 0 1
```

```r
levels(teengamb$sex) <- c('male', 'female')

head(teengamb)
```

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

```r
summary(teengamb)
```

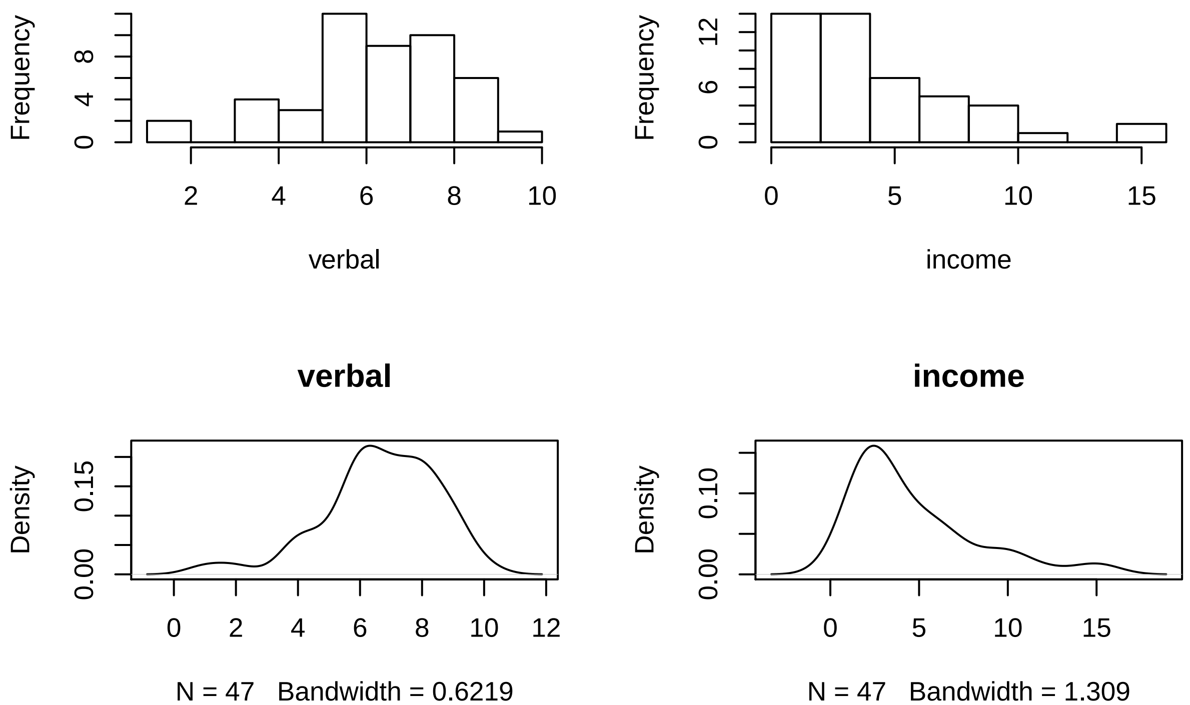
```
##      sex          status          income           verbal          gamble
##  male  :28   Min.   :18.00   Min.   : 0.600   Min.   : 1.00   Min.   :  0.0
##  female:19   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00   1st Qu.:  1.1
##              Median :43.00   Median : 3.250   Median : 7.00   Median :  6.0
##              Mean   :45.23   Mean   : 4.642   Mean   : 6.66   Mean   : 19.3
##              3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00   3rd Qu.: 19.4
##              Max.   :75.00   Max.   :15.000   Max.   :10.00   Max.   :156.0
```

```r
anyNA(teengamb)
```

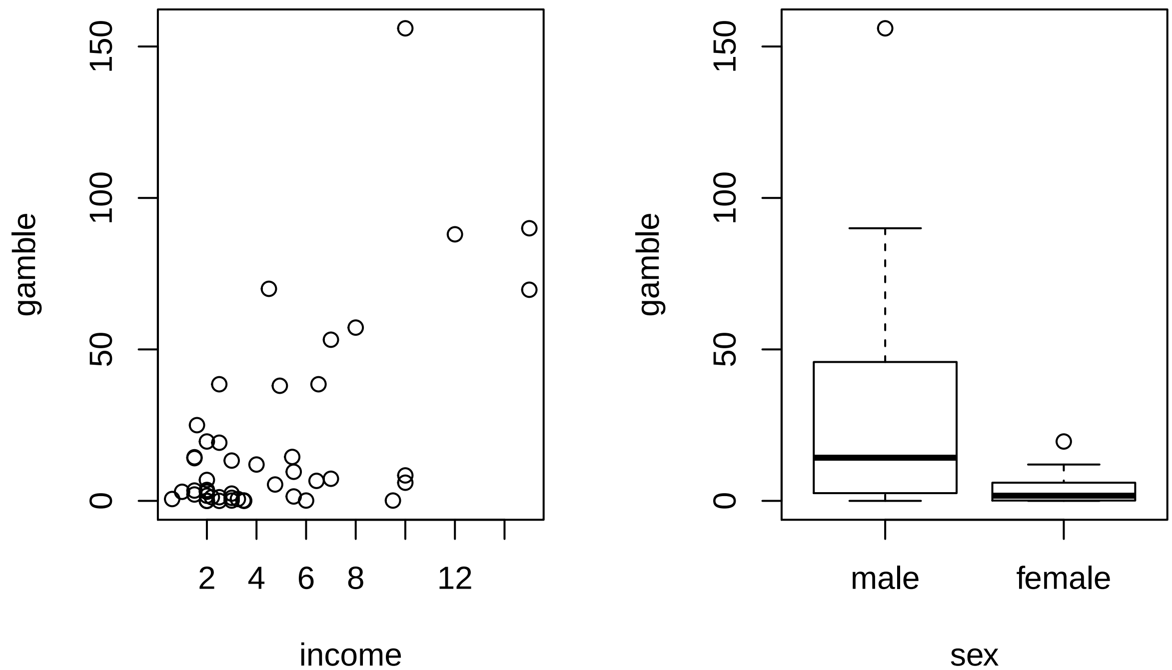
```
## [1] FALSE
```

"The teengamb data frame has 47 rows and 5 columns. A survey was conducted to study teenage gambling in Britain." There appear to be no null values.

```r
par(mfrow = c(2,2))
hist(teengamb$verbal, main = '', xlab = 'verbal')
hist(teengamb$income, main = '', xlab = 'income')
plot(density(teengamb$verbal), main = 'verbal')
plot(density(teengamb$income), main = 'income')
```



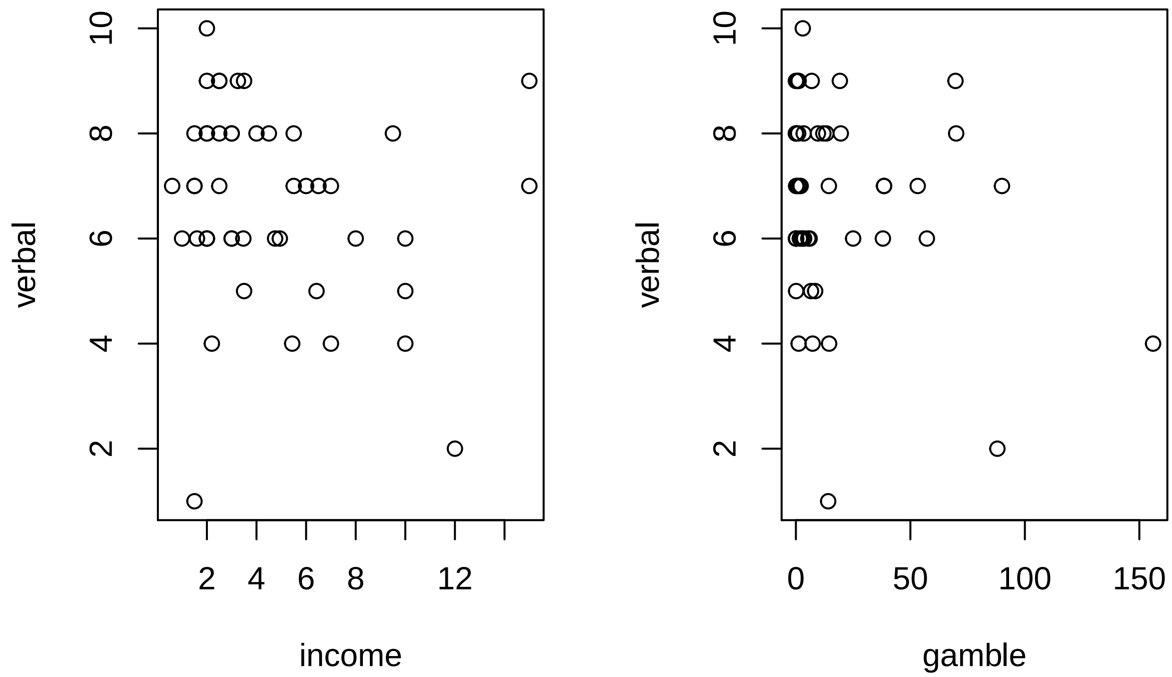

```r
par(mfrow = c(1, 2))
plot(gamble ~ income, teengamb)
plot(gamble ~ sex, teengamb)
```



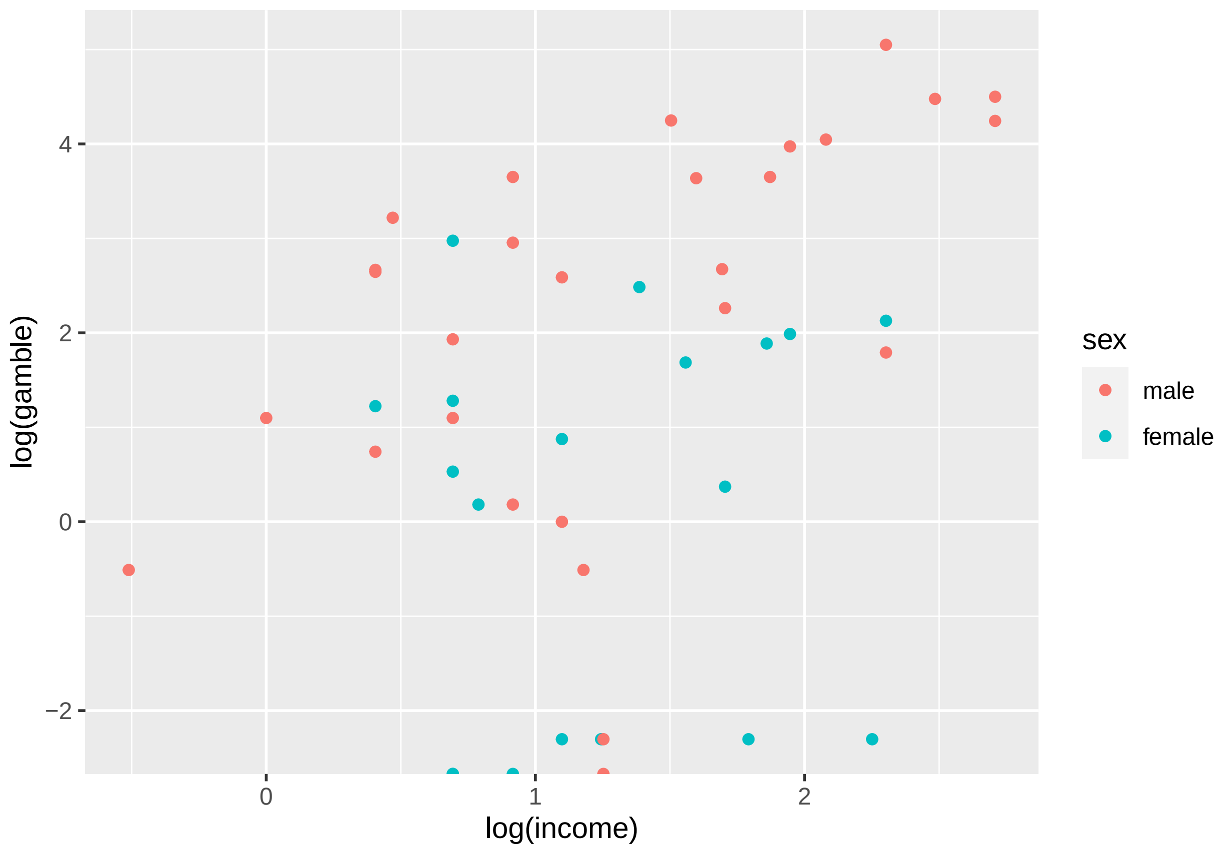
So it seems there are some outliers who make a lot of money and gamble a lot, but it seems that the gamblers are concentrated on the low end of the income spectrum. Men also have a much more prevalent gambling problem than women in the set surveyed.

Out of 2 very rich people represented, one has the biggest gambling habit, and he is male. The median income is fairly low and the spread of verbal scores is centered around 7

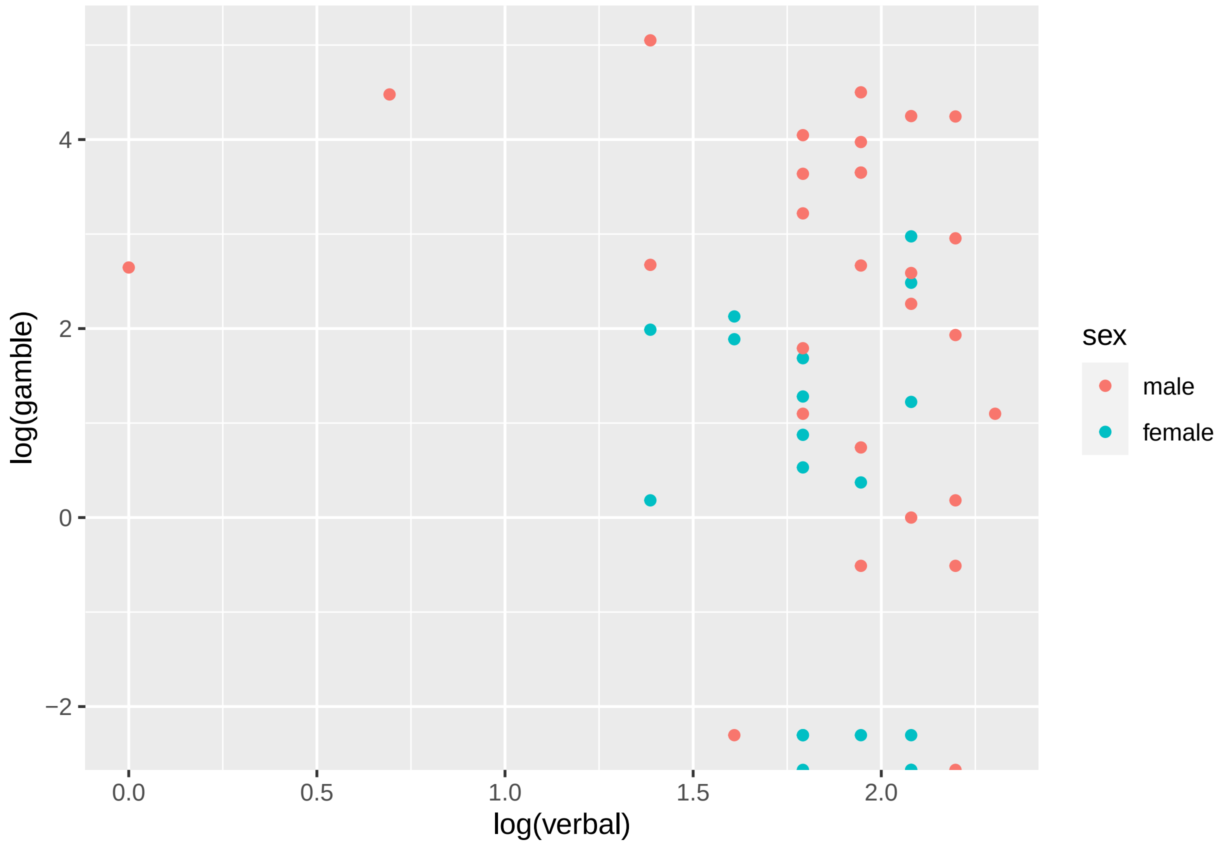
```
par(mfrow = c(1, 2))
plot(verbal ~ income, teengamb)
plot(verbal ~ gamble, teengamb)
```




```
ggplot(teengamb,aes(x=log(income), y = log(gamble), col = sex)) +
  geom_point()
```



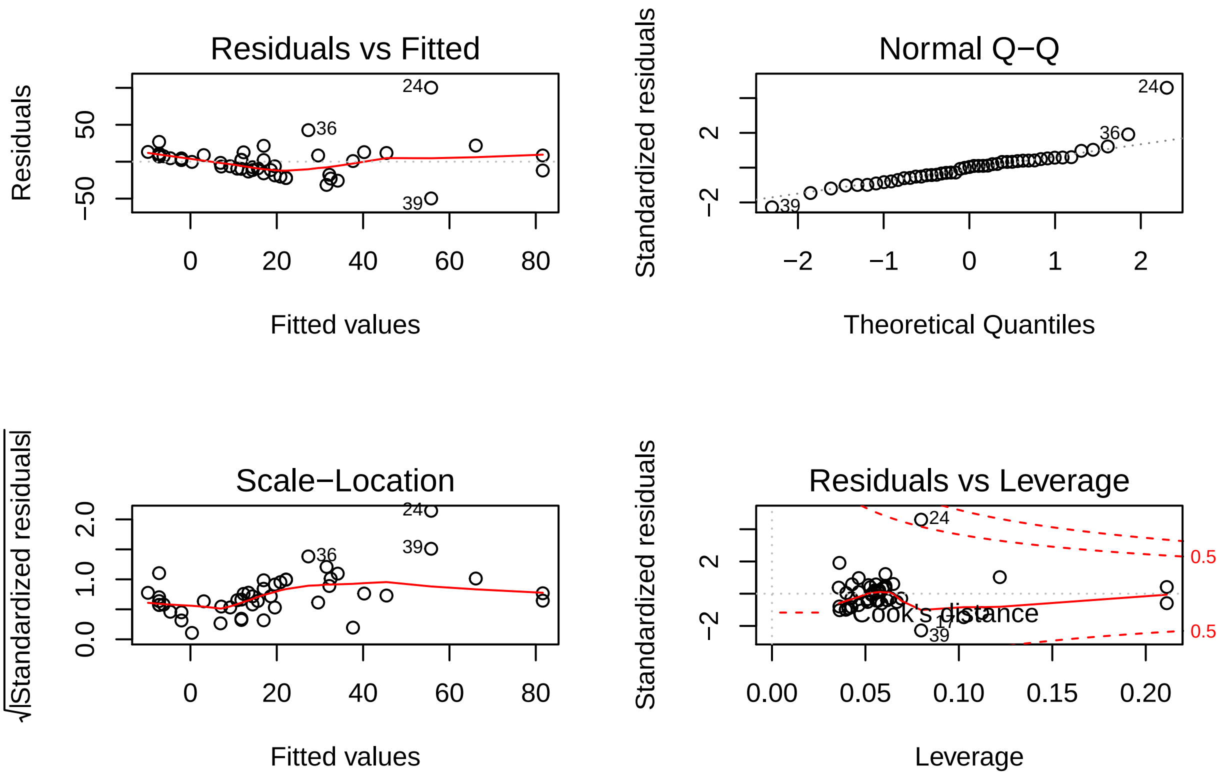

```
ggplot(teengamb,aes(x=log(verbal), y = log(gamble), col = sex)) +
  geom_point()
```



Let's try to fit a model of gambling habits as explained by income and verbal score, adjusted for sex.

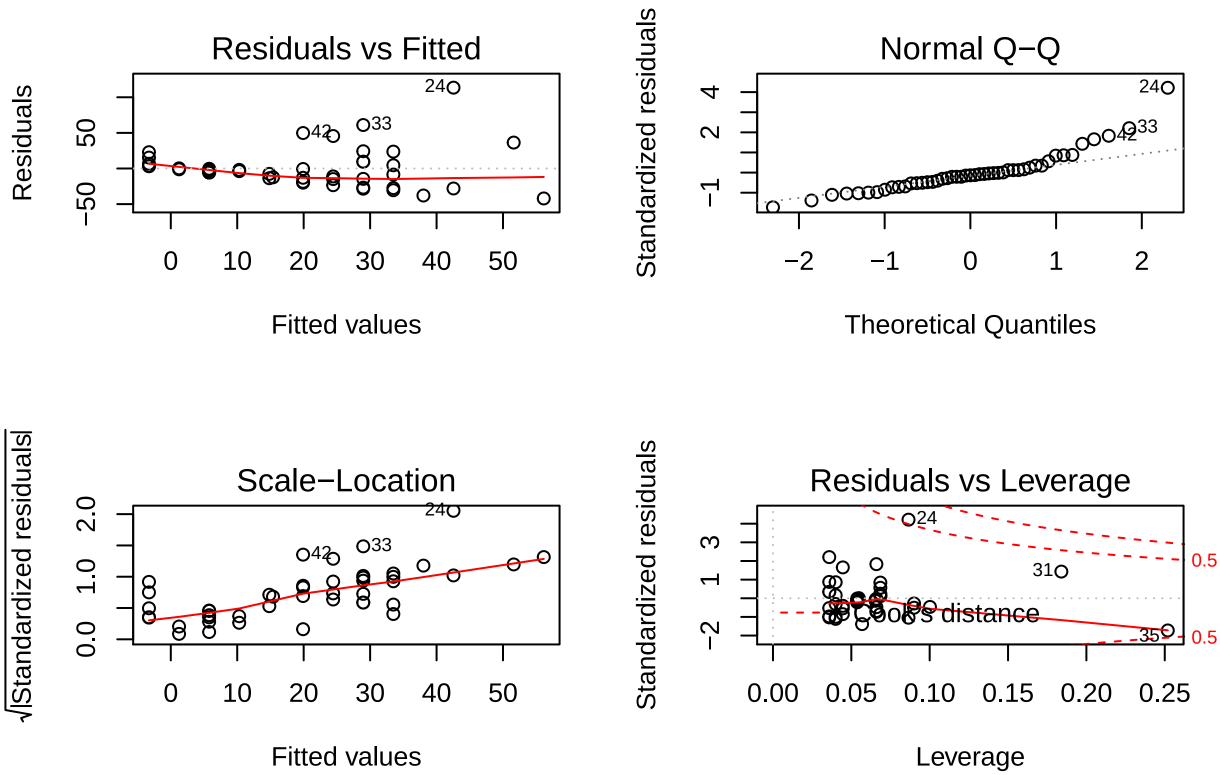
```
lm.inc <- lm(gamble ~ income + sex, teengamb)
lm.verb <- lm(gamble ~ verbal + sex, teengamb)
```

```
par(mfrow = c(2,2))
plot(lm.inc)
```



There is a very strong relationship between income and gambling habits, adjusted for sex.

```
par(mfrow= c(2,2))
plot(lm.verb)
```



Verbal score as a predictor of gambling habits after adjustment for sex, is not as strong an explanatory variable. Still, I would include it in my model.

1.3

The dataset prostate is from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. Make a numerical and graphical summary of the data as in the first question.

```r
data(prostate, package = 'faraway')

head(prostate)
```

```
##       lcavol lweight age      lbph svi      lcp gleason pgg45     lpsa
## 1 -0.5798185  2.7695  50 -1.386294   0 -1.38629       6     0 -0.43078
## 2 -0.9942523  3.3196  58 -1.386294   0 -1.38629       6     0 -0.16252
## 3 -0.5108256  2.6912  74 -1.386294   0 -1.38629       7    20 -0.16252
## 4 -1.2039728  3.2828  58 -1.386294   0 -1.38629       6     0 -0.16252
## 5  0.7514161  3.4324  62 -1.386294   0 -1.38629       6     0  0.37156
## 6 -1.0498221  3.2288  50 -1.386294   0 -1.38629       6     0  0.76547
```

```r
summary(prostate)
```

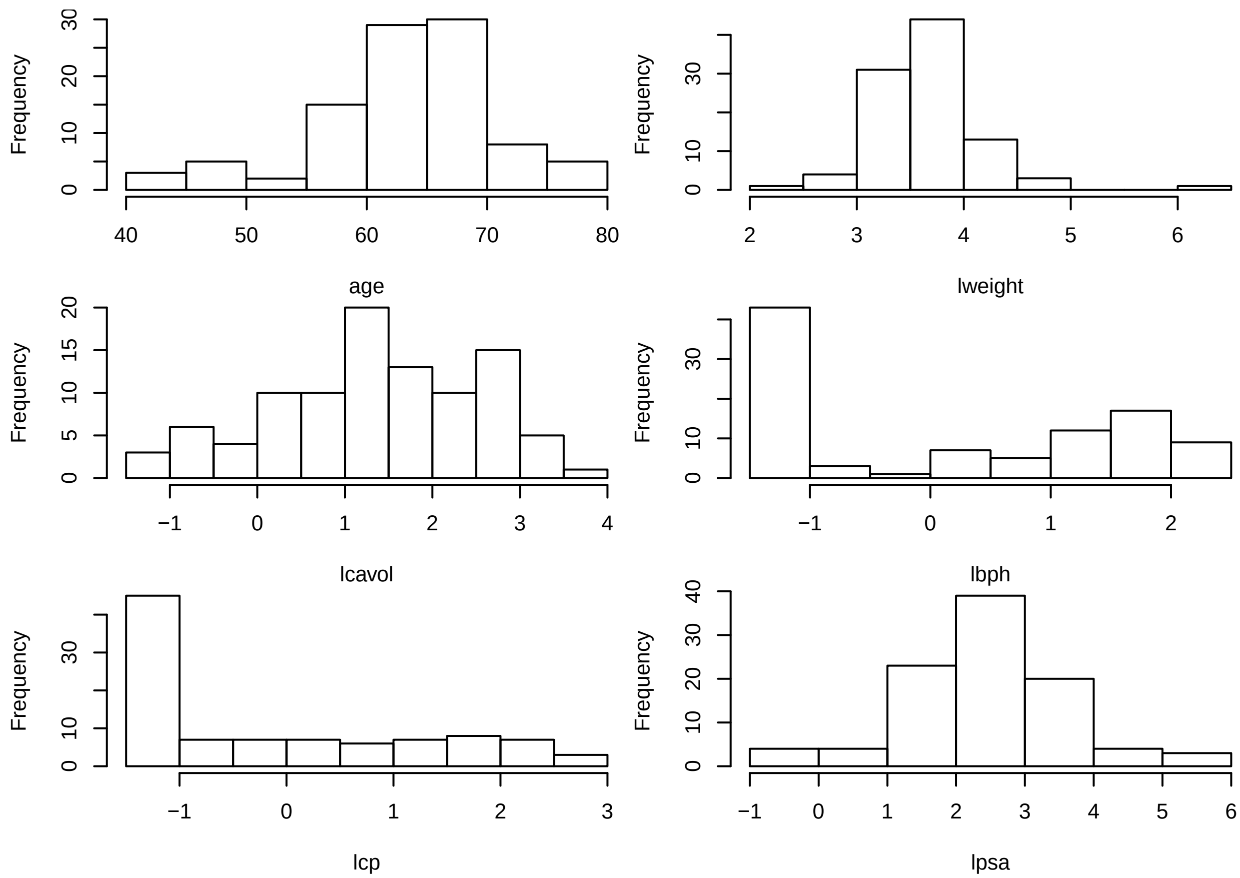
```
##      lcavol           lweight           age             lbph
##  Min.   :-1.3471   Min.   :2.375   Min.   :41.00   Min.   :-1.3863
```

```
##  1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:-1.3863
##  Median : 1.4469   Median :3.623   Median :65.00   Median : 0.3001
##  Mean   : 1.3500   Mean   :3.653   Mean   :63.87   Mean   : 0.1004
##  3rd Qu.: 2.1270   3rd Qu.:3.878   3rd Qu.:68.00   3rd Qu.: 1.5581
##  Max.   : 3.8210   Max.   :6.108   Max.   :79.00   Max.   : 2.3263
##       svi              lcp             gleason          pgg45
##  Min.   :0.0000   Min.   :-1.3863   Min.   :6.000   Min.   :  0.00
##  1st Qu.:0.0000   1st Qu.:-1.3863   1st Qu.:6.000   1st Qu.:  0.00
##  Median :0.0000   Median :-0.7985   Median :7.000   Median : 15.00
##  Mean   :0.2165   Mean   :-0.1794   Mean   :6.753   Mean   : 24.38
##  3rd Qu.:0.0000   3rd Qu.: 1.1786   3rd Qu.:7.000   3rd Qu.: 40.00
##  Max.   :1.0000   Max.   : 2.9042   Max.   :9.000   Max.   :100.00
##       lpsa
##  Min.   :-0.4308
##  1st Qu.: 1.7317
##  Median : 2.5915
##  Mean   : 2.4784
##  3rd Qu.: 3.0564
##  Max.   : 5.5829
```

```r
anyNA(prostate)
```

```
## [1] FALSE
```

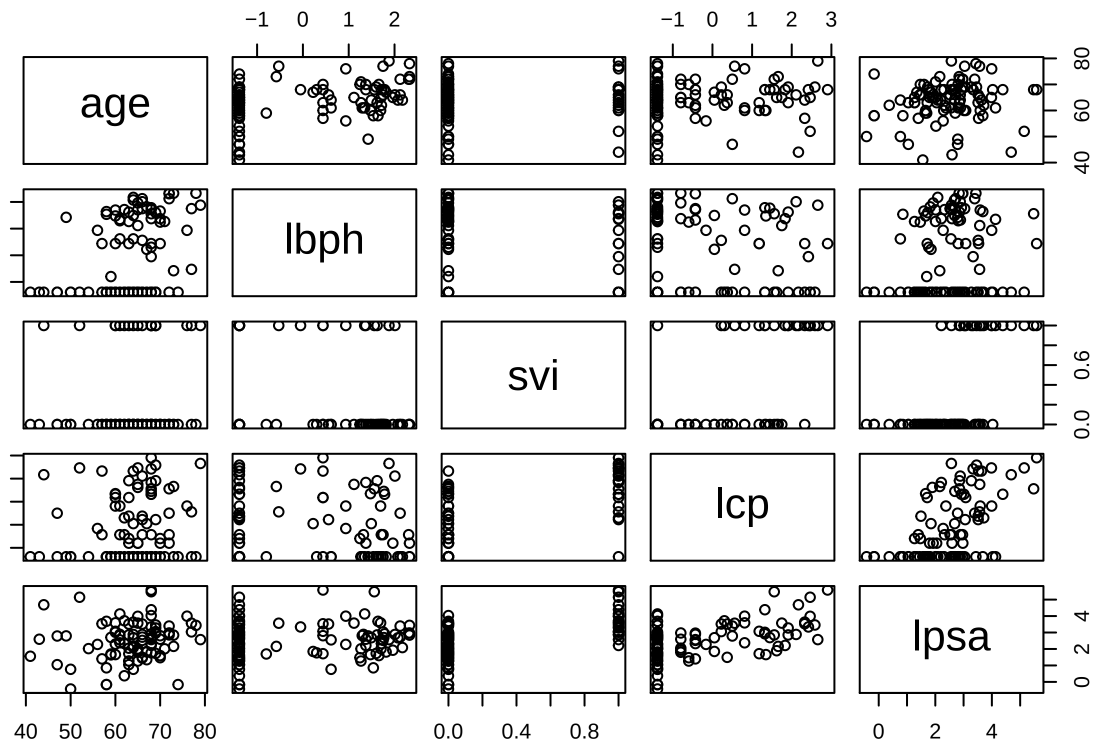
```r
par(mfrow = c(3,2),
    mar = c(4, 4, 0.1, 0.1))
hist(prostate$age, xlab = 'age', main = '')
hist(prostate$lweight, xlab = 'lweight', main = '')
hist(prostate$lcavol, xlab = 'lcavol', main = '')

hist(prostate$lbph, xlab = 'lbph', main = '')
hist(prostate$lcp, xlab = 'lcp', main = '')
hist(prostate$lpsa, xlab = 'lpsa', main = '')
```



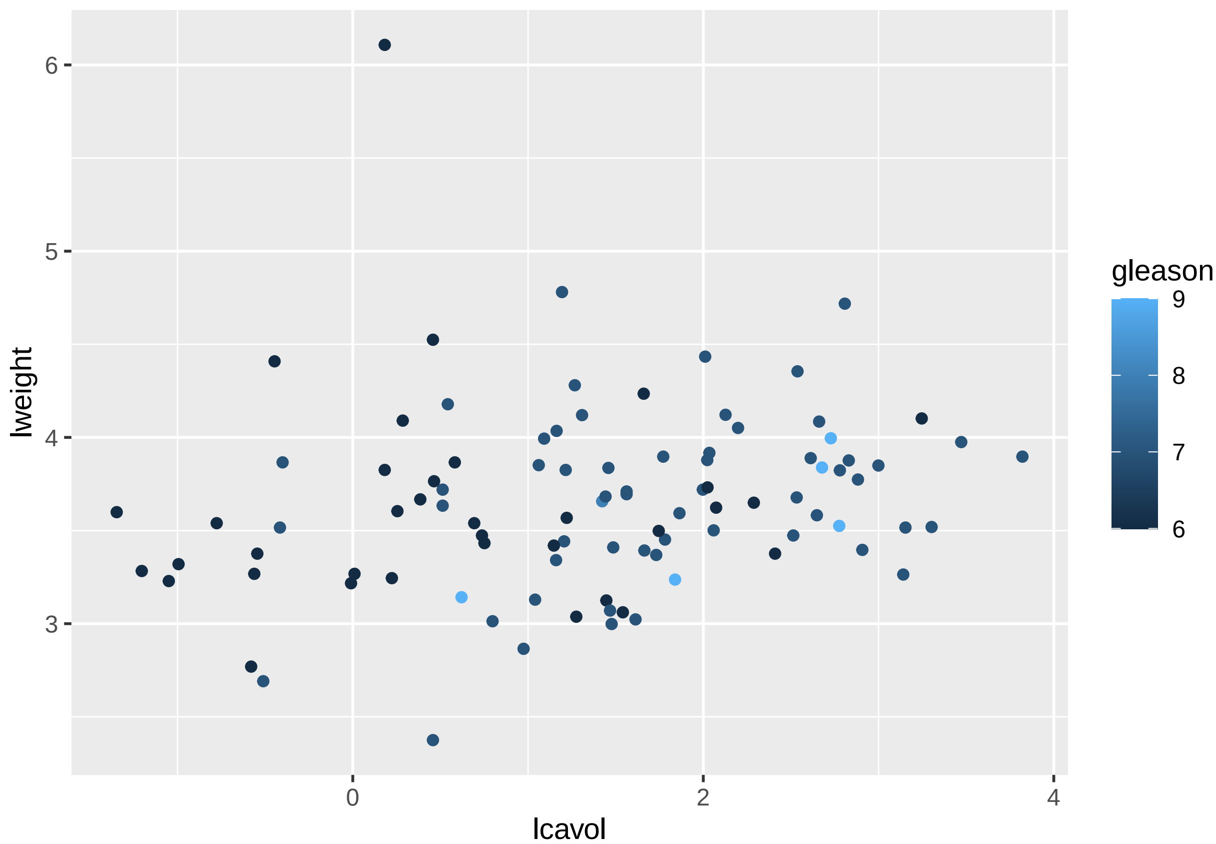
So.. the Gleason Cancer Score is probably what we would like to predict…

lcavol and lweight are logs of cancer volume and weight, respectively. We could perhaps try to find a predictor formula for these, too.

```
pairs(prostate[,c('age', 'lbph', 'svi', 'lcp', 'lpsa')])
```




```
ggplot(prostate, aes(x=lcavol, y = lweight, col = gleason)) +
  geom_point()
```



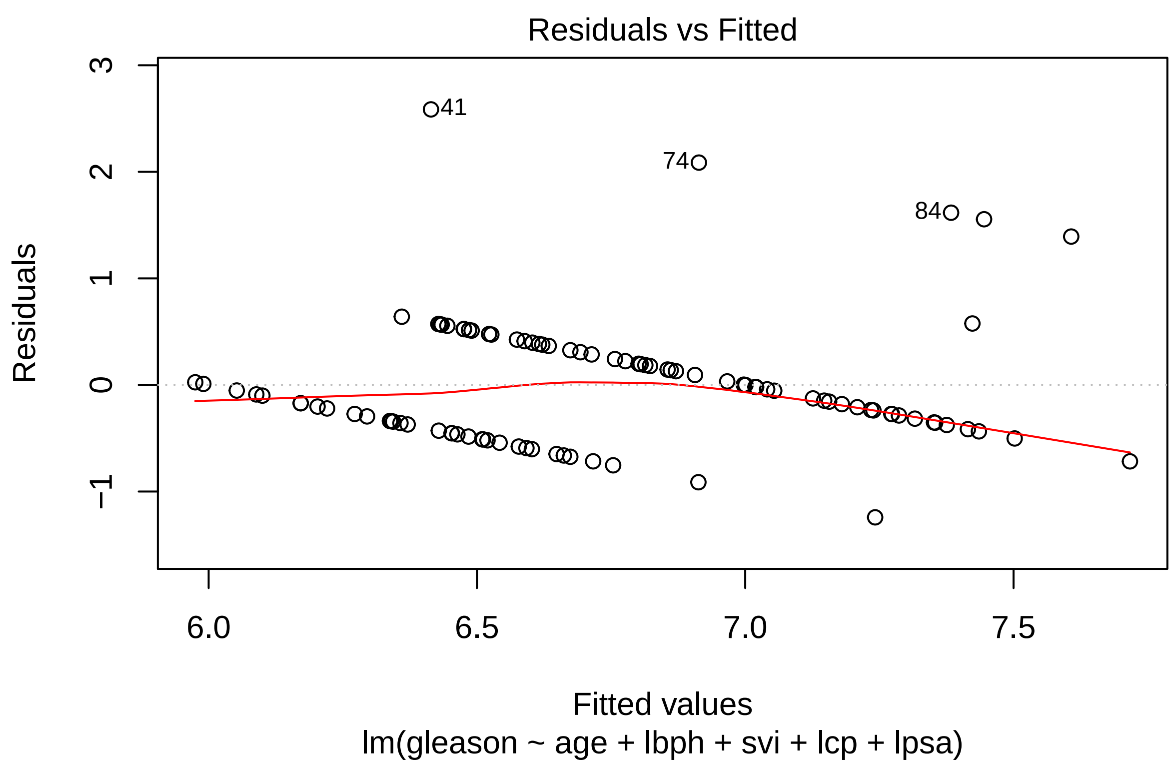
Realistically, we should try to find a model with clinically identified inputs. We can take predictors age, lbph, svi, lcp, and lpsa for gleason. These are the age, log(benign prostatic hyperplasia amount), seminal vesicle invasion, log(capsular penetration), and log(prostate specific antigen).

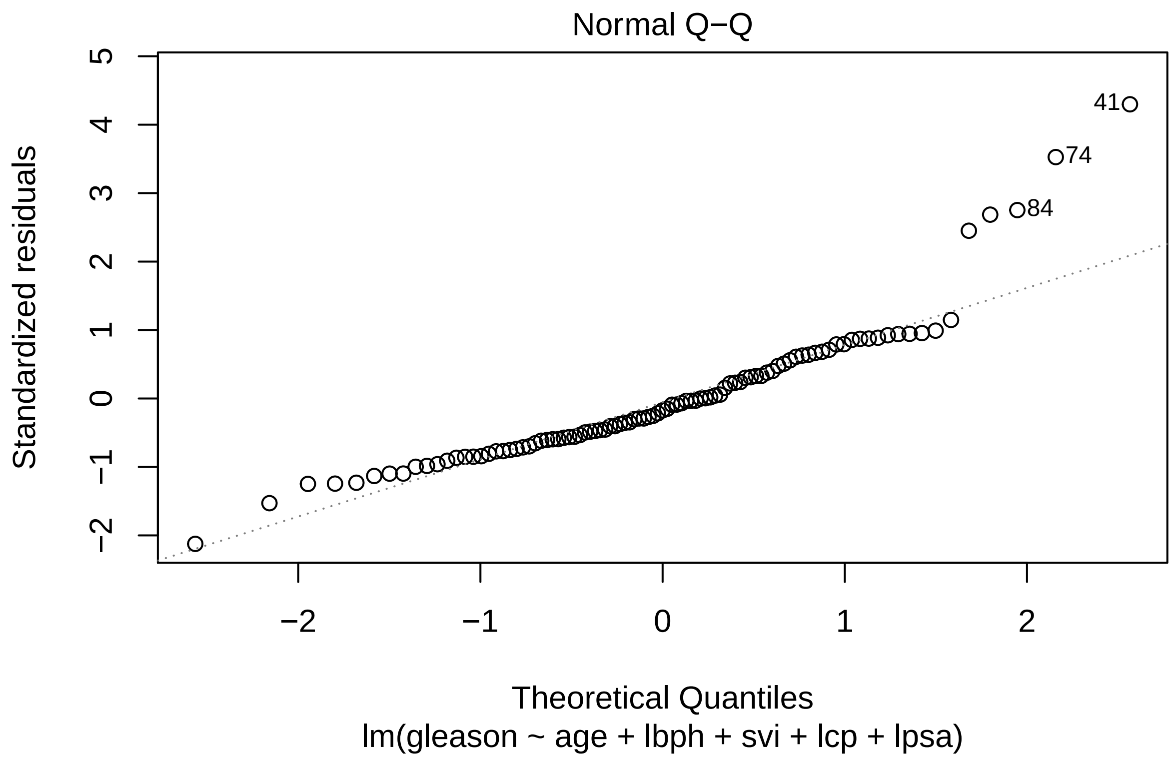
```
lm.gle <- lm(gleason ~ age + lbph + svi + lcp + lpsa, prostate)

summary(lm.gle)
```

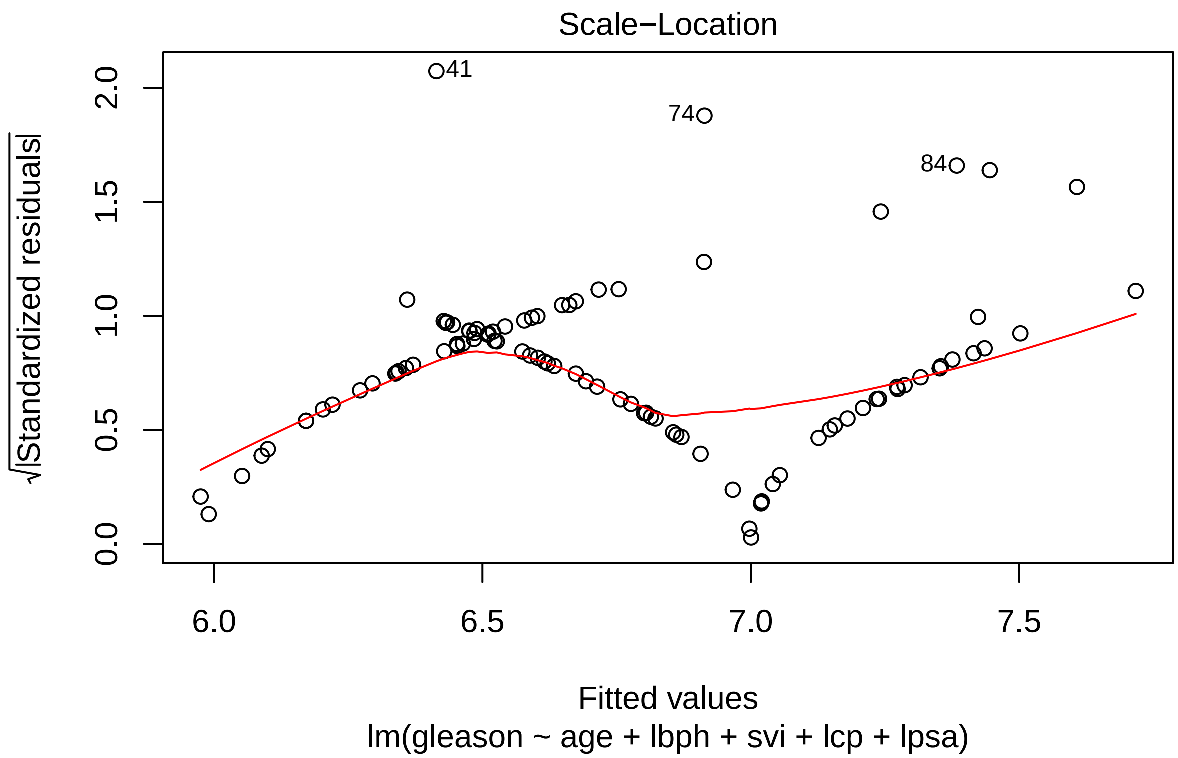
```
##
## Call:
## lm(formula = gleason ~ age + lbph + svi + lcp + lpsa, data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2422 -0.3709 -0.1001  0.3071  2.5857
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.364962   0.599996   8.942 4.18e-14 ***
## age          0.019910   0.009082   2.192 0.030910 *
## lbph        -0.011868   0.047952  -0.248 0.805073
## svi         -0.193051   0.220053  -0.877 0.382636
## lcp          0.253356   0.062879   4.029 0.000116 ***
## lpsa         0.082515   0.070934   1.163 0.247765
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

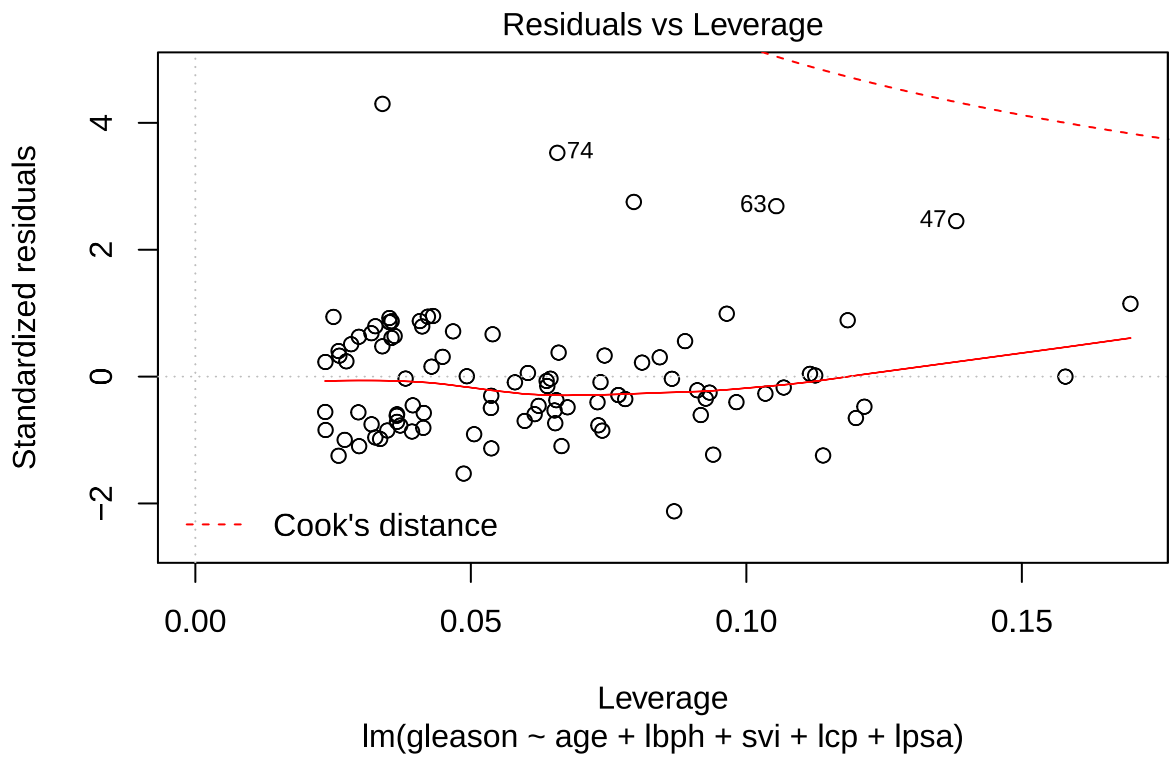
```
##
## Residual standard error: 0.612 on 91 degrees of freedom
## Multiple R-squared:  0.3191, Adjusted R-squared:  0.2817
## F-statistic: 8.529 on 5 and 91 DF,  p-value: 1.179e-06
```

```
plot(lm.gle)
```




Normal Q−Q
Standardized residuals
5
4
3
2
1
0
−1
−2
41
74
84
−2
−1
0
1
2
Theoretical Quantiles
lm(gleason ~ age + lbph + svi + lcp + lpsa)


Scale−Location
41
74
84
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
6.0
6.5
7.0
7.5
Fitted values
lm(gleason ~ age + lbph + svi + lcp + lpsa)



This is not a terrible fit...

1.4

The dataset sat comes from a study entitled "Getting What You Pay For: The Debate Over Equity in Public School Expenditures." Make a numerical and graphical summary of the data as in the first question.

```
data(sat, package = 'faraway')

head(sat)
```

```
##             expend ratio salary takers verbal math total
## Alabama      4.405  17.2 31.144      8    491  538  1029
## Alaska       8.963  17.6 47.951     47    445  489   934
## Arizona      4.778  19.3 32.175     27    448  496   944
## Arkansas     4.459  17.1 28.934      6    482  523  1005
## California   4.992  24.0 41.078     45    417  485   902
## Colorado     5.443  18.4 34.571     29    462  518   980
```

```
summary(sat)
```

```
##      expend           ratio            salary           takers
##  Min.   :3.656   Min.   :13.80   Min.   :25.99   Min.   : 4.00
##  1st Qu.:4.882   1st Qu.:15.22   1st Qu.:30.98   1st Qu.: 9.00
```

```
##  Median :5.768   Median :16.60   Median :33.29   Median :28.00
##  Mean   :5.905   Mean   :16.86   Mean   :34.83   Mean   :35.24
##  3rd Qu.:6.434   3rd Qu.:17.57   3rd Qu.:38.55   3rd Qu.:63.00
##  Max.   :9.774   Max.   :24.30   Max.   :50.05   Max.   :81.00
##      verbal           math           total
##  Min.   :401.0   Min.   :443.0   Min.   : 844.0
##  1st Qu.:427.2   1st Qu.:474.8   1st Qu.: 897.2
##  Median :448.0   Median :497.5   Median : 945.5
##  Mean   :457.1   Mean   :508.8   Mean   : 965.9
##  3rd Qu.:490.2   3rd Qu.:539.5   3rd Qu.:1032.0
##  Max.   :516.0   Max.   :592.0   Max.   :1107.0
```

```r
anyNA(sat)
```

```
## [1] FALSE
```

The sat data frame has 50 rows and 7 columns. Data were collected to study the relationship between expenditures on public education and test results.

expend – Current expenditure per pupil in average daily attendance in public elementary and secondary schools, 1994-95 (in thousands of dollars)

ratio – Average pupil/teacher ratio in public elementary and secondary schools, Fall 1994

salary – Estimated average annual salary of teachers in public elementary and secondary schools, 1994-95 (in thousands of dollars)

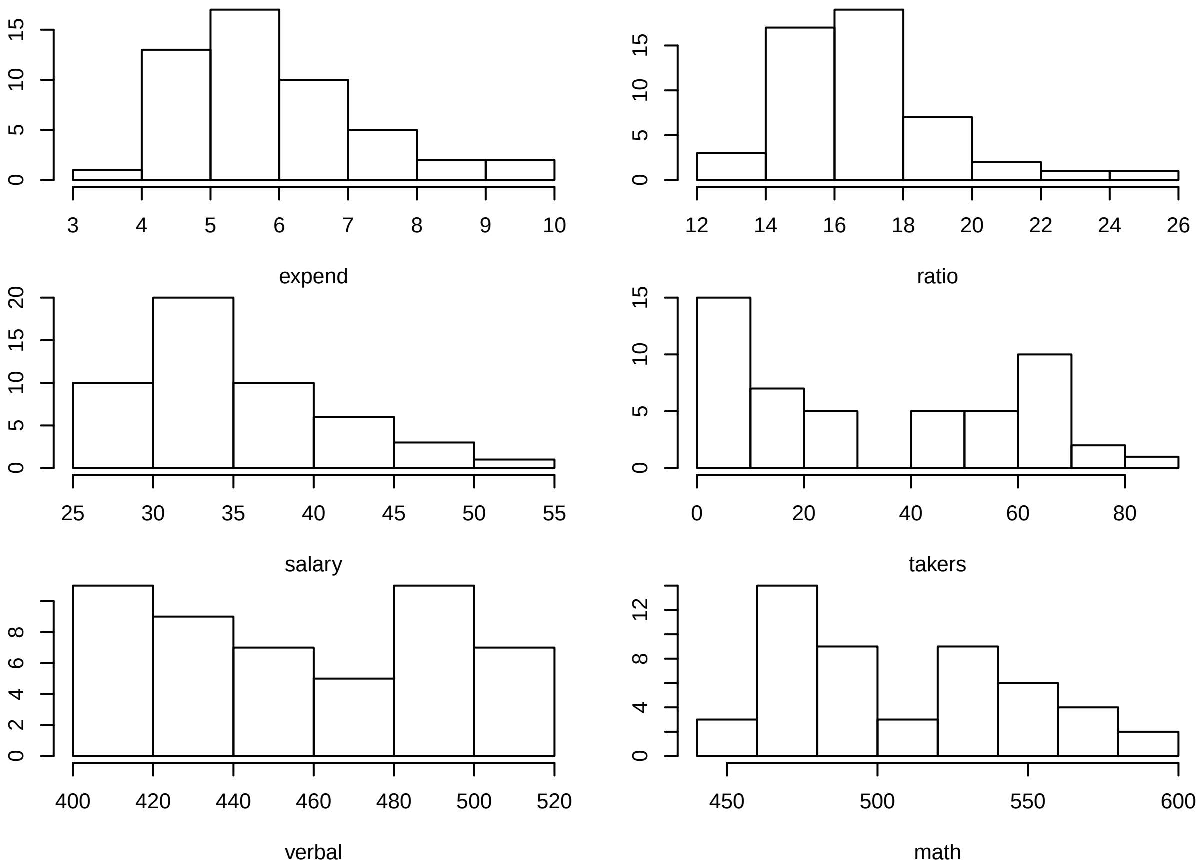
takers – Percentage of all eligible students taking the SAT, 1994-95

verbal – Average verbal SAT score, 1994-95

math – Average math SAT score, 1994-95

total – Average total score on the SAT, 1994-95

```r
par(mfrow = c(3, 2),
    mar = c(4, 4, 0.1, 0.1))

hist(sat$expend, main = '', ylab = '', xlab = 'expend')
hist(sat$ratio, main = '', ylab = '', xlab = 'ratio')
hist(sat$salary, main = '', ylab = '', xlab = 'salary')
hist(sat$takers, main = '', ylab = '', xlab = 'takers')
hist(sat$verbal, main = '', ylab = '', xlab = 'verbal')
hist(sat$math, main = '', ylab = '', xlab = 'math')
```



Math and verbal scores and the number of test takers are all bimodal, and they appear strongly correlated. Expenditures and tecaher/pupil ratio appear skewed in the same way with salary.

We should be trying to predict the math and verbal scores using the other features as inputs. If a model is readily obvious, there is a strong case to be made for the efficacy of education expenditures on math and verbal SAT scores, for this population during this time period.

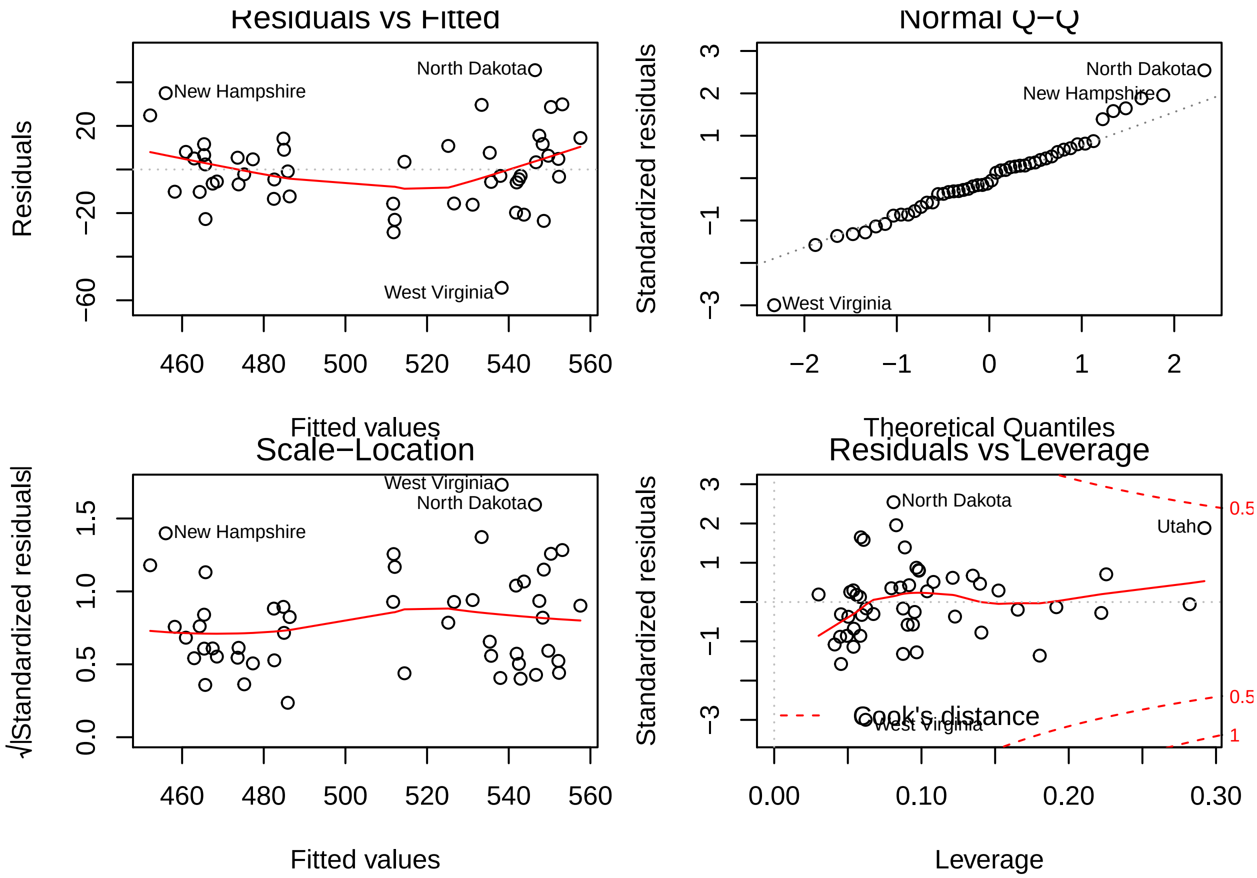
```
lm.math <- lm(math ~ expend + ratio + salary + takers, sat)
lm.verbal <- lm(verbal ~ expend + ratio + salary + takers, sat)
```

```
summary(lm.math)
```

```
##
## Call:
## lm(formula = math ~ expend + ratio + salary + takers, data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.269 -10.282  -1.548   8.797  45.562
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 536.2724    30.2214  17.745  < 2e-16 ***
## expend        3.1560     6.0286   0.524    0.603
## ratio        -1.5428     1.8380  -0.839    0.406
## salary        1.0080     1.3646   0.739    0.464
```

```
## takers       -1.5672     0.1322 -11.855 1.94e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.69 on 45 degrees of freedom
## Multiple R-squared:  0.8015, Adjusted R-squared:  0.7838
## F-statistic: 45.42 on 4 and 45 DF,  p-value: 3.024e-15
```

```r
par(mfrow = c(2, 2),
    mar = c(4, 4, 1, 1))
plot(lm.math)
```

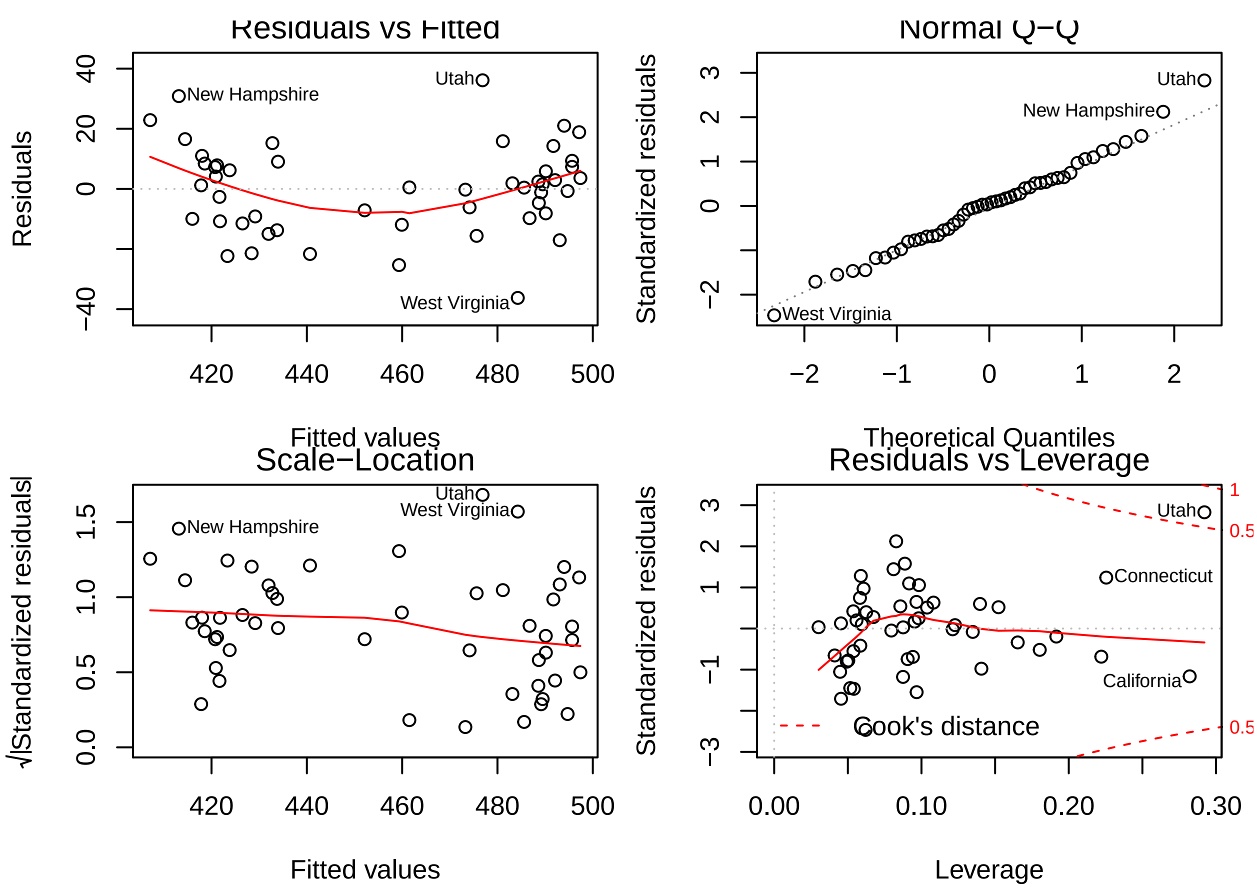



```r
summary(lm.verbal)
```

```
##
## Call:
## lm(formula = verbal ~ expend + ratio + salary + takers, data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.263  -9.915   0.834   8.277  36.131
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 509.6991    24.5539  20.758  < 2e-16 ***
```

```
## expend        1.3066     4.8981   0.267    0.791
## ratio        -2.0814     1.4933  -1.394    0.170
## salary        0.6300     1.1087   0.568    0.573
## takers       -1.3373     0.1074 -12.452 3.53e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.19 on 45 degrees of freedom
## Multiple R-squared:  0.8288, Adjusted R-squared:  0.8136
## F-statistic: 54.46 on 4 and 45 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2, 2),
    mar = c(4, 4, 1, 1))
plot(lm.verbal)
```



Apparently, the observations are labeled by state, and New Hampshire, Utah, and West Virginia are extreme model-influencing outliers. All the rest of the residuals seem appropriate for out model.

1.5

The dataset divusa contains data on divorces in the United States from 1920 to 1996. Make a numerical and graphical summary of the data as in the first question.

```r
data(divusa, package = 'faraway')

head(divusa)
```

```
##   year divorce unemployed femlab marriage birth military
## 1 1920     8.0        5.2  22.70     92.0 117.9   3.2247
## 2 1921     7.2       11.7  22.79     83.0 119.8   3.5614
## 3 1922     6.6        6.7  22.88     79.7 111.2   2.4553
## 4 1923     7.1        2.4  22.97     85.2 110.5   2.2065
## 5 1924     7.2        5.0  23.06     80.3 110.9   2.2889
## 6 1925     7.2        3.2  23.15     79.2 106.6   2.1735
```

```r
summary(divusa)
```

```
##       year         divorce        unemployed         femlab     
##  Min.   :1920   Min.   : 6.10   Min.   : 1.200   Min.   :22.70  
##  1st Qu.:1939   1st Qu.: 8.70   1st Qu.: 4.200   1st Qu.:27.47  
##  Median :1958   Median :10.60   Median : 5.600   Median :37.10  
##  Mean   :1958   Mean   :13.27   Mean   : 7.173   Mean   :38.58  
##  3rd Qu.:1977   3rd Qu.:20.30   3rd Qu.: 7.500   3rd Qu.:47.80  
##  Max.   :1996   Max.   :22.80   Max.   :24.900   Max.   :59.30  
##     marriage          birth           military     
##  Min.   : 49.70   Min.   : 65.30   Min.   : 1.940  
##  1st Qu.: 61.90   1st Qu.: 68.90   1st Qu.: 3.469  
##  Median : 74.10   Median : 85.90   Median : 9.102  
##  Mean   : 72.97   Mean   : 88.89   Mean   :12.365  
##  3rd Qu.: 80.00   3rd Qu.:107.30   3rd Qu.:14.266  
##  Max.   :118.10   Max.   :122.90   Max.   :86.641
```

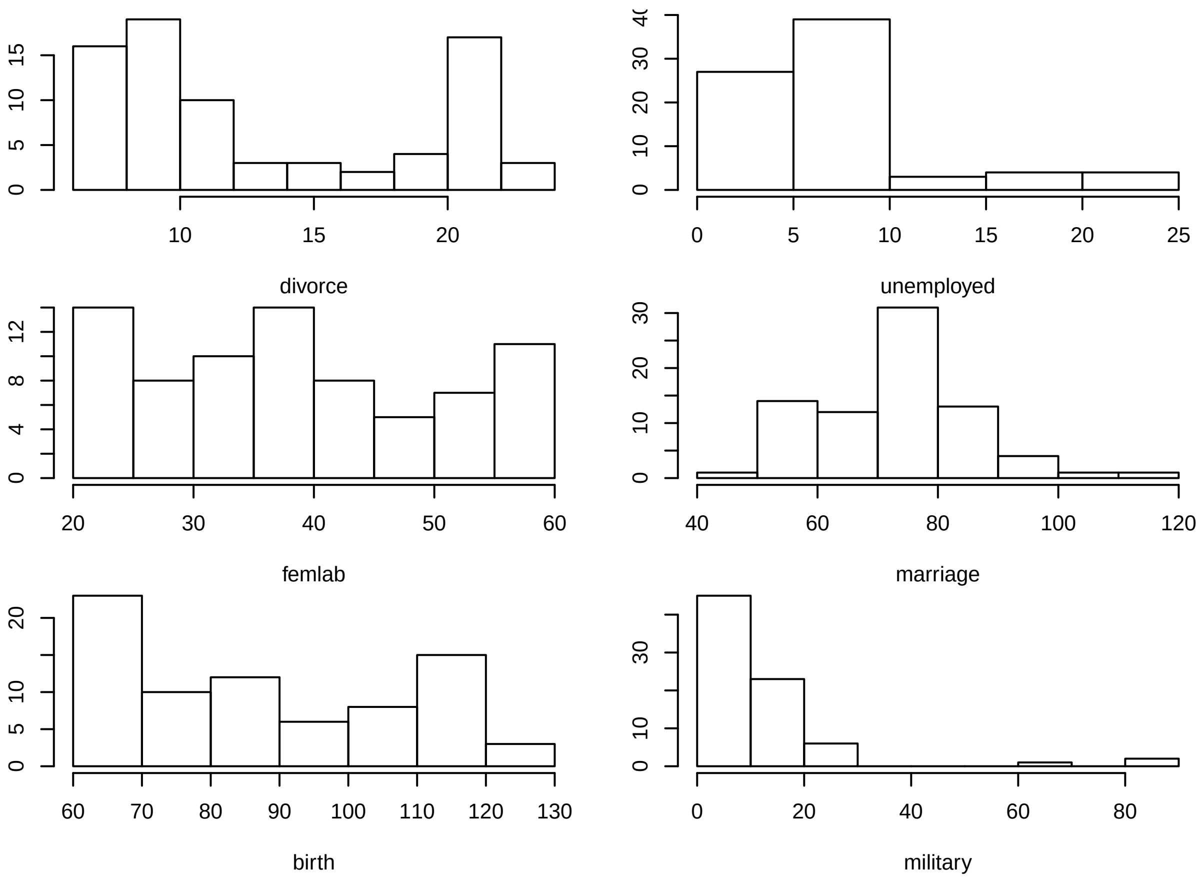
```r
anyNA(divusa)
```

```
## [1] FALSE
```

The seven variables are year (1920 - 1996), divorce per 1000 women aged 15 or more, unemployment rate, percent female participation in labor force aged 16+, births per 1000 women age 15-44, military personnel per 1000 population.

These are some interesting inputs. I guess that the interesting thing would be to predict divorce rates and birthrates from the other features. All of the data is numerical.

```r
par(mfrow = c(3, 2),
    mar = c(4, 4, 0.1, 0.1))

hist(divusa$divorce, main = '', ylab = '', xlab = 'divorce')
hist(divusa$unemployed, main = '', ylab = '', xlab = 'unemployed')
hist(divusa$femlab, main = '', ylab = '', xlab = 'femlab')
hist(divusa$marriage, main = '', ylab = '', xlab = 'marriage')
hist(divusa$birth, main = '', ylab = '', xlab = 'birth')
hist(divusa$military, main = '', ylab = '', xlab = 'military')
```



It looks like the data cross 3 periods of relatively high female involvement in the workforce, or that there were two major movements to increase involvement over the course of the observations. It seems like unemployment hovered around 5% or 6% for most of the timespan. Marriage is normally distributed, but skewed a bit to the right. Military personnel is an exponential distrbution with a couple periods of extremely high involvement… Maybe this is the Vietnam War and WWII? The birth rate appears to be a noisy uniform distribution.

```
lm.birth <- lm(birth ~ year + unemployed +
                 femlab + marriage + military,divusa)

lm.div <- lm(divorce ~ year + unemployed +
               femlab + marriage + military,divusa)
```

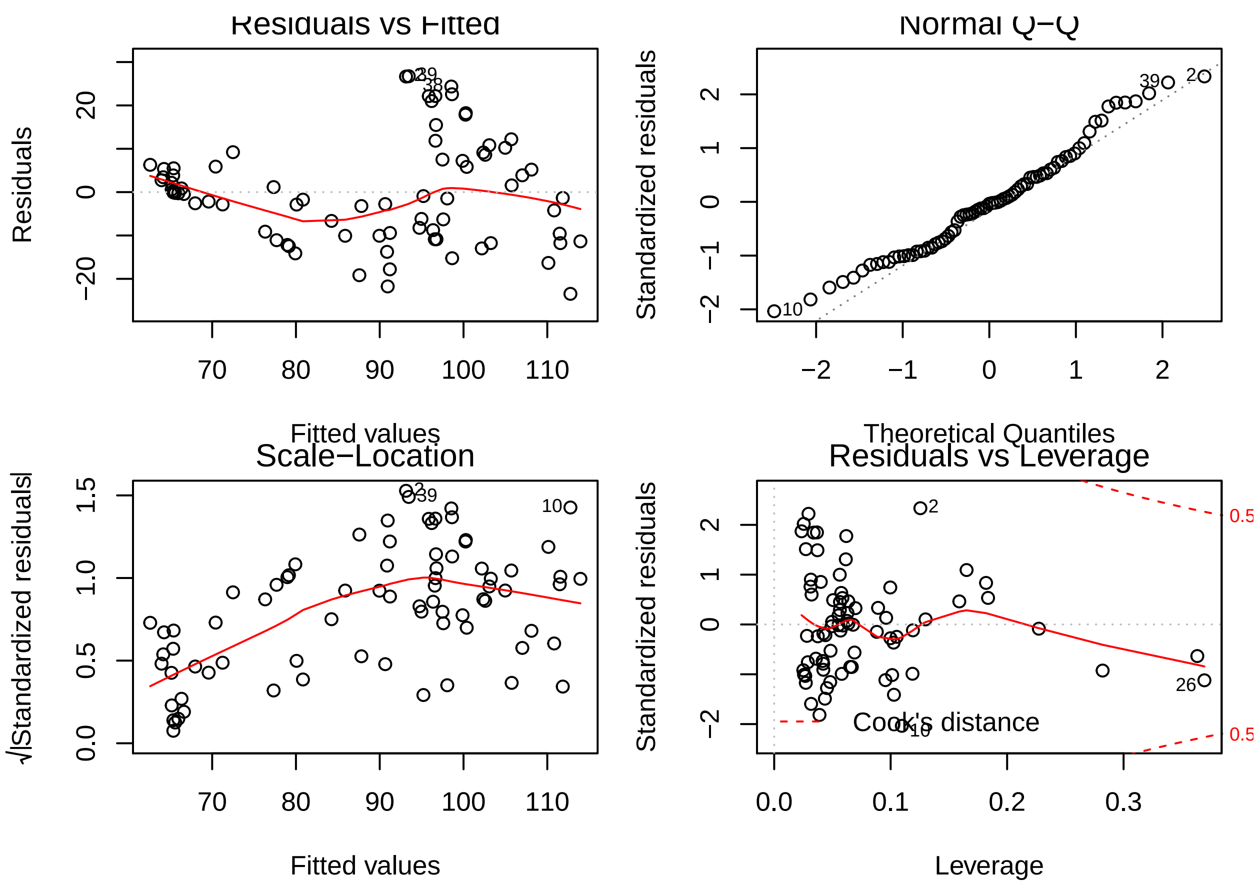
Both features are fitted against the same inputs, unaltered.

```
summary(lm.birth)
```

```
##
## Call:
## lm(formula = birth ~ year + unemployed + femlab + marriage +
##     military, data = divusa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.4618  -9.5837  -0.4316   6.2968  26.7356
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.454e+03  7.822e+02  -1.859  0.06714 .
## year         8.435e-01  4.188e-01   2.014  0.04780 *
## unemployed  -1.854e+00  3.744e-01  -4.951  4.8e-06 ***
## femlab      -2.700e+00  8.704e-01  -3.102  0.00276 **
## marriage     1.244e-01  1.918e-01   0.648  0.51877
## military     3.353e-03  1.108e-01   0.030  0.97594
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.22 on 71 degrees of freedom
## Multiple R-squared:  0.6341, Adjusted R-squared:  0.6084
## F-statistic: 24.61 on 5 and 71 DF,  p-value: 2.728e-14
```

```r
par(mfrow = c(2, 2),
    mar = c(4, 4, 1, 1))
plot(lm.birth)
```

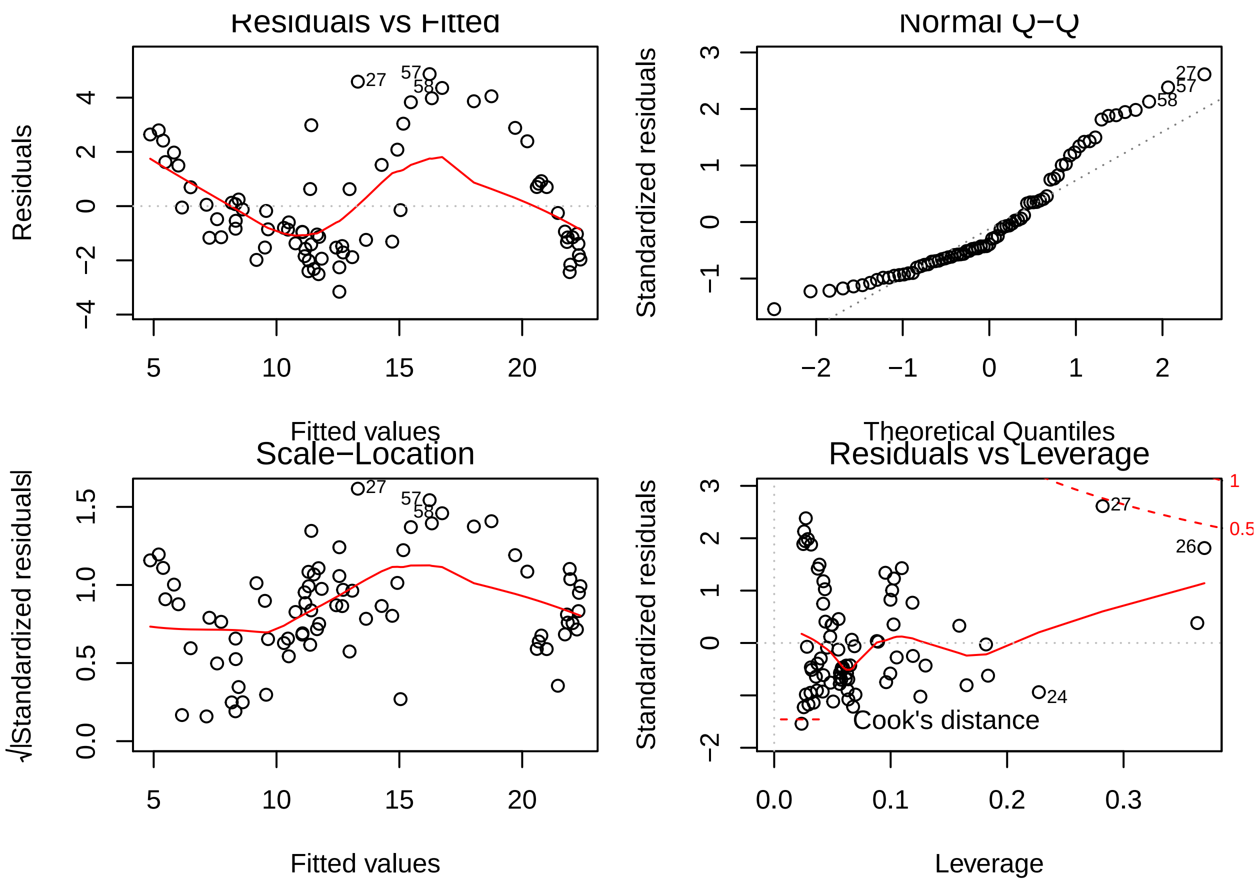



```r
summary(lm.div)
```

```
##
## Call:
## lm(formula = divorce ~ year + unemployed + femlab + marriage +
```

```
##     military, data = divusa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1579 -1.4114 -0.8022  0.9209  4.8680
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 550.23138  132.74378   4.145 9.26e-05 ***
## year         -0.30176    0.07107  -4.246 6.49e-05 ***
## unemployed    0.16746    0.06353   2.636   0.0103 *
## femlab        1.12369    0.14771   7.607 8.93e-11 ***
## marriage      0.13523    0.03255   4.155 8.95e-05 ***
## military     -0.04316    0.01880  -2.295   0.0247 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.073 on 71 degrees of freedom
## Multiple R-squared:  0.8751, Adjusted R-squared:  0.8663
## F-statistic: 99.48 on 5 and 71 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2, 2),
    mar = c(4, 4, 1, 1))
plot(lm.div)
```



The errors on these are very high... Although the models do identify some trends: the marriage rate and the military ratio don't affect birthrates but unemployment and female workplace participation do strongly.

And divorce rates are very hard to predict. It seems there is a strong trend towards divorce over time, and that marriage and female employment rates also have the strongest predictive value.

A Modern Approach to Regression with R

Generally, the linear regression model is written in matrix form as:

$$Y = X\beta + \epsilon$$

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & ... & x_{1p} \\ 1 & x_{21} & ... & x_{2p} \\ ... & ... & & \\ 1 & x_{n1} & ... & x_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ ... \\ \beta_p \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ ... \\ \epsilon_n \end{pmatrix}$$

The least squares estimates are given by:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

We next derive the conditional mean of the least squares estimates:

$$E(\hat{\beta}|X) = E\big((X'X)^{-1}X'Y|X)\big)$$

$$= (X'X)^{-1}X'E(Y|X)$$

$$= (X'X)^{-1}X'X\beta$$

$$= \beta$$