# Notebook 1 Indicator Variable Example

```r
# The first part of the "brick" of lecture notes for this course is about including categorical
# variables as explanatory (X) variables on the right hand side of a linear model. This is an
# additional example of using the simplest form of categorical variable (a Yes/No or 0/1 coded
# variable, often frequently referred to as a "dummy" variable).

# This example is taken from exercise 3.7 on page 80 of Lattin J, Carroll JD and Green, PE (2003)
# Analyzing Mutivariate Data, published by Thomson Brooks/Cole, which was a text for another
# statistics unit (STAT8020 Multivariate Analysis).

# In this example, data was collected in March 1977 on 116 bank employees (of Harry's Trust and
# Savings Bank) hired during 1969 to 1971 as general office trainees. The following variables
# were collected for each employee:
#   EmpID - employee ID number (a three or four digit identifying number)
#   Sex - sex of employee (0 = Male, 1 = Female)
#   Age - age at hire (in months)
#   EducYrs - years of education
#   EducLvl - level of education (1 = Graduate school, 2 = College graduate, 3 = Some college,
#                                 4 = High school graduate, 5 = None of the above)
#   WorkExp - prior work experience (in months)
#   Start - starting salary (in $)
#   Senior - seniority (total number of months of employment)
#   Salary - current salary (in $)

# The data are available in "Bankwages.txt" and can be read into R using any of the methods
# demonstrated in lectures.

# For example, the following approach will work if you downloaded both this example R file
# and the data file to the same directory on your computer and then started RStudio by
# opening the R file with RStudio (with ANU computers, you need to right-click and go
# "Open With" RStudio - if you double-click, then "really helpful: default association is to
# try and open the file with the package Statistica, which is no longer available on the ANU
# computer network):

bankwages <- read.table("BankWages.txt", header=T)

# First examine the data by simply typing the name (or if you are using RStudio, you can
# double-click on the object in the Environment window, which opens the Viewer):

bankwages
```

```
##     EmpID Sex Age EducYrs EducLvl WorkExp Start Senior Salary
## 1     708   0 534      12       4   216.0  6600     70   9000
## 2     712   0 322      15       3    15.0  6300     70  11760
## 3     722   0 334      12       4    40.5  5880     70  10980
## 4     736   0 432      15       3    93.0  6600     94  13920
## 5     745   0 364      15       3    43.0  5700     96  13020
## 6     765   0 378      12       4    75.0  6000     87  11040
## 7     771   0 338      15       3    48.0  6000     72  11340
## 8     785   0 344      15       3    55.0  6300     68  12120
## 9     786   0 354      15       3    34.0  6300     82  32000
## 10    787   0 368      15       3    52.0  6000     81  10620
```

```
## 11     812   0 345       15        3     46.0  6300     67  12480
## 12     816   0 322       14        3     14.5  6000     71  11220
## 13     821   0 364       15        3     55.0  6600     79  11100
## 14     829   0 329       12        4     47.0  6300     66  13416
## 15     843   0 390       15        3     64.0  6300     92  12660
## 16     844   0 383       12        4    106.0  6600     64  13560
## 17     850   0 397       15        3     69.0  6600     83  13260
## 18     857   0 492       12        4    210.0  7200     65  11520
## 19     880   0 352       12        4     46.0  5700     95   9300
## 20     911   0 364       14        3     59.0  6300     91  15960
## 21     931   0 356       12        4     83.0  5700     96  14460
## 22     934   0 394       14        3     60.0  6600     90  15120
## 23     961   0 344       12        4     80.0  6300     66  12108
## 24     968   0 331       15        3     12.0  6300     76  17364
## 25     979   0 386       15        3     68.0  6000     96  16800
## 26     983   0 381       15        3     48.0  6300     91  10980
## 27    1007   0 349       15        3     48.0  6300     70  11940
## 28    1009   0 368       12        4     94.0  6600     79  13560
## 29    1033   0 401       12        4    119.5  6000     94  10800
## 30    1044   0 341       15        3     55.5  6300     70  10140
## 31    1067   0 307       15        3      9.5  5700     65  12660
## 32    1071   0 327       12        4     85.0  6300     63  15660
## 33    1103   0 361       12        4     85.0  6300     82  11940
## 34    1111   0 369       15        3     78.0  6600     67  10860
## 35    1124   0 365       15        3     78.0  6300     94  13560
## 36     644   1 294       12        4      5.0  4500     80   7860
## 37     645   1 276       12        4      0.0  4500     65   8700
## 38     648   1 297        8        5      9.0  4800     81  10980
## 39     651   1 309       15        3      7.0  5400     69   9660
## 40     662   1 289       12        4     12.0  4980     69   9780
## 41     663   1 384       16        2     19.0  6900     85  14280
## 42     666   1 292       12        4      5.0  4500     80   8940
## 43     675   1 279       12        4      0.0  4800     69   8640
## 44     680   1 290       12        4      0.0  4980     69  11160
## 45     682   1 283       12        4      0.0  4500     66   8160
## 46     684   1 680       12        4    139.0  5100     63   8580
## 47     691   1 310       12        4      0.0  4200     93  10560
## 48     698   1 298       12        4      0.0  4500     81   9060
## 49     701   1 299       12        4      7.0  6600     71   8880
## 50     710   1 340       15        3     24.0  6600     86  11760
## 51     714   1 279        8        5     17.0  4980     69  11640
## 52     716   1 297       12        4      5.0  4500     81  10920
## 53     720   1 337       15        3      0.0  4620     93   9660
## 54     724   1 300       12        4     18.0  4500     80   9360
## 55     729   1 281       12        4      2.0  4500     69   9060
## 56     742   1 546       12        4    197.5  5400     64   9360
## 57     743   1 740       12        4    204.5  7500     90  12600
## 58     746   1 301       12        4      7.5  4380     80   9720
## 59     750   1 322       15        3     11.0  4620     93  11700
## 60     753   1 285       12        4      0.0  4500     69   8340
## 61     762   1 301       12        4      3.5  4500     80  11400
## 62     774   1 307       12        4      5.5  4560     93  11040
## 63     780   1 332       15        3     31.5  5340     88  14220
## 64     781   1 315       12        4     24.0  4500     92   8040
```

```
## 65    790   1 288      12        4     2.0  4620     74  10440
## 66    796   1 705      15        3   265.0  7800     82  11100
## 67    803   1 315      15        3     7.0  4800     76  10200
## 68    809   1 298      12        4    24.0  5100     65   9000
## 69    813   1 307      15        3    21.5  5580     73   9180
## 70    818   1 322      16        2    20.0  7800     64  13140
## 71    828   1 311      12        4     9.5  4200     93   9180
## 72    847   1 296      12        4     0.0  4500     77   8520
## 73    863   1 351      16        2    26.0  7200     75  13200
## 74    871   1 294      12        4     0.0  4380     81   9240
## 75    873   1 327      12        4    22.0  6000     76  11664
## 76    876   1 292      12        4     0.0  4500     80   9360
## 77    879   1 284      12        4     4.5  4980     69  12240
## 78    885   1 292       8        5     4.5  4380     81   9000
## 79    888   1 281      12        4     0.0  4500     69   8160
## 80    912   1 295      12        4    18.0  4620     81  10680
## 81    913   1 667      17        1   375.0  5100     73  11640
## 82    919   1 301      12        4     0.0  4500     81   9960
## 83    924   1 303      12        4    11.0  4380     81   7860
## 84    929   1 362      16        2     2.5  6900     80  13800
## 85    937   1 361      16        2    12.0  7200     71  14640
## 86    941   1 299      12        4     8.5  4380     81   8820
## 87    955   1 295      12        4    11.0  4800     81  10200
## 88    970   1 325      15        3    47.0  4440     67  12540
## 89    980   1 350      16        2     4.0  6300     82   9300
## 90    981   1 292      12        4     4.0  4860     77  13800
## 91    994   1 326      15        3    18.5  5100     81  11280
## 92   1004   1 309      12        4    30.0  6120     72  11760
## 93   1011   1 310      12        4    16.0  4500     83   9600
## 94   1012   1 301      12        4     9.0  5700     63  11760
## 95   1015   1 326      12        4    47.0  5400     78   9840
## 96   1028   1 354      15        3     3.0  5400     94  10560
## 97   1039   1 345      15        3     5.0  7200     75  14820
## 98   1043   1 292      12        4     7.5  4800     79   9780
## 99   1050   1 306      12        4     6.0  4200     92  11400
## 100  1054   1 296      12        4     4.5  4380     79   9900
## 101  1056   1 345      15        3    22.0  5400     90  13020
## 102  1059   1 281      12        4     1.5  4500     68   9120
## 103  1062   1 534      12        4     3.0  6000     90  13320
## 104  1063   1 365      16        2     0.0  6600     90  15420
## 105  1064   1 293      12        4    19.5  5100     69   9540
## 106  1076   1 300      12        4     0.0  4380     80   7260
## 107  1078   1 618      12        4   181.0  5400     67   8340
## 108  1035   1 527      16        2   143.0  6600     98  12120
## 109  1090   1 342      15        3     6.0  7200     78  13020
## 110  1091   1 328      12        4    32.0  5400     73  11160
## 111  1099   1 328      12        4    18.0  5580     72  11520
## 112  1102   1 284      12        4     0.0  4500     69   9120
## 113  1108   1 284      12        4     1.5  4500     69   8280
## 114  1110   1 349      15        3    56.5  5580     64  10080
## 115  1125   1 726       8        5   159.0  4800     73   8460
## 116  1129   1 289      12        4    13.0  4500     81   9240
```

```r
names(bankwages)
```

```
## [1] "EmpID"   "Sex"     "Age"     "EducYrs" "EducLvl" "WorkExp" "Start"
## [8] "Senior"  "Salary"
```

```r
# Attach the data for the remainder of this session, so that the variables are available using
# the above column names:

attach(bankwages)

# The question of interest when collecting the data was "is there evidence of any gender (sex)
# discrimination in the employee compensation offered by this bank to general office trainees?"

# We have two measures of compensation in these data - starting and current salary levels
# (Start and Salary) and Sex is a 0/1 variable which indicates the employees gender.
# Firstly, are there apparent differences between the current salaries for males and females:

mean(Salary)
```

```
## [1] 11283.38
```

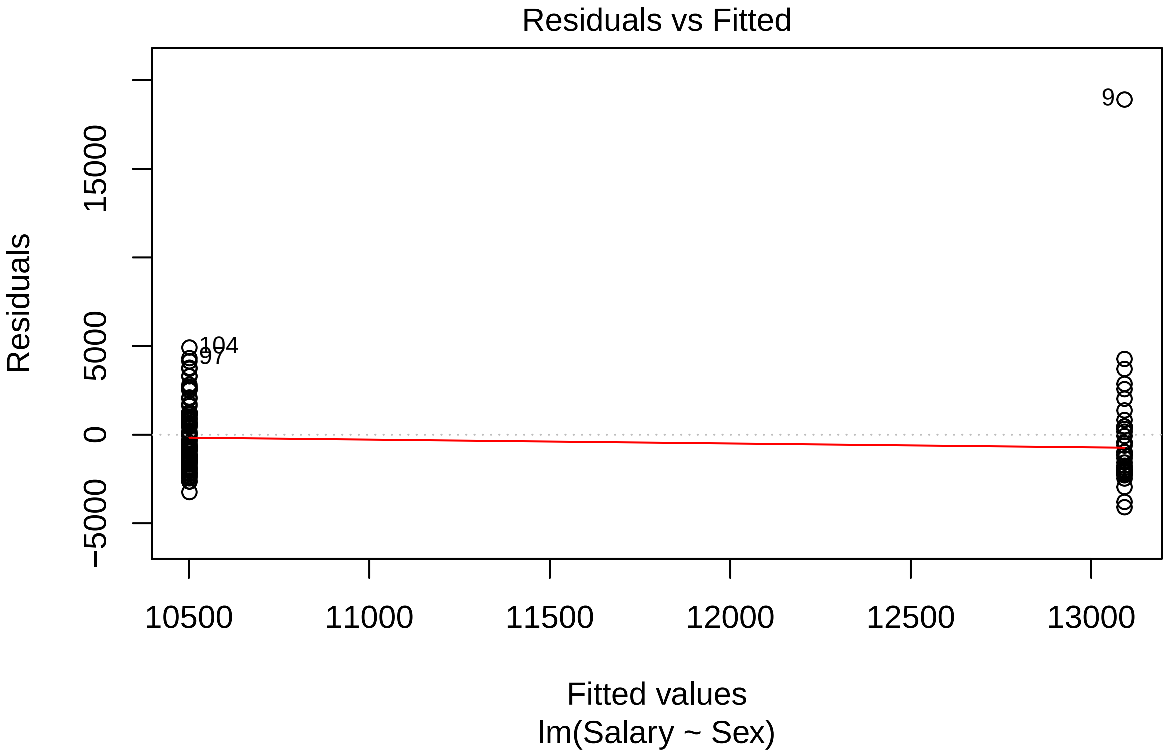
```r
mean(Salary[Sex==0])
```

```
## [1] 13092.23
```

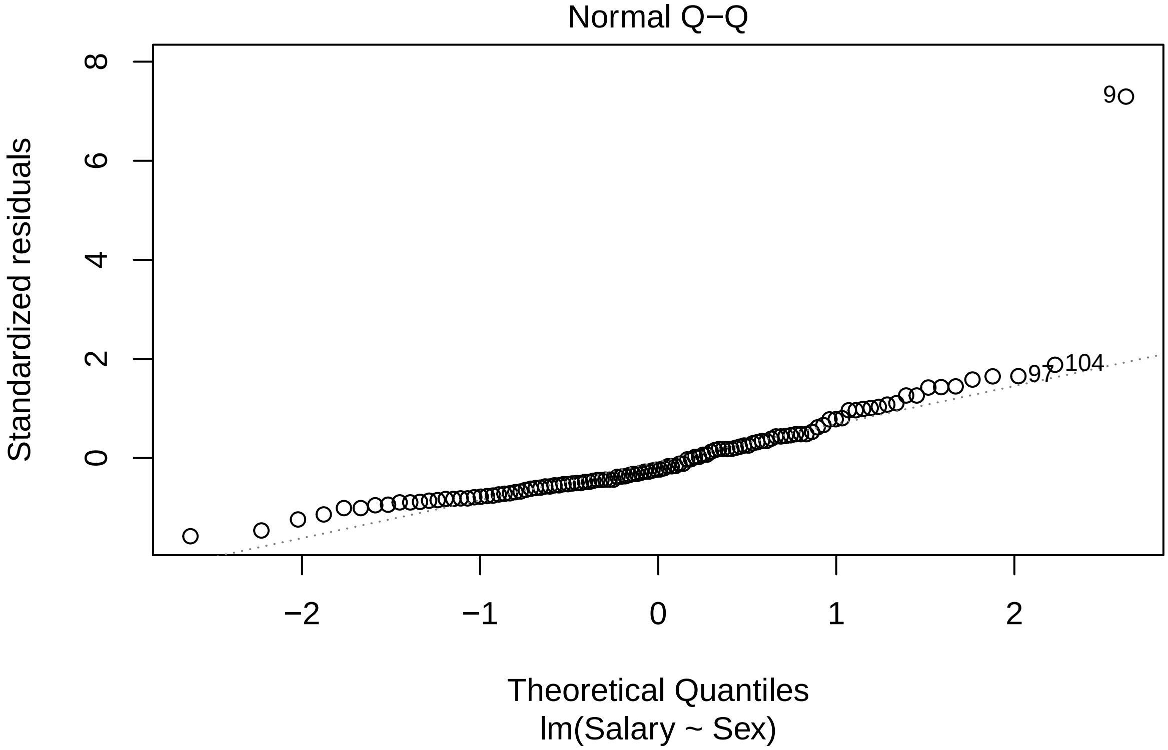
```r
mean(Salary[Sex==1])
```

```
## [1] 10501.78
```

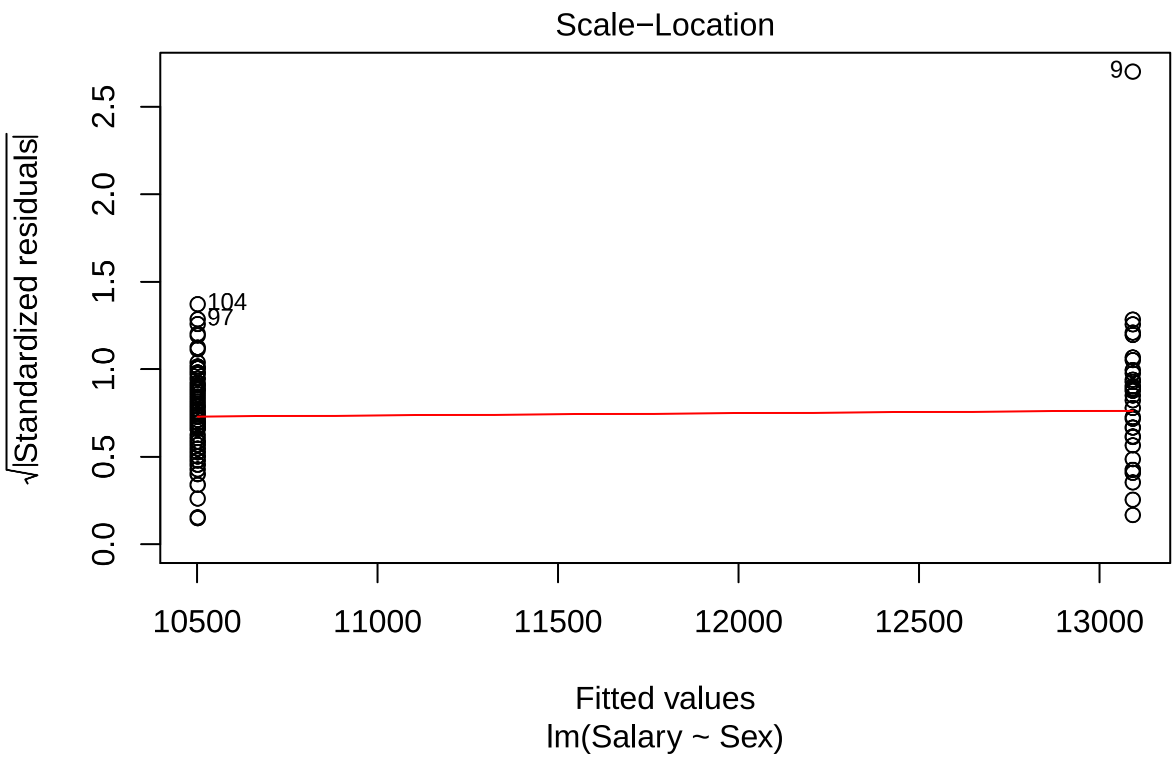
```r
# As Sex is already coded as 0/1 indicator variable, we can decide whether these apparent
# differences in mean current salaries are significant by fitting a simple linear regression
# model:

Salary.lm <- lm(Salary ~ Sex)
plot(Salary.lm)
```

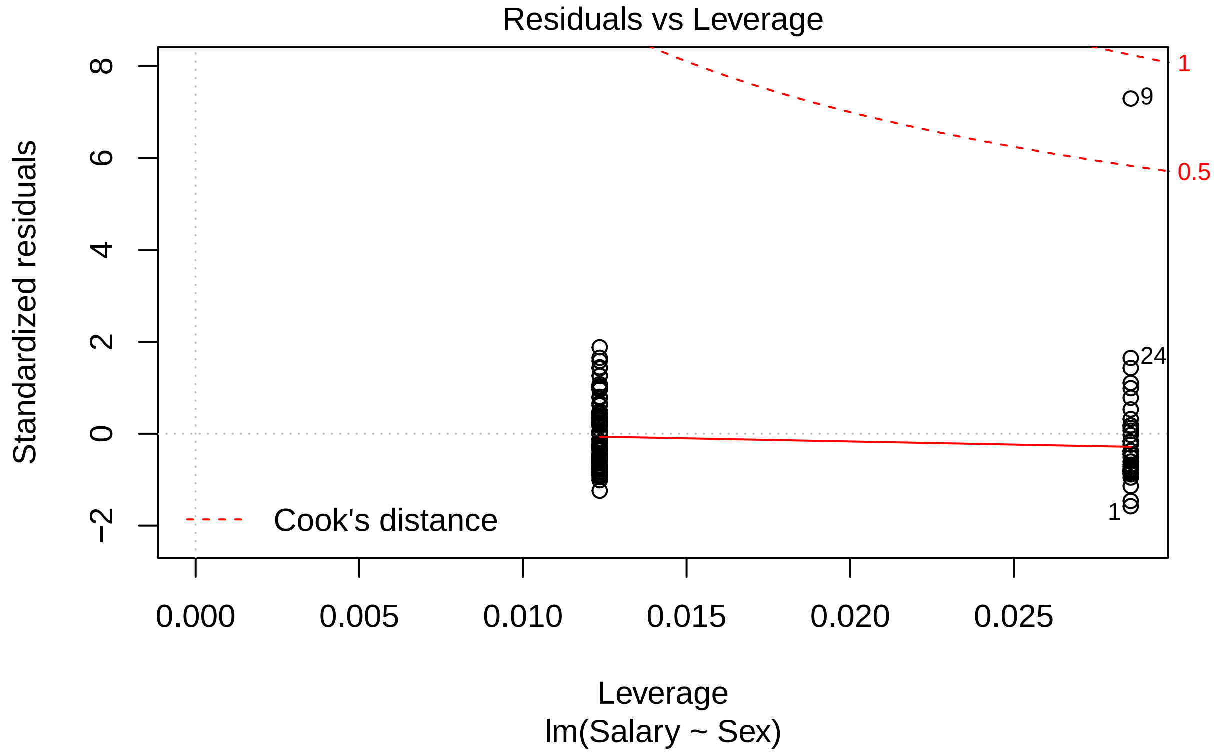


Residuals vs Fitted
Residuals
Fitted values
lm(Salary ~ Sex)

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Salary ~ Sex)




```r
# The plots indicate a definite problem with this model caused by an apparent outlier -
# observation #9, but ignoring this for the moment and looking at the rest of the model output:

anova(Salary.lm)
```

```
## Analysis of Variance Table
```

```
##
## Response: Salary
##            Df    Sum Sq   Mean Sq F value   Pr(>F)
## Sex         1 164000725 164000725  23.714 3.63e-06 ***
## Residuals 114 788398942   6915780
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(Salary.lm)
```

```
##
## Call:
## lm(formula = Salary ~ Sex)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4092  -1564   -607   1138  18908
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13092.2      444.5   29.45  < 2e-16 ***
## Sex          -2590.5      532.0   -4.87 3.63e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2630 on 114 degrees of freedom
## Multiple R-squared:  0.1722, Adjusted R-squared:  0.1649
## F-statistic: 23.71 on 1 and 114 DF,  p-value: 3.63e-06
```

```r
# What are the estimated coefficients of this model? The intercept turns out to be the mean
# salary for males and the slope is the difference between the mean salary for males and the
# mean salary for females:

mean(Salary[Sex==0])
```

```
## [1] 13092.23
```

```r
mean(Salary[Sex==1]) - mean(Salary[Sex==0])
```

```
## [1] -2590.451
```

```r
# Note the t-test on the slope using this parameterisation turns out to be identical to the
# two sample t-test (assuming equal variances), that is typically included in a first year
# introductory statistics course:

t.test(Salary[Sex==1], Salary[Sex==0], var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Salary[Sex == 1] and Salary[Sex == 0]
## t = -4.8697, df = 114, p-value = 3.63e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3644.245 -1536.657
## sample estimates:
## mean of x mean of y
```

```
## 10501.78 13092.23
```

```r
# Note the above "treatment" coding (using 0/1) is not the only way we could have coded sex as
# an indicator variable. Here is an alternative parameterisation of Sex (usually called "sum"
# coding in an experimental design context):

Sex2 <- ifelse(Sex==0,-1,1)
Sex2
```

```
##   [1] -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
##  [24] -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1  1  1  1  1  1  1  1  1  1  1  1
##  [47]  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##  [70]  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
##  [93]  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
## [116]  1
```

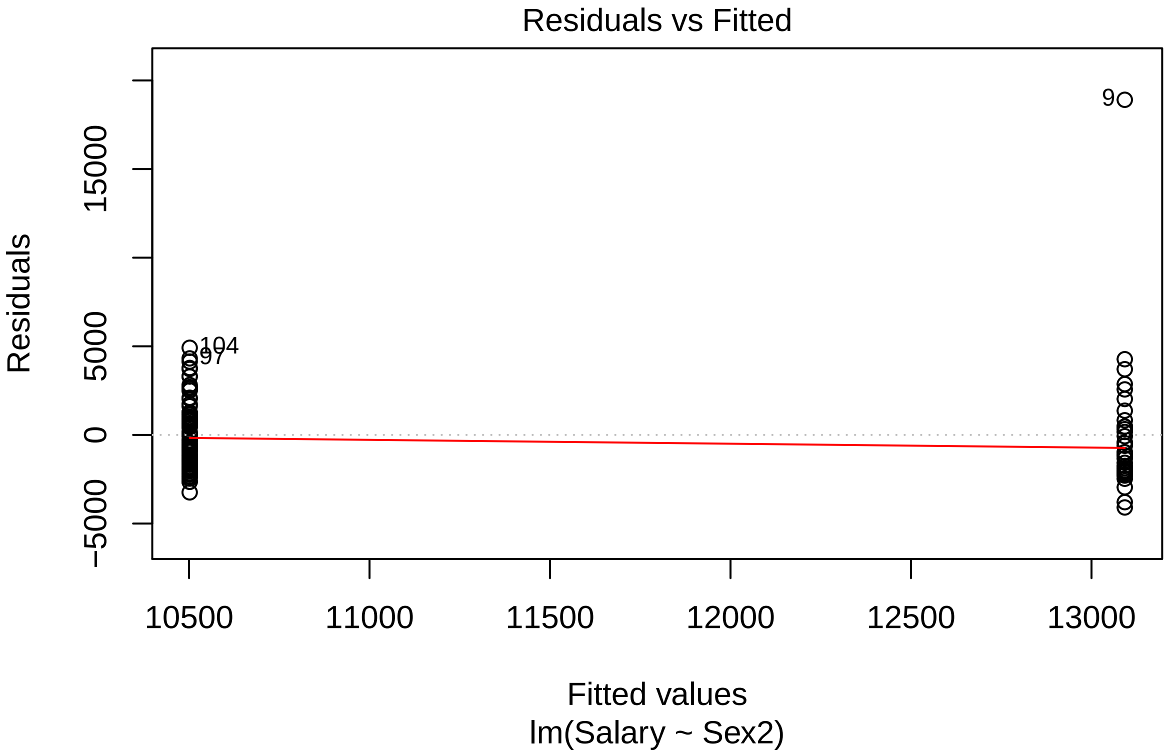
```r
cbind(Sex, Sex2)
```

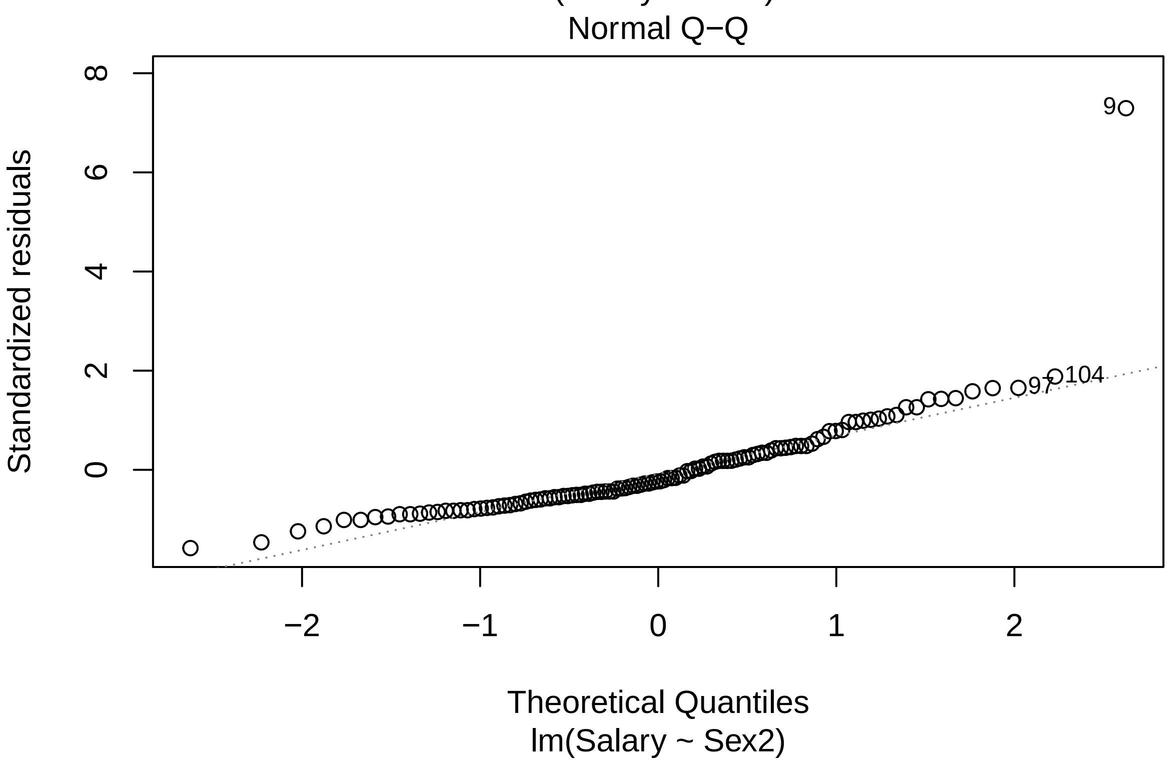
```
##        Sex Sex2
##   [1,]   0   -1
##   [2,]   0   -1
##   [3,]   0   -1
##   [4,]   0   -1
##   [5,]   0   -1
##   [6,]   0   -1
##   [7,]   0   -1
##   [8,]   0   -1
##   [9,]   0   -1
##  [10,]   0   -1
##  [11,]   0   -1
##  [12,]   0   -1
##  [13,]   0   -1
##  [14,]   0   -1
##  [15,]   0   -1
##  [16,]   0   -1
##  [17,]   0   -1
##  [18,]   0   -1
##  [19,]   0   -1
##  [20,]   0   -1
##  [21,]   0   -1
##  [22,]   0   -1
##  [23,]   0   -1
##  [24,]   0   -1
##  [25,]   0   -1
##  [26,]   0   -1
##  [27,]   0   -1
##  [28,]   0   -1
##  [29,]   0   -1
##  [30,]   0   -1
##  [31,]   0   -1
##  [32,]   0   -1
##  [33,]   0   -1
##  [34,]   0   -1
##  [35,]   0   -1
##  [36,]   1    1
##  [37,]   1    1
```

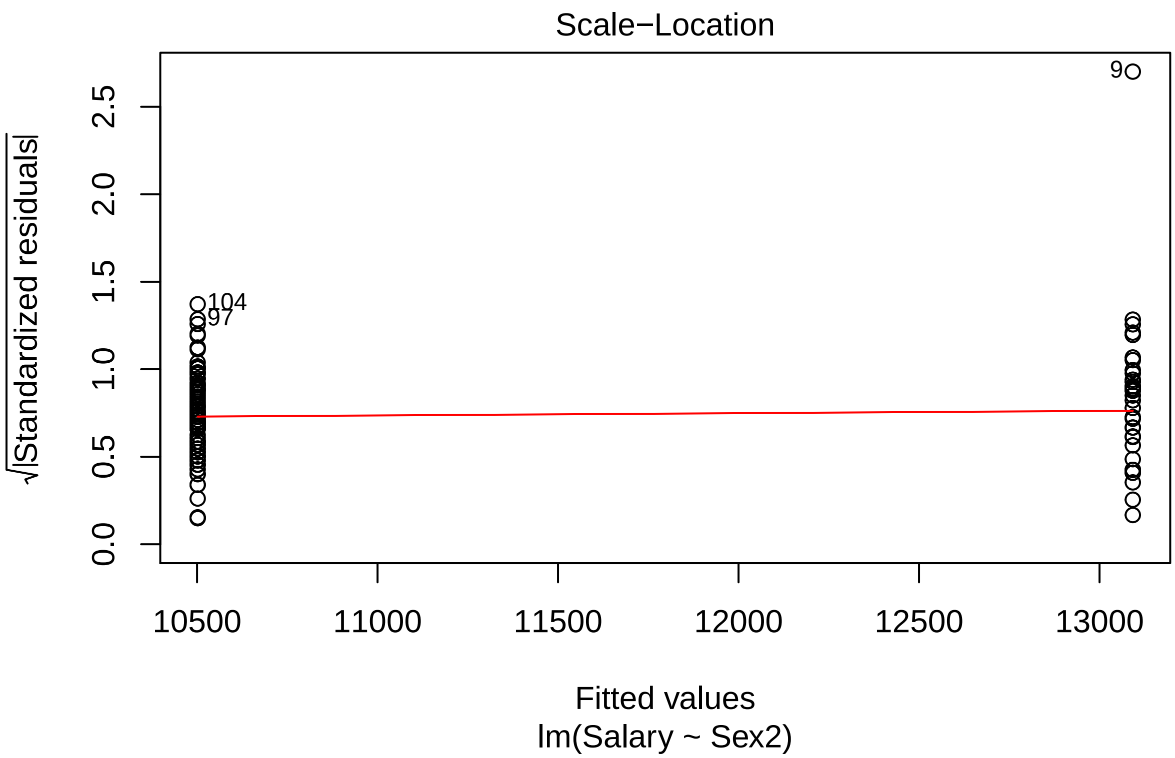
```
## [38,]    1    1
## [39,]    1    1
## [40,]    1    1
## [41,]    1    1
## [42,]    1    1
## [43,]    1    1
## [44,]    1    1
## [45,]    1    1
## [46,]    1    1
## [47,]    1    1
## [48,]    1    1
## [49,]    1    1
## [50,]    1    1
## [51,]    1    1
## [52,]    1    1
## [53,]    1    1
## [54,]    1    1
## [55,]    1    1
## [56,]    1    1
## [57,]    1    1
## [58,]    1    1
## [59,]    1    1
## [60,]    1    1
## [61,]    1    1
## [62,]    1    1
## [63,]    1    1
## [64,]    1    1
## [65,]    1    1
## [66,]    1    1
## [67,]    1    1
## [68,]    1    1
## [69,]    1    1
## [70,]    1    1
## [71,]    1    1
## [72,]    1    1
## [73,]    1    1
## [74,]    1    1
## [75,]    1    1
## [76,]    1    1
## [77,]    1    1
## [78,]    1    1
## [79,]    1    1
## [80,]    1    1
## [81,]    1    1
## [82,]    1    1
## [83,]    1    1
## [84,]    1    1
## [85,]    1    1
## [86,]    1    1
## [87,]    1    1
## [88,]    1    1
## [89,]    1    1
## [90,]    1    1
## [91,]    1    1
```

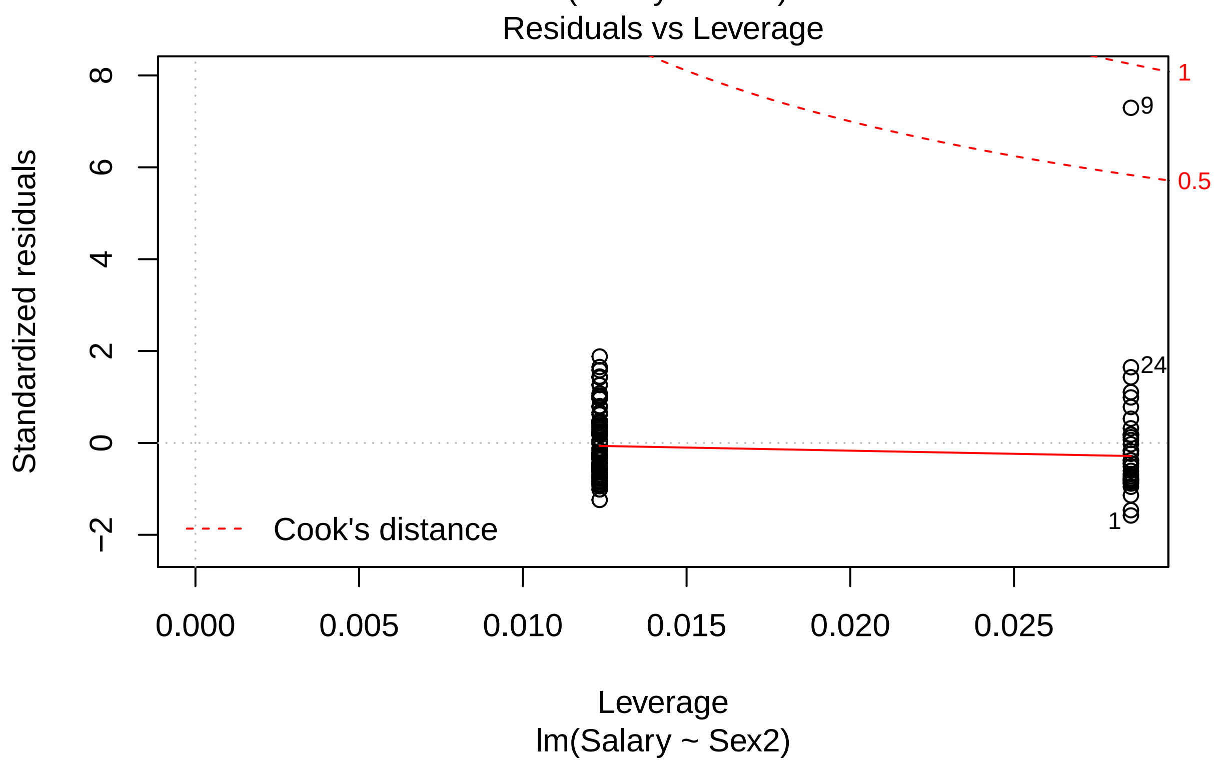
```
##  [92,]    1     1
##  [93,]    1     1
##  [94,]    1     1
##  [95,]    1     1
##  [96,]    1     1
##  [97,]    1     1
##  [98,]    1     1
##  [99,]    1     1
## [100,]    1     1
## [101,]    1     1
## [102,]    1     1
## [103,]    1     1
## [104,]    1     1
## [105,]    1     1
## [106,]    1     1
## [107,]    1     1
## [108,]    1     1
## [109,]    1     1
## [110,]    1     1
## [111,]    1     1
## [112,]    1     1
## [113,]    1     1
## [114,]    1     1
## [115,]    1     1
## [116,]    1     1
```

```r
# However, the model using this new variable gives output that is almost identical to the
# original model:

Salary.lm2 <- lm(Salary ~ Sex2)
plot(Salary.lm2)
```


Residuals vs Fitted
Residuals
15000
5000
0
−5000
9
104
97
10500
11000
11500
12000
12500
13000
Fitted values
lm(Salary ~ Sex2)


Normal Q−Q
Standardized residuals
8
6
4
2
0
9
104
97
−2
−1
0
1
2
Theoretical Quantiles
lm(Salary ~ Sex2)




```r
anova(Salary.lm2)
```

```
## Analysis of Variance Table
##
## Response: Salary
##             Df    Sum Sq   Mean Sq F value    Pr(>F)
```

```
## Sex2        1 164000725 164000725  23.714 3.63e-06 ***
## Residuals 114 788398942   6915780
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(Salary.lm2)
```

```
##
## Call:
## lm(formula = Salary ~ Sex2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4092  -1564   -607   1138  18908
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    11797        266   44.35  < 2e-16 ***
## Sex2           -1295        266   -4.87 3.63e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2630 on 114 degrees of freedom
## Multiple R-squared:  0.1722, Adjusted R-squared:  0.1649
## F-statistic: 23.71 on 1 and 114 DF,  p-value: 3.63e-06
```

```r
# The only difference is in the estimated coefficients of the model (the model parameters).
# The new intercept turns out to be, not the overall mean salary (which is what would have
# happened had if we had the same number of males and females - called a balanced design),
# but instead it is the simple (rather than weighted) average of the mean salary for males
# and the mean salary for females:

mean(Salary)
```

```
## [1] 11283.38
```

```r
(mean(Salary[Sex==0]) + mean(Salary[Sex==1]))/2
```

```
## [1] 11797
```

```r
coef(Salary.lm2)
```

```
## (Intercept)        Sex2
##   11797.003   -1295.225
```

```r
# If we add the slope coefficient to from this new intercept (i.e. when Sex2 = +1),
# we arrive at the mean salary for males:

coef(Salary.lm2)[1] - coef(Salary.lm2)[2]
```

```
## (Intercept)
##    13092.23
```

```r
mean(Salary[Sex==0])
```

```
## [1] 13092.23
```

```r
# And if we subtract the slope coefficient from the intercept (i.e. when Sex2 = +1),
# we get the mean salary for females:
```

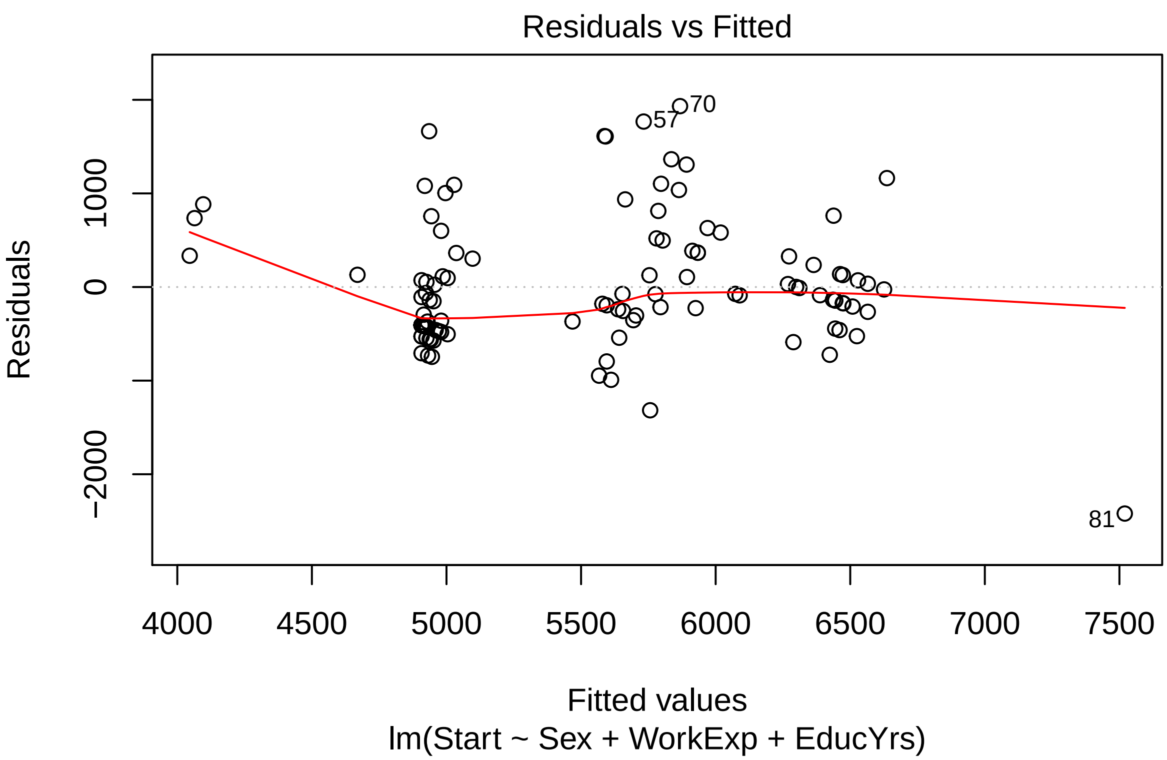
```r
coef(Salary.lm2)[1] + coef(Salary.lm2)[2]
```

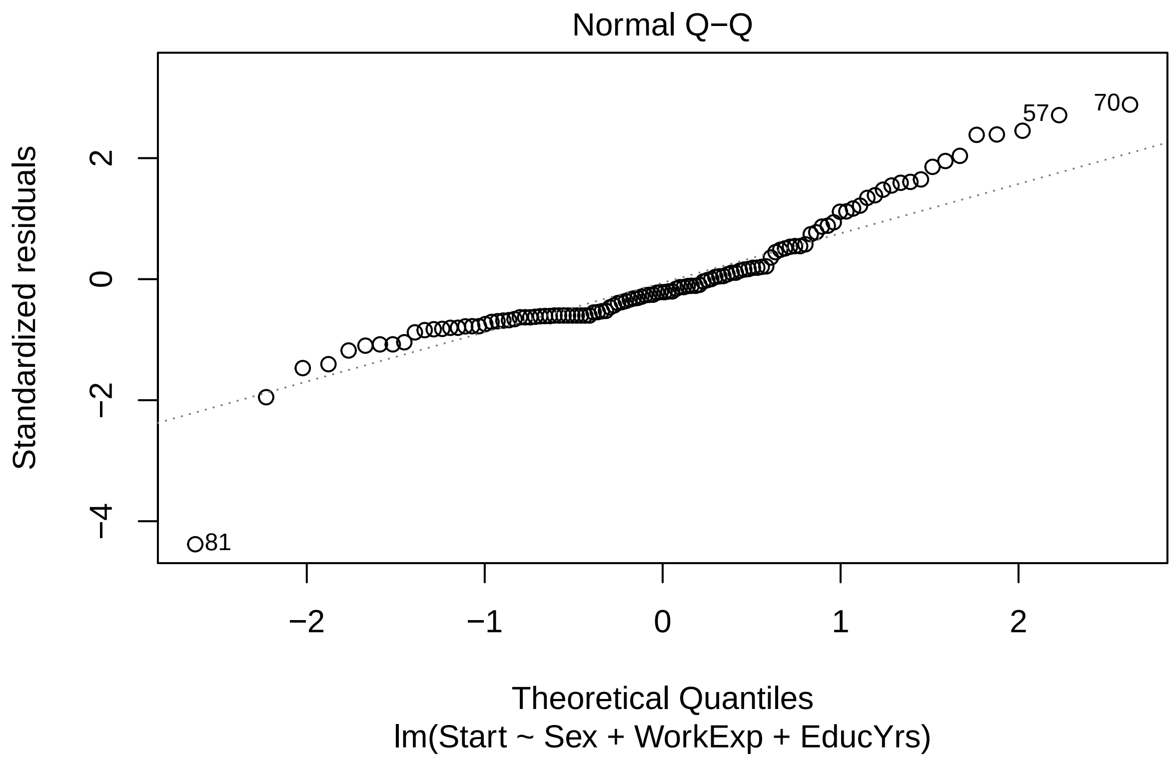
```
## (Intercept)
##    10501.78
```

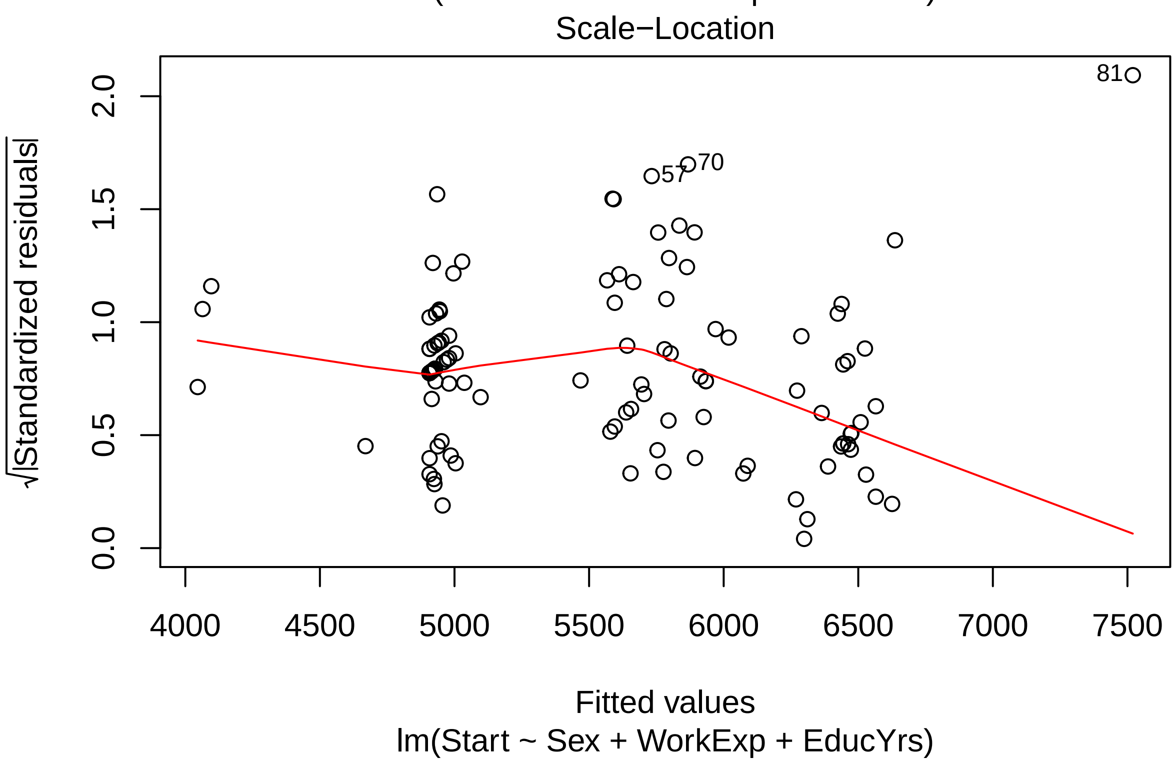
```r
mean(Salary[Sex==1])
```

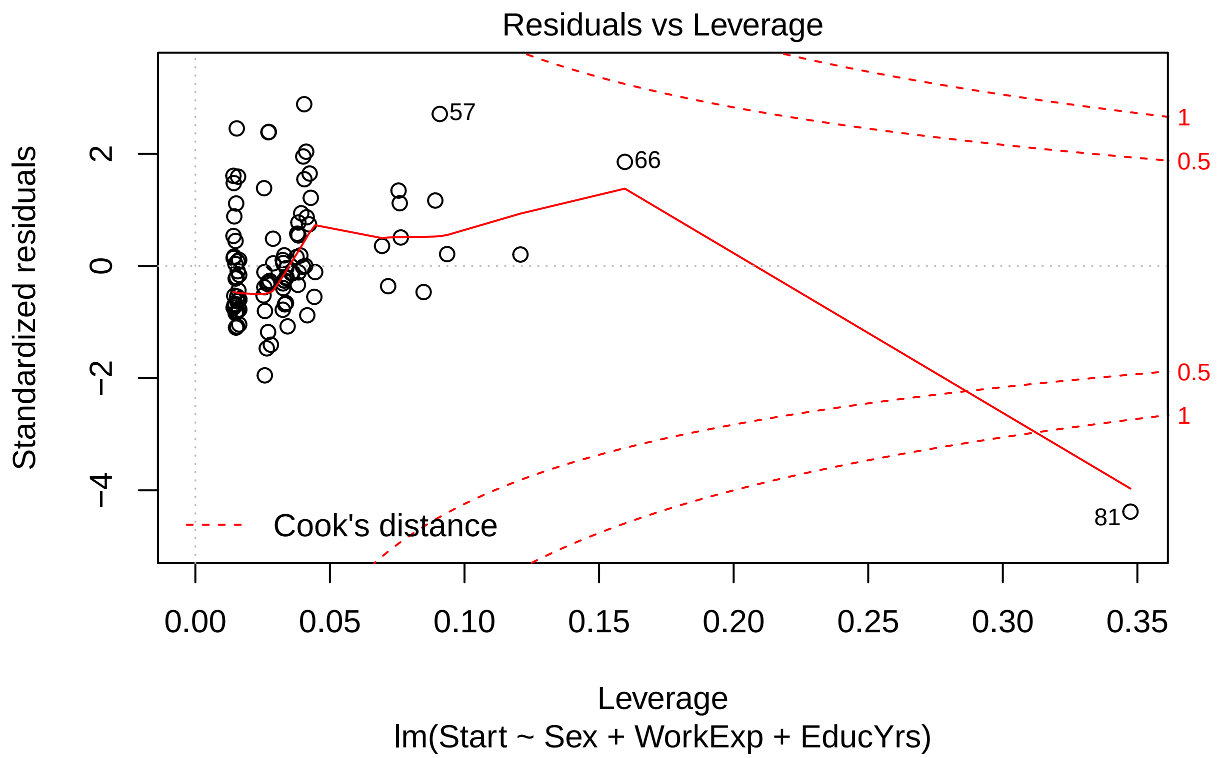
```
## [1] 10501.78
```

```r
# We will discuss these alternative ways of parameterising a model later in this part of the
# course, but before we leave this example, let's examine the models that Lattin, Carroll and
# Green fit in their solutions to this exercise. First a model for examining the differences
# in starting salaries, controlling for the effects of work experience and years of education,
# which they do by including these additional variables in the model:

Start.lm <- lm(Start ~ Sex + WorkExp + EducYrs)
plot(Start.lm)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Start ~ Sex + WorkExp + EducYrs)
57
70
81


Scale−Location
√|Standardized residuals|
Fitted values
lm(Start ~ Sex + WorkExp + EducYrs)
81
57
70




```r
anova(Start.lm)
```

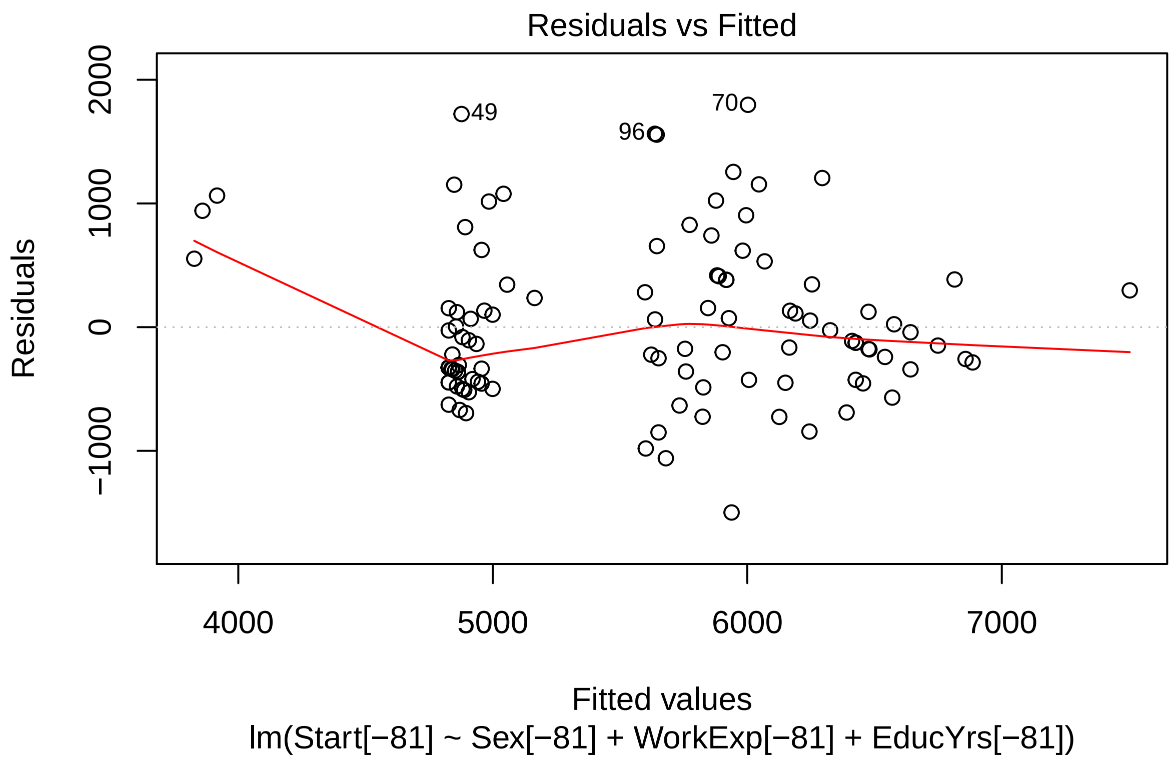
```
## Analysis of Variance Table
##
## Response: Start
##            Df   Sum Sq  Mean Sq F value    Pr(>F)
## Sex         1 27119338 27119338  57.983 9.044e-12 ***
## WorkExp     1  7395592  7395592  15.812 0.0001243 ***
## EducYrs     1 17758902 17758902  37.970 1.159e-08 ***
## Residuals 112 52383723   467712
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

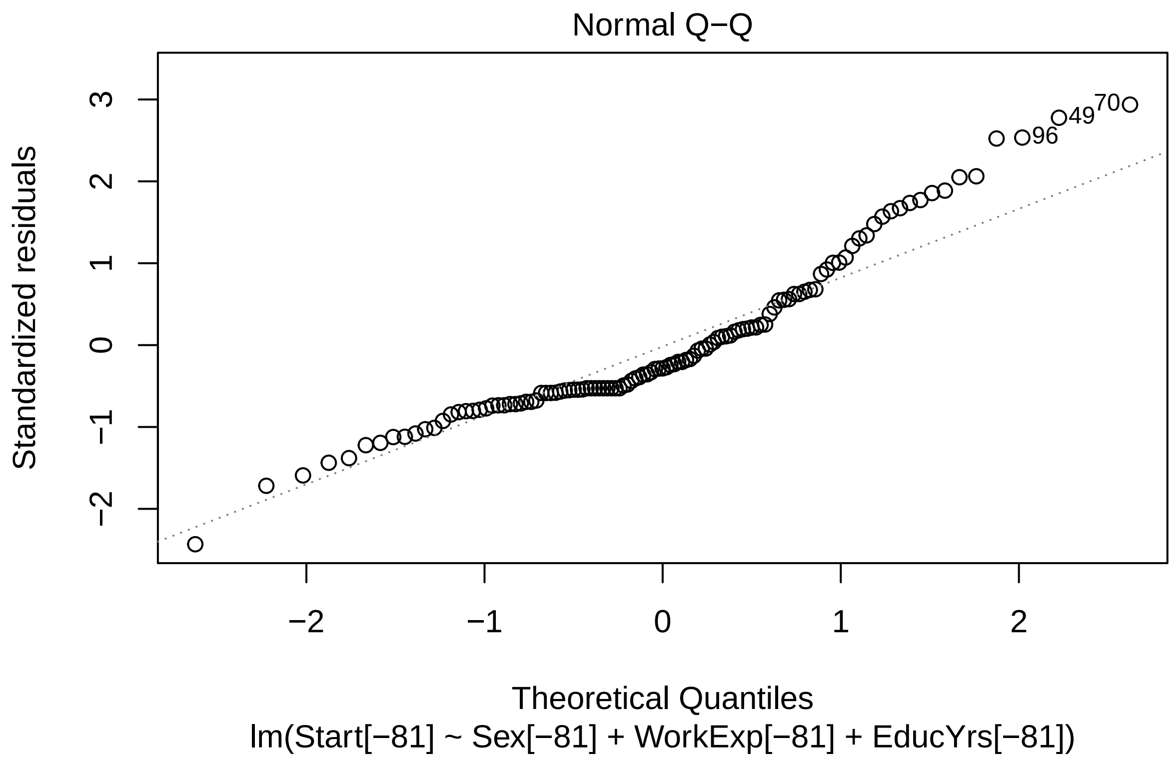
```r
summary(Start.lm)
```

```
##
## Call:
## lm(formula = Start ~ Sex + WorkExp + EducYrs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2420.0  -413.9  -144.1   329.0  1932.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2952.017    508.440   5.806 6.07e-08 ***
## Sex         -683.376    148.151  -4.613 1.06e-05 ***
## WorkExp        4.035      1.081   3.734 0.000298 ***
## EducYrs      219.896     35.686   6.162 1.16e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

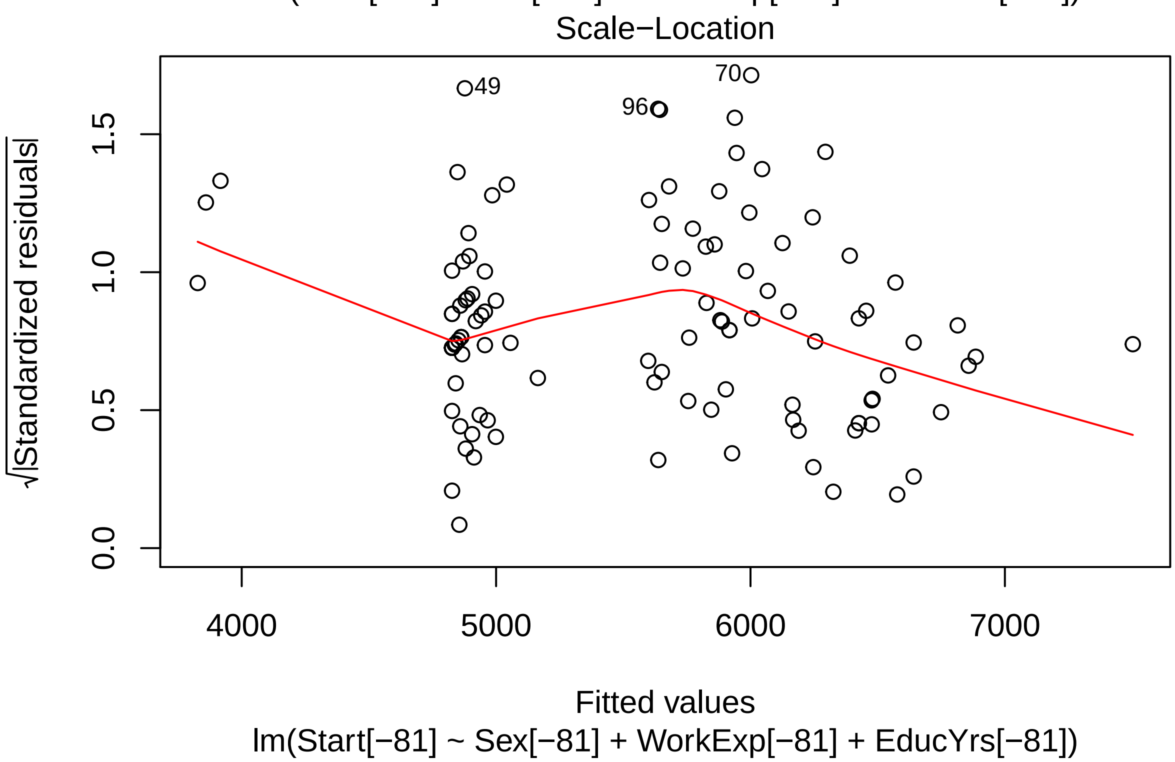
```
##
## Residual standard error: 683.9 on 112 degrees of freedom
## Multiple R-squared:  0.4995, Adjusted R-squared:  0.4861
## F-statistic: 37.26 on 3 and 112 DF,  p-value: < 2.2e-16
```

```
# There are problems with this model, notably that observation #81, a female who turns out to
# be the only college graduate out of all 116 employees, appears to have a large negative
# residual - i.e. she has a much lower starting salary than her additional education would
# suggest. We could treat her as a special case and exclude her from the analysis (one way of
# dealing with potential outliers, but not necessarily the best way in this instance):

Start.lm2 <- lm(Start[-81] ~ Sex[-81] + WorkExp[-81] + EducYrs[-81])
plot(Start.lm2)
```

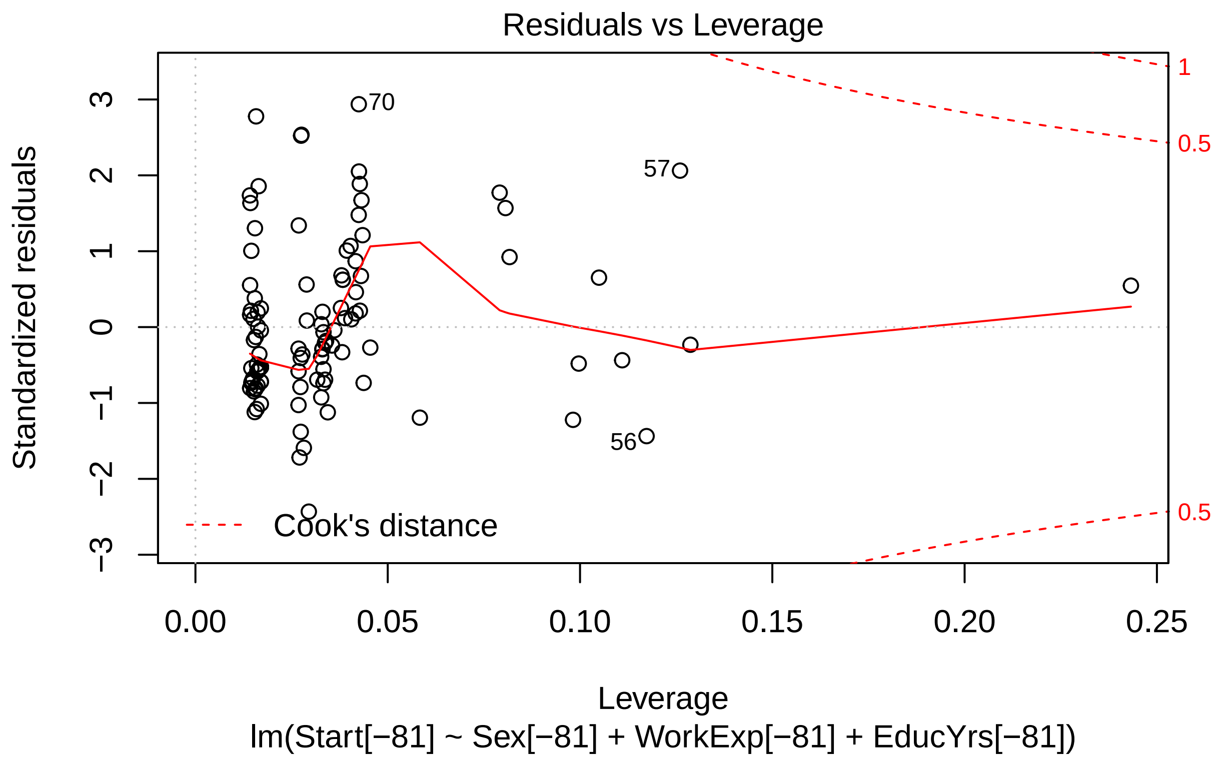




Normal Q−Q
Standardized residuals
−2
−1
0
1
2
3
70
49
96
Theoretical Quantiles
lm(Start[−81] ~ Sex[−81] + WorkExp[−81] + EducYrs[−81])

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
49
96
70
4000
5000
6000
7000
Fitted values
lm(Start[−81] ~ Sex[−81] + WorkExp[−81] + EducYrs[−81])




```r
anova(Start.lm2)
```

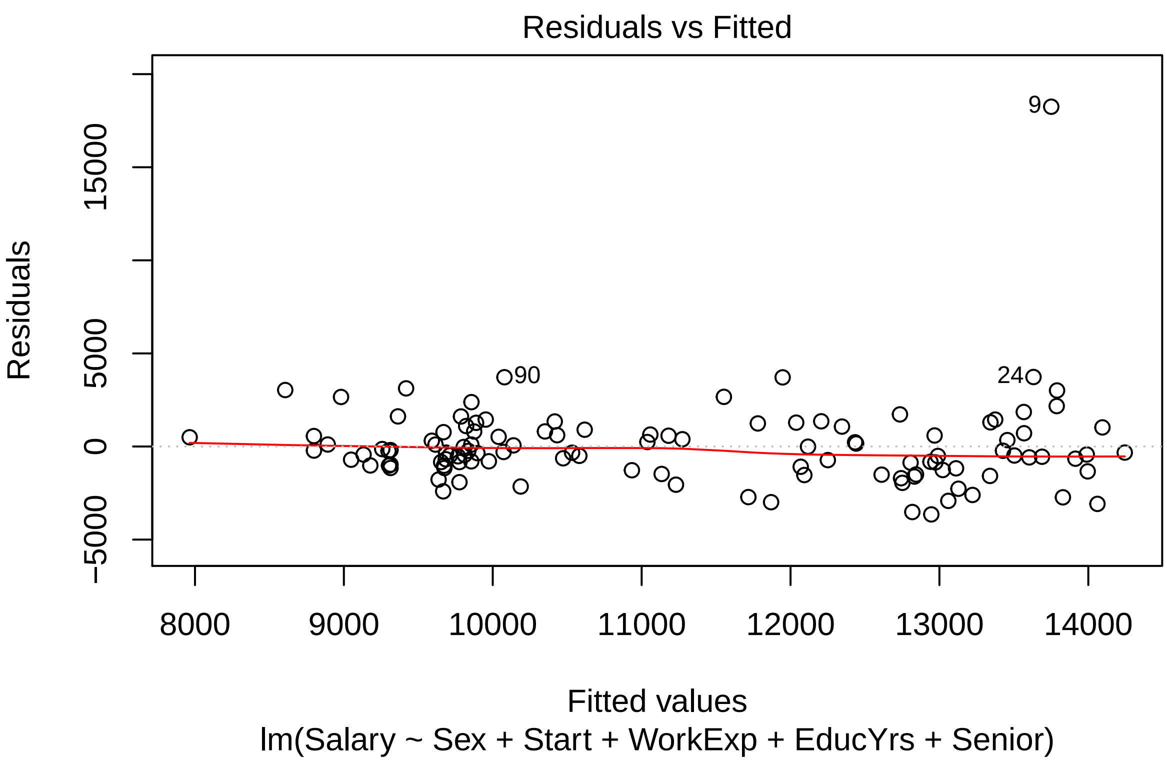
```
## Analysis of Variance Table
##
## Response: Start[-81]
##                Df   Sum Sq  Mean Sq F value    Pr(>F)
## Sex[-81]        1 26947666 26947666  68.906 2.761e-13 ***
## WorkExp[-81]    1 11034978 11034978  28.217 5.642e-07 ***
## EducYrs[-81]    1 23081686 23081686  59.021 6.663e-12 ***
## Residuals     111 43409548   391077
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

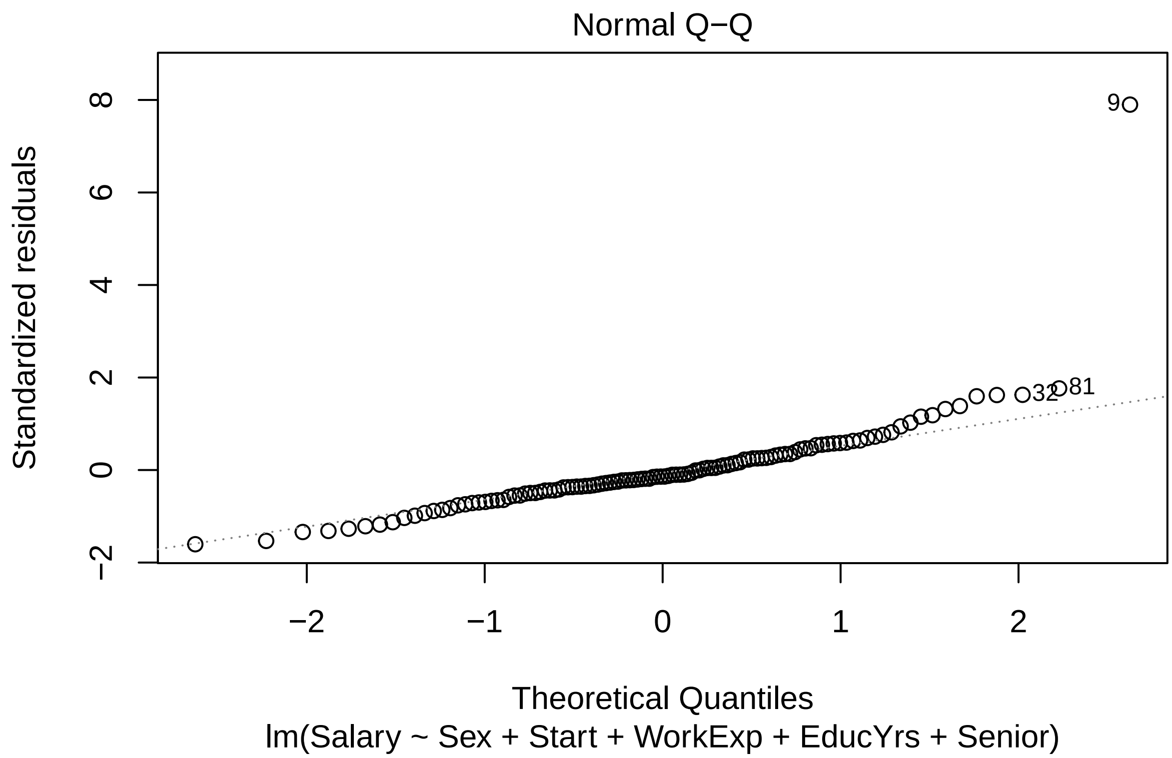
```r
summary(Start.lm2)
```

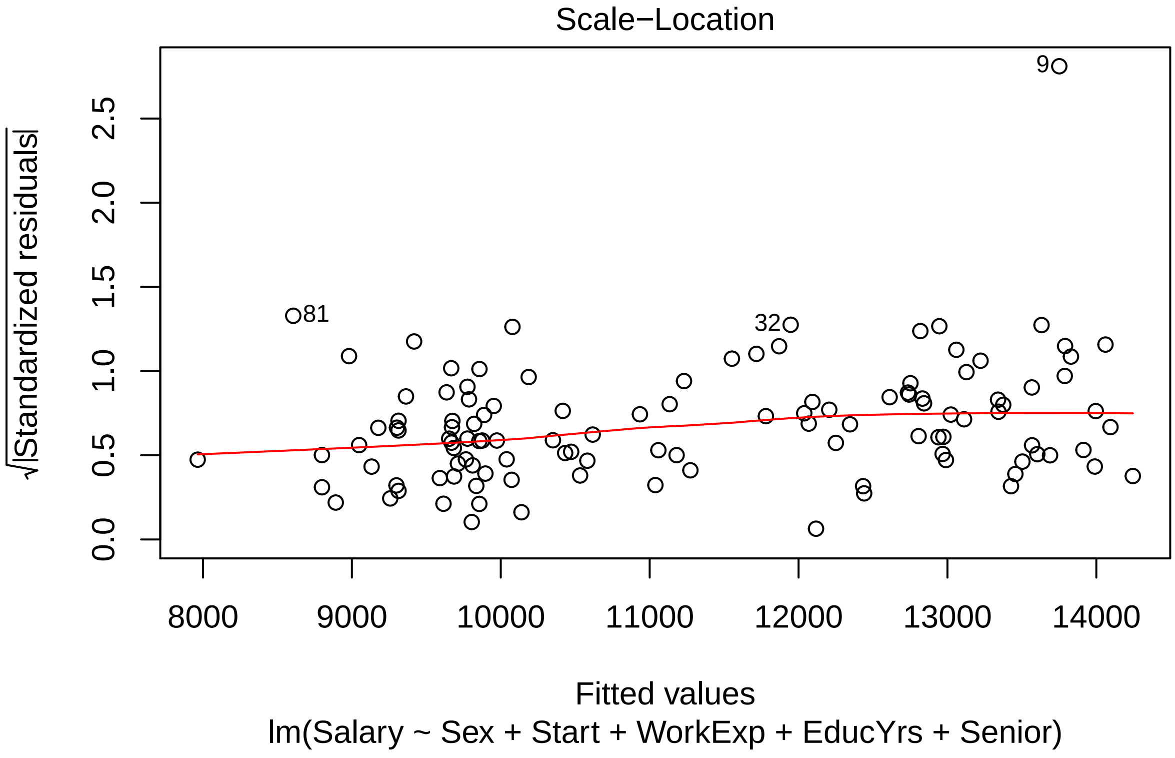
```
##
## Call:
## lm(formula = Start[-81] ~ Sex[-81] + WorkExp[-81] + EducYrs[-81])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1498.3  -362.7  -175.3   320.3  1797.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2210.749    489.999   4.512 1.61e-05 ***
## Sex[-81]     -480.568    141.932  -3.386 0.000982 ***
## WorkExp[-81]    7.177      1.186   6.051 1.99e-08 ***
## EducYrs[-81]  258.055     33.590   7.683 6.66e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

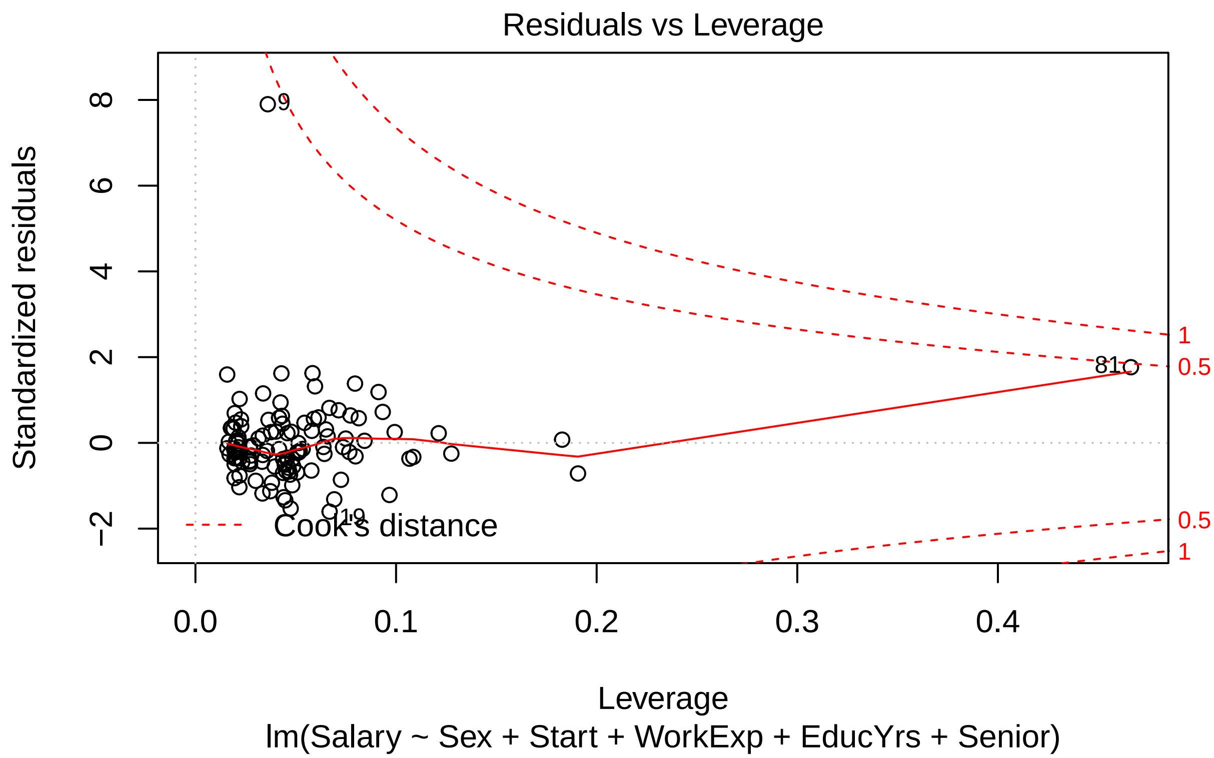
```
##
## Residual standard error: 625.4 on 111 degrees of freedom
## Multiple R-squared:  0.5845, Adjusted R-squared:  0.5733
## F-statistic: 52.05 on 3 and 111 DF,  p-value: < 2.2e-16
```

```
# Compare the ANOVA tables and estimated coefficients of Start.lm and Start.lm2 - both models
# suggest there is significant discrimination against females in starting salaries, controlling
# for work experience and years of education, but the size of the mean difference in starting
# salaries, b(Sex) = -$683 in Start.lm compared to b(Sex) = -$481 in Start.lm2, suggests that
# observation 81 has had a significant effect on this difference!

# Finally, Lattin, Carroll and Green proposed this model for current salaries:

Salary.lm3 <- lm(Salary ~ Sex + Start + WorkExp + EducYrs + Senior)
plot(Salary.lm3)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Salary ~ Sex + Start + WorkExp + EducYrs + Senior)
9
32
81
−2
0
2
4
6
8
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
Fitted values
lm(Salary ~ Sex + Start + WorkExp + EducYrs + Senior)
9
81
32
0.0
0.5
1.0
1.5
2.0
2.5
8000
9000
10000
11000
12000
13000
14000




```r
anova(Salary.lm3)
```

```
## Analysis of Variance Table
##
## Response: Salary
##            Df    Sum Sq   Mean Sq F value    Pr(>F)
## Sex         1 164000725 164000725 29.6240 3.216e-07 ***
## Start       1 116820146 116820146 21.1016 1.166e-05 ***
## WorkExp     1  23570694  23570694  4.2577   0.04143 *
## EducYrs     1  17697856  17697856  3.1968   0.07653 .
## Senior      1  21341708  21341708  3.8550   0.05212 .
## Residuals 110 608968538   5536078
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary(Salary.lm3)
```

```
##
## Call:
## lm(formula = Salary ~ Sex + Start + WorkExp + EducYrs + Senior)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3645.7 -1046.5  -325.7   779.3 18248.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -225.5811  2615.8534  -0.086   0.9314
## Sex         -1337.9169   558.2854  -2.396   0.0182 *
## Start           1.1972     0.3278   3.652   0.0004 ***
## WorkExp        -6.9152     3.9437  -1.753   0.0823 .
```

```
## EducYrs       197.3786   144.9695   1.362   0.1761
## Senior         45.2260    23.0343   1.963   0.0521 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2353 on 110 degrees of freedom
## Multiple R-squared:  0.3606, Adjusted R-squared:  0.3315
## F-statistic: 12.41 on 5 and 110 DF,  p-value: 1.497e-09
```

```
# This model also has problems with two potential outliers: observations #9 (a male with a very
# large current salary) and #81 (again). We will discuss how to deal with outliers (again),
# later in the course.
```