STAT 8020 R Lab 9: Multiple Linear Regression V

Whitney

September 16, 2020

Contents

Species diversity on the Galapagos Islands

```
#install.packages("faraway")
library(faraway)
data(gala)
gala
```

```
##              Species Endemics    Area Elevation Nearest Scruz Adjacent
## Baltra            58       23   25.09       346     0.6   0.6     1.84
## Bartolome         31       21    1.24       109     0.6  26.3   572.33
## Caldwell           3        3    0.21       114     2.8  58.7     0.78
## Champion          25        9    0.10        46     1.9  47.4     0.18
## Coamano            2        1    0.05        77     1.9   1.9   903.82
## Daphne.Major      18       11    0.34       119     8.0   8.0     1.84
## Daphne.Minor      24        0    0.08        93     6.0  12.0     0.34
## Darwin            10        7    2.33       168    34.1 290.2     2.85
## Eden               8        4    0.03        71     0.4   0.4    17.95
## Enderby            2        2    0.18       112     2.6  50.2     0.10
## Espanola          97       26   58.27       198     1.1  88.3     0.57
## Fernandina        93       35  634.49      1494     4.3  95.3  4669.32
## Gardner1          58       17    0.57        49     1.1  93.1    58.27
## Gardner2           5        4    0.78       227     4.6  62.2     0.21
## Genovesa          40       19   17.35        76    47.4  92.2   129.49
## Isabela          347       89 4669.32      1707     0.7  28.1   634.49
## Marchena          51       23  129.49       343    29.1  85.9    59.56
## Onslow             2        2    0.01        25     3.3  45.9     0.10
## Pinta            104       37   59.56       777    29.1 119.6   129.49
## Pinzon           108       33   17.95       458    10.7  10.7     0.03
## Las.Plazas        12        9    0.23        94     0.5   0.6    25.09
## Rabida            70       30    4.89       367     4.4  24.4   572.33
## SanCristobal     280       65  551.62       716    45.2  66.6     0.57
## SanSalvador      237       81  572.33       906     0.2  19.8     4.89
## SantaCruz        444       95  903.82       864     0.6   0.0     0.52
## SantaFe           62       28   24.08       259    16.5  16.5     0.52
## SantaMaria       285       73  170.92       640     2.6  49.2     0.10
```

```
## Seymour           44       16    1.84       147     0.6   9.6    25.09
## Tortuga           16        8    1.24       186     6.8  50.9    17.95
## Wolf              21       12    2.85       253    34.1 254.7     2.33
```

```
galaNew <- gala[, -2]
galaNew
```

```
##              Species    Area Elevation Nearest Scruz Adjacent
## Baltra            58   25.09       346     0.6   0.6     1.84
## Bartolome         31    1.24       109     0.6  26.3   572.33
## Caldwell           3    0.21       114     2.8  58.7     0.78
## Champion          25    0.10        46     1.9  47.4     0.18
## Coamano            2    0.05        77     1.9   1.9   903.82
## Daphne.Major      18    0.34       119     8.0   8.0     1.84
## Daphne.Minor      24    0.08        93     6.0  12.0     0.34
## Darwin            10    2.33       168    34.1 290.2     2.85
## Eden               8    0.03        71     0.4   0.4    17.95
## Enderby            2    0.18       112     2.6  50.2     0.10
## Espanola          97   58.27       198     1.1  88.3     0.57
## Fernandina        93  634.49      1494     4.3  95.3  4669.32
## Gardner1          58    0.57        49     1.1  93.1    58.27
## Gardner2           5    0.78       227     4.6  62.2     0.21
## Genovesa          40   17.35        76    47.4  92.2   129.49
## Isabela          347 4669.32      1707     0.7  28.1   634.49
## Marchena          51  129.49       343    29.1  85.9    59.56
## Onslow             2    0.01        25     3.3  45.9     0.10
## Pinta            104   59.56       777    29.1 119.6   129.49
## Pinzon           108   17.95       458    10.7  10.7     0.03
## Las.Plazas        12    0.23        94     0.5   0.6    25.09
## Rabida            70    4.89       367     4.4  24.4   572.33
## SanCristobal     280  551.62       716    45.2  66.6     0.57
## SanSalvador      237  572.33       906     0.2  19.8     4.89
## SantaCruz        444  903.82       864     0.6   0.0     0.52
## SantaFe           62   24.08       259    16.5  16.5     0.52
## SantaMaria       285  170.92       640     2.6  49.2     0.10
## Seymour           44    1.84       147     0.6   9.6    25.09
## Tortuga           16    1.24       186     6.8  50.9    17.95
## Wolf              21    2.85       253    34.1 254.7     2.33
```

Diagnostics in Multiple Linear Regression

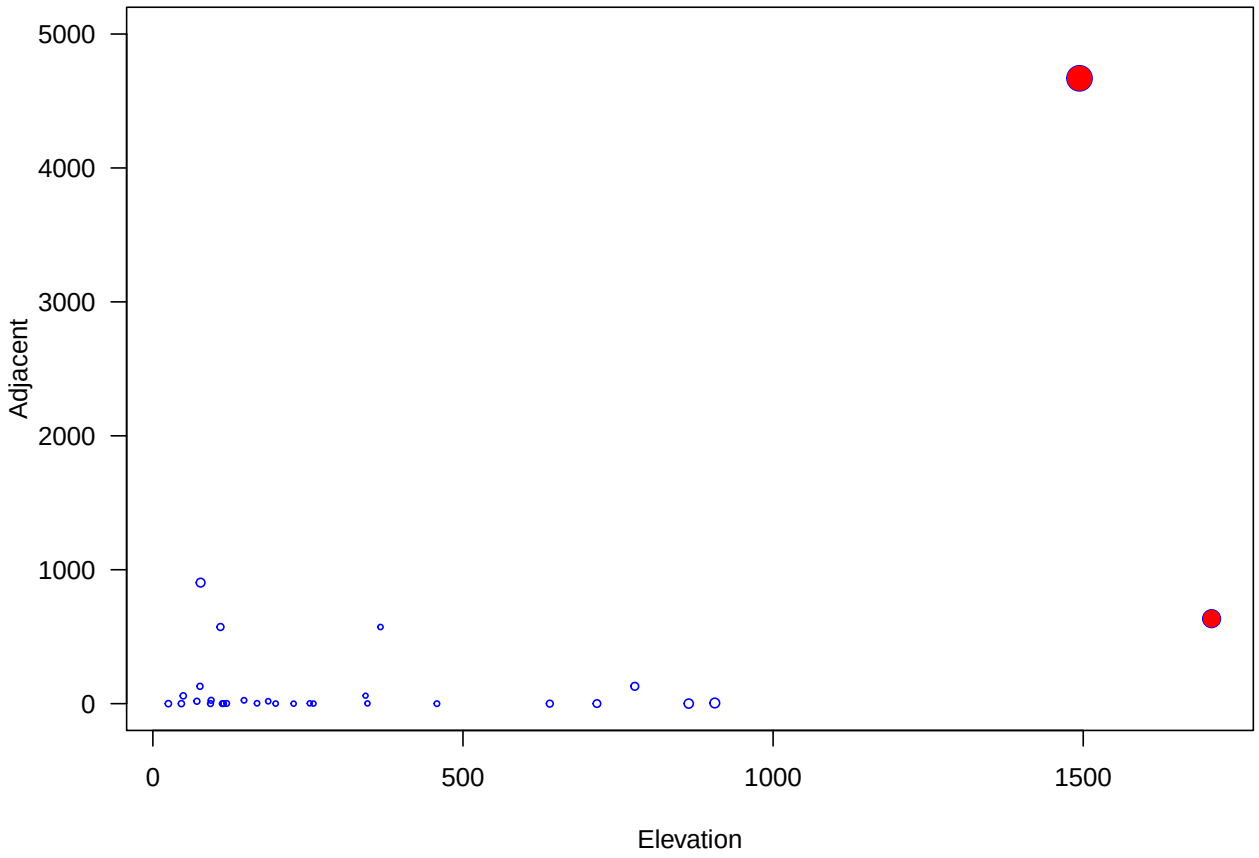
Leverage

```
full <- lm(Species ~ ., data = galaNew)
step_gala <- step(full)
```

```
## Start:  AIC=251.93
## Species ~ Area + Elevation + Nearest + Scruz + Adjacent
##
##             Df Sum of Sq    RSS    AIC
## - Nearest    1         0  89232 249.93
## - Area       1      4238  93469 251.33
## - Scruz      1      4636  93867 251.45
## <none>                    89231 251.93
## - Adjacent   1     66406 155638 266.62
```

```
## - Elevation  1    131767 220998 277.14
##
## Step:  AIC=249.93
## Species ~ Area + Elevation + Scruz + Adjacent
##
##             Df Sum of Sq    RSS    AIC
## - Area       1      4436  93667 249.39
## <none>                    89232 249.93
## - Scruz      1      7544  96776 250.37
## - Adjacent   1     72312 161544 265.74
## - Elevation  1    139445 228677 276.17
##
## Step:  AIC=249.39
## Species ~ Elevation + Scruz + Adjacent
##
##             Df Sum of Sq    RSS    AIC
## - Scruz      1      6336 100003 249.35
## <none>                    93667 249.39
## - Adjacent   1     69860 163527 264.11
## - Elevation  1    275784 369451 288.56
##
## Step:  AIC=249.35
## Species ~ Elevation + Adjacent
##
##             Df Sum of Sq    RSS    AIC
## <none>                   100003 249.35
## - Adjacent   1     73251 173254 263.84
## - Elevation  1    280817 380820 287.47
```

```
X <- model.matrix(step_gala)
H <- X %*% solve((t(X) %*% X)) %*% t(X)
lev <- hat(X)
high_lev <- which(lev >= 2 * 3 / 30)
attach(gala)

par(las = 1)
plot(Elevation, Adjacent,
     cex = sqrt(5 * lev),
     col = "blue", ylim = c(0, 5000))
points(Elevation[high_lev],
       Adjacent[high_lev], col = "red",
       pch = 16,
       cex = sqrt(5 *lev[high_lev]))
```

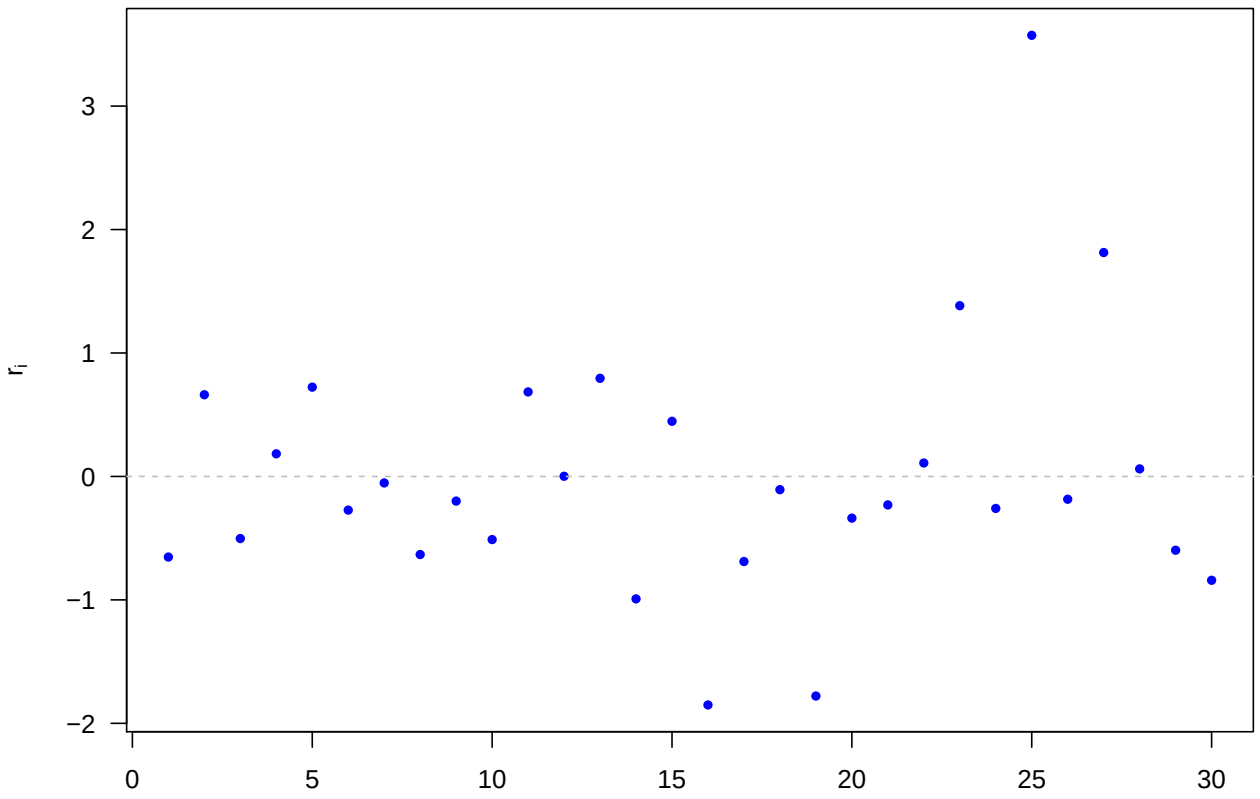


Studentized Residuals

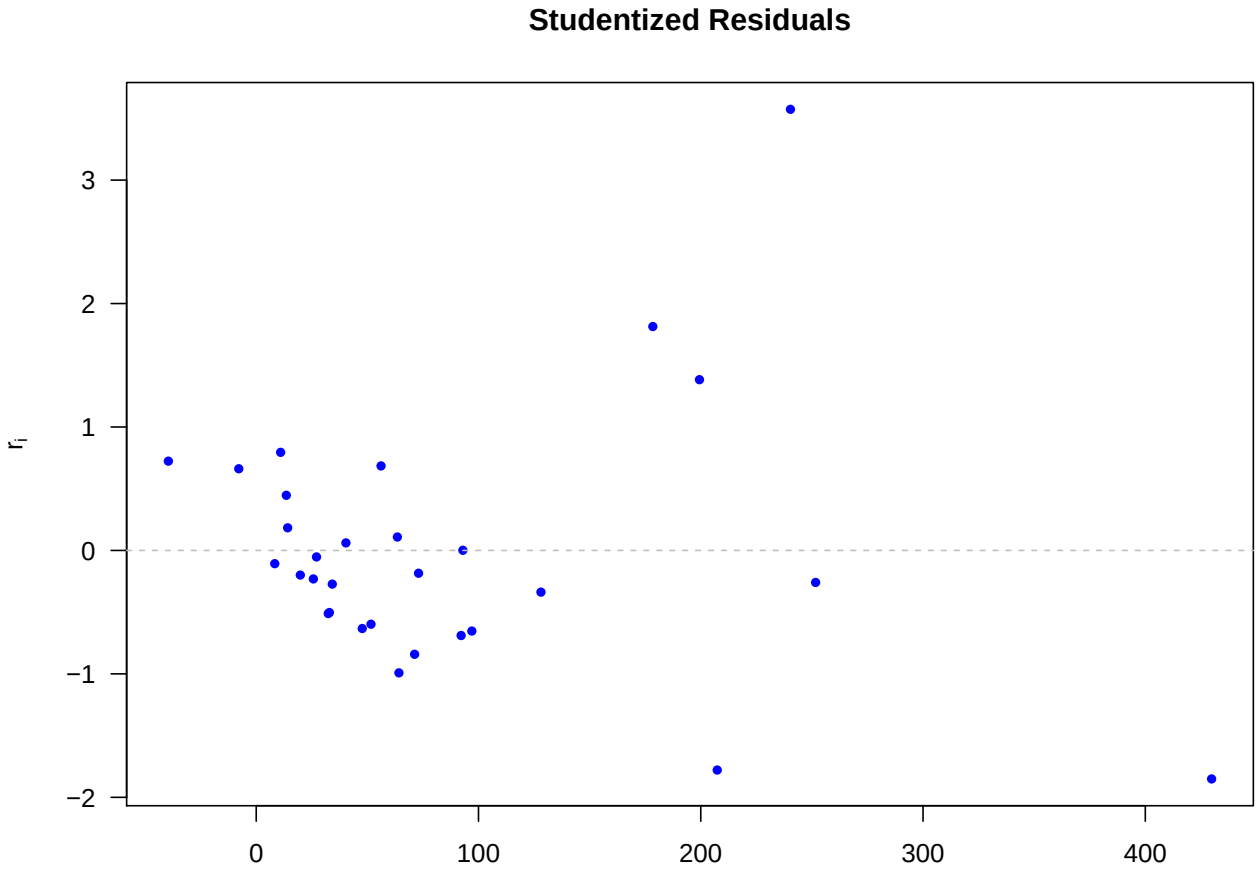
```
gs <- summary(step_gala)
gs$sig
```

```
## [1] 60.85898
```

```
studRes <- gs$res / (gs$sig * sqrt(1 - lev))

par(las = 1)
plot(studRes, pch = 16,
     cex = 0.8, col = "blue",
     ylab = expression(r[i]), main = "Studentized Residuals", xlab = "")
abline(h = 0, lty = 2, col = "gray")
```

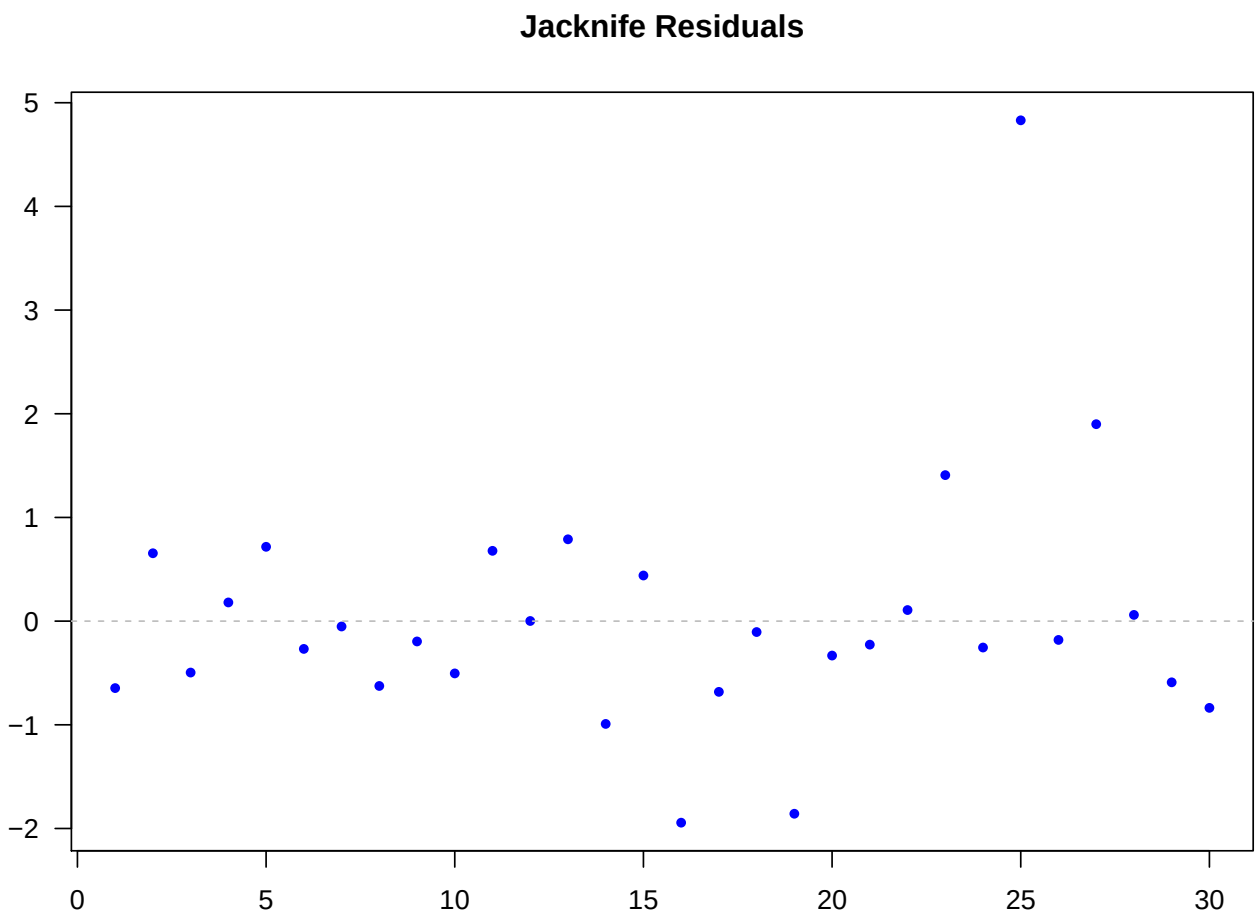



```
par(las = 1)
plot(step_gala$fitted.values, studRes, pch = 16,
     cex = 0.8, col = "blue",
     ylab = expression(r[i]), main = "Studentized Residuals", xlab = "")
abline(h = 0, lty = 2, col = "gray")
```



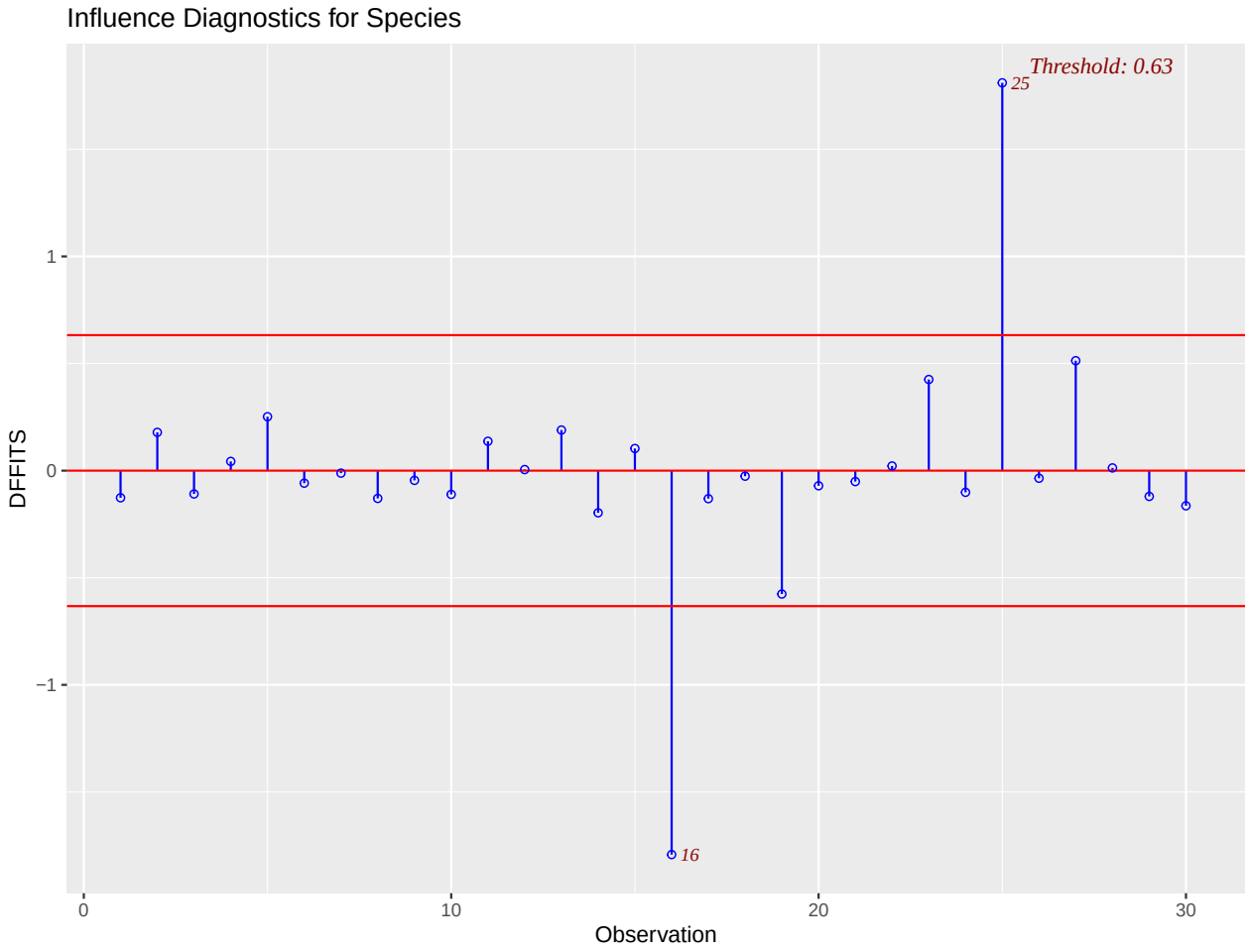
Jackknife Residuals

```
jack <- rstudent(step_gala)

par(las = 1)
plot(jack, pch = 16,
     cex = 0.8, col = "blue", main =" Jacknife Residuals ", xlab = "",
     ylab = "")
abline(h = 0, lty = 2, col = "gray")
```



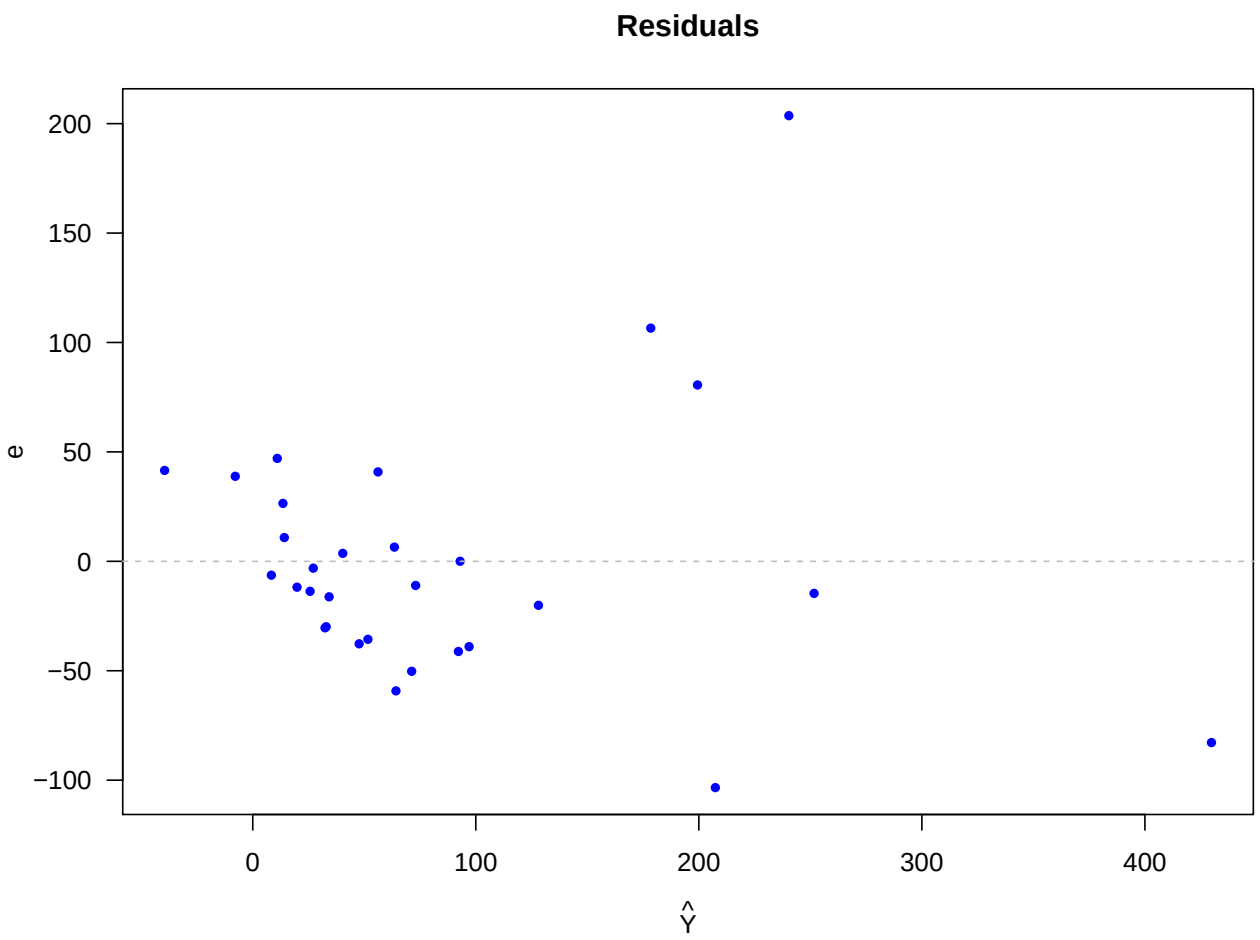
Identifying Influential Observations: DFFITS

```
library(olsrr)
ols_plot_dffits(step_gala)
```

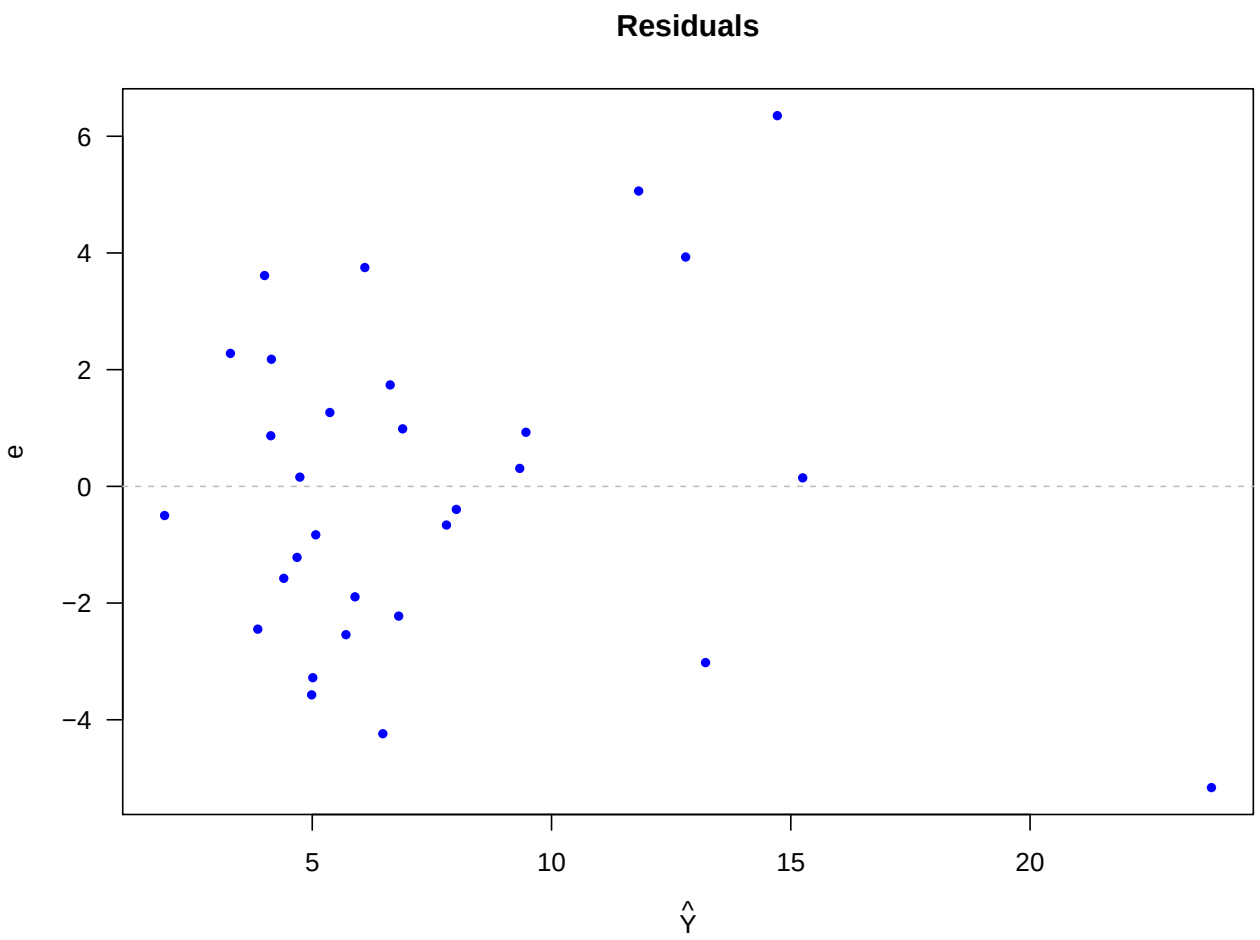


Residual Plot

```
par(las = 1)
plot(step_gala$fitted.values,
     step_gala$residuals,
     pch = 16, cex = 0.8, col = "blue", main =" Residuals ",
     xlab = expression(hat(Y)), ylab = expression(e))
abline(h = 0, lty = 2, col = "gray")
```




```r
sqrt_fit <- lm(sqrt(Species) ~ Elevation + Adjacent)

par(las = 1)
plot(sqrt_fit$fitted.values,
     sqrt_fit$residuals,
     pch = 16, cex = 0.8, col = "blue", main =" Residuals ",
     xlab = expression(hat(Y)), ylab = expression(e))
abline(h = 0, lty = 2, col = "gray")
```



Regression with Both Quantitative and Qualitative Predictors: Salaries for Professors Data Set

The 2008-09 nine-month academic salary for Assistant Professors, Associate Professors and Professors in a college in the U.S. The data were collected as part of the on-going effort of the college's administration to monitor salary differences between male and female faculty members.

```
library(carData)
```

```
## Warning: package 'carData' was built under R version 3.6.2
```

```
data("Salaries")
head(Salaries)
```

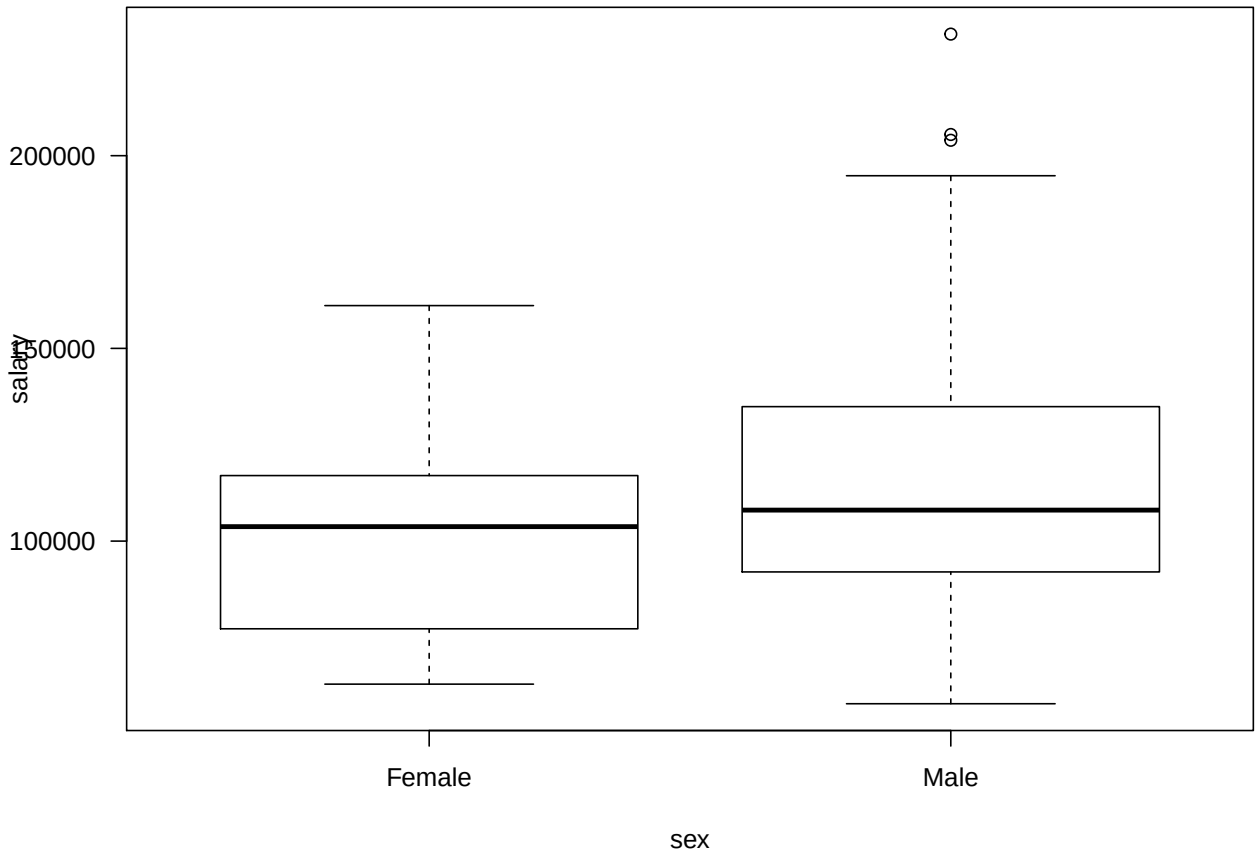
```
##        rank discipline yrs.since.phd yrs.service  sex salary
## 1      Prof          B            19          18 Male 139750
## 2      Prof          B            20          16 Male 173200
## 3  AsstProf          B             4           3 Male  79750
## 4      Prof          B            45          39 Male 115000
## 5      Prof          B            40          41 Male 141500
## 6 AssocProf          B             6           6 Male  97000
```

```
summary(Salaries)
```

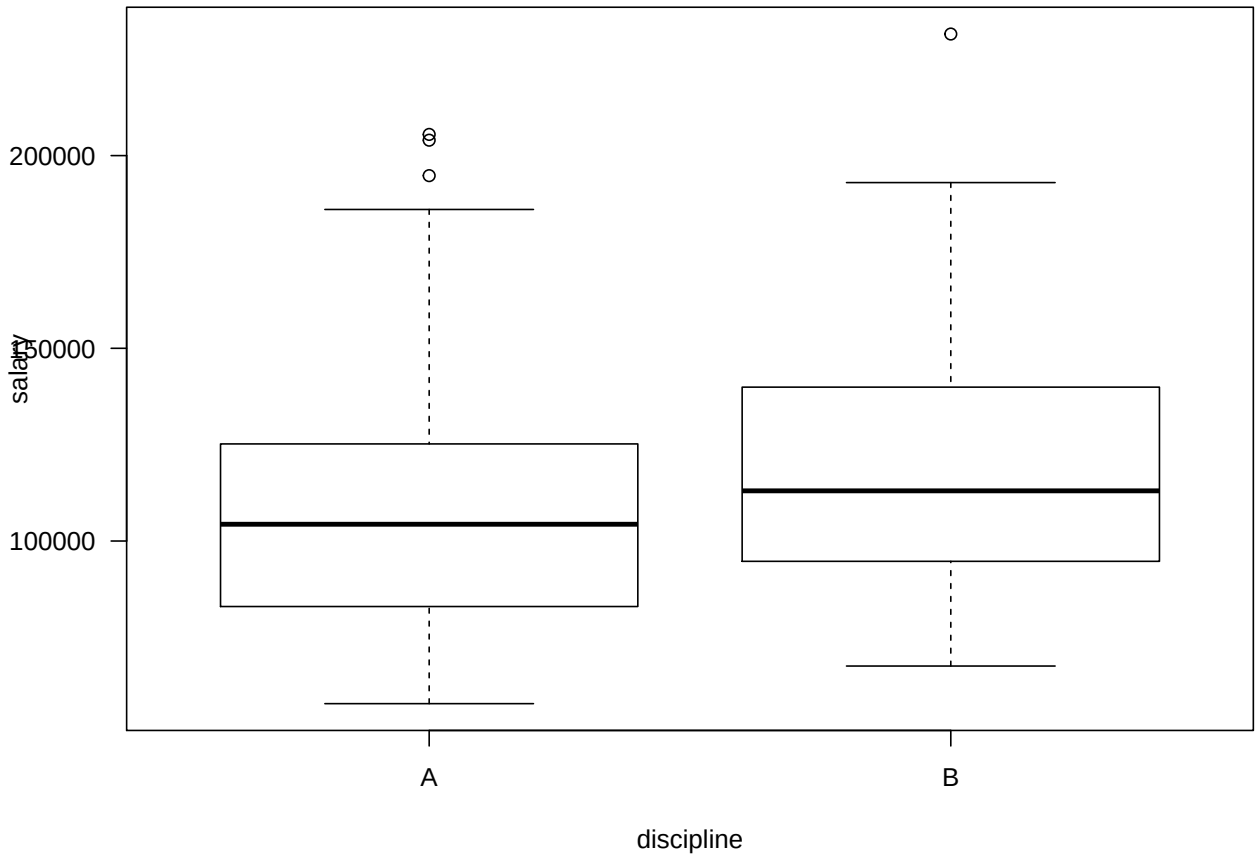
```
##         rank     discipline yrs.since.phd    yrs.service        sex
##  AsstProf : 67   A:181      Min.   : 1.00   Min.   : 0.00   Female: 39
```

```
##  AssocProf: 64   B:216      1st Qu.:12.00   1st Qu.: 7.00   Male  :358
##  Prof     :266             Median :21.00   Median :16.00
##                            Mean   :22.31   Mean   :17.61
##                            3rd Qu.:32.00   3rd Qu.:27.00
##                            Max.   :56.00   Max.   :60.00
##      salary
##  Min.   : 57800
##  1st Qu.: 91000
##  Median :107300
##  Mean   :113706
##  3rd Qu.:134185
##  Max.   :231545
```

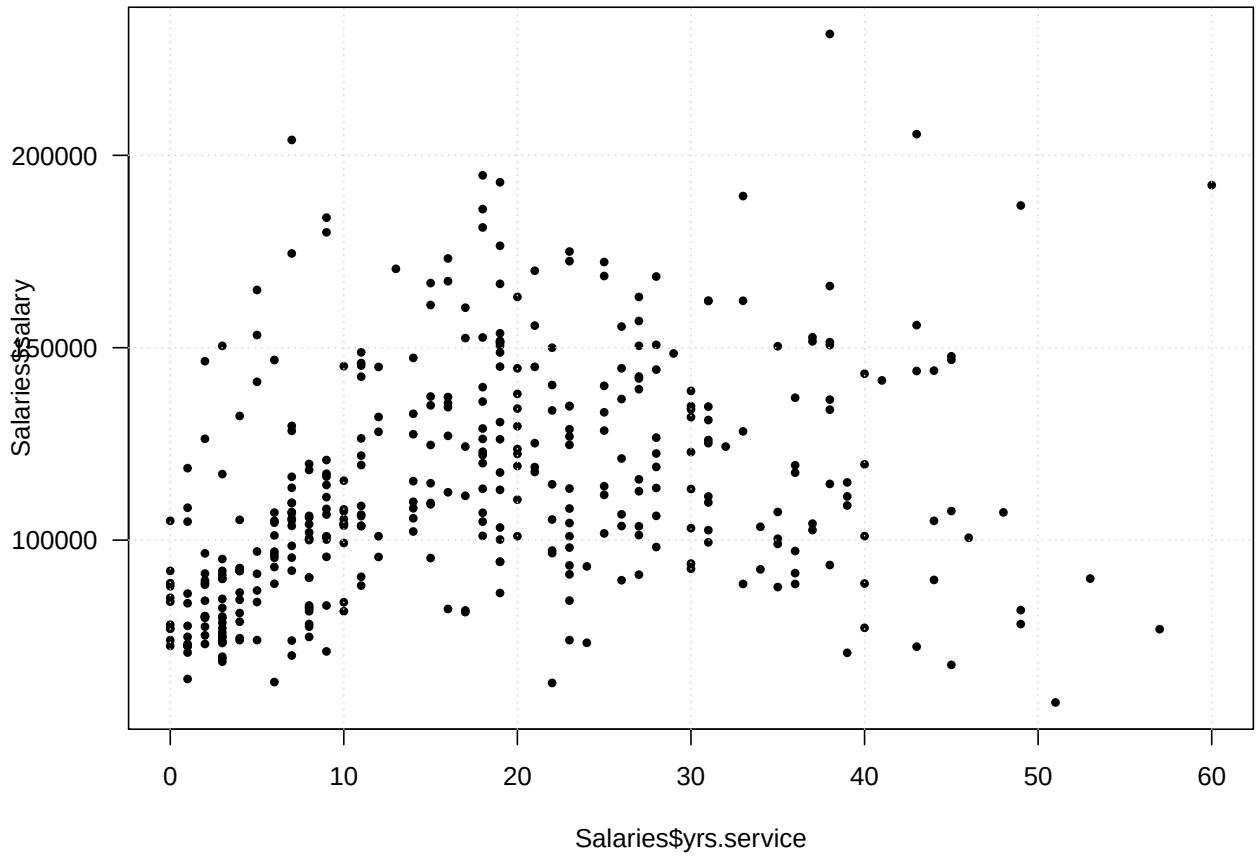
```
boxplot(salary ~ sex, data = Salaries, las = 1)
```




```
boxplot(salary ~ discipline, data = Salaries, las = 1)
```




```
plot(Salaries$yrs.service, Salaries$salary, las = 1, pch = 16, cex = 0.75)
grid()
```

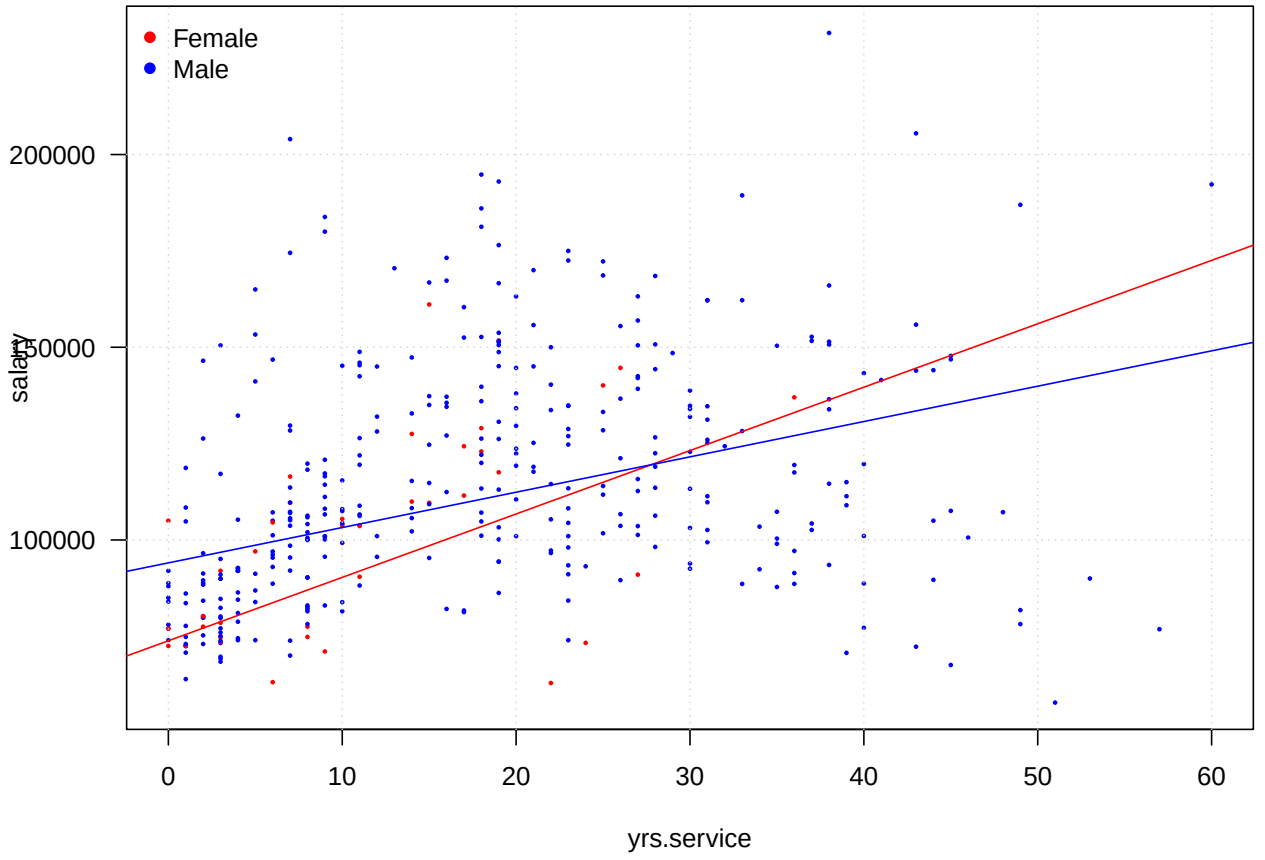



```
model <- lm(salary ~ discipline + rank + sex + yrs.service, data = Salaries)
X <- model.matrix(model)

attach(Salaries)
sex.col <- ifelse(sex == "Male", "blue", "red")
plot(yrs.service, salary, pch = 16, cex = 0.4,
     col = sex.col, las = 1)
grid()

m1 <- lm(salary ~ sex * yrs.since.phd)
summary(m1)
```

```
##
## Call:
## lm(formula = salary ~ sex * yrs.since.phd)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -83012 -19442  -2988  15059 102652
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            73840.8     8696.7   8.491 4.27e-16 ***
## sexMale                20209.6     9179.2   2.202 0.028269 *
## yrs.since.phd           1644.9      454.6   3.618 0.000335 ***
## sexMale:yrs.since.phd   -728.0      468.0  -1.555 0.120665
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27420 on 393 degrees of freedom
## Multiple R-squared:  0.1867, Adjusted R-squared:  0.1805
## F-statistic: 30.07 on 3 and 393 DF,  p-value: < 2.2e-16
```

```
coeff <- m1$coefficients
abline(coeff[1], coeff[3], col = "red")
abline(coeff[1] + coeff[2], coeff[3] + coeff[4],
       col = "blue")
legend("toplef", legend = c("Female", "Male"),
       pch = 16, col = c("red", "blue"),
       bty = "n")
```



Polynomial regression: Housing Values in Suburbs of Boston

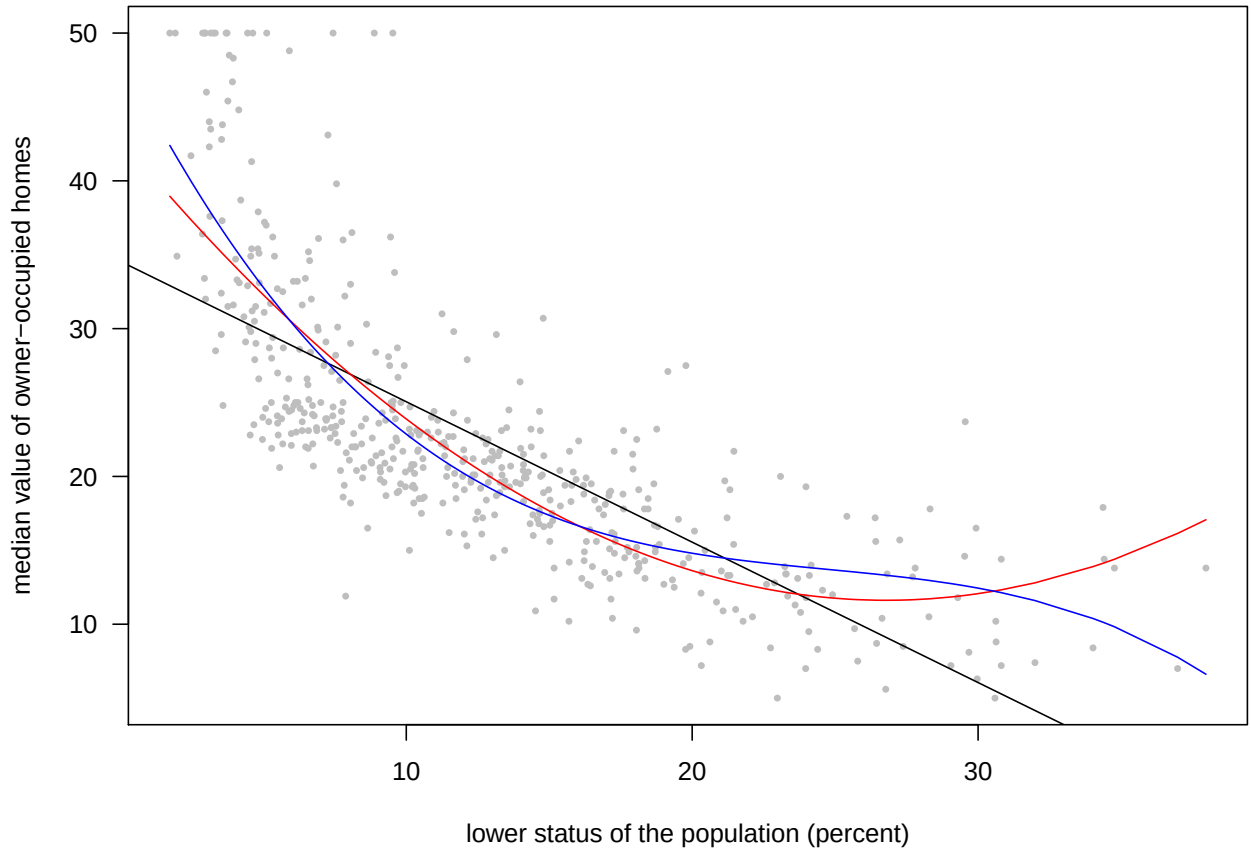
```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:olsrr':
##
##     cement
```

```
data(Boston)

plot(Boston$lstat, Boston$medv, col = "gray", pch = 16,
```

```
     cex = 0.6, las = 1, xlab = "lower status of the population (percent)", ylab = "median value of owne

m1 <- lm(medv ~ lstat, data = Boston)
abline(m1)

m2 <- lm(medv ~ lstat + I(lstat^2), data = Boston)
lines(sort(Boston$lstat), m2$fitted.values[order(Boston$lstat)], col = "red")

m3 <- lm(medv ~ lstat + I(lstat^2)+ I(lstat^3), data = Boston)
lines(sort(Boston$lstat), m3$fitted.values[order(Boston$lstat)], col = "blue")
```




```
m3new <- lm(medv ~ poly(lstat, 3), data = Boston)
summary(m3new)
```

```
##
## Call:
## lm(formula = medv ~ poly(lstat, 3), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.5441  -3.7122  -0.5145   2.4846  26.4153
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       22.5328     0.2399  93.937  < 2e-16 ***
## poly(lstat, 3)1 -152.4595     5.3958 -28.255  < 2e-16 ***
## poly(lstat, 3)2   64.2272     5.3958  11.903  < 2e-16 ***
```

```
## poly(lstat, 3)3  -27.0511     5.3958  -5.013 7.43e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.396 on 502 degrees of freedom
## Multiple R-squared:  0.6578, Adjusted R-squared:  0.6558
## F-statistic: 321.7 on 3 and 502 DF,  p-value: < 2.2e-16
```