Linear Regression

A Short Course on Data Analysis Using R Software (2017)

Wan Nor Arifin (wnarifin@usm.my), Universiti Sains Malaysia
Website: sites.google.com/site/wnarifin

Contents

1 Introduction

Multiple Linear Regression is given by

$$E(Y|\mathbf{X}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} = \beta_0 + \sum \beta_{p-1} X_{p-1}$$

where the $\mathbf{X}$ (in bold) denotes a collection of Xs. p is the number of estimated parameters.

2 Preliminaries

2.1 Load libraries

```
library(car)
library(psych)
```

2.2 Load data set

```
salary = Salaries  # data from `car`, Salaries for Professors...
`?`(Salaries)
str(salary)
```

```
## 'data.frame':    397 obs. of  6 variables:
##  $ rank         : Factor w/ 3 levels "AsstProf","AssocProf",..: 3 3 1 3 3 2 3 3 3 3 ...
##  $ discipline   : Factor w/ 2 levels "A","B": 2 2 2 2 2 2 2 2 2 2 ...
##  $ yrs.since.phd: int  19 20 4 45 40 6 30 45 21 18 ...
##  $ yrs.service  : int  18 16 3 39 41 6 23 45 20 18 ...
##  $ sex          : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 1 ...
##  $ salary       : int  139750 173200 79750 115000 141500 97000 175000 147765 119250 129000 ...
```

```
names(salary)
```

```
## [1] "rank"          "discipline"    "yrs.since.phd" "yrs.service"   "sex"
## [6] "salary"
```

```
# - View the levels of categorical variables
lapply(salary[c("rank", "discipline", "sex")], levels)
```

```
## $rank
## [1] "AsstProf"  "AssocProf" "Prof"
##
## $discipline
## [1] "A" "B"
##
## $sex
## [1] "Female" "Male"
```

3 Linear Regression

3.1 Data exploration

3.1.1 Descriptive statistics

```
describe(salary[c(3, 4, 6)])  # var 3, 4, 6 are numbers
```

```
##               vars   n      mean       sd median   trimmed      mad   min    max  range
## yrs.since.phd    1 397     22.31    12.89     21     21.83    14.83     1     56     55
## yrs.service      2 397     17.61    13.01     16     16.51    14.83     0     60     60
## salary           3 397 113706.46 30289.04 107300 111401.61 29355.48 57800 231545 173745
##               skew kurtosis      se
## yrs.since.phd 0.30    -0.81    0.65
## yrs.service   0.65    -0.34    0.65
## salary        0.71     0.18 1520.16
```

```
summary(salary[c(1, 2, 5)])  # var 1, 2, 5 are factors
```

```
##         rank     discipline     sex
##  AsstProf : 67   A:181      Female: 39
##  AssocProf: 64   B:216      Male  :358
##  Prof     :266
```

```r
lapply(salary[c(1, 2, 5)], function(x) summary(x)/length(x) * 100)  # in percent
```

```
## $rank
##  AsstProf AssocProf      Prof
##  16.87657  16.12091  67.00252
##
## $discipline
##        A        B
## 45.59194 54.40806
##
## $sex
##    Female      Male
##  9.823678 90.176322
```

```r
# - Salary by groups
describeBy(salary$salary, salary$rank)
```

```
##
##  Descriptive statistics by group
## group: AsstProf
##    vars  n     mean      sd median trimmed     mad   min   max range skew kurtosis     se
## X1    1 67 80775.99 8174.11  79800 80825.6 9340.38 63100 97032 33932 0.08       -1 998.63
## ------------------------------------------------------------------------
## group: AssocProf
##    vars  n     mean      sd  median  trimmed      mad   min    max range  skew kurtosis
## X1    1 64 93876.44 13831.7 95626.5 93937.38 14624.37 62884 126431 63547 -0.08    -0.71
##         se
## X1 1728.96
## ------------------------------------------------------------------------
## group: Prof
##    vars   n     mean       sd   median  trimmed      mad   min    max  range skew
## X1    1 266 126772.1 27718.67 123321.5 125080.8 28409.58 57800 231545 173745 0.58
##    kurtosis      se
## X1     0.32 1699.54
```

```r
describeBy(salary$salary, salary$discipline)
```

```
##
##  Descriptive statistics by group
## group: A
##    vars   n     mean       sd median  trimmed      mad   min    max  range skew kurtosis
## X1    1 181 108548.4 30538.15 104350 105515.3 31653.51 57800 205500 147700 0.84     0.34
##         se
## X1 2269.88
## ------------------------------------------------------------------------
## group: B
##    vars   n     mean       sd   median  trimmed      mad   min    max  range skew
## X1    1 216 118028.7 29459.14 113018.5 116020.4 31162.03 67559 231545 163986 0.67
##    kurtosis      se
## X1     0.16 2004.44
```

```r
describeBy(salary$salary, salary$sex)
```

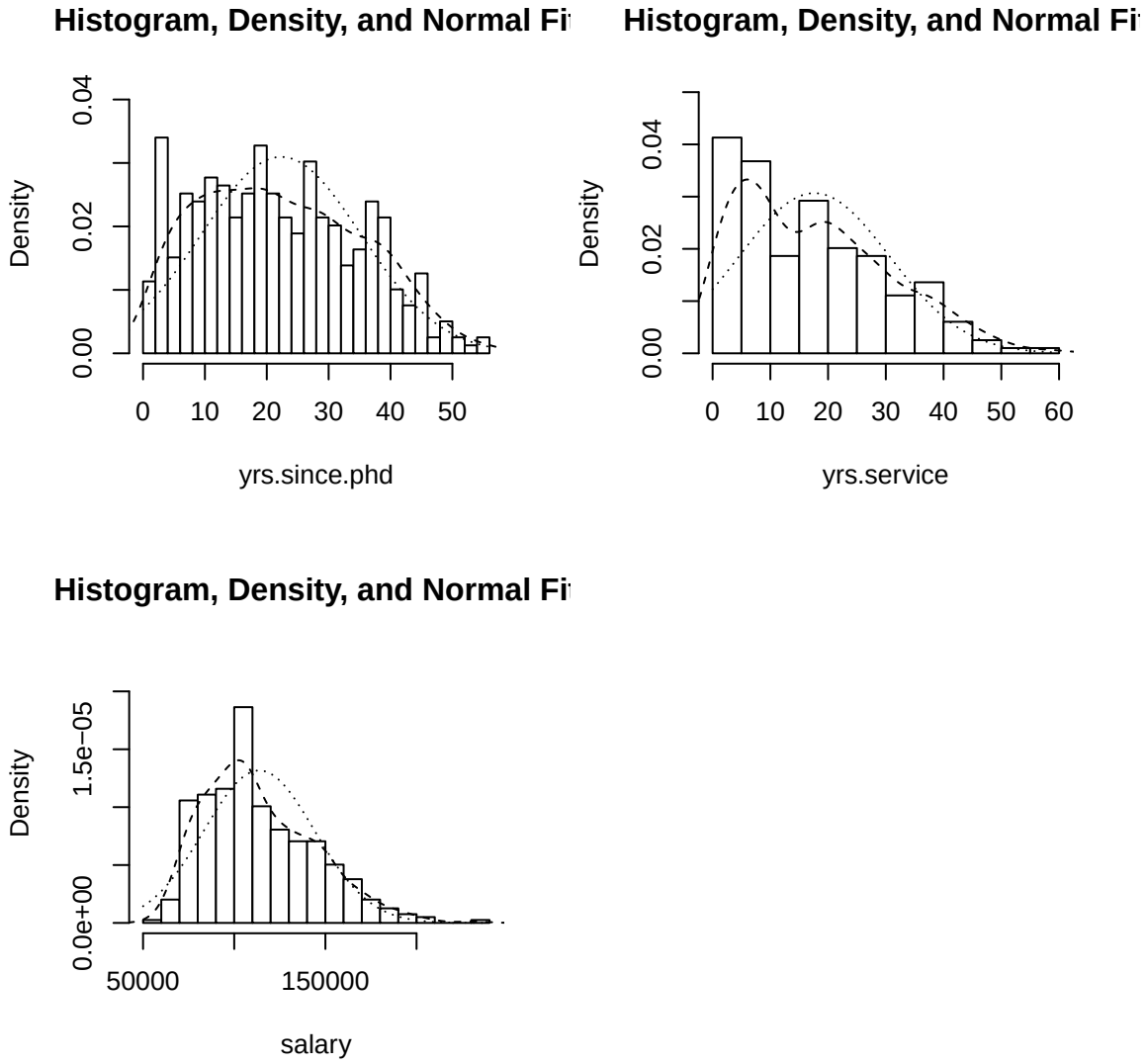
```
##
##  Descriptive statistics by group
## group: Female
```

```
##    vars  n     mean       sd median  trimmed      mad   min    max range skew kurtosis
## X1    1 39 101002.4 25952.13 103750 99531.06 35229.54 62884 161101 98217 0.42     -0.8
##         se
## X1 4155.67
## ------------------------------------------------------------------------
## group: Male
##    vars   n      mean       sd median  trimmed      mad   min    max  range skew kurtosis
## X1    1 358 115090.4 30436.93 108043 112748.1 29586.02 57800 231545 173745 0.71     0.15
##         se
## X1 1608.64
```

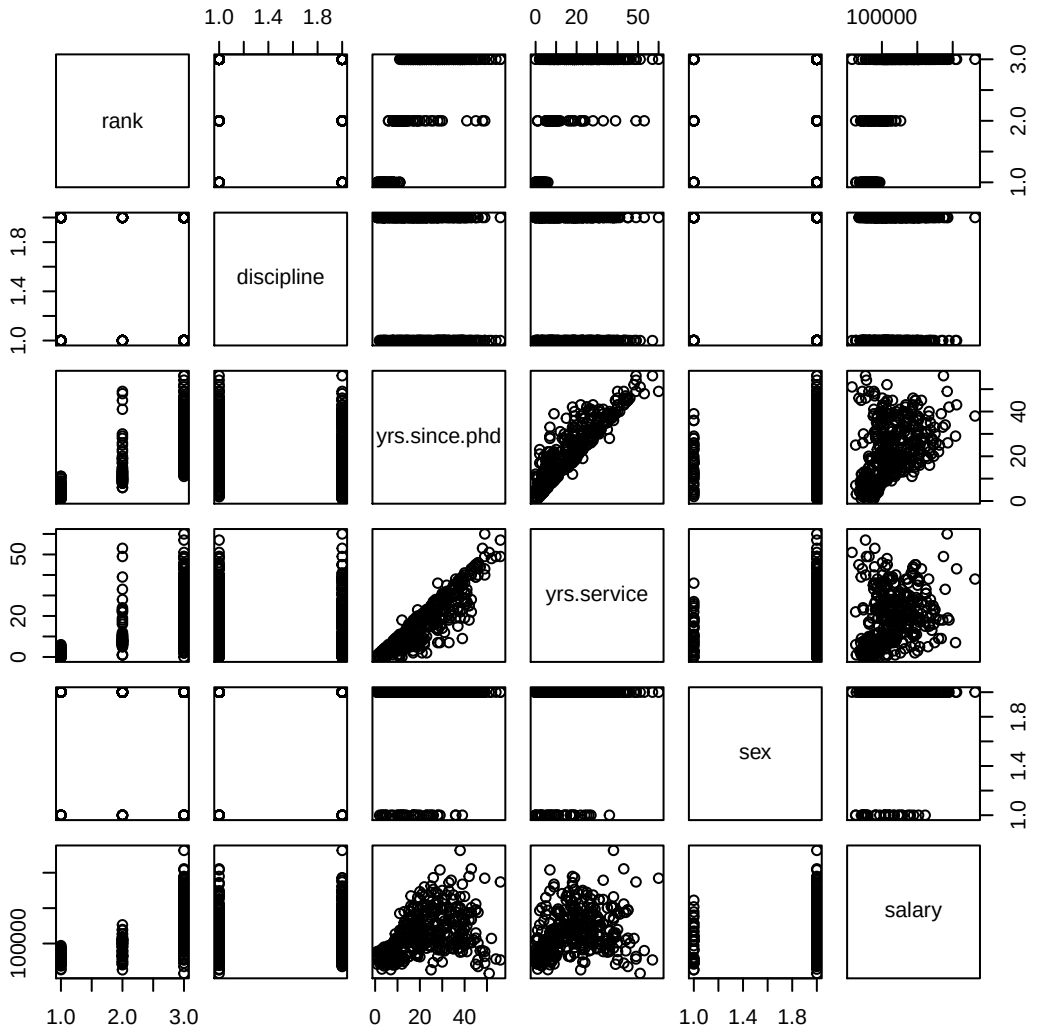
```
# lapply(salary[c(1,2,5)], function(x) describeBy(salary$salary, x)) # one line code
```

3.1.2 Plots

```
multi.hist(salary[c(3, 4, 6)])
```




```
plot(salary)
```



3.2 Univariable

```
str(salary)
```

```
## 'data.frame':    397 obs. of  6 variables:
##  $ rank         : Factor w/ 3 levels "AsstProf","AssocProf",..: 3 3 1 3 3 2 3 3 3 3 ...
##  $ discipline   : Factor w/ 2 levels "A","B": 2 2 2 2 2 2 2 2 2 2 ...
##  $ yrs.since.phd: int  19 20 4 45 40 6 30 45 21 18 ...
##  $ yrs.service  : int  18 16 3 39 41 6 23 45 20 18 ...
##  $ sex          : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 1 ...
##  $ salary       : int  139750 173200 79750 115000 141500 97000 175000 147765 119250 129000 ...
```

```
# - Years since PhD,
linear.u.phd = glm(salary ~ yrs.since.phd, data = salary)
summary(linear.u.phd)
```

```
##
## Call:
## glm(formula = salary ~ yrs.since.phd, data = salary)
##
## Deviance Residuals:
```

```
##    Min      1Q  Median      3Q     Max
## -84171  -19432   -2858   16086  102383
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    91718.7     2765.8  33.162   <2e-16 ***
## yrs.since.phd    985.3      107.4   9.177   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 758098328)
##
##     Null deviance: 3.6330e+11  on 396  degrees of freedom
## Residual deviance: 2.9945e+11  on 395  degrees of freedom
## AIC: 9247.8
##
## Number of Fisher Scoring iterations: 2
```

```
# - Years in service,
linear.u.ser = glm(salary ~ yrs.service, data = salary)
summary(linear.u.ser)
```

```
##
## Call:
## glm(formula = salary ~ yrs.service, data = salary)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -81933  -20511   -3776   16417  101947
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  99974.7     2416.6   41.37  < 2e-16 ***
## yrs.service    779.6      110.4    7.06 7.53e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 816686970)
##
##     Null deviance: 3.6330e+11  on 396  degrees of freedom
## Residual deviance: 3.2259e+11  on 395  degrees of freedom
## AIC: 9277.4
##
## Number of Fisher Scoring iterations: 2
```

```
# - Rank,
linear.u.ran = glm(salary ~ rank, data = salary)
summary(linear.u.ran)
```

```
##
## Call:
## glm(formula = salary ~ rank, data = salary)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -68972  -16376   -1580   11755  104773
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      80776       2887  27.976  < 2e-16 ***
## rankAssocProf    13100       4131   3.171  0.00164 **
## rankProf         45996       3230  14.238  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 558550449)
##
##     Null deviance: 3.6330e+11  on 396  degrees of freedom
## Residual deviance: 2.2007e+11  on 394  degrees of freedom
## AIC: 9127.5
##
## Number of Fisher Scoring iterations: 2
```

```
# - Discipline,
linear.u.dis = glm(salary ~ discipline, data = salary)
summary(linear.u.dis)
```

```
##
## Call:
## glm(formula = salary ~ discipline, data = salary)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -50748  -24611   -4429   19138  113516
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   108548       2227  48.751  < 2e-16 ***
## disciplineB     9480       3019   3.141  0.00181 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 897341368)
##
##     Null deviance: 3.6330e+11  on 396  degrees of freedom
## Residual deviance: 3.5445e+11  on 395  degrees of freedom
## AIC: 9314.8
##
## Number of Fisher Scoring iterations: 2
```

```
# - Sex,
linear.u.sex = glm(salary ~ sex, data = salary)
summary(linear.u.sex)
```

```
##
## Call:
## glm(formula = salary ~ sex, data = salary)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -57290  -23502   -6828   19710  116455
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   101002       4809  21.001  < 2e-16 ***
## sexMale        14088       5065   2.782  0.00567 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 902077538)
##
##     Null deviance: 3.6330e+11  on 396  degrees of freedom
## Residual deviance: 3.5632e+11  on 395  degrees of freedom
## AIC: 9316.9
##
## Number of Fisher Scoring iterations: 2
```

```
# - LR test
linear.u0 = glm(salary ~ 1, data = salary)
summary(linear.u0)
```

```
##
## Call:
## glm(formula = salary ~ 1, data = salary)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -55906  -22706   -6406   20479  117839
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   113706       1520    74.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 917425865)
##
##     Null deviance: 3.633e+11  on 396  degrees of freedom
## Residual deviance: 3.633e+11  on 396  degrees of freedom
## AIC: 9322.6
##
## Number of Fisher Scoring iterations: 2
```

```
cat(names(salary), sep = " + ")
```

```
## rank + discipline + yrs.since.phd + yrs.service + sex + salary
```

```
add1(linear.u0, scope = ~rank + discipline + yrs.since.phd + yrs.service + sex, test = "LRT")
```

```
## Single term additions
##
## Model:
## salary ~ 1
##               Df   Deviance    AIC scaled dev.  Pr(>Chi)
## <none>           3.6330e+11 9322.6
## rank           2 2.2007e+11 9127.5     199.012 < 2.2e-16 ***
## discipline     1 3.5445e+11 9314.8       9.792  0.001753 **
## yrs.since.phd  1 2.9945e+11 9247.8      76.735 < 2.2e-16 ***
```

```
## yrs.service    1 3.2259e+11 9277.4      47.181 6.472e-12 ***
## sex            1 3.5632e+11 9316.9       7.702  0.005517 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# - p on adding that var = univar
```

3.3 Multivariable

```
# - All
linear.m.all = glm(salary ~ rank + discipline + yrs.since.phd + yrs.service + sex, data = salary)
summary(linear.m.all)
```

```
##
## Call:
## glm(formula = salary ~ rank + discipline + yrs.since.phd + yrs.service +
##     sex, data = salary)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -65248  -13211   -1775   10384   99592
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    65955.2     4588.6  14.374  < 2e-16 ***
## rankAssocProf  12907.6     4145.3   3.114  0.00198 **
## rankProf       45066.0     4237.5  10.635  < 2e-16 ***
## disciplineB    14417.6     2342.9   6.154 1.88e-09 ***
## yrs.since.phd    535.1      241.0   2.220  0.02698 *
## yrs.service     -489.5      211.9  -2.310  0.02143 *
## sexMale         4783.5     3858.7   1.240  0.21584
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 507990599)
##
##     Null deviance: 3.6330e+11  on 396  degrees of freedom
## Residual deviance: 1.9812e+11  on 390  degrees of freedom
## AIC: 9093.8
##
## Number of Fisher Scoring iterations: 2
```

```
drop1(linear.m.all, test = "LRT")  # p on rmv that var
```

```
## Single term deletions
##
## Model:
## salary ~ rank + discipline + yrs.since.phd + yrs.service + sex
##               Df   Deviance    AIC scaled dev.  Pr(>Chi)
## <none>           1.9812e+11 9093.8
## rank           2 2.6762e+11 9209.2     119.389 < 2.2e-16 ***
## discipline     1 2.1735e+11 9128.6      36.791 1.315e-09 ***
## yrs.since.phd  1 2.0062e+11 9096.8       4.986   0.02555 *
## yrs.service    1 2.0083e+11 9097.2       5.394   0.02021 *
```

```
## sex            1 1.9890e+11 9093.4       1.561   0.21147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# - Stepwise
linear.m.step = step(linear.m.all, direction = "both")
```

```
## Start:  AIC=9093.83
## salary ~ rank + discipline + yrs.since.phd + yrs.service + sex
##
##                 Df   Deviance    AIC
## - sex            1 1.9890e+11 9093.4
## <none>             1.9812e+11 9093.8
## - yrs.since.phd  1 2.0062e+11 9096.8
## - yrs.service    1 2.0083e+11 9097.2
## - discipline     1 2.1735e+11 9128.6
## - rank           2 2.6762e+11 9209.2
##
## Step:  AIC=9093.39
## salary ~ rank + discipline + yrs.since.phd + yrs.service
##
##                 Df   Deviance    AIC
## <none>             1.9890e+11 9093.4
## + sex            1 1.9812e+11 9093.8
## - yrs.since.phd  1 2.0140e+11 9096.3
## - yrs.service    1 2.0147e+11 9096.5
## - discipline     1 2.1839e+11 9128.5
## - rank           2 2.6958e+11 9210.1
```

```
summary(linear.m.step)
```

```
##
## Call:
## glm(formula = salary ~ rank + discipline + yrs.since.phd + yrs.service,
##     data = salary)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -65244  -13498   -1455    9638   99682
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    69869.0     3332.1  20.968  < 2e-16 ***
## rankAssocProf  12831.5     4147.7   3.094  0.00212 **
## rankProf       45287.7     4236.7  10.689  < 2e-16 ***
## disciplineB    14505.2     2343.4   6.190 1.52e-09 ***
## yrs.since.phd    534.6      241.2   2.217  0.02720 *
## yrs.service     -476.7      211.8  -2.250  0.02497 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 508688005)
##
##     Null deviance: 3.633e+11  on 396  degrees of freedom
## Residual deviance: 1.989e+11  on 391  degrees of freedom
```

```
## AIC: 9093.4
##
## Number of Fisher Scoring iterations: 2
```

```
linear.m.step$anova
```

```
##    Step Df  Deviance Resid. Df   Resid. Dev      AIC
## 1       NA        NA       390 198116333525 9093.826
## 2 - sex  1 780676354       391 198897009879 9093.388
```

```
# - Chosen model
linear.m1 = glm(salary ~ rank + discipline + yrs.since.phd + yrs.service, data = salary)
summary(linear.m1)
```

```
##
## Call:
## glm(formula = salary ~ rank + discipline + yrs.since.phd + yrs.service,
##     data = salary)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -65244  -13498   -1455    9638   99682
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    69869.0     3332.1  20.968  < 2e-16 ***
## rankAssocProf  12831.5     4147.7   3.094  0.00212 **
## rankProf       45287.7     4236.7  10.689  < 2e-16 ***
## disciplineB    14505.2     2343.4   6.190 1.52e-09 ***
## yrs.since.phd    534.6      241.2   2.217  0.02720 *
## yrs.service     -476.7      211.8  -2.250  0.02497 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 508688005)
##
##     Null deviance: 3.633e+11  on 396  degrees of freedom
## Residual deviance: 1.989e+11  on 391  degrees of freedom
## AIC: 9093.4
##
## Number of Fisher Scoring iterations: 2
```

```
# - LR test
drop1(linear.m1, test = "LRT")  # p on rmv that var
```

```
## Single term deletions
##
## Model:
## salary ~ rank + discipline + yrs.since.phd + yrs.service
##               Df   Deviance    AIC scaled dev.  Pr(>Chi)
## <none>           1.9890e+11 9093.4
## rank           2 2.6958e+11 9210.1     120.713 < 2.2e-16 ***
## discipline     1 2.1839e+11 9128.5      37.111 1.116e-09 ***
## yrs.since.phd  1 2.0140e+11 9096.3       4.959   0.02595 *
## yrs.service    1 2.0147e+11 9096.5       5.109   0.02380 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.4 Multicollinearity

```
cbind(summary(linear.m1)$coefficients[, 1:2])  # SE
```

```
##                 Estimate Std. Error
## (Intercept)   69869.0110  3332.1448
## rankAssocProf 12831.5375  4147.6685
## rankProf      45287.6890  4236.6534
## disciplineB   14505.1514  2343.4181
## yrs.since.phd   534.6313   241.1593
## yrs.service    -476.7179   211.8312
```

```
vif(linear.m1)  # VIF
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## rank           2.003562  2        1.189736
## discipline     1.063139  1        1.031086
## yrs.since.phd 7.518920  1        2.742065
## yrs.service   5.908984  1        2.430840
```

3.5 Interaction

```
add1(linear.m1, scope = ~. + rank * discipline * yrs.since.phd * yrs.service, test = "LRT")
```

```
## Single term additions
##
## Model:
## salary ~ rank + discipline + yrs.since.phd + yrs.service
##                            Df   Deviance    AIC scaled dev. Pr(>Chi)
## <none>                        1.9890e+11 9093.4
## rank:discipline             2 1.9838e+11 9096.4      1.0300 0.597506
## rank:yrs.since.phd          2 1.9800e+11 9095.6      1.8025 0.406066
## discipline:yrs.since.phd    1 1.9879e+11 9095.2      0.2231 0.636696
## rank:yrs.service            2 1.9808e+11 9095.8      1.6264 0.443440
## discipline:yrs.service      1 1.9623e+11 9090.0      5.3563 0.020648 *
## yrs.since.phd:yrs.service   1 1.9554e+11 9088.6      6.7650 0.009296 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# - two interactions: discipline:yrs.service; yrs.since.phd:yrs.service
```

3.6 Revised models

```
linear.m2 = glm(salary ~ rank + discipline + yrs.since.phd + yrs.service + yrs.since.phd:yrs.service +
    discipline:yrs.service, data = salary)
summary(linear.m2)  # interractions included
```

```
##
## Call:
## glm(formula = salary ~ rank + discipline + yrs.since.phd + yrs.service +
```

```
##     yrs.since.phd:yrs.service + discipline:yrs.service, data = salary)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -66219  -12814   -1483    9640   95308
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               70067.114   4211.523  16.637  < 2e-16 ***
## rankAssocProf              6358.223   4814.292   1.321   0.1874
## rankProf                  34988.186   5771.198   6.063 3.17e-09 ***
## disciplineB                8222.623   3905.270   2.106   0.0359 *
## yrs.since.phd               979.652    302.345   3.240   0.0013 **
## yrs.service                  82.678    396.800   0.208   0.8351
## yrs.since.phd:yrs.service   -21.301      9.266  -2.299   0.0220 *
## disciplineB:yrs.service     351.296    178.164   1.972   0.0493 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 497690316)
##
##     Null deviance: 3.633e+11  on 396  degrees of freedom
## Residual deviance: 1.936e+11  on 389  degrees of freedom
## AIC: 9086.7
##
## Number of Fisher Scoring iterations: 2
```

```
vif(linear.m2)  # very large VIF
```

```
##                                GVIF Df GVIF^(1/(2*Df))
## rank                       3.800437  2        1.396235
## discipline                 3.017760  1        1.737170
## yrs.since.phd             12.079364  1        3.475538
## yrs.service               21.191824  1        4.603458
## yrs.since.phd:yrs.service 25.255181  1        5.025453
## discipline:yrs.service     3.548516  1        1.883750
```

```
# - remove yrs.since.phd, yrs.service
linear.m1.1 = glm(salary ~ rank + discipline, data = salary)
summary(linear.m1.1)
```

```
##
## Call:
## glm(formula = salary ~ rank + discipline, data = salary)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -65990  -14049   -1288   10760   97996
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      71944       3135  22.948  < 2e-16 ***
## rankAssocProf    13762       3961   3.475 0.000569 ***
## rankProf         47844       3112  15.376  < 2e-16 ***
## disciplineB      13761       2296   5.993 4.65e-09 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 513076201)
##
##     Null deviance: 3.6330e+11  on 396  degrees of freedom
## Residual deviance: 2.0164e+11  on 393  degrees of freedom
## AIC: 9094.8
##
## Number of Fisher Scoring iterations: 2
```

```
# effect of adding them
add1(linear.m1.1, scope = ~. + yrs.since.phd + yrs.service, test = "LRT")
```

```
## Single term additions
##
## Model:
## salary ~ rank + discipline
##               Df   Deviance    AIC scaled dev. Pr(>Chi)
## <none>           2.0164e+11 9094.8
## yrs.since.phd  1 2.0147e+11 9096.5     0.32628   0.5679
## yrs.service    1 2.0140e+11 9096.3     0.47649   0.4900
```

```
# - add yrs.since.phd
linear.m1.2 = glm(salary ~ rank + discipline + yrs.since.phd, data = salary)
summary(linear.m1.2)
```

```
##
## Call:
## glm(formula = salary ~ rank + discipline + yrs.since.phd, data = salary)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -67395  -13480   -1536   10416   97166
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   71405.40    3278.32  21.781  < 2e-16 ***
## rankAssocProf 13030.16    4168.17   3.126   0.0019 **
## rankProf      46211.57    4238.52  10.903  < 2e-16 ***
## disciplineB   14028.68    2345.90   5.980 5.03e-09 ***
## yrs.since.phd    71.92     126.68   0.568   0.5706
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 513962494)
##
##     Null deviance: 3.6330e+11  on 396  degrees of freedom
## Residual deviance: 2.0147e+11  on 392  degrees of freedom
## AIC: 9096.5
##
## Number of Fisher Scoring iterations: 2
```

```
# - add yrs.service
linear.m1.3 = glm(salary ~ rank + discipline + yrs.service, data = salary)
summary(linear.m1.3)
```

```
##
## Call:
## glm(formula = salary ~ rank + discipline + yrs.service, data = salary)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -64198  -14040   -1299   10724   99253
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   72253.53    3169.48  22.797  < 2e-16 ***
## rankAssocProf 14483.23    4100.53   3.532 0.000461 ***
## rankProf      49377.50    3832.90  12.883  < 2e-16 ***
## disciplineB   13561.43    2315.91   5.856 1.01e-08 ***
## yrs.service     -76.33     111.25  -0.686 0.493039
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 513768063)
##
##     Null deviance: 3.633e+11  on 396  degrees of freedom
## Residual deviance: 2.014e+11  on 392  degrees of freedom
## AIC: 9096.3
##
## Number of Fisher Scoring iterations: 2
```

```
summary(linear.m1)  # too much discrepancy between model w & w/out yrs.since.phd, yrs.service
```

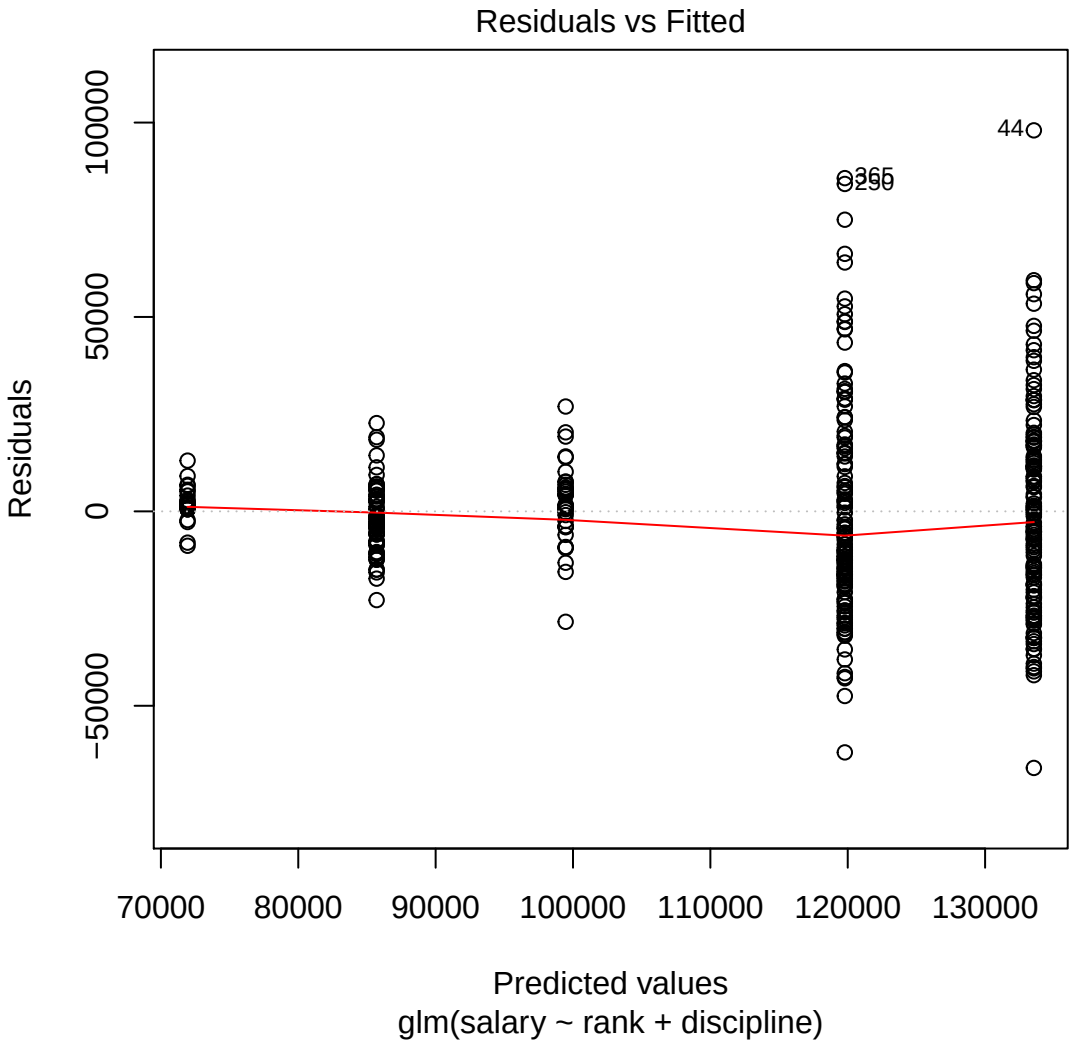
```
##
## Call:
## glm(formula = salary ~ rank + discipline + yrs.since.phd + yrs.service,
##     data = salary)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -65244  -13498   -1455    9638   99682
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    69869.0     3332.1  20.968  < 2e-16 ***
## rankAssocProf  12831.5     4147.7   3.094  0.00212 **
## rankProf       45287.7     4236.7  10.689  < 2e-16 ***
## disciplineB    14505.2     2343.4   6.190 1.52e-09 ***
## yrs.since.phd    534.6      241.2   2.217  0.02720 *
## yrs.service     -476.7      211.8  -2.250  0.02497 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 508688005)
##
##     Null deviance: 3.633e+11  on 396  degrees of freedom
## Residual deviance: 1.989e+11  on 391  degrees of freedom
## AIC: 9093.4
##
## Number of Fisher Scoring iterations: 2
```

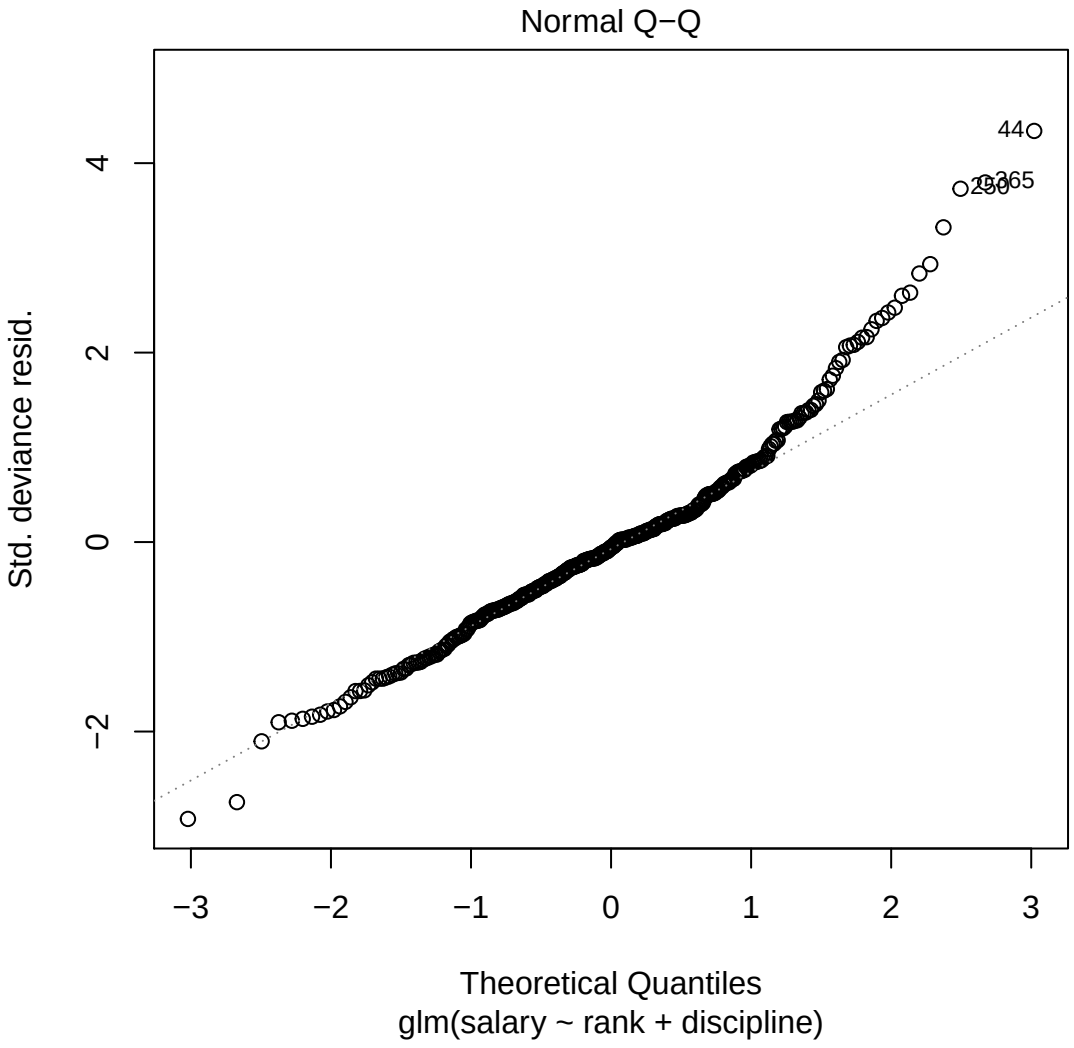
```r
# - the chosen one
linear.m3 = linear.m1.1  # salary ~ rank + discipline
summary(linear.m3)
```

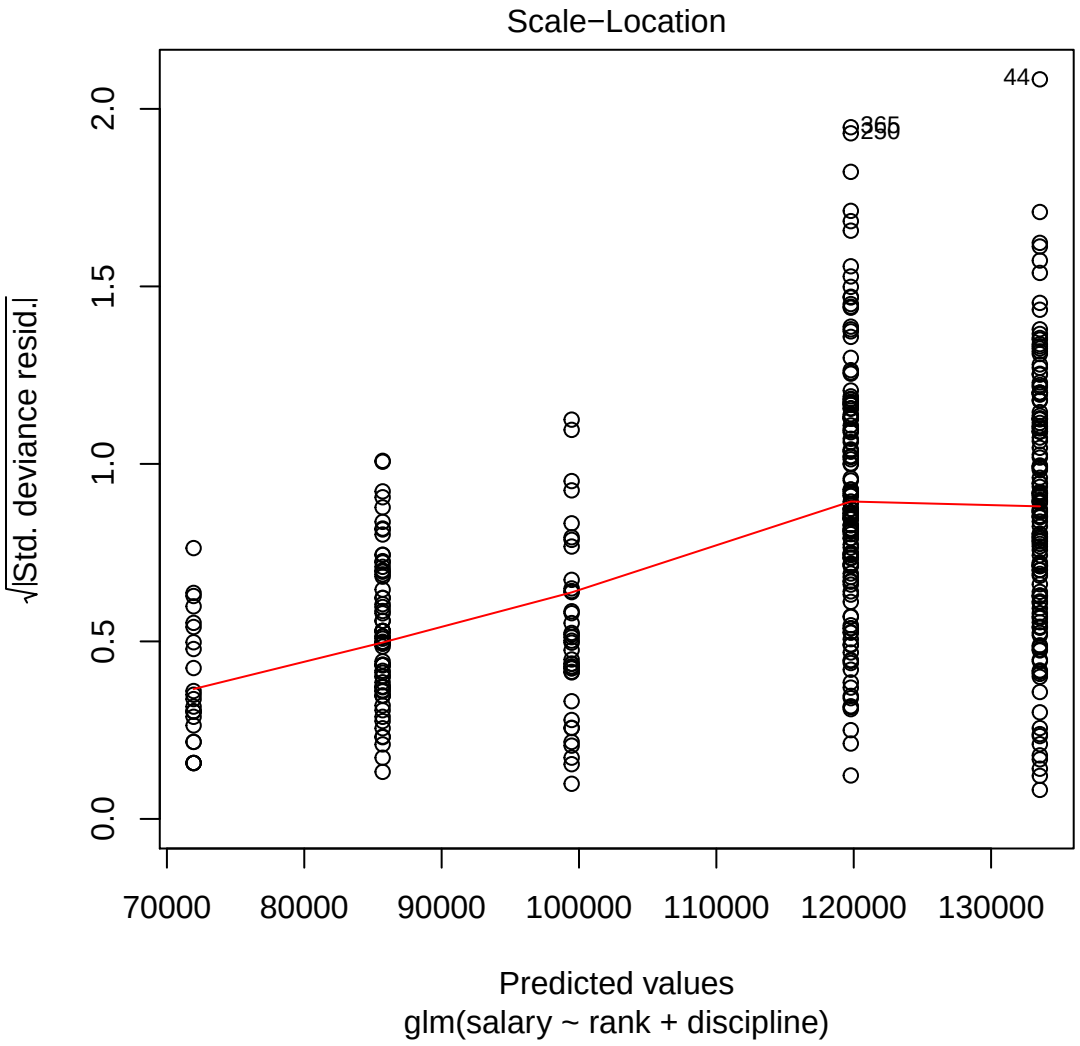
```
##
## Call:
## glm(formula = salary ~ rank + discipline, data = salary)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -65990  -14049   -1288   10760   97996
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      71944       3135  22.948  < 2e-16 ***
## rankAssocProf    13762       3961   3.475 0.000569 ***
## rankProf         47844       3112  15.376  < 2e-16 ***
## disciplineB      13761       2296   5.993 4.65e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 513076201)
##
##     Null deviance: 3.6330e+11  on 396  degrees of freedom
## Residual deviance: 2.0164e+11  on 393  degrees of freedom
## AIC: 9094.8
##
## Number of Fisher Scoring iterations: 2
```

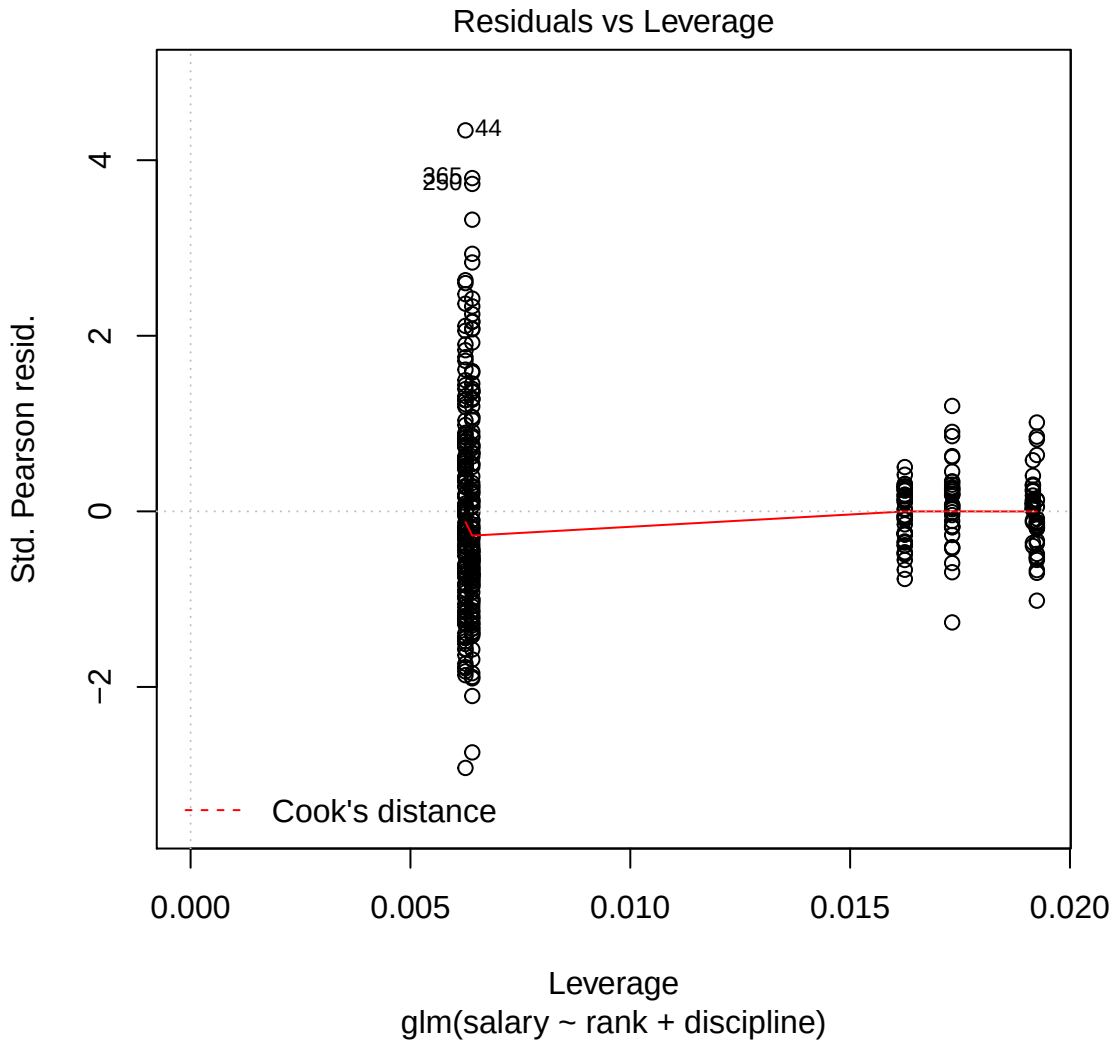
3.7 Residuals & Influentials

```r
plot(linear.m3)  # all defaults 1:4
```

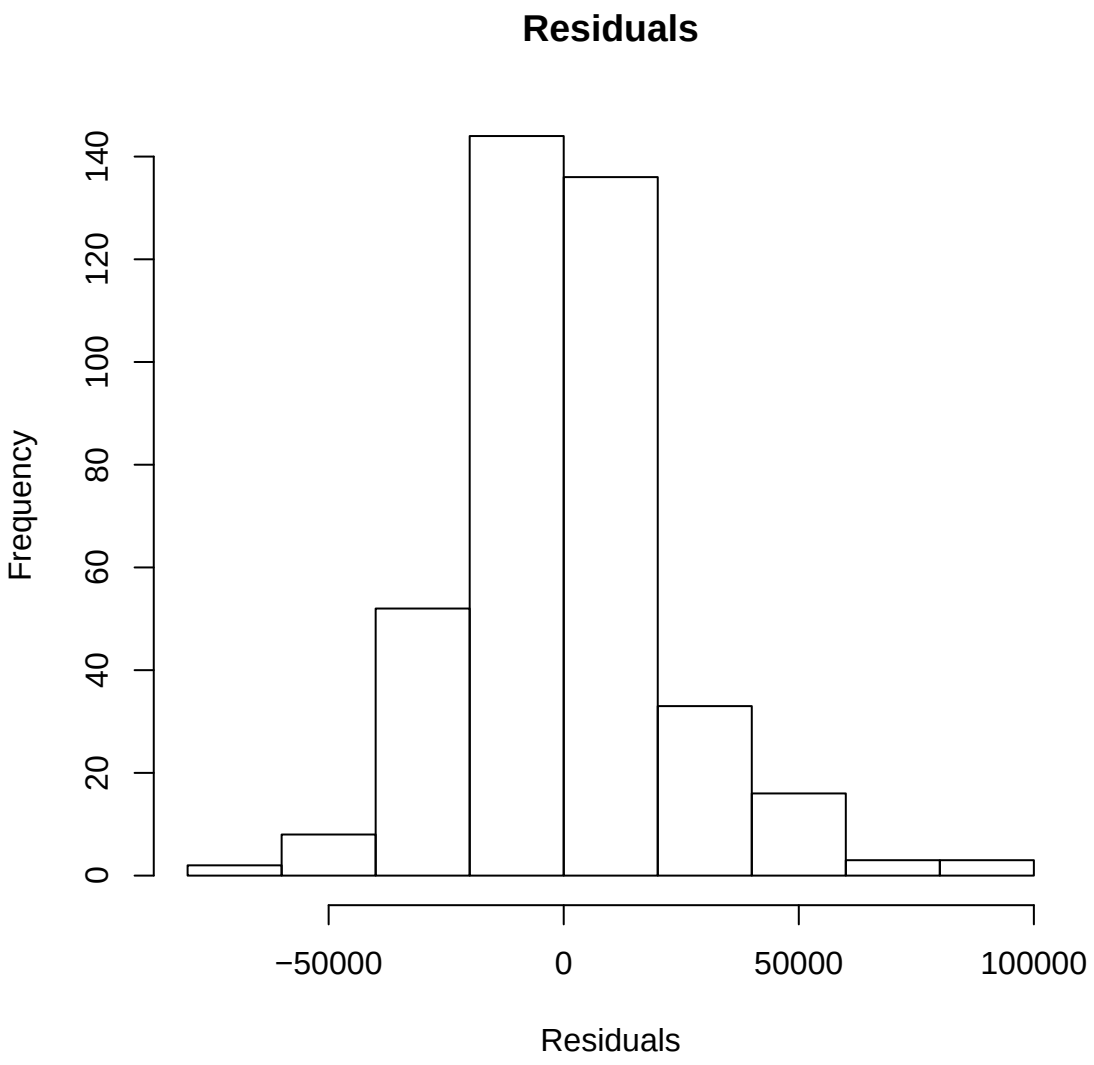

Residuals vs Fitted
Residuals
100000
50000
0
−50000
44
365
250
70000
80000
90000
100000
110000
120000
130000
Predicted values
glm(salary ~ rank + discipline)


Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
glm(salary ~ rank + discipline)
44
365
−3
−2
−1
0
1
2
3
4

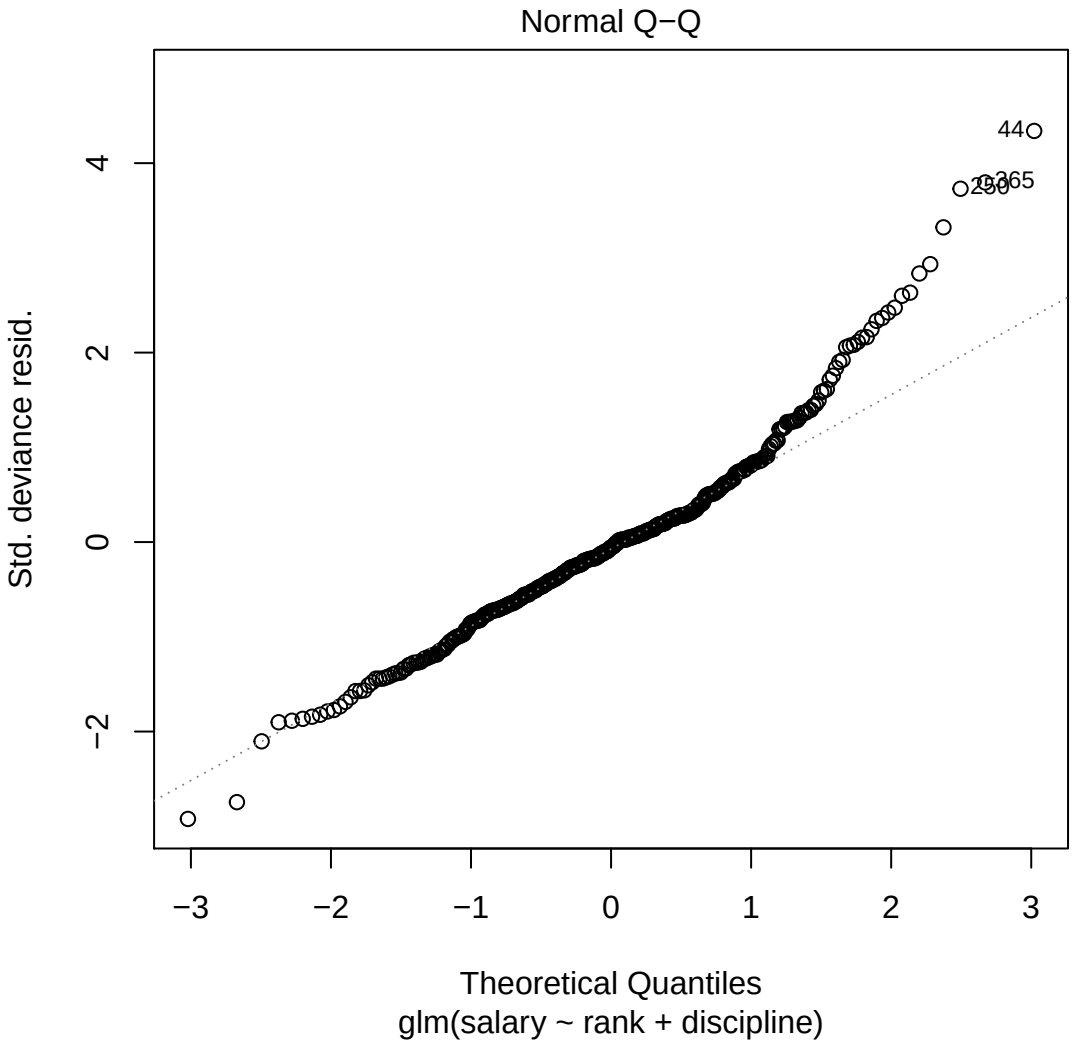

Scale−Location
√|Std. deviance resid.|
0.0
0.5
1.0
1.5
2.0
44
365
290
70000
80000
90000
100000
110000
120000
130000
Predicted values
glm(salary ~ rank + discipline)



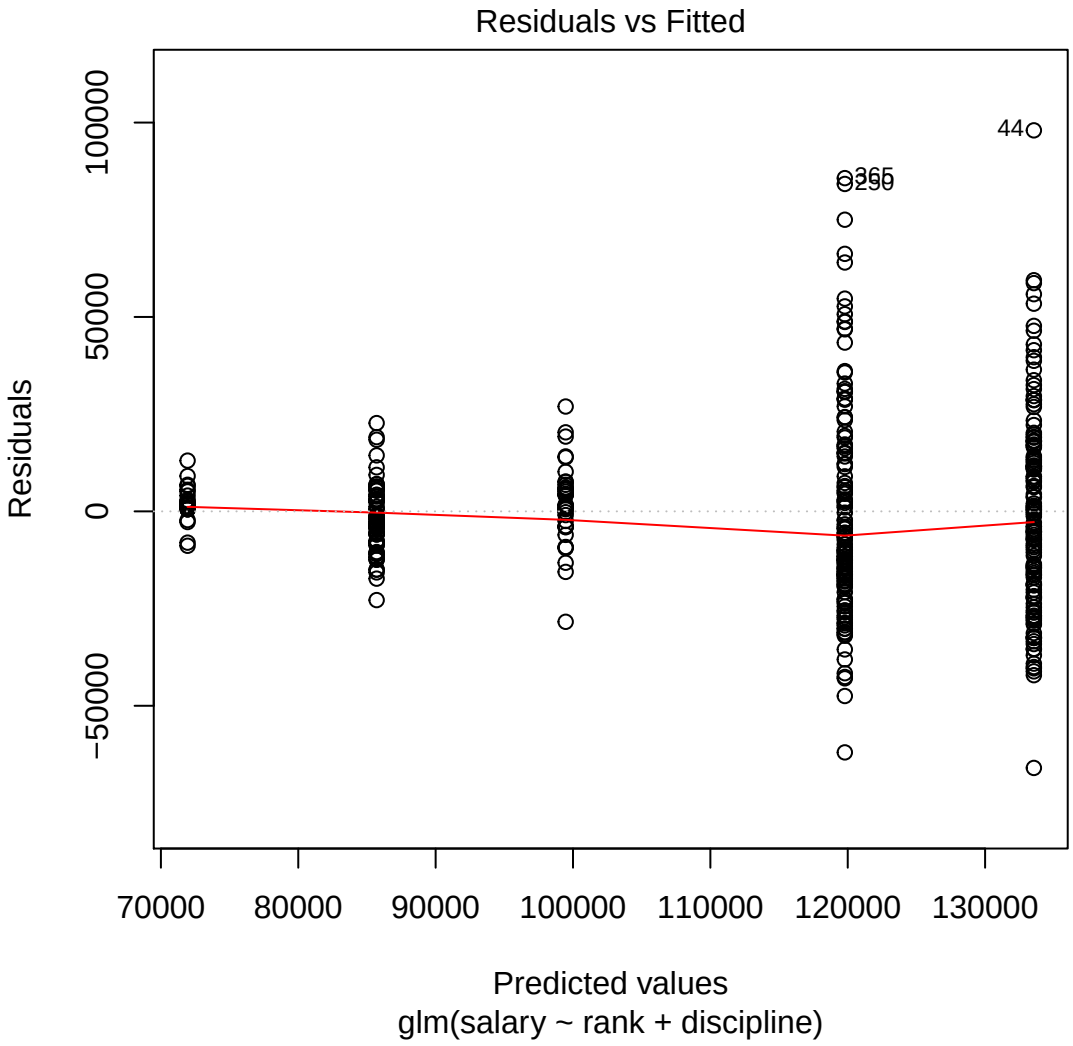

```
# - Normality
hist(resid(linear.m3), main = "Residuals", xlab = "Residuals", ylab = "Frequency")
```



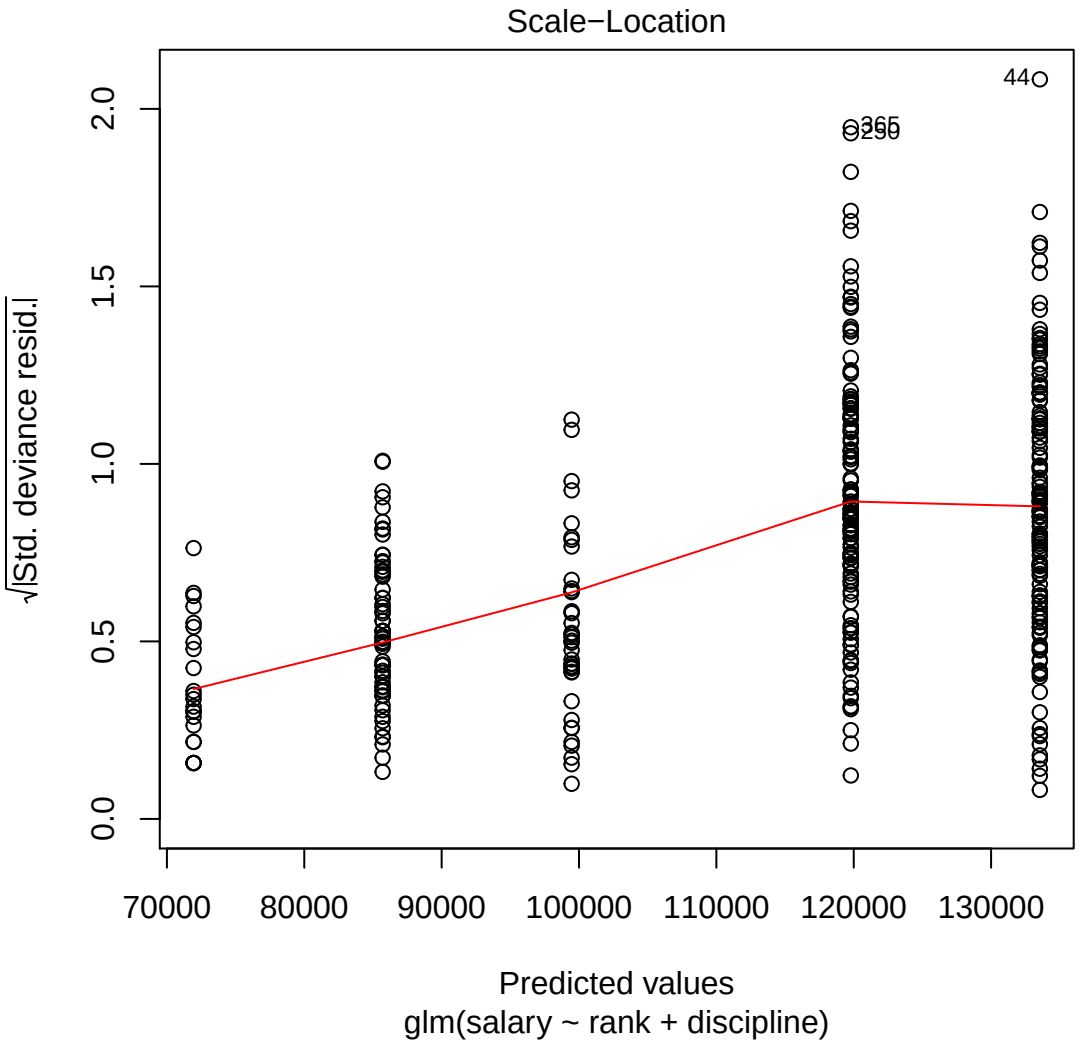

```
plot(linear.m3, which = 2)
```



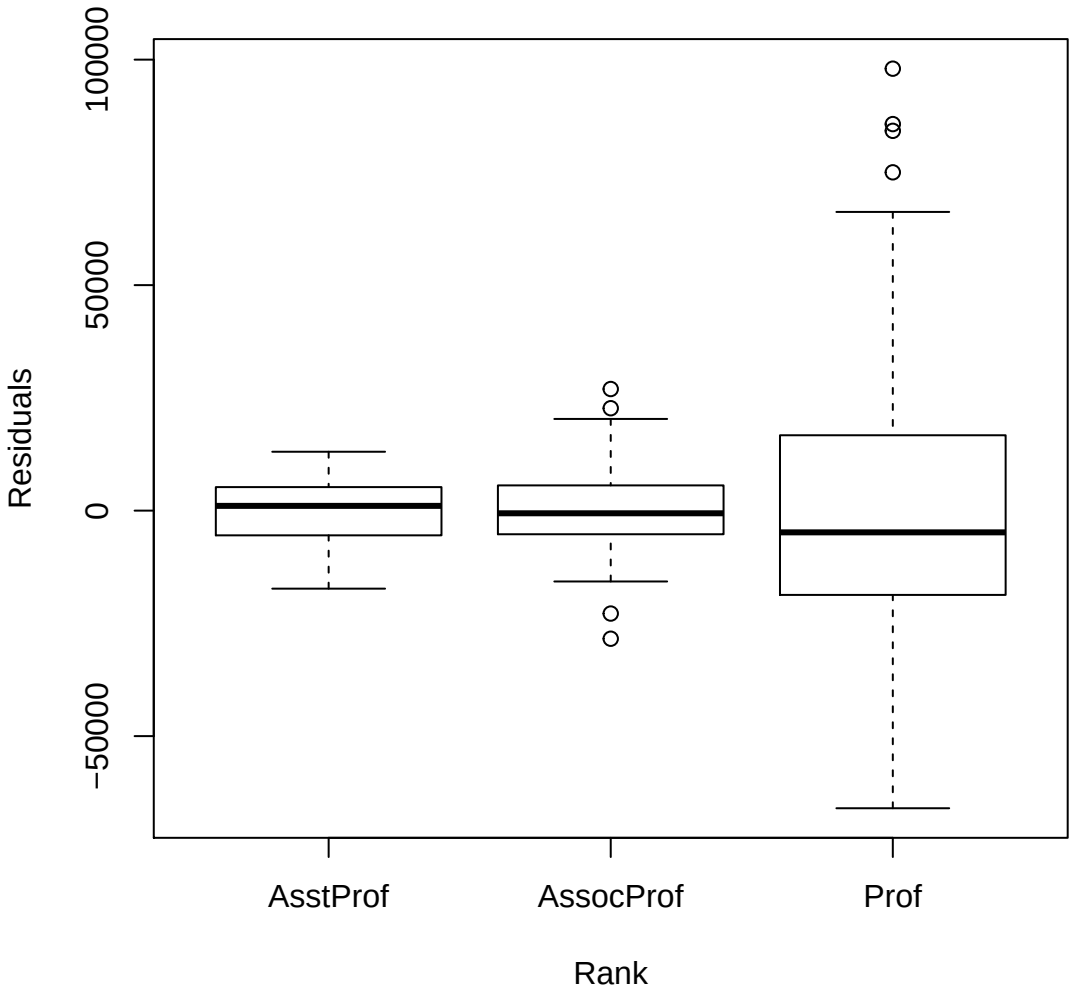

```
# - Linearity
plot(linear.m3, which = 1)  # residuals vs predicted
```



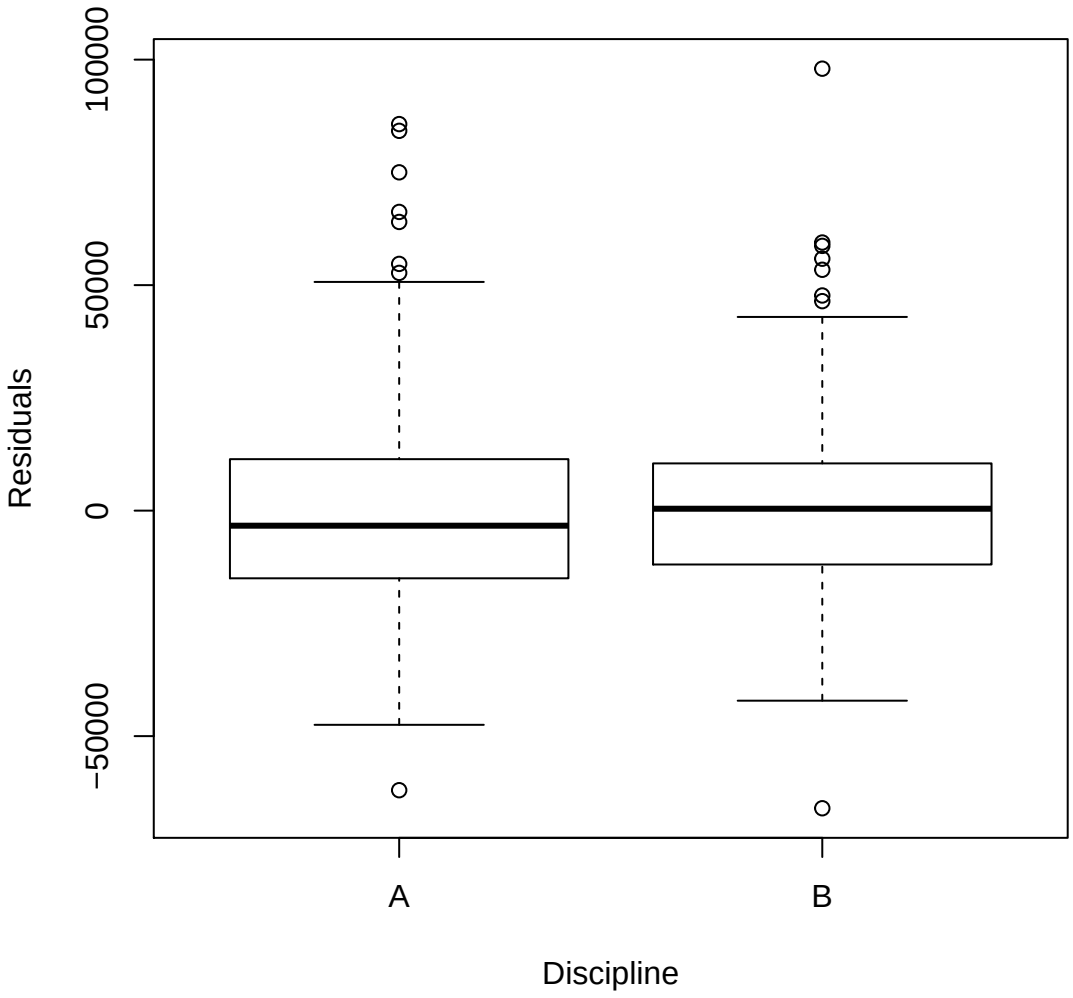

```
plot(linear.m3, which = 3)
```



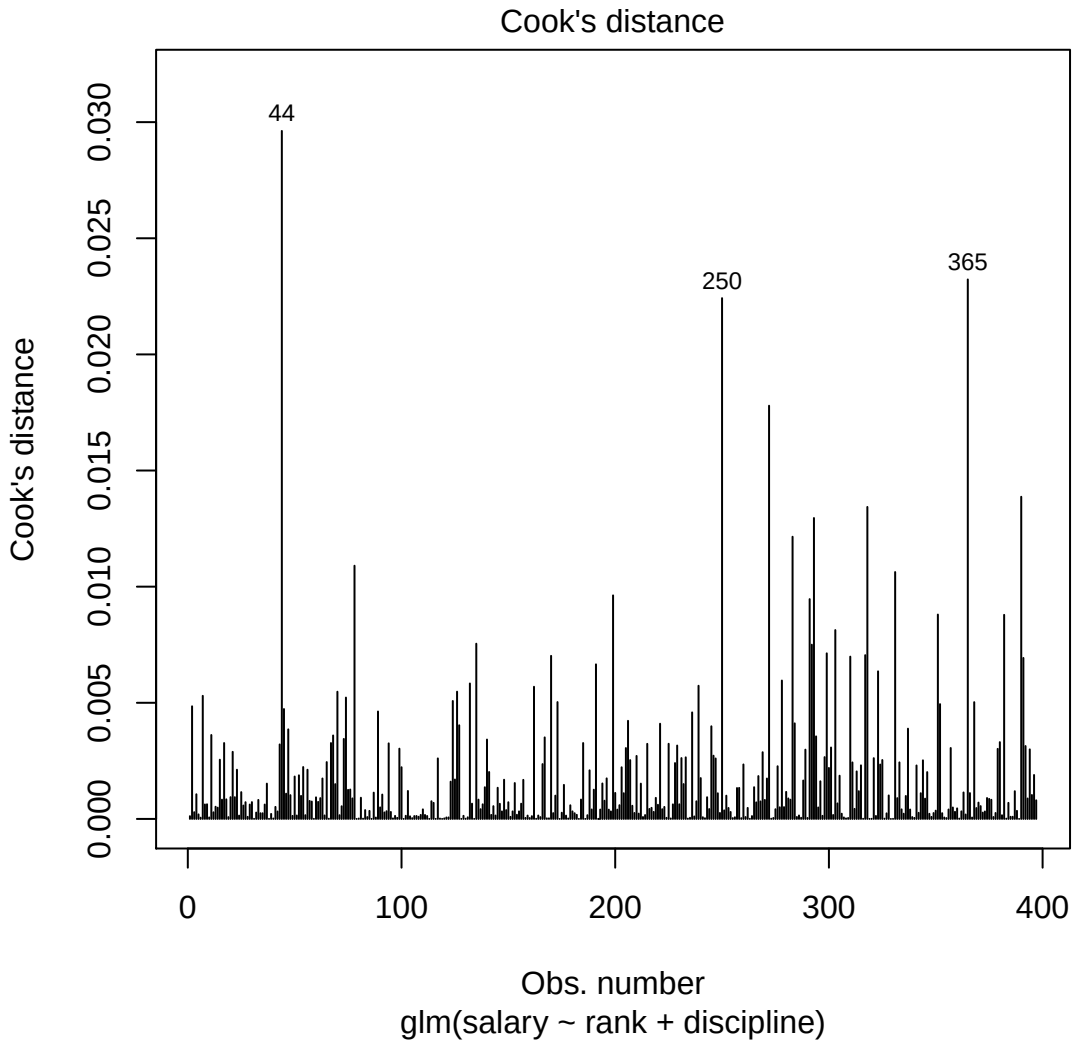

```
plot(linear.m3$residuals ~ salary$rank, ylab = "Residuals", xlab = "Rank")  # prof. variance is big
```



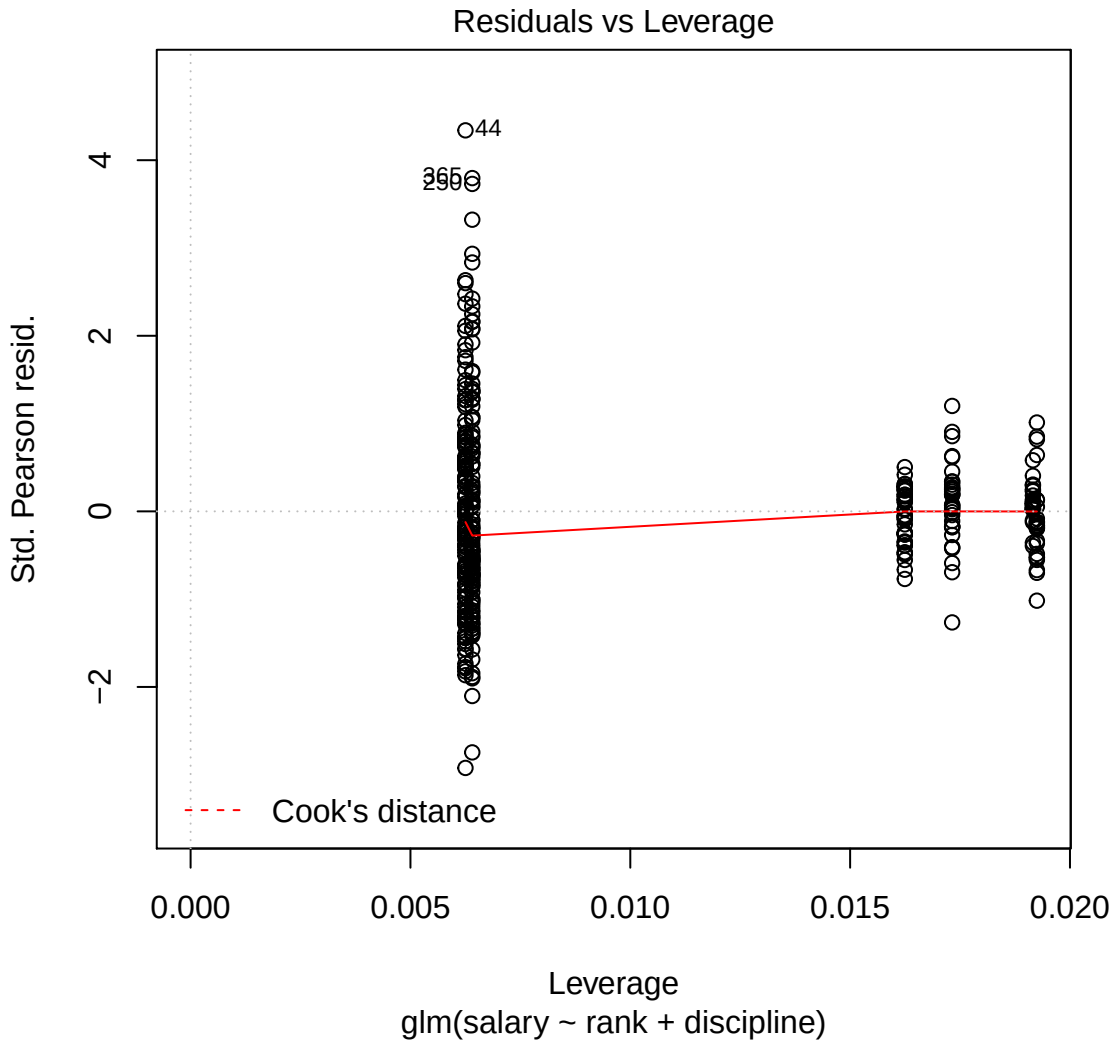

```
plot(linear.m3$residuals ~ salary$discipline, ylab = "Residuals", xlab = "Discipline")
```



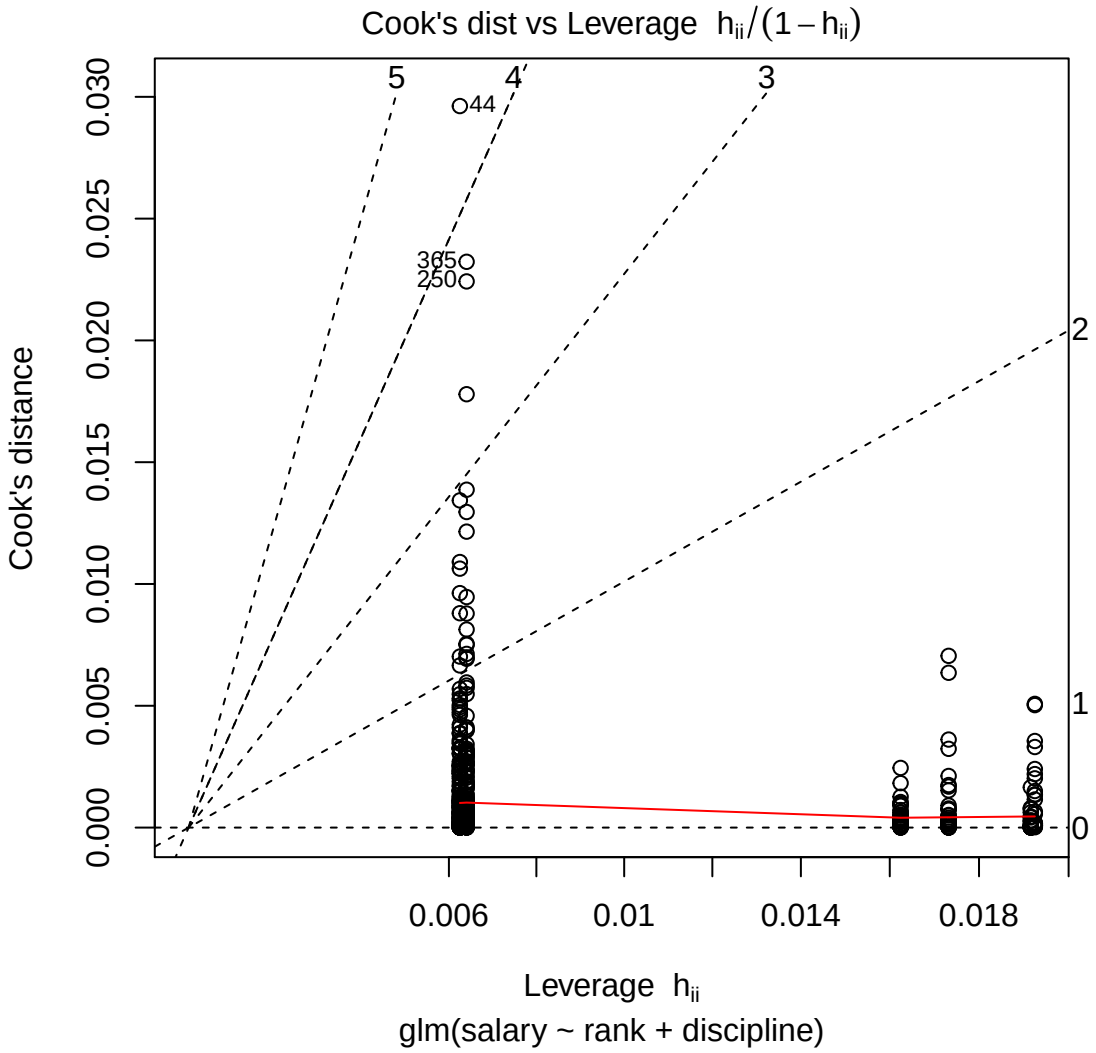

```
# - Influentials
plot(linear.m3, which = 4)  # all D < 1
```



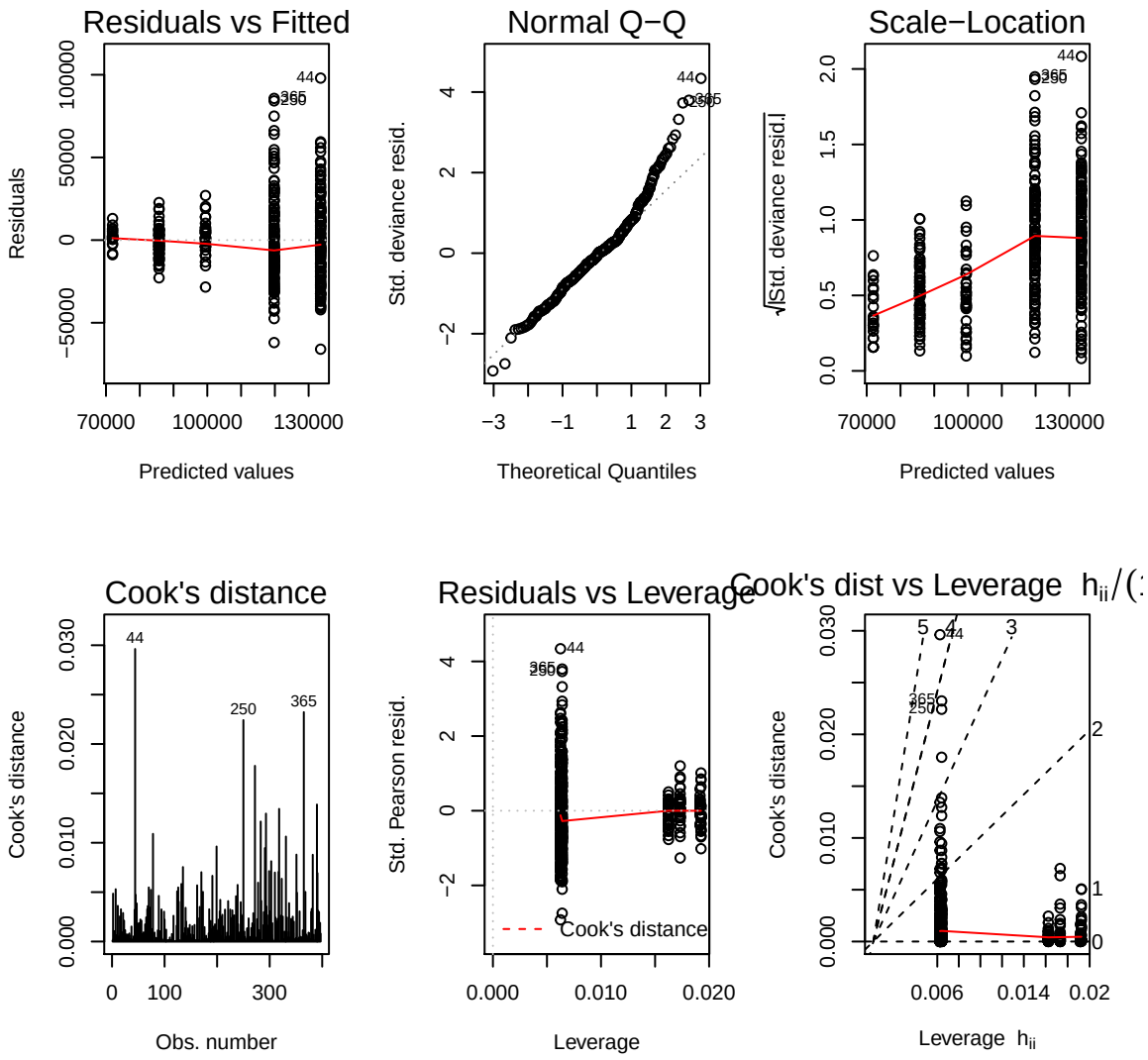

```
plot(linear.m3, which = 5)  # leverage < 0.5
```




```
plot(linear.m3, which = 6)
```




```
par(mfrow = c(2, 3))
plot(linear.m3, which = 1:6)
```




```
par(mfrow = c(1, 1))  # reset
# - May need to handle these influential cases, but beyond the context of this workshop -
# Somehow, ended up with only cat var, basically an ANOVA
summary(aov(linear.m3))
```

```
##              Df    Sum Sq   Mean Sq F value   Pr(>F)
## rank          2 1.432e+11 7.162e+10  139.58  < 2e-16 ***
## discipline    1 1.843e+10 1.843e+10   35.92 4.65e-09 ***
## Residuals   393 2.016e+11 5.131e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# - But it depends on your obj. of analysis, predict / compare groups
```

3.8 Final model

```
# - Accept linear.m3
summary(linear.m3)
```

```
##
## Call:
## glm(formula = salary ~ rank + discipline, data = salary)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -65990  -14049   -1288   10760   97996
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      71944       3135  22.948  < 2e-16 ***
## rankAssocProf    13762       3961   3.475 0.000569 ***
## rankProf         47844       3112  15.376  < 2e-16 ***
## disciplineB      13761       2296   5.993 4.65e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 513076201)
##
##     Null deviance: 3.6330e+11  on 396  degrees of freedom
## Residual deviance: 2.0164e+11  on 393  degrees of freedom
## AIC: 9094.8
##
## Number of Fisher Scoring iterations: 2
```

```
library(rsq)  # R^2 for GLM
rsq(linear.m3)
```

```
## [1] 0.4449805
```

```
# - salary ~ rank + discipline
final = cbind(salary[c("rank", "discipline", "salary")], predicted_salary = predict(linear.m3))
final_ranked = final[order(final$rank), ]
head(final_ranked)
```

```
##        rank discipline salary predicted_salary
## 3  AsstProf          B  79750         85705.28
## 12 AsstProf          B  79800         85705.28
## 13 AsstProf          B  77700         85705.28
## 14 AsstProf          B  78000         85705.28
## 28 AsstProf          B  82379         85705.28
## 29 AsstProf          B  77000         85705.28
```

```
tail(final_ranked)
```

```
##     rank discipline salary predicted_salary
## 391 Prof          A 166605         119788.2
## 392 Prof          A 151292         119788.2
## 393 Prof          A 103106         119788.2
## 394 Prof          A 150564         119788.2
## 395 Prof          A 101738         119788.2
## 396 Prof          A  95329         119788.2
```

```
# - review back levels/var
levels(salary$rank)
```

```
## [1] "AsstProf"  "AssocProf" "Prof"
```

```r
levels(salary$discipline)
```

```
## [1] "A" "B"
```

```r
# - if rank = 'Prof', discipline = 'B'
predict(linear.m3, list(rank = "Prof", discipline = "B"), se.fit = T)
```

```
## $fit
##        1
## 133549.1
##
## $se.fit
## [1] 1790.938
##
## $residual.scale
## [1] 22651.19
```

```r
head(salary[salary$rank == "Prof" & salary$discipline == "B", c("rank", "discipline", "salary")])
```

```
##   rank discipline salary
## 1 Prof          B 139750
## 2 Prof          B 173200
## 4 Prof          B 115000
## 5 Prof          B 141500
## 7 Prof          B 175000
## 8 Prof          B 147765
```

```r
mean(salary[salary$rank == "Prof" & salary$discipline == "B", "salary"])
```

```
## [1] 133393.8
```

```r
# - if rank = 'AsstProf', discipline = 'B'
predict(linear.m3, list(rank = "AsstProf", discipline = "B"), se.fit = T)
```

```
## $fit
##        1
## 85705.28
##
## $se.fit
## [1] 2886.917
##
## $residual.scale
## [1] 22651.19
```

```r
head(salary[salary$rank == "AsstProf" & salary$discipline == "B", c("rank", "discipline", "salary")])
```

```
##        rank discipline salary
## 3  AsstProf          B  79750
## 12 AsstProf          B  79800
## 13 AsstProf          B  77700
## 14 AsstProf          B  78000
## 28 AsstProf          B  82379
## 29 AsstProf          B  77000
```

```r
mean(salary[salary$rank == "AsstProf" & salary$discipline == "B", "salary"])
```

```
## [1] 84593.91
```

References

Fox, J., & Weisberg, S. (2017). *Car: Companion to applied regression.* Retrieved from https://CRAN.R-project.org/package=car

Kutner, M. H., Nachtscheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical model (*5th ed. Singapore: McGraw-Hill.

Revelle, W. (2017). *Psych: Procedures for psychological, psychometric, and personality research.* Retrieved from https://CRAN.R-project.org/package=psych

Zhang, D. (2017). *Rsq: R-squared and related measures.* Retrieved from https://CRAN.R-project.org/package=rsq