R Computation: Relaxing the model assumptions

Outlier We recall from previous results that for residual random vectors

$$\mathbf{Y} - \widehat{\mathbf{Y}}$$

we have (under correct specification)

$$\mathrm{Var}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y} - \widehat{\mathbf{Y}}|\mathbf{X}] = \sigma^2(\mathbf{I}_n - \mathbf{H}).$$

Taking the diagonal elements, this implies that the variance of the ith residual is

$$\sigma^2(1 - h_{ii})$$

which we may use as a means to process the residual so that it appears on a standard scale.

An **outlier** is a point for which the residual (or standardized residual) is large.

- Such points need to be considered carefully as they may exert a lot of influence on the fit.
- Outliers may need to be deleted from the data set.

Using standard large sample arguments, a data point may be considered an outlier if

$$\left| \frac{y_i - \widehat{y}_i}{\sqrt{\sigma^2(1 - h_{ii})}} \right| > 2$$

Leverage From standard theory, we have that

$$\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

where $\mathbf{H}$ is the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. Thus

$$\widehat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$$

The coefficients h_{ij} measure the importance of each of the original data $y_1, \ldots, y_n$ in predicting y_i. h_{ij} is termed the **leverage** of point j on point i. We have

$$h_{ii} = \mathbf{x}_i(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i^\top$$

and if this value is large, the data point i is considered influential.

Influence Consider the fit of a regression model to data indexed $i = 1, \ldots, n$, and consider refitting the model with the ith point deleted. Let

- $\mathbf{y}_{(i)}$ be the response vector with the ith response deleted;
- $\mathbf{X}_{(i)}$ be the $\mathbf{X}$ matrix with the ith row deleted.

The least squares estimate when point i is deleted is

$$\widehat{\beta}_{(i)} = (\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)})^{-1} \mathbf{X}_{(i)}^\top \mathbf{y}_{(i)}.$$

We then have the prediction at $\mathbf{x} = \mathbf{x}_i$ as

$$\widehat{y}_{(i)} = \mathbf{x}_i \widehat{\beta}_{(i)}.$$

We attempt to assess model validity using this 'out-of-sample' prediction. We compare estimates

- $\widehat{\beta}$ from the full data set
- $\widehat{\beta}_{(i)}$ when the ith data point is removed.

As well as the regression estimates, we also have the estimates of σ^2:

- $\widehat{\sigma}^2$ from the full data set
- $\widehat{\sigma}^2_{(i)}$ when the ith data point is removed.

We might use **Cook's distance** D_i for data point i

$$D_i = \frac{(\widehat{\beta}_{(i)} - \widehat{\beta})^\top (\mathbf{X}^\top \mathbf{X}) (\widehat{\beta}_{(i)} - \widehat{\beta})}{p\text{MS}_{Res}} = \frac{(\widehat{\mathbf{y}}_{(i)} - \widehat{\mathbf{y}})^\top (\widehat{\mathbf{y}}_{(i)} - \widehat{\mathbf{y}})}{p\text{MS}_{Res}}$$

as a global measure of influence on inference on a standardized scale.

Example: Life Cycle Data Under the life-cycle savings hypothesis as developed by Franco Modigliani, the savings ratio is explained by per-capita disposable income, the percentage rate of change in per-capita disposable income, and two demographic variables: the percentage of population less than 15 years old and the percentage of the population over 75 years old. The data are averaged over the decade 1960--1970 to remove the business cycle or other short-term fluctuations.

- predictor `pop15` -- % of population under 15

- predictor `pop75` -- % of population over 75
- predictor `dpi` -- real per-capita disposable income
- predictor `ddpi` -- % growth rate of dpi
- response `sr` -- savings ratio (aggregate personal saving divided by disposable income)

```
LifeCycleSavings

##                  sr pop15 pop75  dpi  ddpi
## Australia      11.4    29  2.87 2330  2.87
## Austria        12.1    23  4.41 1508  3.93
## Belgium        13.2    24  4.43 2108  3.82
## Bolivia         5.8    42  1.67  189  0.22
## Brazil         12.9    42  0.83  728  4.56
## Canada          8.8    32  2.85 2983  2.43
## Chile           0.6    40  1.34  663  2.67
## China          11.9    45  0.67  290  6.51
## Colombia        5.0    47  1.06  277  3.08
## Costa Rica     10.8    48  1.14  471  2.80
## Denmark        16.9    24  3.93 2497  3.99
## Ecuador         3.6    46  1.19  288  2.19
## Finland        11.2    28  2.37 1681  4.32
## France         12.6    25  4.70 2214  4.52
## Germany        12.6    23  3.35 2457  3.44
## Greece         10.7    26  3.10  871  6.28
## Guatamala       3.0    46  0.87  290  1.48
## Honduras        7.7    47  0.58  232  3.19
## Iceland         1.3    34  3.08 1900  1.12
## India           9.0    41  0.96   89  1.54
## Ireland        11.3    31  4.19 1140  2.99
## Italy          14.3    25  3.48 1390  3.54
## Japan          21.1    27  1.91 1257  8.21
## Korea           4.0    42  0.91  208  5.81
## Luxembourg     10.3    22  3.73 2449  1.57
## Malta          15.5    33  2.47  601  8.12
## Norway         10.2    26  3.67 2231  3.62
## Netherlands    14.7    25  3.25 1741  7.66
## New Zealand    10.7    33  3.17 1488  1.76
```

```
## Nicaragua       7.3    45  1.21  326  2.48
## Panama          4.4    44  1.20  569  3.61
## Paraguay        2.0    41  1.05  221  1.03
## Peru           12.7    44  1.28  400  0.67
## Philippines    12.8    46  1.12  152  2.00
## Portugal       12.5    29  2.85  580  7.48
## South Africa   11.1    32  2.28  651  2.19
## South Rhodesia 13.3    32  1.52  251  2.00
## Spain          11.8    28  2.87  769  4.35
## Sweden          6.9    21  4.54 3299  3.01
## Switzerland    14.1    23  3.73 2631  2.70
## Turkey          5.1    43  1.08  390  2.96
## Tunisia         2.8    46  1.21  250  1.13
## United Kingdom  7.8    23  4.46 1814  2.01
## United States   7.6    30  3.43 4002  2.45
## Venezuela       9.2    46  0.90  813  0.53
## Zambia         18.6    45  0.56  138  5.14
## Jamaica         7.7    41  1.73  380 10.23
## Uruguay         9.2    28  2.72  767  1.88
## Libya           8.9    44  2.07  124 16.71
## Malaysia        4.7    47  0.66  243  5.08
```

```r
str(LifeCycleSavings)
```

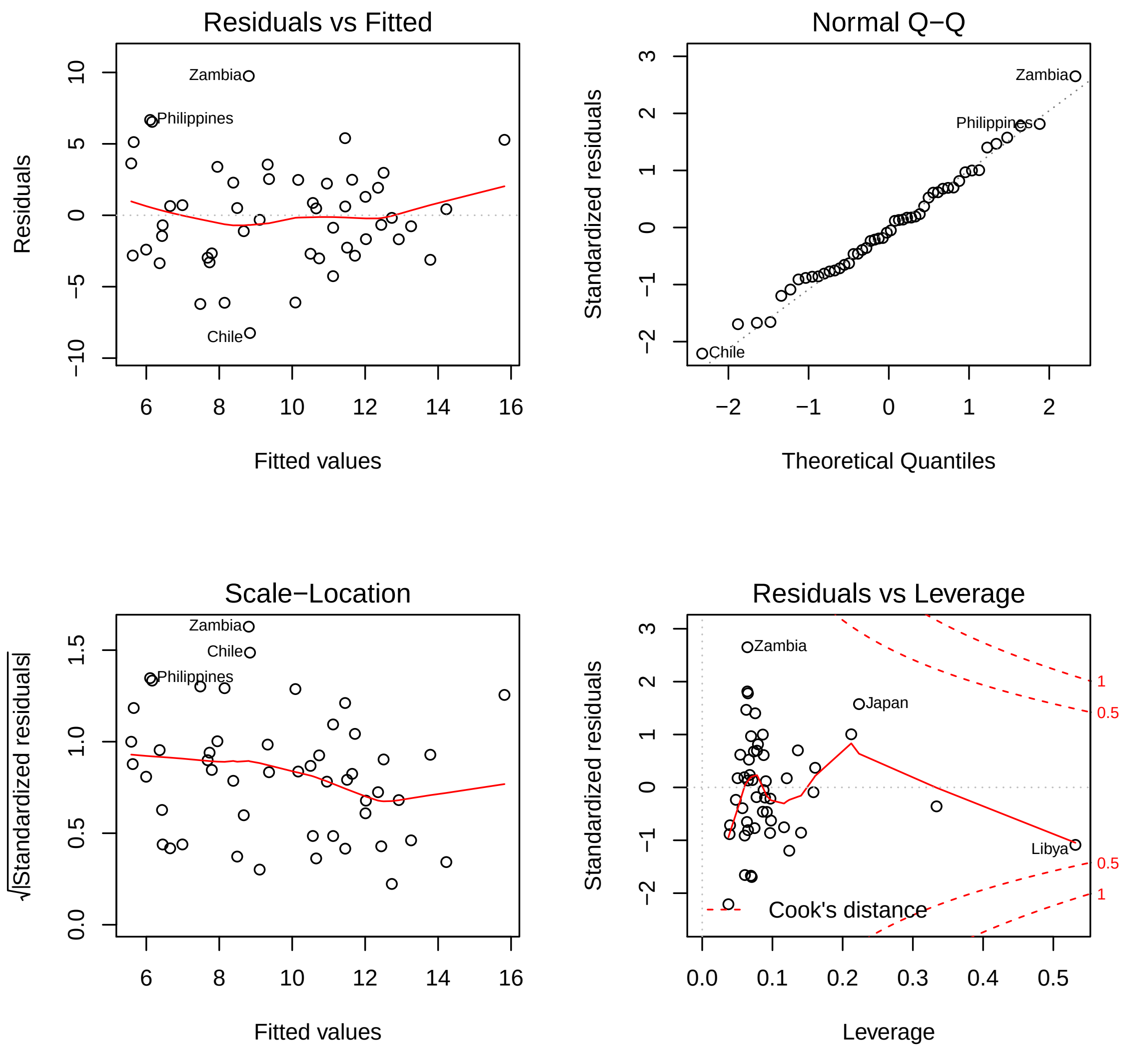
```
## 'data.frame': 50 obs. of  5 variables:
##  $ sr   : num  11.43 12.07 13.17 5.75 12.88 ...
##  $ pop15: num  29.4 23.3 23.8 41.9 42.2 ...
##  $ pop75: num  2.87 4.41 4.43 1.67 0.83 2.85 1.34 0.67 1.06 1.14 ...
##  $ dpi  : num  2330 1508 2108 189 728 ...
##  $ ddpi : num  2.87 3.93 3.82 0.22 4.56 2.43 2.67 6.51 3.08 2.8 ...
```

```r
fit1 <- lm(sr ~ pop15 + pop75 + dpi + ddpi,
    data = LifeCycleSavings)
summary(fit1)
```

```
##
## Call:
## lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = LifeCycleSavings)
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

```
## -8.242 -2.686 -0.249  2.428  9.751
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.566087   7.354516    3.88  0.00033 ***
## pop15       -0.461193   0.144642   -3.19  0.00260 **
## pop75       -1.691498   1.083599   -1.56  0.12553
## dpi         -0.000337   0.000931   -0.36  0.71917
## ddpi         0.409695   0.196197    2.09  0.04247 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.8 on 45 degrees of freedom
## Multiple R-squared:  0.338,Adjusted R-squared:  0.28
## F-statistic: 5.76 on 4 and 45 DF,  p-value: 0.00079
```

```r
par(mfrow = c(2, 2))
plot(fit1)
```



We can also compute the diagonals of the hat matrix (h_{ii})

```
inf.diags <- lm.influence(fit1)
data.frame(hat = inf.diags$hat[c(1:7, 42:50)])

##                    hat
## Australia        0.068
## Austria          0.120
## Belgium          0.087
```

```
## Bolivia        0.089
## Brazil         0.070
## Canada         0.158
## Chile          0.037
## Tunisia        0.075
## United Kingdom 0.117
## United States  0.334
## Venezuela      0.086
## Zambia         0.064
## Jamaica        0.141
## Uruguay        0.098
## Libya          0.531
## Malaysia       0.065
```

The deletion change in $\boldsymbol{\beta}$ estimates $\widehat{\beta}_{(i)} - \widehat{\beta}$ are

```
data.frame(signif(inf.diags$coef[c(1:7, 42:50),
    ], 4))
```

```
##                X.Intercept.    pop15  pop75     dpi
## Australia             0.092 -0.00153 -0.029  4.3e-05
## Austria              -0.075  0.00087  0.045 -3.5e-05
## Belgium              -0.475  0.00750  0.132 -3.3e-05
## Bolivia               0.043 -0.00186 -0.025  3.0e-05
## Brazil                0.660 -0.00892 -0.194  1.1e-04
## Canada                0.040 -0.00099  0.011 -3.3e-05
## Chile                -1.400  0.01832  0.227 -1.8e-05
## Tunisia               0.545 -0.01526 -0.084  4.2e-05
## United Kingdom        0.345 -0.00521 -0.186  1.2e-04
## United States         0.513 -0.01065  0.041 -2.2e-04
## Venezuela            -0.374  0.01458 -0.036  1.1e-04
## Zambia                1.118 -0.01064 -0.341  8.1e-05
## Jamaica               0.808 -0.01454 -0.062 -6.6e-06
## Uruguay              -0.993  0.01876  0.032  1.2e-04
## Libya                 4.042 -0.06975 -0.411 -1.8e-05
## Malaysia              0.272 -0.00888  0.035 -4.6e-05
##                    ddpi
## Australia      -3.2e-05
## Austria        -1.6e-03
## Belgium        -1.4e-03
```

```
## Bolivia         8.1e-03
## Brazil          1.3e-02
## Canada         -5.3e-04
## Chile           2.2e-02
## Tunisia         2.0e-02
## United Kingdom  2.0e-02
## United States  -6.5e-03
## Venezuela      -2.4e-02
## Zambia          4.2e-02
## Jamaica        -5.8e-02
## Uruguay         2.0e-02
## Libya          -2.0e-01
## Malaysia       -1.4e-02
```

The Cook's distance and Leverage can be computed by

```
influence.measures(fit1)

## Influence measures of
##   lm(formula = sr ~ pop15 + pop75 + dpi + ddpi, data = LifeCycleSavings) :
##
##                  dfb.1_ dfb.pp15 dfb.pp75  dfb.dpi
## Australia       0.01232 -0.01044 -0.02653  0.04534
## Austria        -0.01005  0.00594  0.04084 -0.03672
## Belgium        -0.06416  0.05150  0.12070 -0.03472
## Bolivia         0.00578 -0.01270 -0.02253  0.03185
## Brazil          0.08973 -0.06163 -0.17907  0.11997
## Canada          0.00541 -0.00675  0.01021 -0.03531
## Chile          -0.19941  0.13265  0.21979 -0.01998
## China           0.02112 -0.00573 -0.08311  0.05180
## Colombia        0.03910 -0.05226 -0.02464  0.00168
## Costa Rica     -0.23367  0.28428  0.14243  0.05638
## Denmark        -0.04051  0.02093  0.04653  0.15220
## Ecuador         0.07176 -0.09524 -0.06067  0.01950
## Finland        -0.11350  0.11133  0.11695 -0.04364
## France         -0.16600  0.14705  0.21900 -0.02942
## Germany        -0.00802  0.00822  0.00835 -0.00697
## Greece         -0.14820  0.16394  0.02861  0.15713
## Guatamala       0.01552 -0.05485  0.00614  0.00585
## Honduras       -0.00226  0.00984 -0.01020  0.00812
```

```
## Iceland         0.24789 -0.27355 -0.23265 -0.12555
## India           0.02105 -0.01577 -0.01439 -0.01374
## Ireland        -0.31001  0.29624  0.48156 -0.25733
## Italy           0.06619 -0.07097  0.00307 -0.06999
## Japan           0.63987 -0.65614 -0.67390  0.14610
## Korea          -0.16897  0.13509  0.21895  0.00511
## Luxembourg     -0.06827  0.06888  0.04380 -0.02797
## Malta           0.03652 -0.04876  0.00791 -0.08659
## Norway          0.00222 -0.00035 -0.00611 -0.01594
## Netherlands     0.01395 -0.01674 -0.01186  0.00433
## New Zealand    -0.06002  0.06510  0.09412 -0.02638
## Nicaragua      -0.01209  0.01790  0.00972 -0.00474
## Panama          0.02828 -0.05334  0.01446 -0.03467
## Paraguay       -0.23227  0.16416  0.15826  0.14361
## Peru           -0.07182  0.14669  0.09148 -0.08585
## Philippines    -0.15707  0.22681  0.15743 -0.11140
## Portugal       -0.02140  0.02551 -0.00380  0.03991
## South Africa    0.02218 -0.02030 -0.00672 -0.02049
## South Rhodesia  0.14390 -0.13472 -0.09245 -0.06956
## Spain          -0.03035  0.03131  0.00394  0.03512
## Sweden          0.10098 -0.08162 -0.06166 -0.25528
## Switzerland     0.04323 -0.04649 -0.04364  0.09093
## Turkey         -0.01092 -0.01198  0.02645  0.00161
## Tunisia         0.07377 -0.10500 -0.07727  0.04439
## United Kingdom  0.04671 -0.03584 -0.17129  0.12554
## United States   0.06910 -0.07289  0.03745 -0.23312
## Venezuela      -0.05083  0.10080 -0.03366  0.11366
## Zambia          0.16361 -0.07917 -0.33899  0.09406
## Jamaica         0.10958 -0.10022 -0.05722 -0.00703
## Uruguay        -0.13403  0.12880  0.02953  0.13132
## Libya           0.55074 -0.48324 -0.37974 -0.01937
## Malaysia        0.03684 -0.06113  0.03235 -0.04956
##                 dfb.ddpi   dffit cov.r   cook.d    hat inf
## Australia      -0.000159  0.0627 1.193 8.04e-04 0.0677
## Austria        -0.008182  0.0632 1.268 8.18e-04 0.1204
## Belgium        -0.007265  0.1878 1.176 7.15e-03 0.0875
## Bolivia         0.040642 -0.0597 1.224 7.28e-04 0.0895
## Brazil          0.068457  0.2646 1.082 1.40e-02 0.0696
## Canada         -0.002649 -0.0390 1.328 3.11e-04 0.1584
## Chile           0.120007 -0.4554 0.655 3.78e-02 0.0373   *
```

```
## China           0.110627  0.2008 1.150 8.16e-03 0.0780
## Colombia        0.009084 -0.0960 1.167 1.88e-03 0.0573
## Costa Rica     -0.032824  0.4049 0.968 3.21e-02 0.0755
## Denmark         0.048854  0.3845 0.934 2.88e-02 0.0627
## Ecuador         0.047786 -0.1695 1.139 5.82e-03 0.0637
## Finland        -0.017132 -0.1464 1.203 4.36e-03 0.0920
## France          0.023952  0.2765 1.226 1.55e-02 0.1362
## Germany        -0.000293 -0.0152 1.226 4.74e-05 0.0874
## Greece         -0.059599 -0.2811 1.140 1.59e-02 0.0966
## Guatamala       0.097217 -0.2305 1.085 1.07e-02 0.0605
## Honduras       -0.001887  0.0482 1.186 4.74e-04 0.0601
## Iceland         0.184698 -0.4768 0.866 4.35e-02 0.0705
## India          -0.018958  0.0381 1.202 2.97e-04 0.0715
## Ireland        -0.093317  0.5216 1.268 5.44e-02 0.2122
## Italy          -0.028648  0.1388 1.162 3.92e-03 0.0665
## Japan           0.388603  0.8597 1.085 1.43e-01 0.2233
## Korea          -0.169492 -0.4303 0.870 3.56e-02 0.0608
## Luxembourg      0.049134 -0.1401 1.196 3.99e-03 0.0863
## Malta           0.153014  0.2386 1.128 1.15e-02 0.0794
## Norway         -0.001462 -0.0522 1.168 5.56e-04 0.0479
## Netherlands     0.022591  0.0366 1.229 2.74e-04 0.0906
## New Zealand    -0.064740  0.1469 1.134 4.38e-03 0.0542
## Nicaragua      -0.010467  0.0397 1.174 3.23e-04 0.0504
## Panama         -0.007889 -0.1775 1.067 6.33e-03 0.0390
## Paraguay        0.270478 -0.4655 0.873 4.16e-02 0.0694
## Peru           -0.287184  0.4811 0.831 4.40e-02 0.0650
## Philippines    -0.170674  0.4884 0.818 4.52e-02 0.0643
## Portugal       -0.028011 -0.0690 1.233 9.73e-04 0.0971
## South Africa   -0.016326  0.0343 1.195 2.41e-04 0.0651
## South Rhodesia -0.057920  0.1607 1.313 5.27e-03 0.1608
## Spain           0.005340 -0.0526 1.208 5.66e-04 0.0773
## Sweden         -0.013316 -0.4526 1.086 4.06e-02 0.1240
## Switzerland    -0.018828  0.1903 1.147 7.33e-03 0.0736
## Turkey          0.025138 -0.1445 1.100 4.22e-03 0.0396
## Tunisia         0.103058 -0.2177 1.131 9.56e-03 0.0746
## United Kingdom  0.100314 -0.2722 1.189 1.50e-02 0.1165
## United States  -0.032729 -0.2510 1.655 1.28e-02 0.3337   *
## Venezuela      -0.124486  0.3071 1.095 1.89e-02 0.0863
## Zambia          0.228232  0.7482 0.512 9.66e-02 0.0643   *
## Jamaica        -0.295461 -0.3456 1.200 2.40e-02 0.1408
```

```
## Uruguay          0.099591 -0.2051 1.187 8.53e-03 0.0979
## Libya           -1.024477 -1.1601 2.091 2.68e-01 0.5315   *
## Malaysia        -0.072294 -0.2126 1.113 9.11e-03 0.0652
```

Assessing Normality via Probability Plots As a further assessment of the model assumptions in linear regression model, we may use probability plots to assess distributional assumptions concerning the residual errors ϵ_i.

Recall that we assume

$$\mathbb{E}[\mathbf{Y}|\mathbf{X}] = \mathbf{X}\beta \qquad \mathrm{Var}[\mathbf{Y}|\mathbf{X}] = \sigma^2\mathbf{I}_n$$

that is

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

where

$$\mathbb{E}[\epsilon|\mathbf{X}] = \mathbf{0} \qquad \mathrm{Var}[\epsilon|\mathbf{X}] = \sigma^2\mathbf{I}_n$$

In a parametric analysis, we presume $\boldsymbol{\epsilon} \sim \mathrm{Normal}(\mathbf{0}, \sigma^2\mathbf{I}_n)$.

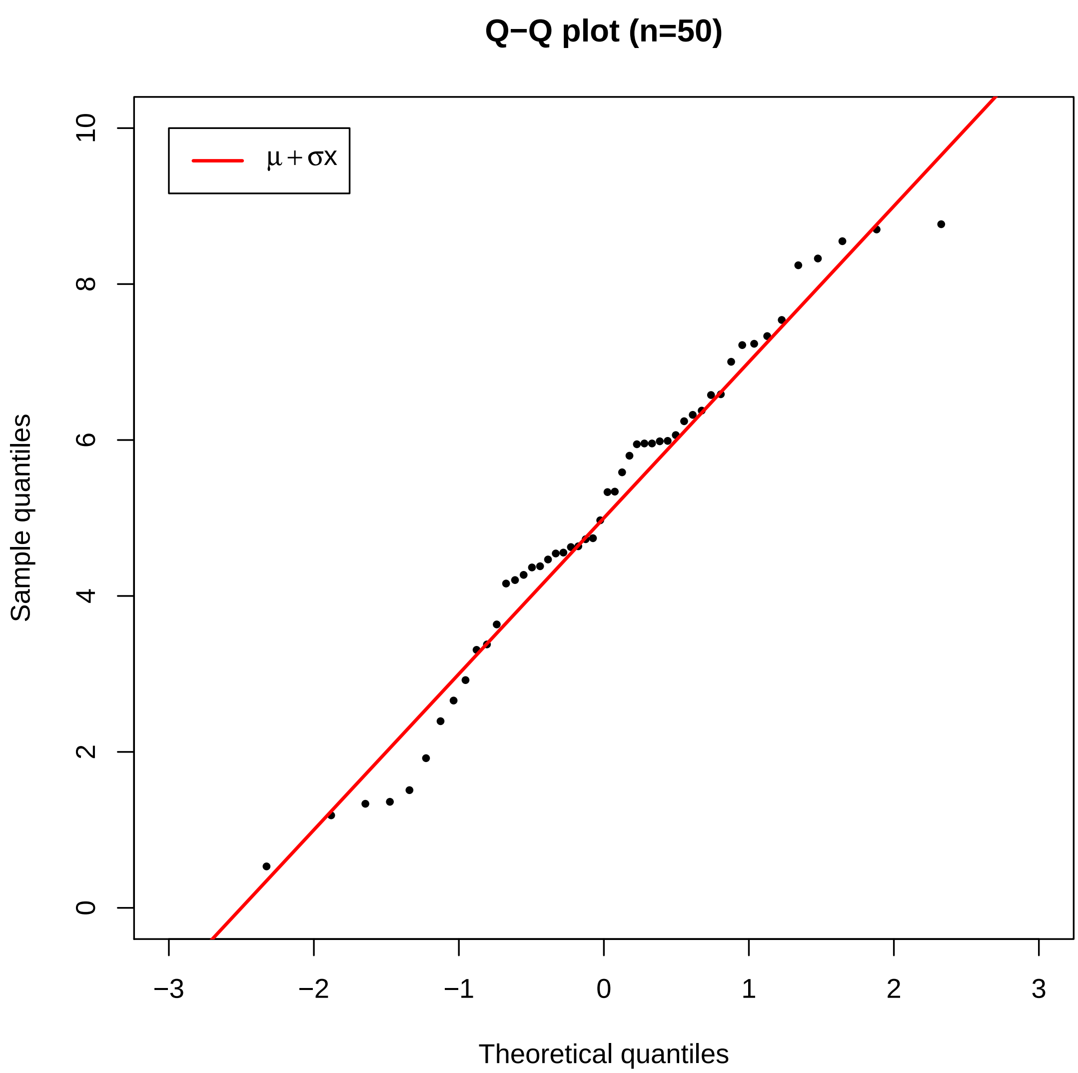
For a collection of residuals $e_i, i = 1, \ldots, n$, we may check whether the Normality assumption is violated using probability plotting.

Q-Q plot:

- x-axis: the values
$$\Phi^{-1}\left(\frac{i - 1/2}{n}\right) \qquad i = 1, \ldots, n$$
where $\Phi^{-1}(.)$ is the standard normal inverse cdf;
- y-axis: the residuals sorted into ascending order.

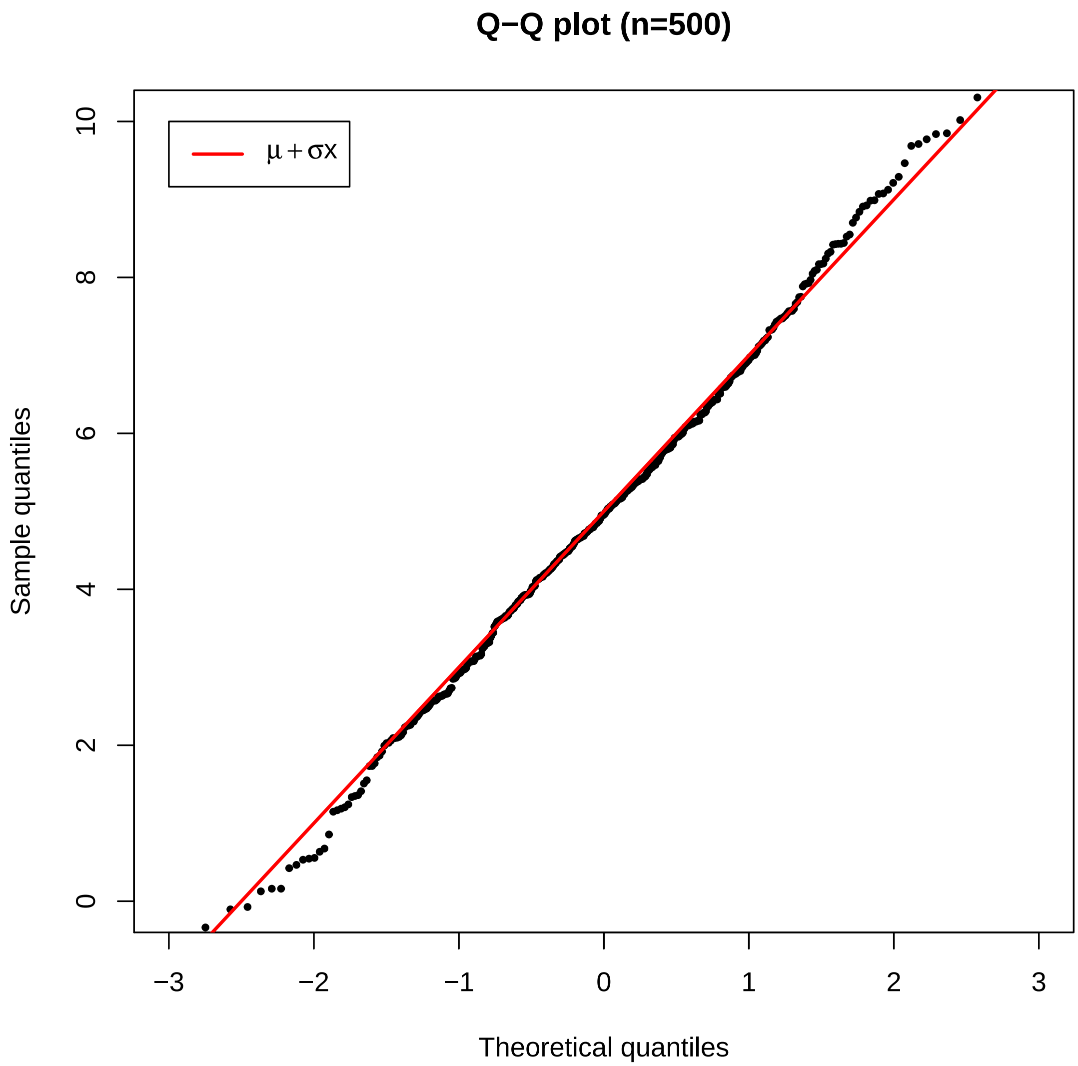
```r
# Q-Q n=50
set.seed(54)
n <- 50
X <- rnorm(n, 5, 2)
xvec.qq <- qnorm((c(1:n) - 0.5)/n)
yvec.qq <- sort(X)
plot(xvec.qq, yvec.qq, pch = 19, cex = 0.5,
    xlim = range(-3, 3), ylim = range(0,
        10), xlab = "Theoretical quantiles",
    ylab = "Sample quantiles")
```

```r
abline(5, 2, col = "red", lwd = 2)
title("Q-Q plot (n=50)")
legend(-3, 10, c(expression(mu + sigma *
    x)), lwd = 2, col = "red")
```



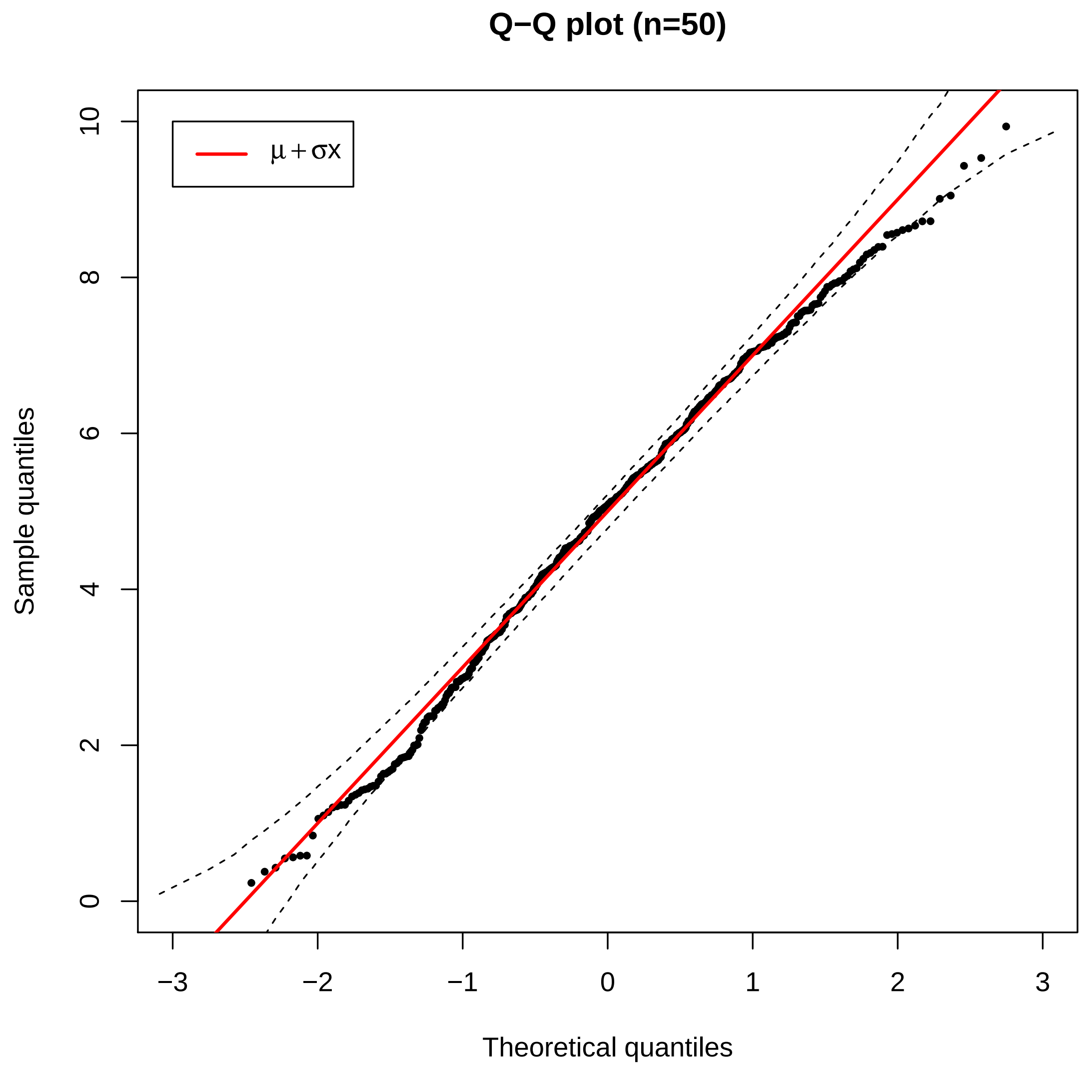

```r
# Q-Q n=500
set.seed(54)
n <- 500
```

```r
X <- rnorm(n, 5, 2)
xvec.qq <- qnorm((c(1:n) - 0.5)/n)
yvec.qq <- sort(X)
plot(xvec.qq, yvec.qq, pch = 19, cex = 0.5,
    xlim = range(-3, 3), ylim = range(0,
        10), xlab = "Theoretical quantiles",
    ylab = "Sample quantiles")
abline(5, 2, col = "red", lwd = 2)
title("Q-Q plot (n=500)")
legend(-3, 10, c(expression(mu + sigma *
    x)), lwd = 2, col = "red")
```




```
nreps <- 5000
set.seed(2332)
qq.vals50 <- matrix(0, nrow = nreps, ncol = 50)
qq.vals500 <- matrix(0, nrow = nreps, ncol = 500)
for (irep in 1:nreps) {
    n <- 50
    X <- rnorm(n, 5, 2)
    qq.vals50[irep, ] <- sort(X)
```

```r
    n <- 500
    X <- rnorm(n, 5, 2)
    qq.vals500[irep, ] <- sort(X)
}
n <- 50
xvec.qq.50 <- qnorm((c(1:n) - 0.5)/n)
n <- 500
xvec.qq.500 <- qnorm((c(1:n) - 0.5)/n)
qq.ci.50 <- apply(qq.vals50, 2, quantile,
    probs = c(0.025, 0.975))
qq.ci.500 <- apply(qq.vals500, 2, quantile,
    probs = c(0.025, 0.975))
# Q-Q
xvec.qq <- qnorm((c(1:n) - 0.5)/n)
yvec.qq <- sort(X)
plot(xvec.qq.500, qq.ci.500[1, ], ylim = range(0,
    10), xlim = range(-3, 3), xlab = "Theoretical quantiles",
    ylab = "Sample quantiles", type = "l",
    lty = 2)
lines(xvec.qq.500, qq.ci.500[2, ], lty = 2)
points(xvec.qq, yvec.qq, pch = 19, cex = 0.5)
abline(5, 2, col = "red", lwd = 2)
title("Q-Q plot (n=50)")
legend(-3, 10, c(expression(mu + sigma *
    x)), lwd = 2, col = "red")
```




```
# Q-Q
xvec.qq <- qnorm((c(1:n) - 0.5)/n)
yvec.qq <- sort(X)
plot(xvec.qq.50, qq.ci.50[1, ], ylim = range(0,
    10), xlim = range(-3, 3), xlab = "Theoretical quantiles",
    ylab = "Sample quantiles", type = "l",
    lty = 2)
lines(xvec.qq.50, qq.ci.50[2, ], lty = 2)
```

```r
points(xvec.qq, yvec.qq, pch = 19, cex = 0.5)
abline(5, 2, col = "red", lwd = 2)
title("Q-Q plot (n=500)")
legend(-3, 10, c(expression(mu + sigma *
    x)), lwd = 2, col = "red")
```