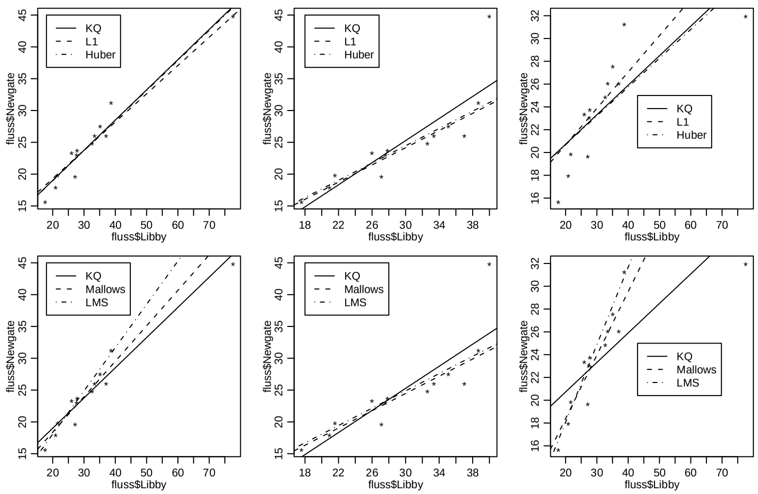
Lösungsskizze Serie 5

1. a)

	1.		2.		3.	
Beob. (x, y)	(77.6, 44.9)		(40.0, 44.9)		(77.6, 32.0)	
	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
KQ	9.51	0.47	-0.76	0.87	15.54	0.26
L_1	10.41	0.44	3.87	0.68	14.29	0.32
Huber	9.52	0.47	4.10	0.68	15.62	0.25
Mallows	7.24	0.56	4.13	0.68	7.24	0.56
LMS	4.47	0.68	4.47	0.68	4.47	0.68

b)



2. a) Gemäss Skript ist die asymptotische Kovarianzmatrix der Huber-Regression gegeben durch

$$\frac{\mathcal{E}[\psi_c(\varepsilon_i/\sigma)^2]}{P[|\varepsilon_i| \le c\sigma]^2}\sigma^2\,\mathcal{E}\left[x_i x_i^T\right]^{-1}.$$

Es gilt: $P[|\varepsilon_i| \le c\sigma] \approx 2f_\varepsilon(0)c\sigma$ für c klein und $\lim_{c\downarrow 0}\frac{1}{c^2}\,\mathcal{E}[\psi_c(\varepsilon_i/\sigma)^2] = 1$.
Also gilt:

$$\lim_{c\downarrow 0}\frac{\mathcal{E}[\psi_c(\varepsilon_i/\sigma)^2]}{P[|\varepsilon_i| \le c\sigma]^2}\sigma^2\,\mathcal{E}\left[x_i x_i^T\right]^{-1} = \frac{1}{(2f_\varepsilon(0)\sigma)^2}\sigma^2\,\mathcal{E}\left[x_i x_i^T\right]^{-1} = \frac{1}{4(f_\varepsilon(0))^2}\,\mathcal{E}\left[x_i x_i^T\right]^{-1}.$$

b) Die asymptotische Kovarianz des KQ-Schätzers ist $V_{KQ} = \sigma^2\,\mathcal{E}\left[x_i x_i^T\right]^{-1}$. Bezeichnen wir die asymptotische Kovarianz des L_1-Schätzers mit V_{L_1}, so ist das Verhältnis der Genauigkeiten der beiden Schätzer durch den folgenden Ausdruck gegeben:

$$\frac{V_{KQ}}{V_{L_1}} = 4\sigma^2(f_\varepsilon(0))^2$$

(Dieser Ausdruck wird auch asymptotische relative Effizienz des L_1-Schätzers bezüglich dem KQ-Schätzer genannt.)
Für die beiden Fällen gilt nun:

1) ε normalverteilt: $\frac{V_{KQ}}{V_{L_1}} = \frac{2}{\pi} \approx 0.64$.
Die L_1-Regression schneidet also wesentlich schlechter ab.
2) ε t_3-verteilt: $\frac{V_{KQ}}{V_{L_1}} = 4 * 3 * \frac{4}{3\pi^2} \approx 1.62$.
Hier wäre die L_1-Regression vorzuziehen.

3. Bemerkung: Der vollständige R-Code, für diese Aufgabe befindet sich am Ende der Musterlösung.

a) Kleinste Quadrate Schätzer:

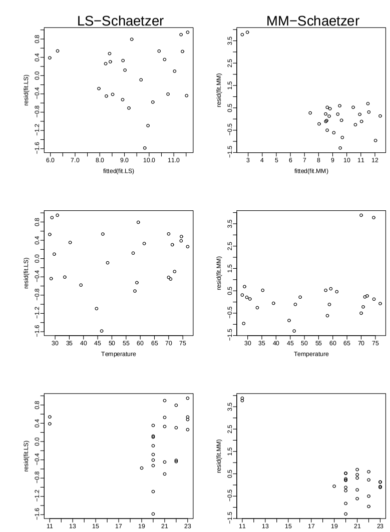
```
> fit.LS <- lm(Steam ~ Operating.Days + Temperature)
> summary(fit.LS)
...
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      9.126885   1.102801   8.276 3.35e-08 ***
Operating.Days   0.202815   0.045768   4.431 0.000210 ***
Temperature     -0.072393   0.007999  -9.050 7.19e-09 ***
...
```

MM-Schätzer:

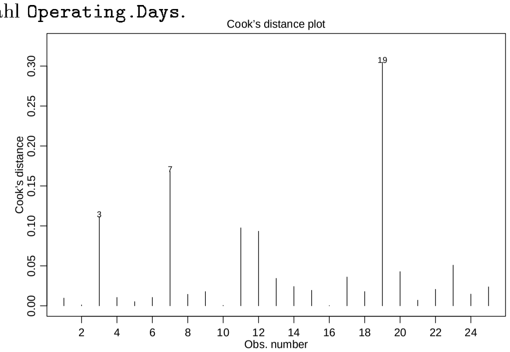
```
> fit.MM <- rlm(Steam ~ Operating.Days+Temperature, method="MM")
> summary(fit.MM)
...
Coefficients:
                 Value    Std. Error t value
(Intercept)       3.0872   0.9692     3.1855
Operating.Days    0.5147   0.0402    12.7971
Temperature      -0.0829   0.0070   -11.7867
...
```

Bemerkung: Ausser bei der erklärenden Variable `Temperatur` unterscheiden sich die Koeffizienten (gemessen an der Standardabweichung), die man mit dem Kleinste Quadrate resp. dem MM-Schätzer erhält, sehr stark.

b) Die beiden Beobachtungen mit sehr kleiner Anzahl `Operating.Days` werden bei der MM-Schätzung als Ausreisser / Hebelpunkte erkannt und dementsprechend behandelt. Die restlichen Residuen zeigen beim MM-Schätzer keine unerwünschten Strukturen mehr. Bei der gewöhnlichen Regression werden die beiden Hebelpunkte nicht als "bösartig" erkannt.



c) Vorallem die Beobachtung 19, aber auch die Beobachtungen 7 und 3 haben eine grossen Einfluss auf die Parameterschätzungen. Die Beobachtungen 7 und 19 sind diejenigen mit einer kleinen Anzahl `Operating.Days`.



d) **Kleinste Quadrate Schätzer:**

```
> fit2.LS <- lm(Steam ~ Operating.Days + Temperature + Working.Holidays)
> summary(fit2.LS)
...
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.812157   1.970691   1.427  0.16828
Operating.Days   0.524734   0.096959   5.412 2.28e-05 ***
Temperature     -0.082215   0.007002 -11.742 1.09e-10 ***
Working.Holidays 3.946691   1.099371   3.590  0.00172 **
...
```

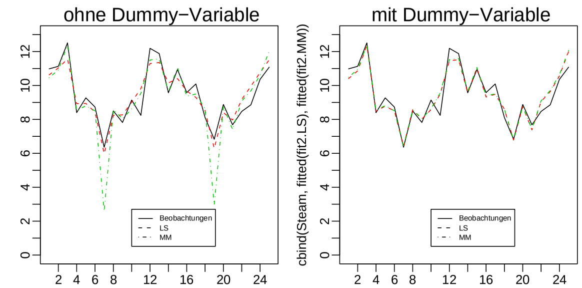
MM-Schätzer:

```
> fit2.MM <- rlm(Steam ~ Operating.Days + Temperature + Working.Holidays , method="MM")
> summary(fit2.MM)
...
Coefficients:
                Value    Std. Error t value
(Intercept)       3.2108   1.9843     1.6181
Operating.Days    0.5107   0.0976     5.2315
Temperature      -0.0833   0.0071   -11.8176
Working.Holidays  3.7816   1.1070     3.4162
...
```

Bemerkung: Die Koeffizienten-Schätzungen stimmen bedeutend besser überein. Die Residuen-Plots (hier nicht eingefügt) zeigen keine unerwünschten Strukturen mehr. Die Residuen an den beiden Hebelpunkten sind nun auch beim MM-Schätzer sehr klein, da die Dummy-Variable `Working.Holidays` ins Modell eingebaut wurde.

e) Ohne Dummy-Variable `Working.Holidays`:
Der robuste Fit zeigt deutlich, dass das Modell für die extrem tiefen Werte schlecht passt.

Mit Dummy-Variable `Working.Holidays`:
Die "fitted values" unterscheiden sich kaum voneinander.



R-Code:

```
## Daten einlesen
D.steam <- read.table.url("http://stat.ethz.ch/Teaching/Datasets/dsteam.dat",header=T)
library("MASS"); library("lqs")
attach(D.steam)

## a)
fit.LS <- lm(Steam ~ Operating.Days + Temperature)
summary(fit.LS)
fit.MM <- rlm(Steam ~ Operating.Days+Temperature, method="MM")
summary(fit.MM)
## b)
par(mfcol=c(3,2), mar=c(5,5,3,1), mgp=c(2,0.5,0)) #Grafikparameter setzen
plot(fitted(fit.LS), resid(fit.LS), cex=1)
mtext("LS-Schaetzer", side=3, line=0.1, cex=1.5)
plot(Temperature, resid(fit.LS), cex=1)
plot(Operating.Days, resid(fit.LS), cex=1)
plot(fitted(fit.MM), resid(fit.MM), cex=1)
mtext("MM-Schaetzer", side=3, line=0.1, cex=1.5)
plot(Temperature, resid(fit.MM), cex=1)
plot(Operating.Days, resid(fit.MM), cex=1)
## c)
par(mfrow=c(2,2))
plot(fit.LS)
## d)
fit2.LS <- lm(Steam ~ Operating.Days + Temperature + Working.Holidays)
summary(fit2.LS)
fit2.MM <- rlm(Steam ~ Operating.Days + Temperature + Working.Holidays , method="MM")
summary(fit2.MM)
par(mfrow=c(1,2), mar=c(5,5,3,1), mgp=c(2,0.5,0))
plot(fitted(fit2.LS), resid(fit2.LS), cex=1)
mtext("LS-Schaetzer", side=3, line=0.1, cex=1.5)
plot(fitted(fit2.MM), resid(fit2.MM), cex=1)
mtext("MM-Schaetzer", side=3, line=0.1, cex=1.5)
## e)
par(mfrow=c(1,2), mar=c(4,3,3,1), mgp=c(2,0.5,0), pty="s")
matplot(cbind(Steam, fitted(fit.LS), fitted(fit.MM)), type="l",ylim=c(0,13), lty=c(1,2,4), cex=1,ylab="")
legend(10,2.7, c("Beobachtungen", "LS", "MM"), lty=c(1,2,4), cex=0.6)
mtext("ohne Dummy-Variable", side=3, line=0.1, cex=1.5)
matplot(cbind(Steam, fitted(fit2.LS), fitted(fit2.MM)), type="l",ylim=c(0,13), lty=c(1,2,4), cex=1)
legend(10,2.7, c("Beobachtungen", "LS", "MM"), lty=c(1,2,4), cex=0.6)
mtext("mit Dummy-Variable", side=3, line=0.1, cex=1.5)

detach('D.steam')
```