n|w

Multiple Linear Regression by Medians[1]

Beat Hulliger, FHNW School of Business

SST-2018, Zürich

[1] Partly financed by Hasler-Stiftung

Introduction

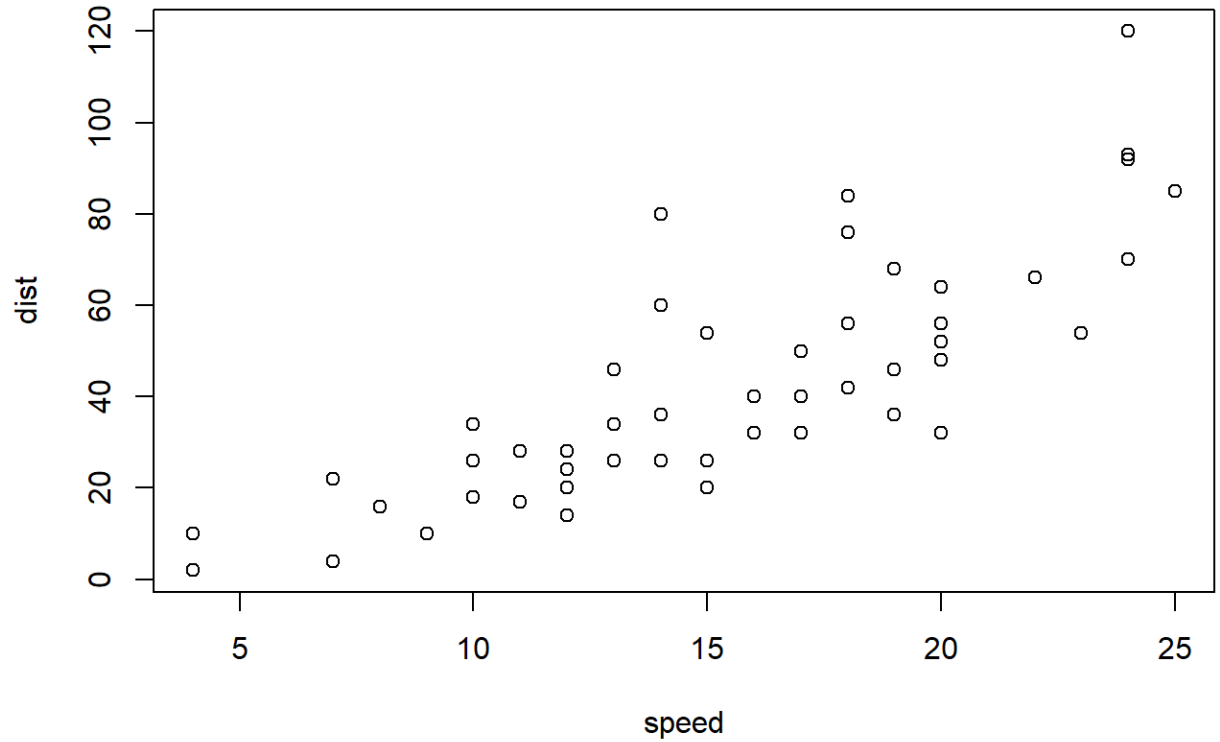


Figure: cars dataset in R (Ezekiel, M. (1930) Methods of Correlation Analysis. Wiley.)

Motivation

- Developing a simple method for the survey on added value in Switzerland: need for a simple to implement robust method: One-step from median regression.
- Separation of checking outliers and then downweighting them explicitly (editing and imputation) (Hulliger, 1999).
- Starting value for more sophisticated robust regression methods based on M-estimation (Andrews, 1974).
- Hand calculation of a resistant line (Tukey, 1977).

Where is the Problem?

- How to extend the simple linear regression ideas to multiple regression?
- Idea: Only rough outlier detection needed, then submit to clerical work.
- Which of the proposals for a simple robust line fit should be used?
- How to enhance initial solutions?

Notation

Sample $(x_i, y_i), i = 1, \ldots, n$.
Simple linear regression: $y_i = \beta_0 + \beta_1 x_i + E_i$ with $E[E_i] = 0$, $V[E_i] = t_i \sigma_E^2$ (with $t_i = x_i^q$ for $q = 0, 1, 2$).
For survey sampling assume weights $w_i, i = 1, \ldots, n$.
Multiple linear regression: $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + E_i$
Median of part L of sample: $M(y_L)$.
Weighted median of part L of sample: $M(y_L, w_L)$.
Centering at median: $x_i' = x_i - M(x_i)$

Simple linear regression

median-median-line : Split sample into L, M, R part and use

$$b_1 = \frac{M(y_R) - M(y_L)}{M(x_R) - M(x_L)}$$

$$b_0 = M(y_L) - b_1 M(x_L) + \tfrac{1}{3}(M(y_M) - b_1 M(x_M))$$

- Split on x: Andrews (1974)
- Split on x and y: Cotton (stackoverflow) referring to Tukey ("ninther").
- For $x > 0$ and $y > 0$ and ratio estimation $b = M(y, w)/M(x, w)$ (Hulliger, 1999).

Resistant line: Split sample into "equal" L, M, R parts based on x. Define $e_i = y_i - bx_i$ for arbitrary b. Resistant slope b_{RL} solves

$$M(e_L(b_{RL})) = M(e_R(b_{RL})),$$

Solution for b_{RL} can be found by iteration of the median-median-line (split on x) but a better solution is given by (Johnstone and Velleman, 1985) using the root of a function depending on the slope.

R function `line()` implements the algorithm of Wood for the resistant line, non-iterated in its default form thus mimicking the median-median-line but using $M(y - bx)$ as the intercept[2].

[2] Thanks to Tobias Schoch for looking into the C-Code!

From LS to medians

- Least Squares:

$$b_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})/t_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2/t_i} \text{ and } b_0 = \bar{y} - b_1\bar{x}$$

- `medyxline()`: Replace the sums in the LS estimator by medians:

$$b_1 = \frac{M((y - M(y))(x - M(x))/t)}{M((x - M(x))^2/t)}$$

- Treatment of zero's needed for $t_i = (x_i - M(x))^2$.

- `medline()`: LS estimator as weighted mean of slopes:

$$b_1 = \frac{\sum_{i=1}^{n} \frac{(y_i - \bar{y})}{(x_i - \bar{x})}(x_i - \bar{x})^2/t_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2/t_i}$$

and replace sums by medians:

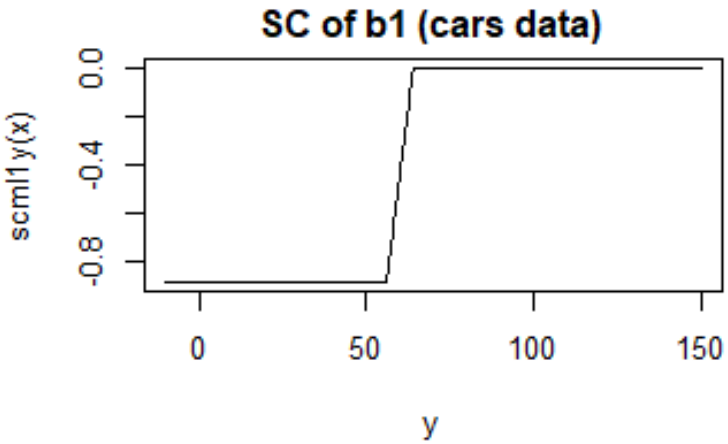
$$b_1 = M\left(\frac{y - M(y)}{x - M(x)}, (x - M(x))^2/t\right)$$

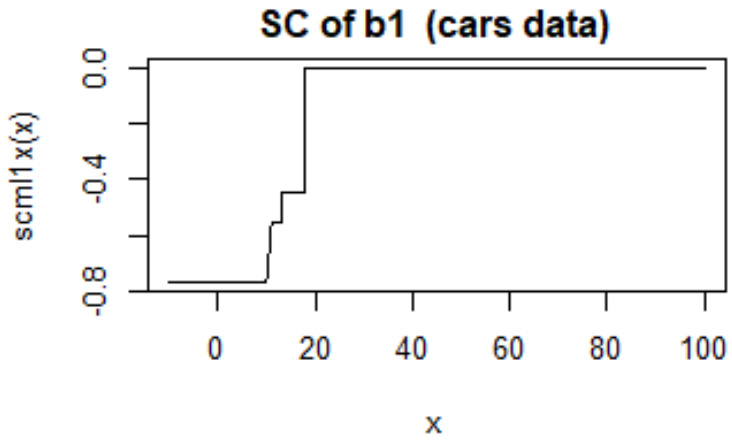
Results for cars

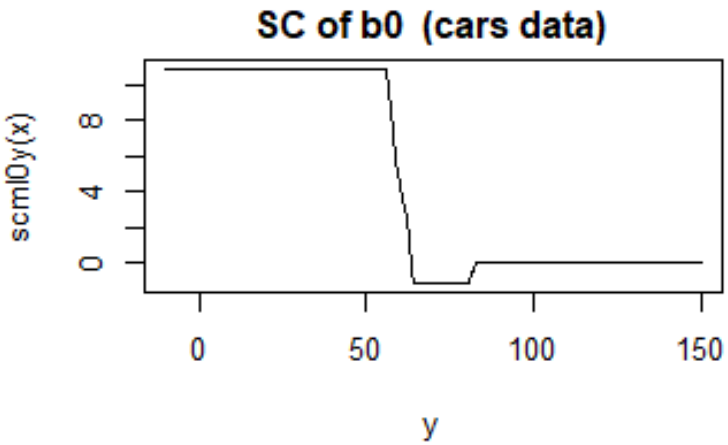
	intercept0	slope0	intercepty	slopey	interceptx	slopex
ls	-17.579	3.932	-35.980	5.439	9.609	2.040
line	-29.333	4.667	-29.333	4.667	-30.500	4.667
medmedline	-27.000	4.600	-27.000	4.600	-27.000	4.600
medyxline0	-9.938	3.062	-9.938	3.062	-9.938	3.062
medyxline1	-12.750	3.250	-12.750	3.250	-12.750	3.250
medyxline2	-11.182	3.145	-11.182	3.145	-8.610	2.974
medline0	-20.667	3.778	-20.667	3.778	13.895	1.474
medline1	-20.667	3.778	-20.667	3.778	-1.500	2.500
medline2	-11.182	3.145	-11.182	3.145	-8.610	2.974

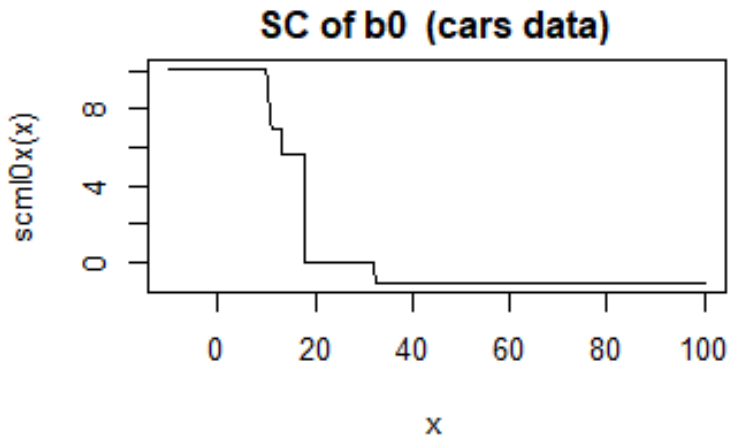
y-outlier: y[49]=360, x-outlier: x[49]=72 (original value $\times 3$)

Sensitivity Curve for `line()`

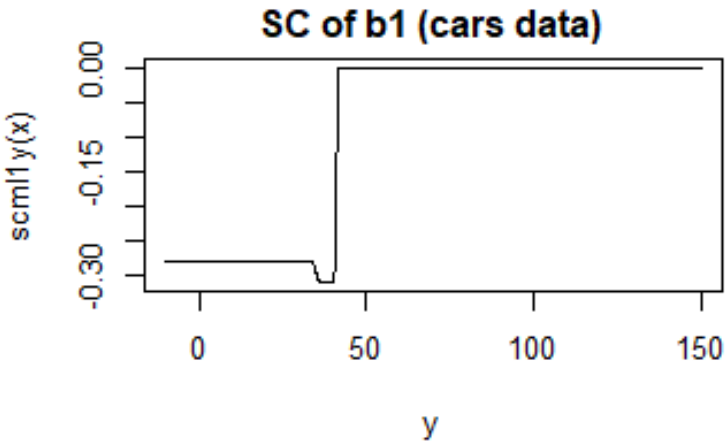


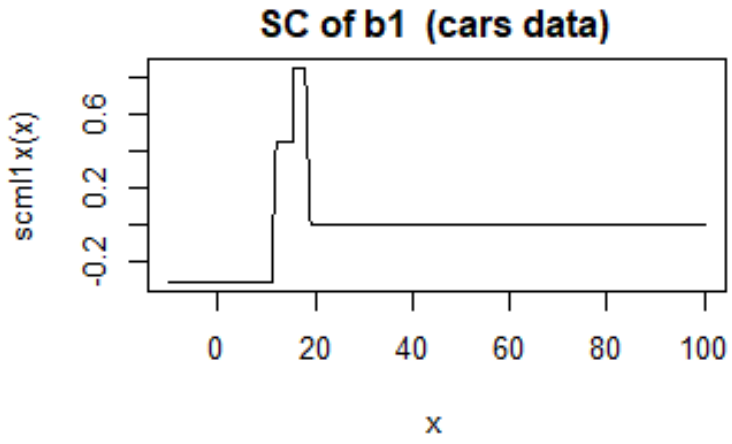


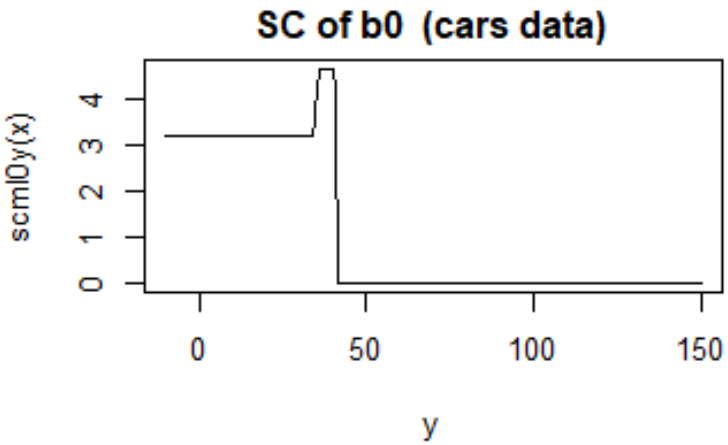


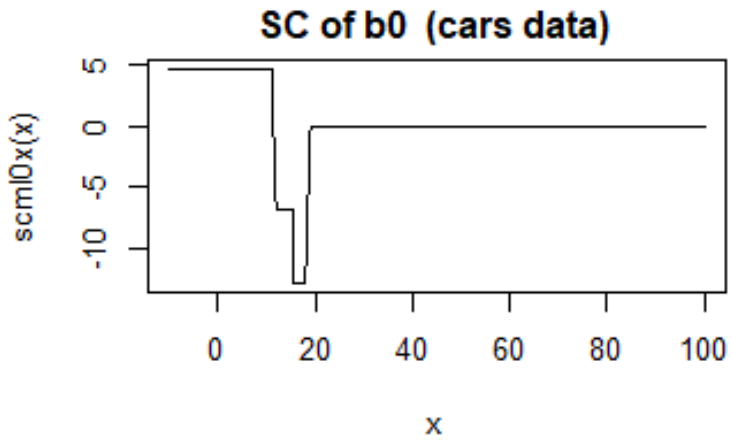


Sensitivity Curve for `medyxline()` with $q = 0$

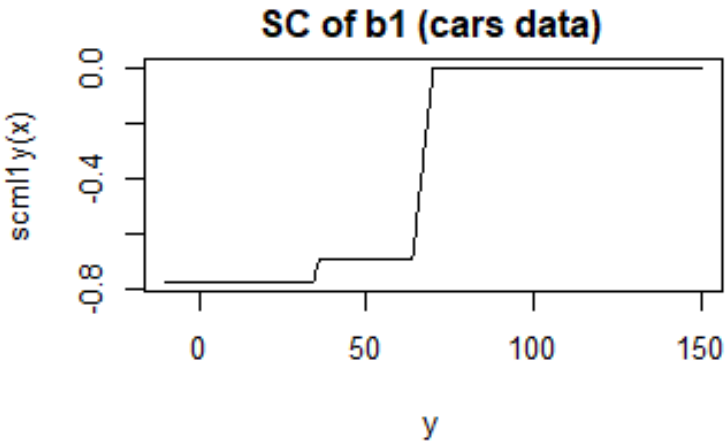


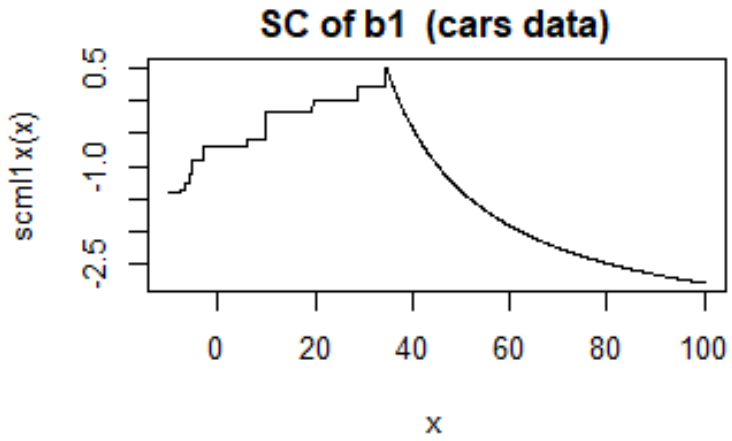


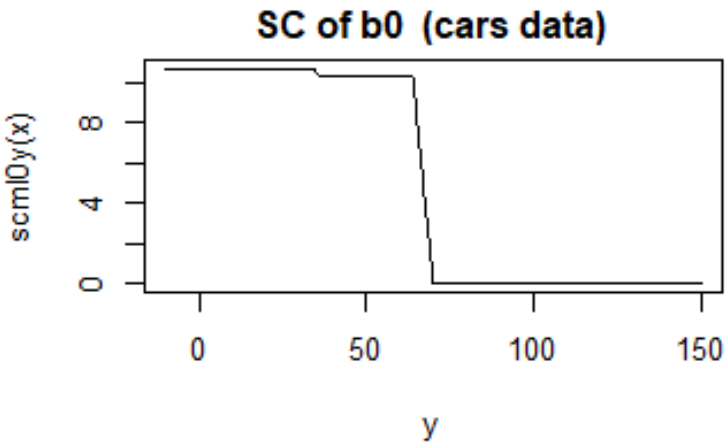


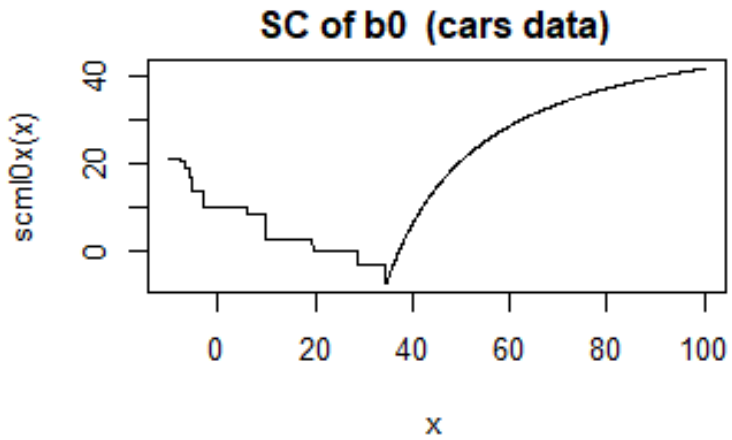


Sensitivity Curve for `medline()` with $q = 0$

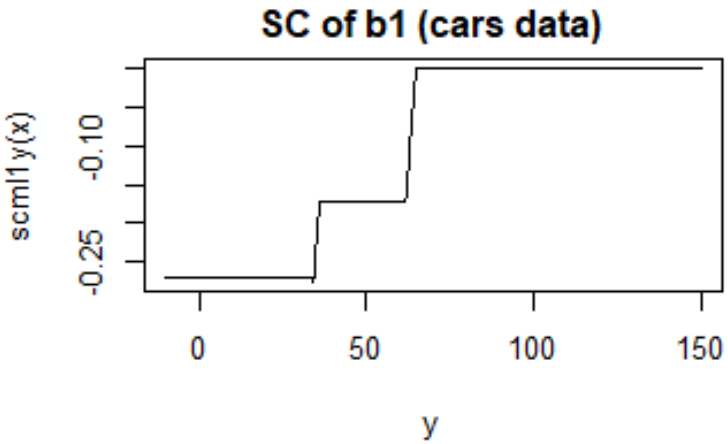


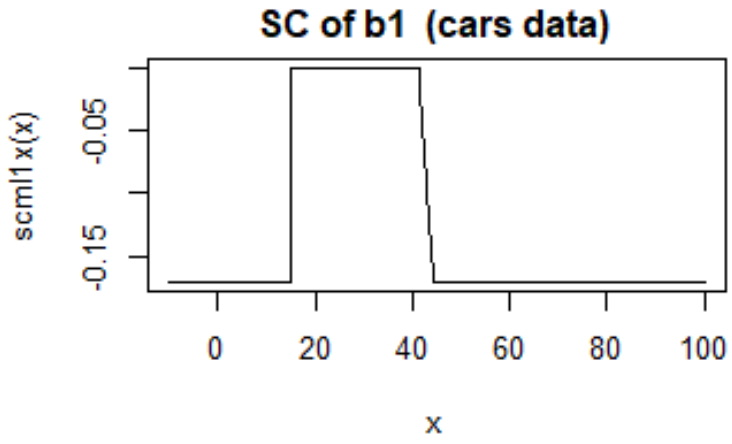


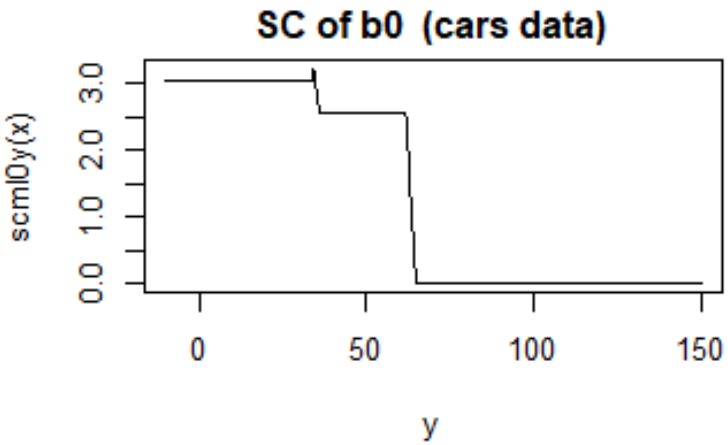


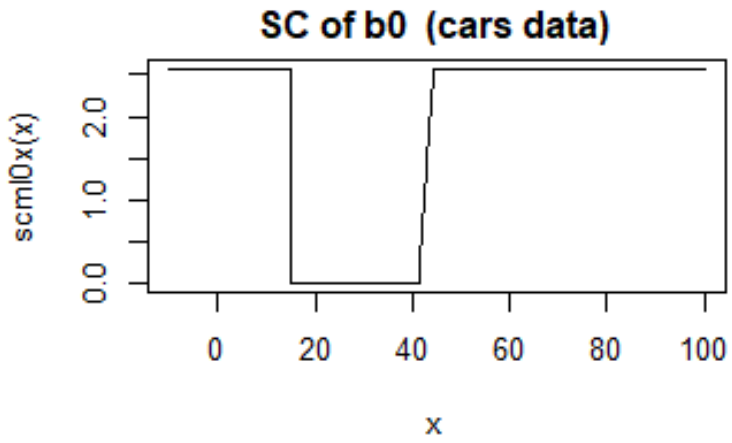


Sensitivity Curve for `medline()` with $q = 2$









Remarks from limited exploration

- `medline()` is not robust against x with $q = 0$ or $q = 1$.
- `line()`, `medline(q=2)` and `medyxline()` seem to be robust for y and x.
- `medline(q=2)` and `medyxline()` may have a bias under asymmetric distribution of x and y.
- `medmedline()` may be more stable than `line()`.

Sequential partial regression

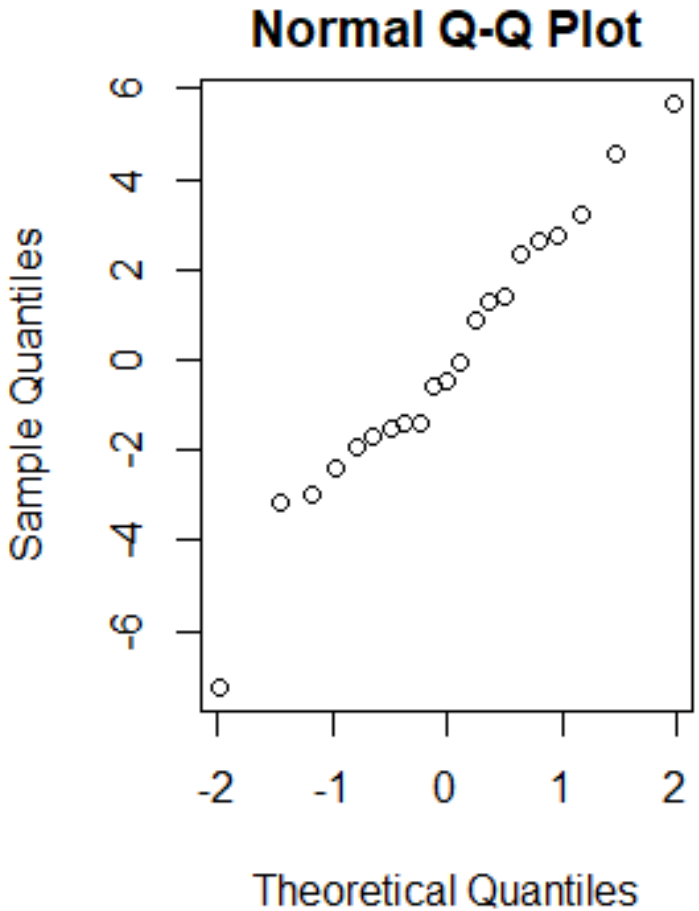
Well known procedure to reduce multiple linear regression least squares to a sequence of simple linear regressions by partial residuals and partial predictors. Use $x_0 = 1$.

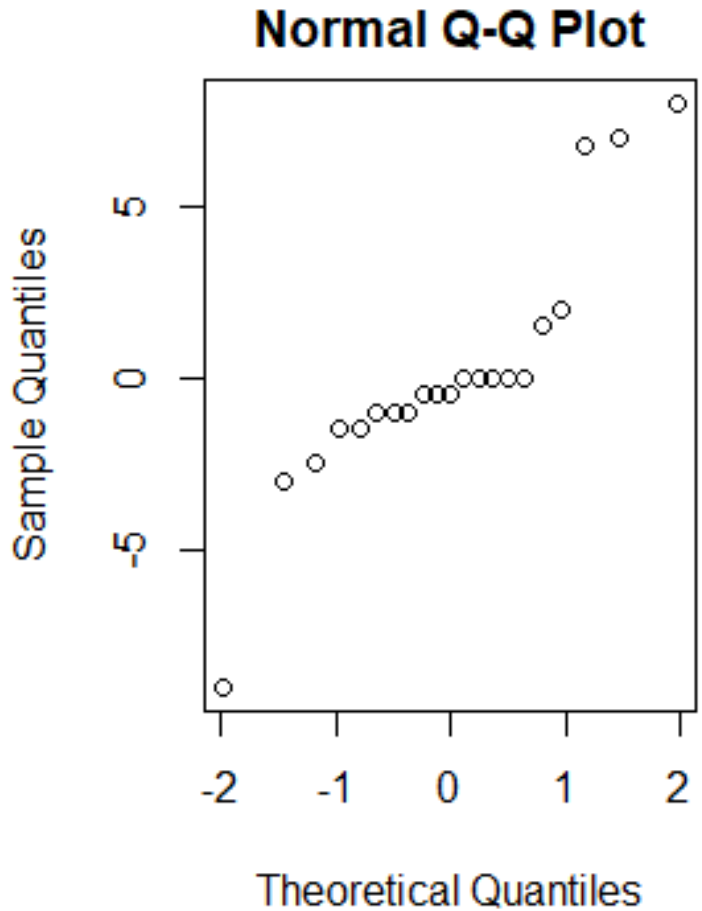
1. Regress $x_1, \ldots, x_p$ and y on x_0 and replace their values by the residuals $x_{10}, \ldots, x_{p0}$ and y_0. Retain the slope for the response y.
2. Regress $x_{20}, \ldots, x_{p0}$ and y_0 on x_{10} and replace their values by the residuals $x_{21}, \ldots, x_{p1}$ and y_1. Retain the slope for the response y_0.
3. $\vdots$
4. Regress y_{p-1} on $x_{p,p-1}$. Retin the slope for y_{p-1}.

Robust versions

- Note that the set of resulting coefficients depends in general on the sequence of the predictors.
- (Andrews, 1974) used the median-median-line (with tuning parameters for the splits) with sequential partial regression as the starting value for M-type robust regression.
- Instead of the median-median-line any robust estimator for simple linear regression may be used.
- Sequential partial regression with intercept implemented with `medline()`: `medreg()`.

Stackloss data: left `lm()`, right `medreg()`

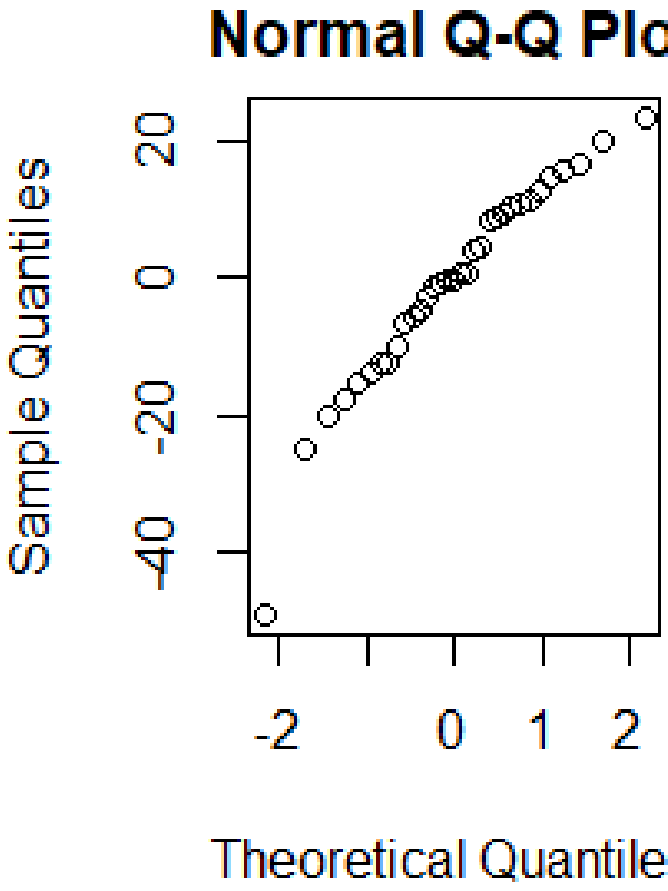


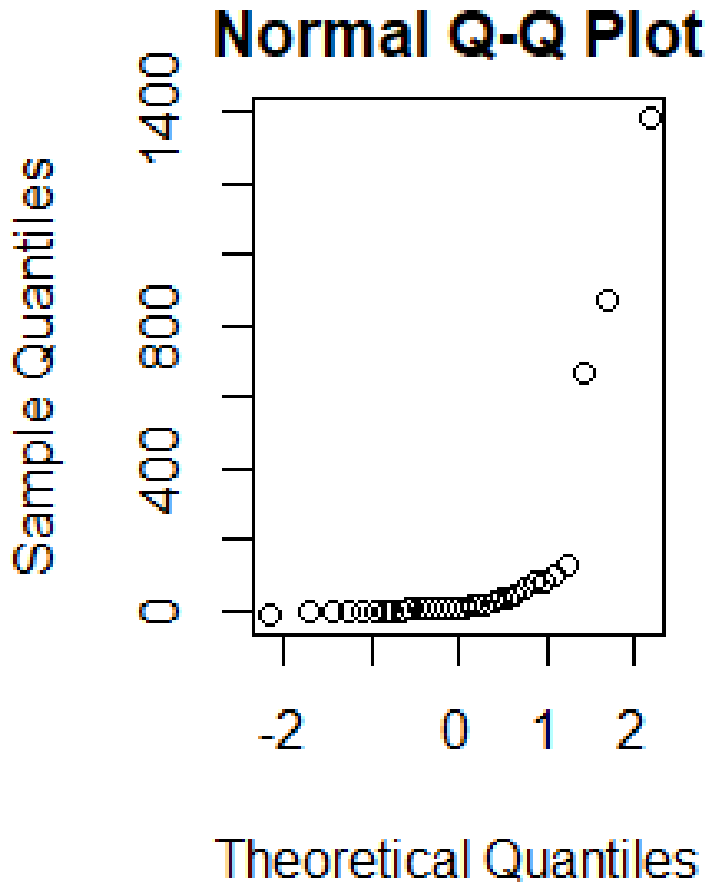


MU284

- Stratified sample of size 33 from MU284 population (Swedish Municipalities) (Särndahl et al. 1992) with 3 strata according population size in 1975 and sampling fractions 0.08, 0.30 and 1.
- 3 largest cities have inclusion probability 1.
- Received municipal taxes as linear model on population and number of municipal employees.

MU284: left `svyglm()`, right `medreg()`





Final remarks

- `line()` should be updated and allow for `medmedline()`.
- `medline()` and `medyxline()` are robust.
- Multiple linear regression with medians (`medreg()` based on `medline()`) works and seems useful for outlier detection and one-step M-Regression.
- Unclear how far needed or useful when fully iterated M-estimator or MM-estimator follows.
- Further research on simple linear regression needed: Consistency and efficiency.

Andrews, D. F. (1974). A robust method for multiple linear regression. *Technometrics 16*(4), 523–531.

Hulliger, B. (1999). Simple and robust estimators for sampling. In *ASA Proceedings of the Section on Survey Research Methods*, pp. 54–63. American Statistical Association.

Johnstone, I. M. and P. F. Velleman (1985). The resistant line and related regression methods. *Journal of the American Statistical Association 80*(392), 1041–1054.

Tukey, J. (1977). *Exploratory Data Analysis*. Addison-Wesley series in behavioral science. Addison-Wesley Publishing Company.