Zurich University
of Applied Sciences

School of
Engineering

IDP Institute of
Data Analysis and
Process Design

WBL Statistik 2024 — Robust Fitting

Half-Day 2: Robust Regression Estimation

Andreas Ruckstuhl
Institut für Datenanalyse und Prozessdesign
Zürcher Hochschule für Angewandte Wissenschaften

Outline:

Half-Day 1
- Regression Model and the Outlier Problem
- Measuring Robustness
- Location M-Estimation
- Inference
- Regression M-Estimation
- Example from Molecular Spectroscopy

Half-Day 2
- General Regression M-Estimation
- Regression MM-Estimation
- Example from Finance
- Robust Inference
- Robust Estimation with GLM

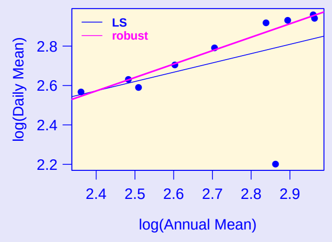
Half-Day 3
- Robust Estimation of the Covariance Matrix
- Principal Component Analysis
- Linear Discriminant Analysis
- Baseline Removal: An application of robust fitting beyond theory

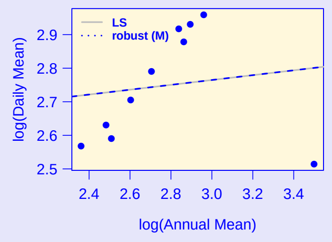
2.3 General Regression M-Estimation

The regression M-estimators will fail at the presence of leverage points.

To see that,

- compare the following examples (modified Air Quality data):





- or check the influence function:

$$IF\left\langle \underline{x}, y; \underline{\widehat{\theta}}_{\mathrm{M}}, \mathcal{N} \right\rangle = \psi\left\langle \frac{r}{\sigma} \right\rangle \underbrace{\boldsymbol{M}\,\underline{x}}_{\text{unbounded}} .$$

☞ **Bound the total influence function!**

A First Workaround: GM-Estimation

An simple modification of the Huber's ψ-function can remedy:

Either (Mallows)

$$\sum_{i=1}^{n} \psi_c \left\langle \frac{r_i \left\langle \widehat{\theta} \right\rangle}{\sigma} \right\rangle x_i^{(k)} w \left\langle d \left\langle \underline{x}_i \right\rangle \right\rangle = 0 \,, \qquad k = 1, \ldots, p,$$

or (Schweppe)

$$\sum_{i=1}^{n} \psi_c \left\langle \frac{r_i \left\langle \widehat{\theta} \right\rangle}{\sigma \cdot w \left\langle d \left\langle \underline{x}_i \right\rangle \right\rangle} \right\rangle x_i^{(k)} w \left\langle d \left\langle \underline{x}_i \right\rangle \right\rangle = \sum_{i=1}^{n} \psi_{c \cdot w \left\langle d \left\langle \underline{x}_i \right\rangle \right\rangle} \left\langle \frac{r_i \left\langle \widehat{\theta} \right\rangle}{\sigma} \right\rangle x_i^{(k)} = 0 \,,$$

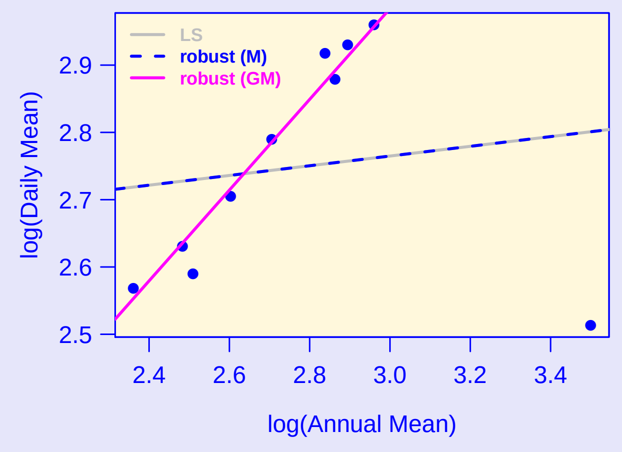
where $w\langle\rangle$ is a suitable weight function and $d\langle \underline{x}_i \rangle$ is some measure of the "outlyingness" of $\underline{x}_i$.

Examples for $w\langle\rangle$ and $d\langle \underline{x}_i \rangle$:

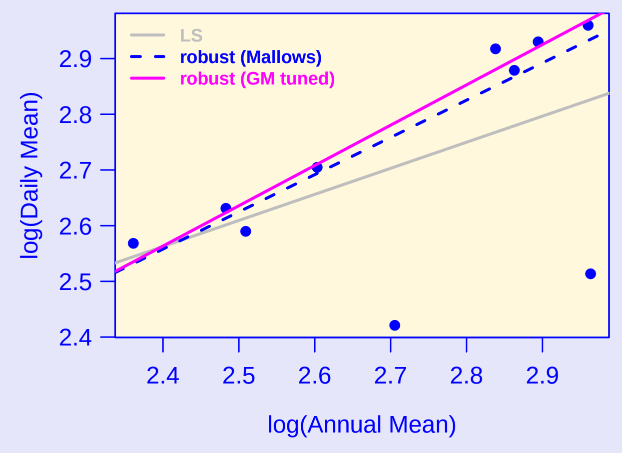
- $d\langle x_i \rangle = \dfrac{(x_i - median\langle x_k \rangle)}{MAD\langle x_k \rangle}$ or Mahalanobis distance and $w\langle d\langle x_i \rangle \rangle =$ Huber's weight function (cf. LN 2.3.b)
- $w\langle x_i \rangle = 1 - H_{ii}$ or $w\langle \underline{x}_i \rangle = \sqrt{1 - H_{ii}}$, where H_{ii} is the leverage

Example Air Quality (modified)

```
x.h <- 1-hat(model.matrix(y ~ x, AQ))
AQ.GMfit <- rlm(y ~ x, data=AQ, weights=x.h, wt.method="case")
```



Example Air Quality (Initial Data)



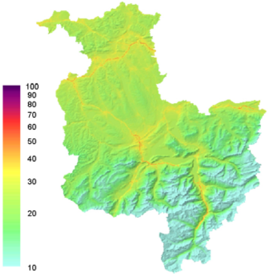
Example Air Quality: The Map

By using robust estimation methods, we

- are able to run a regression analysis **every hour automatically** and
- obtain **reliable** estimates each time

Hence robust methods provide a sound basis for the false colour map!

Note:
The outliers are identified and analyse separately. The result of this outlier analysis is part of the false colour map.

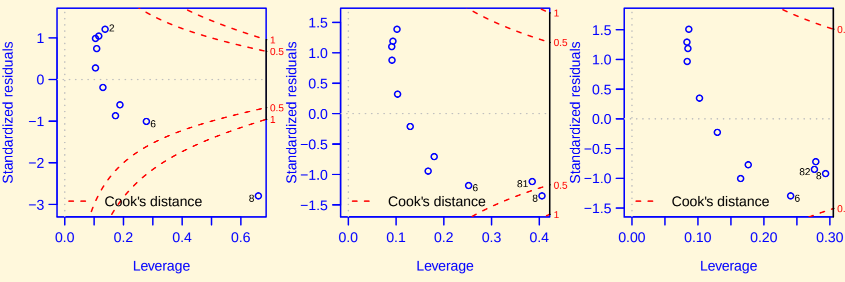


Breakdown point of GM-Estimators

However, the maximum breakdown point of general regression M-estimators cannot exceed $1/p$! (p is the number of variables)

Hence, it is not possible to detect clusters of leverage points with the projection matrix $\boldsymbol{H}$ in residual analysis.

The following example gives some insight why this happens: Look at the “Residuals vs Leverage” plots, where one, two and three outliers are put at the outlying leverage point:

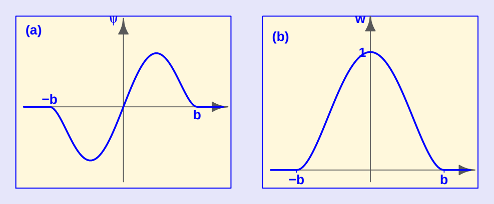


2.4 Robust Regression MM-estimation

Regressions M-Estimator with Redescending ψ

- Computational Experiments show:
  Regression M-estimators are robust if distant outliers are rejected completely!
- Theoretically and computationally more convenient: Ignore influence of distant outliers gradually

  For example by a so-called `redescending` ψ functions like Tukey's biweight function (ψ function (left) and corresponding weight function (right))



- But the equation defining the M-estimator has **many solutions** and only **one may identify the outliers correctly**.
- Solution depends on starting value! - **Good initial values are required!**

Robust Estimator With High Breakdown Point

Regression estimators with high breakdown point are e.g. the **S-estimator**.

Instead of

$$\sum_{i=1}^{n} \left(\frac{r_i \langle \underline{\theta} \rangle}{\sigma} \right)^2 \stackrel{!}{=} \min_{\underline{\theta}} \qquad \text{solve} \qquad \frac{1}{n-p} \sum_{i=1}^{n} \rho \left\langle \frac{y_i - \underline{x}_i^T \underline{\theta}}{s} \right\rangle = 0.5 \,,$$

where s **must be as small as possible**

(i.e. the equation should have a solution in $\underline{\theta}$).

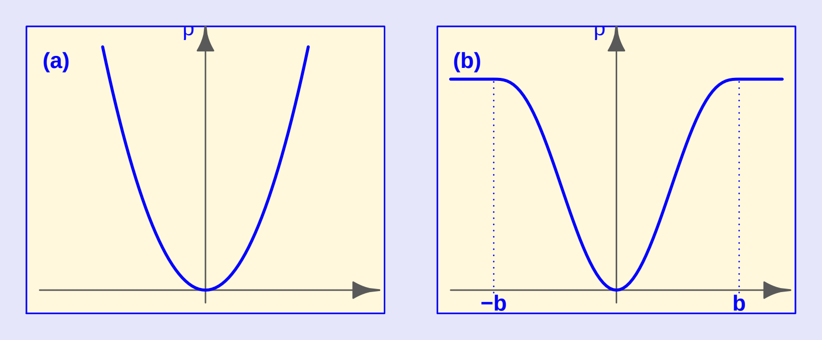
A high breakdown point implies that $\rho \langle \cdot \rangle$ must be symmetric and bounded.

The function $\rho \langle \cdot \rangle$ can be

$$\rho \langle \cdot \rangle = \rho_{b_o} \langle u \rangle := \begin{cases} 1 - \left(1 - \left(\frac{u}{b_o} \right)^2 \right)^3 & \text{if } |u| < b_o \\ 1 & \text{otherwise} \end{cases}$$

(its derivative is Tukey‘s bisquare function). To get a breakdown point of 0.5 the tuning constant b_o must be 1.548.

ρ function of the least squares estimator (left) and of Tukey's bisquare function (right)



Pros:

- high breakdown point of (about) 0.5
- computable
  (at least approximately for small data set, i.e. a few thousand observations and about 20 – 30 predictor variables).

Cons:

- efficiency of just 28.7% at the Gaussian distribution!
- challenging computation – basically done by a random resampling algorithm.
  Such an approach may result in different solutions when the algorithm is run twice or more times with the same data except but different seeds.

Regression MM-Estimator

We have

- a redescending M-etimator which is highly resistant and **highly efficient** but requires suitable starting values
- an S-estimator which is **highly resistant** but very inefficient

Combining the strength of both estimators yields the **regression MM-estimator** (**m**odified **M**-estimator):

- An S-estimator with breakdown point $\varepsilon^* = 1/2$ is used as initial estimator
  Tukey's bisquare function with $\rho_{b_o=1.548}$: ☞ $\widehat{\underline{\theta}}^{(o)}$ and s_o
- The redescending regression M-estimator is applied
  using Tukey's bisquare ψ-function $\psi_{b_1=4.687}$ and fixed scale parameter $\sigma = s_o$ from the initial estimation; starting value is $\widehat{\underline{\theta}}^{(o)}$.

The regression MM-estimator has a breakdown point of $\varepsilon^* = 1/2$ and an efficiency and an asymptotic distribution like the regression M-estimator.

Example from Finance

Return-Based Style Analysis of Fund of Hedge Funds

(Joint work with P. Meier and his group)

A fund of hedge funds (FoHF) is a fund that invests in a portfolio of different hedge funds to diversify the risks associated with a single hedge fund.

A hedge fund is an investment instrument that undertakes a wider range of investment and trading activities in addition to traditional long-only investment funds.

One of the difficulties in risk monitoring of Fund of Hedge Funds (FoHF) is their limited transparency.

- Many FoHF will only disclose partial information on their underlying portfolio
- The underlying investment strategy (style of FoHF), which is the crucial characterisation of FoHF, is self-declared

A return-based style analysis searches for the combination of indices of sub-styles of hedge fund that would most closely replicate the actual performance of the FoHF over a specified time period.

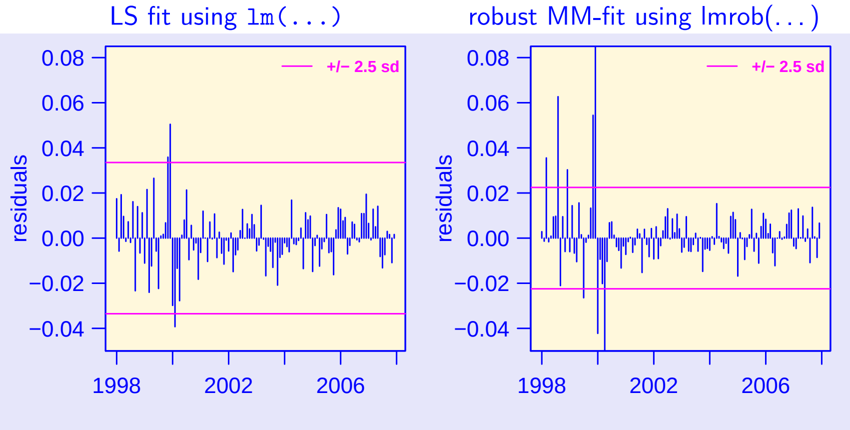
Such a style analysis is done basically by fitting a (in Finance) so-called multifactor model:

$$R_t = \alpha + \sum_{k=1}^{p} \beta_k I_{k,t} + E_t$$

where

$$\begin{aligned} R_t &= \text{return on the FoHF at time } t \\ \alpha &= \text{the excess return (a constant) of the FoHF} \\ I_{k,t} &= \text{the index return of sub-style } k \text{ (= factor) at time } t \\ \beta_k &= \text{the change in the return on the FoHF per unit change in factor } k \\ p &= \text{the number of used sub-indices} \\ E_t &= \text{residual (error) which cannot be explained by the factors} \end{aligned}$$

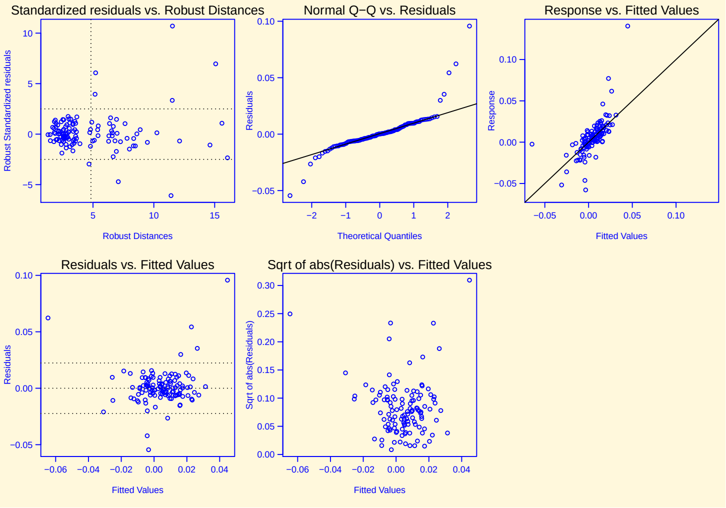
Residuals VS Time



Robust MM-fit identifies clearly two different investment periods: one before April 2000 and one afterwards.

Residual Analysis with MM-Fit

```
plot(FoHF.rlm)
```



2.5 Robust Inference and Variable Selection

- Outliers may also influence the result of a classical test crucially.
  It might happen that the null hypothesis is rejected, because an interfering alternative H_I (outliers) is present. That is, the rejection of the null hypothesis is justified but accepting the actual alternative H_A is unjustified.

- To understand such situations better, one can explore the effect of contamination on the level and power of hypothesis tests.

- Heritier and Ronchetti (1994) showed that the effects of contamination on both the level and power of a test are inherited from the underlying estimator (= test statistic).

  That means that

the test is robust if its test statistic is based on a robust estimator.

Asymptotic Distribution of the MM-estimator

The Regression MM-estimator is **asymptotically Gaussian distributed** with

- mean (expectation) $\underline{\theta}$ and
- covariance matrix $\sigma^2 \tau \boldsymbol{C}^{-1}$, where $\boldsymbol{C} = (1/n) \sum_i \underline{x}_i \underline{x}_i^T$.

The covariance matrix of $\widehat{\underline{\theta}}$ is estimated by

$$\widehat{\boldsymbol{V}} = \frac{s_o^2}{n} \widehat{\tau} \widehat{\boldsymbol{C}}^{-1}$$

where $\widehat{\boldsymbol{C}} = \frac{\frac{1}{n} \sum_i w_i \underline{x}_i \underline{x}_i^T}{\frac{1}{n} \sum_i w_i}$, $\widehat{\tau} = \frac{\frac{1}{n} \sum_{i=1}^n \left(\psi_{b_1} \left\langle \widetilde{r}_i \right\rangle \right)^2}{\left(\frac{1}{n} \sum_{i=1}^n \psi' \left\langle \widetilde{r}_i \right\rangle \right)^2}$

with $\widetilde{r}_i := \frac{y_i - \underline{x}_i^T \widehat{\underline{\theta}}^{(o)}}{s_o}$, $w_i := \frac{\psi_{b_1} \langle \tilde{r}_i \rangle}{\tilde{r}_i}$

Note that $\widehat{\underline{\theta}}^{(o)}$ and s_o come from the initial estimation.

A Further Modification to the MM-Estimator

- The estimated covariance matrix $\widehat{\boldsymbol{V}}$ depends on quantities ($\widehat{\underline{\theta}}^{(o)}$ and s_o) from the initial estimator
- The initial S-estimator, however, is known to be very inefficient.
- Koller and Stahel (2011, 2014) investigated the effects of this construction on the efficiency of the estimated confidence intervals ...

  ... and, as a consequence, came up with an additional modification:

  Extend the current regression MM-estimator by two additional steps:
  - replaces s_o by a more efficient scale estimator
  - Then apply another M-estimator but with a more slowly redescending *psi*-function.

They called this estimation procedure **regression SMDM-estimator** and it is implemented in `lmrob(..., setting="KS2014")`.

Example from Finance with another target FoHF

Return-Based Style Analysis of Fund of Hedge Funds - RBSA2

	lm(FoHF ∼ . , data=FoHF2)			lmrob(FoHF ∼ ., data=FoHF2, setting="KS2014")		
	Estimate	se	$\Pr(>\|t\|)$	Estimate	se	$\Pr(>\|t\|)$
(I)	-0.0019	0.0017	0.2610	-0.0030	0.0014	**0.0377**
RV	0.0062	0.3306	0.9850	0.3194	0.2803	0.2564
CA	-0.0926	0.1658	0.5780	-0.0671	0.1383	0.6280
FIA	0.0757	0.1472	0.6083	-0.0204	0.1279	0.8730
EMN	0.1970	0.1558	0.2094	0.2721	0.1328	**0.0430**
ED	-0.3010	0.1614	0.0655	-0.4763	0.1389	**0.0009**
EDD	0.0687	0.1301	0.5986	0.1019	0.1112	0.3611
EDRA	0.0735	0.1882	0.6971	0.0903	0.1583	0.5689
LSE	0.4407	0.1521	**0.0047**	0.5813	0.1295	**2.05e-05**
GM	0.1723	0.0822	**0.0390**	-0.0159	0.0747	0.8319
EM	0.1527	0.0667	**0.0245**	0.1968	0.0562	**0.0007**
SS	0.0282	0.0414	0.4973	0.0749	0.0356	**0.0378**
	Residual standard error: 0.009315			Residual standard error: 0.007723		

The 95% confidence interval of β_{SS} is

$0.028 \pm 1.99 \cdot 0.041 = [-0.054, 0.110]$	$0.075 \pm 1.96 \cdot 0.036 = [0.004, 0.146]$

Example from Finance: RBSA2 (cont.)

A fund of hedge funds (FoHF) may be classified by the style of their target funds into *focussed directional, focussed non-directional* or *diversified*.

If our considered FoHF is a *focussed non-directional* FoHF, then it should be invested in `LSE`, `GM`, `EM`, `SS` and hence the other parameter should be zero.

Goal: We want to test the hypothesis that $q < p$ of the p elements of the parameter vector $\underline{\theta}$ are zero.

First, let's introduce some notation to express the results more easily:

- There is no real loss of generality if we suppose that the model is parameterized so that the null hypothesis can be expressed as $H_0 : \underline{\theta}_1 = \underline{0}$ where $\underline{\theta} = (\underline{\theta}_1^T, \underline{\theta}_2^T)^T$.
- Further, let $\widehat{\boldsymbol{V}}_{11}$ be the quadratic submatrix containing the first q rows and columns of $\widehat{\boldsymbol{V}}$.

The so-called **Wald-type test statistic** can now be expressed as

$$W = \underline{\theta}_1^T \, (\widehat{\boldsymbol{V}}_{11})^{-1} \, \underline{\theta}_1 \, .$$

It can be shown, that this test statistic is asymptotically χ^2 distributed with q degrees of freedom.

This test statistic also provides the basis for confidence intervals of a single parameter θ_k:

$$\widehat{\theta}_k \pm q^{\mathcal{N}}_{1-\alpha/2} \cdot \sqrt{\widehat{V}_{kk}}$$

where $q^{\mathcal{N}}_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantil of the standard Gaussian distribution.

Comments:

- Since we have estimated the covariance matrix $\mathrm{Cov}\left\langle \widehat{\underline{\theta}} \right\rangle$ by $\widehat{\boldsymbol{V}}$, it seems obvious from classical regression inference that replacing χ^2_q by $F_{q,\,n-p}$ is a reasonable adjustment in the distribution of the test statistic for estimating the variance σ^2.
- However, to do so has no formal justification and it would be better to avoid too small sample sizes because all results are asymptotical in their nature anyway (and then $\frac{1}{q} \cdot \chi^2_q \approx F_{q,\,n-p}$).

For an MM-estimator, we may define a **robust deviance** by

$$D\left\langle \underline{y}, \widehat{\underline{\theta}}_{\mathrm{MM}} \right\rangle := 2 \cdot s_o^2 \cdot \sum_{i=1}^{n} \rho\left\langle \frac{y_i - \underline{x}_i^T \widehat{\underline{\theta}}_{\mathrm{MM}}}{s_o} \right\rangle \quad \text{with } s_o^2 \text{ from the initial estimator}\,.$$

Similar to generalised linear models: The robust generalisation of the F-test,

$$\frac{\left(SS_{reduced} - SS_{full}\right)/q}{SS_{full}/(n-p)} = \frac{\left(SS_{reduced} - SS_{full}\right)/q}{\widehat{\sigma}^2}$$

is to replace the sum of squares by the robust deviance so that we obtain the test statistic

$$\Delta^* = \tau^* \cdot \frac{D\left\langle \underline{y}, \widehat{\underline{\theta}}^{r}_{\mathrm{MM}} \right\rangle - D\left\langle \underline{y}, \widehat{\underline{\theta}}^{f}_{\mathrm{MM}} \right\rangle}{s_o^2}$$

$$\text{with} \quad \tau^* = \left(\frac{1}{n}\sum_{i=1}^{n} \psi'_{b_1}\langle \tilde{r}_i \rangle\right) \Big/ \left(\frac{1}{n}\sum_{i=1}^{n} \left(\psi_{b_1}\langle \tilde{r}_i \rangle\right)^2\right).$$

Then $\Delta^* \overset{a}{\sim} \chi^2_q$ **under the null hypothesis.**

Example from Finance - RBSA2

 # Least squares estimator:

```
> FoHF2.lm1 <- lm(FoHF ~ ., data=FoHF2)
> FoHF2.lm2 <- lm(FoHF ~ LSE + GM + EM + SS, data=FoHF2)
> anova(FoHF2.lm2, FoHF2.lm1)
```

Analysis of Variance Table

Model 1: FoHF ∼ LSE + GM + EM + SS
Model 2: FoHF ∼ RV + CA + FIA + EMN + ED + EDD + EDRA + LSE + GM + EM + SS

	Res.Df	RSS	Df	Sum of Sq	F	$Pr(>F)$
1	96	0.0085024				
2	89	0.0077231	7	0.00077937	1.2831	0.2679

 # Robust with SMDM-estimator (i.e., `setting="KS2014"`)

```
> FoHF2.rlm1 <- lmrob(FoHF ~ ., data=FoHF2, setting="KS2014")
> anova(FoHF.rlm1, FoHF ~ LSE + GM + EM + SS, test="Wald")
```

Robust Wald Test Table

Model 1: FoHF ∼ RV + CA + FIA + EMN + ED + EDD + EDRA + LSE + GM + EM + SS
Model 2: FoHF ∼ LSE + GM + EM + SS
Largest model fitted by lmrob(), i.e. SMDM

	pseudoDf	Test.Stat	Df	$Pr(> chisq)$	
1	89				
2	96	25.066	7	0.0007388	***

Example from Finance - RBSA2 (cont.)

```
# Robust with SMDM-estimator and Wald test
> FoHF2.rlm1 <- lmrob(FoHF ~ ., data=FoHF2, setting="KS2014")
> anova(FoHF2.rlm1, FoHF ~ LSE + GM + EM + SS, test="Wald")
```

Robust Wald Test Table

Model 1: FoHF ~ RV + CA + FIA + EMN + ED + EDD + EDRA + LSE + GM + EM + SS
Model 2: FoHF ~ LSE + GM + EM + SS

Largest model fitted by lmrob(), i.e. SMDM

	pseudoDf	Test.Stat	Df	Pr(> *chisq*)	
1	89				
2	96	24.956	7	0.0007388	***

```
# Robust with SMDM-estimator and Deviance test
> anova(FoHF2.rlm1, FoHF ~ LSE + GM + EM + SS, test="Deviance")
```

Robust Deviance Table

Model 1: FoHF ~ RV + CA + FIA + EMN + ED + EDD + EDRA + LSE + GM + EM + SS
Model 2: FoHF ~ LSE + GM + EM + SS

Largest model fitted by lmrob(), i.e. SMDM

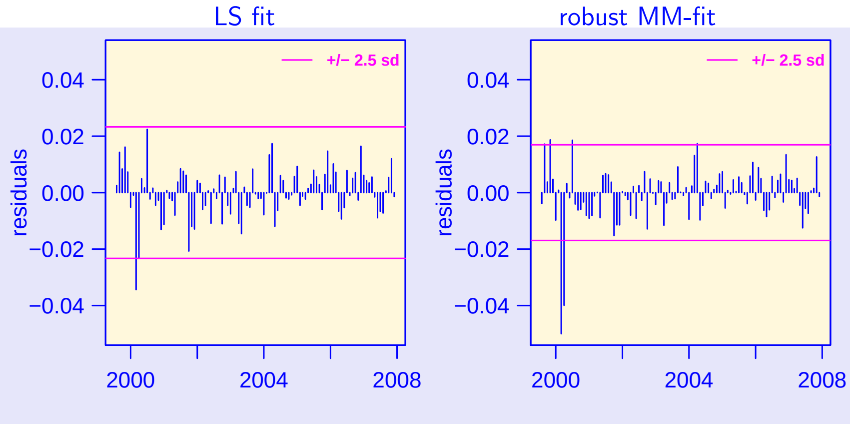
	pseudoDf	Test.Stat	Df	Pr(> *chisq*)	
1	89				
2	96	25.089	7	0.0007318	***

Example from Finance - RBSA2 (i.e., SLIDE 21 again)

Return-Based Style Analysis of Fund of Hedge Funds

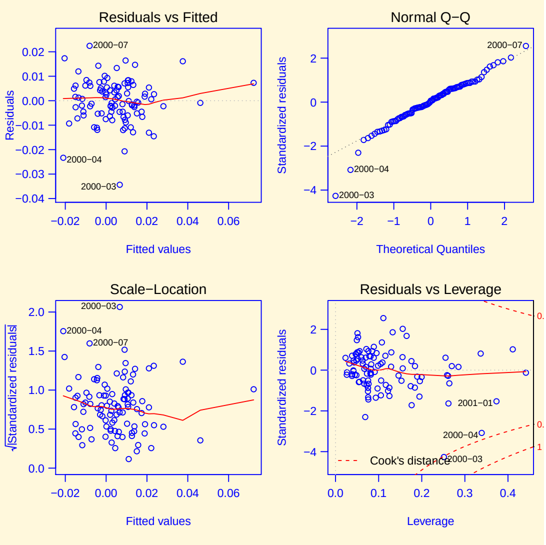
	lm(FoHF ∼ . , data=FoHF)			lmrob(FoHF ∼ ., data=FoHF, setting="KS2014")		
	Estimate	se	$\Pr(>\|t\|)$	Estimate	se	$\Pr(>\|t\|)$
(I)	-0.0019	0.0017	0.2610	-0.0030	0.0014	**0.0377**
RV	0.0062	0.3306	0.9850	0.3194	0.2803	0.2564
CA	-0.0926	0.1658	0.5780	-0.0671	0.1383	0.6280
FIA	0.0757	0.1472	0.6083	-0.0204	0.1279	0.8730
EMN	0.1970	0.1558	0.2094	0.2721	0.1328	**0.0430**
ED	-0.3010	0.1614	0.0655	-0.4763	0.1389	**0.0009**
EDD	0.0687	0.1301	0.5986	0.1019	0.1112	0.3611
EDRA	0.0735	0.1882	0.6971	0.0903	0.1583	0.5689
LSE	0.4407	0.1521	**0.0047**	0.5813	0.1295	**2.05e-05**
GM	0.1723	0.0822	**0.0390**	-0.0159	0.0747	0.8319
EM	0.1527	0.0667	**0.0245**	0.1968	0.0562	**0.0007**
SS	0.0282	0.0414	0.4973	0.0749	0.0356	**0.0378**
	Residual standard error: 0.009315			Residual standard error: 0.007723		

Example from Finance - RBSA2: Residuals VS Time

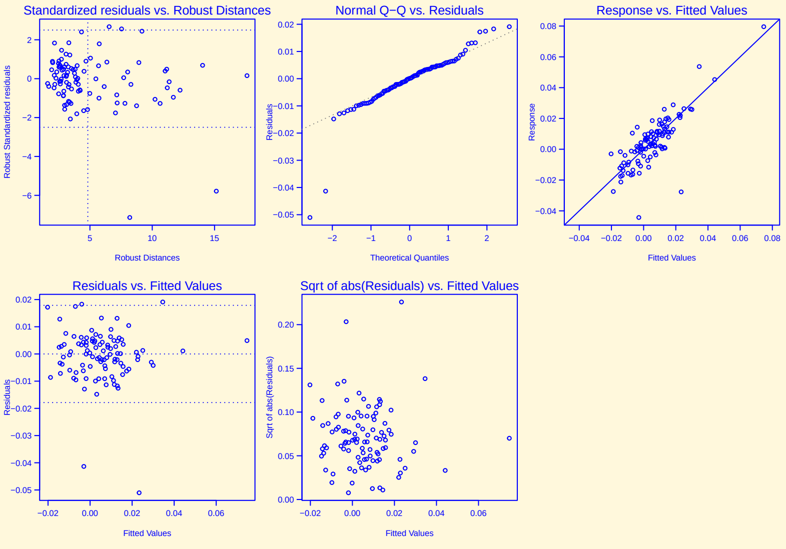


Robust MM-fit identifies clearly two outliers.

Example from Finance - RBSA2: Residual Analysis with LS-Fit



Example from Finance - RBSA2: Residual Analysis with Robust Fit



Variable Selection

A variable selection criterion that is based on robust principles is the robustified version of Akaike's "final prediction error" criterion (not to be confused with AIC):

$$\text{RFPE}\langle C\rangle := \frac{1}{n}\sum_{i=1}^{n}\rho\left\langle\frac{r_i^C}{\widehat{\sigma}_f}\right\rangle + \frac{q}{n}\widehat{\tau}$$

with C the set of variables considered, q the number of variables considered, and $\widehat{\sigma}_f$ the robust scale estimate based on the full set of variables. The correction factor $\widehat{\tau}$ is calculated as on Slide 26, HD1, except that $\widehat{\sigma}$ is replaced by $\widehat{\sigma}_f$.

Example Fund of Hedge Funds II

```
## Classical variable selection with AIC
> step(FoHF2.lm1)  # ...output shortened ...
FoHF ~ EMN + ED + LSE + GM + EM
          Df   Sum of Sq        RSS     AIC
 <none>                   0.0078582  -943.59
 - EMN     1  0.00032529  0.0081835  -941.50
 - ED      1  0.00043180  0.0082900  -940.19
 - GM      1  0.00049510  0.0083533  -939.42
 - EM      1  0.00059036  0.0084486  -938.28
 - LSE     1  0.00123463  0.0090928  -930.85

## Robust variable selection criterion
> library(RobStatTM)
> h.cont <- lmrobdet.control(bb=0.5, efficiency=0.85, family="bisquare")
> FoHF2.rlm1 <- lmrobdetMM(FoHF ~ . , data=FoHF2, control=h.cont)
> step.lmrobdetMM(FoHF2.rlm1)   # ...output shortened ...
Model: FoHF ~ RV + EMN + ED + LSE + EM + SS
scale: 0.007999167
          Df     RFPE
 <none>       0.20127
 RV        1  0.20384
 EMN       1  0.20371
 ED        1  0.22507
 LSE       1  0.23117
 EM        1  0.22365
 SS        1  0.20724
```

This two procedures do not select the same variables, but agree on 4 variables.

3 Generalised Linear Models

3.1 Unified Model Formulation

Generalized linear models were formulated by John Nelder and Robert Wedderburn as a way of unifying various statistical regression models, including linear regression, logistic regression, Poisson regression and Gamma regression.

The generalization is based on a refomulation of the linear regression model. Instead of

$$Y_i = \theta_0 + \theta_1 \cdot x_i^{(1)} + \ldots + \theta_p \cdot x_i^{(p)} + E_i\,,\; i = 1, \ldots n, \quad \text{with } E_i \text{ ind. } \sim \mathcal{N}\left\langle 0, \sigma^2 \right\rangle$$

use

$$Y_i \text{ ind. } \sim \mathcal{N}\left\langle \mu_i, \sigma^2 \right\rangle \qquad \text{with } \mu_i = \theta_0 + \theta_1 \cdot x_i^{(1)} + \ldots + \theta_p \cdot x_i^{(p)}$$

The expectation μ_i may be linked to the linear predictor $\eta_i = \theta_0 + \theta_1 \cdot x_i^{(1)} + \ldots + \theta_p \cdot x_i^{(p)}$ by another function than the identity function. In general, we assume

$$g\left\langle \mu_i \right\rangle = \eta_i$$

Discrete Generalised Linear Models

The two discrete generalised linear models are the **binary / binomial regression model** and **Poisson regression model**.

Let Y_i, $i = 1, \ldots, n$ be the response and $\eta_i = \underline{x}_i^T \underline{\theta} = \sum_{j=1}^{p} x_i^{(j)} \cdot \theta_j$ its linear predictor. Then

Binary / Binomial Regression	**Poisson Regression**
Y_i indep. $\sim \mathcal{B}\langle \pi_i, m_i \rangle$ with	Y_i indep. $\sim \mathcal{P}\langle \lambda_i \rangle$ with
$\mathrm{E}\left\langle \frac{Y_i}{m_i} \middle| \underline{x}_i \right\rangle = \pi_i$	$\mathrm{E}\langle Y_i \mid \underline{x}_i \rangle = \lambda_i$
$\mathrm{Var}\left\langle \frac{Y_i}{m_i} \middle| \underline{x}_i \right\rangle = \frac{\pi_i \cdot (1-\pi_i)}{m_i}$	$\mathrm{Var}\langle Y_i \mid \underline{x}_i \rangle = \lambda_i$

The mean response π_i and λ_i, respectively, are related to the linear predictor η_i by the link function $g\langle \cdot \rangle$: $g\langle \pi_i \rangle = \eta_i$. The canonical links are

$g\langle \pi_i \rangle = \log\langle \pi_i/(1-\pi_i) \rangle$ (Logit model)

$g\langle \lambda_i \rangle = \log\langle \lambda_i \rangle$ (log-linear model)

Other choices for the link functions are possible.

Gamma Regression

Let Y_i, $i = 1, \ldots, n$ be the response and $\eta_i = \underline{x}_i^T \underline{\theta} = \sum_{j=1}^{p} x_i^{(j)} \cdot \theta_j$ its linear predictor. Then

Y_i indep. $\sim \texttt{Gamma}\langle \alpha_i, \beta_i \rangle$ with

$$\mathrm{E}\langle Y_i | \underline{x}_i \rangle = \frac{\alpha_i}{\beta_i}$$
$$\mathrm{Var}\langle Y_i | \underline{x}_i \rangle = \frac{\alpha_i}{\beta_i^2}$$

Common link functions are

$$g\langle \mu_i \rangle = \frac{1}{\mu_i} \quad \text{inverse (canonical)}$$
$$g\langle \mu_i \rangle = \log\langle \mu_i \rangle$$
$$g\langle \mu_i \rangle = \mu_i \quad \text{identity.}$$

In GLM it is assumed that

- the response Y_i is independently distributed according to a distribution from the exponential family with expectation $\mathrm{E}\langle Y_i \rangle = \mu_i$.
- The expectation μ_i is linked by a function g to the linear predictor $\underline{x}_i^T \underline{\beta}$: $g\langle \mu_i \rangle = \underline{x}_i^T \underline{\beta}$
- The variance of the response depends on $\mu_i = \mathrm{E}\langle Y_i \rangle$: $\mathrm{Var}\langle Y_i \rangle = \phi\, V\langle \mu_i \rangle$.

  The so-called **variance function** $V\langle \mu_i \rangle$ is determined by the distribution. ϕ is the dispersion parameter.

Estimating Equation

The estimating equations of GLM can be written in a unified form,

$$0 = \sum_{i=1}^{n} \frac{y_i - \mu_i}{V\langle\mu_i\rangle}\, \mu_i'\, \underline{x}_i = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\sqrt{V\langle\mu_i\rangle}}\, \frac{\mu_i'}{\sqrt{V\langle\mu_i\rangle}}\, \underline{x}_i$$

where $\mu_i' = \partial\mu\langle\eta_i\rangle / \partial\eta_i$ is the derivative of the inverse link function.

$\frac{y_i - \mu_i}{\sqrt{V\langle\mu_i\rangle}}$ are called **Pearson residuals**.

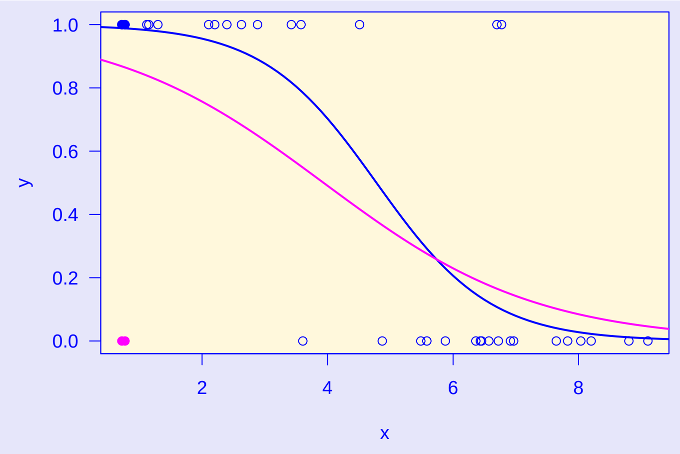
If there are no leverage points, their variance is approximately constant,

$$\text{that is } \mathrm{Var}\left\langle \frac{y_i - \mu_i}{\sqrt{V\langle\mu_i\rangle}} \right\rangle \approx \phi\sqrt{1 - H_{ii}}\,.$$

In R use `glm(Y ~ ..., family=..., data=...)` to fit a GLM to data.

In GLM, we face similar problems with the standard estimator as in linear regression problems at the presence of contaminated data.

If only two observations are misclassified (magenta points) ...



Mallows type (robust) quasi-likelihood estimator (Mqle)

Cantoni and Ronchetti (2001) suggested a Mallows type robustification of the estimating equation of the GLM-estimator:

$$\underline{0} = \sum_{i=1}^{n} \left(\psi_c \langle r_i \rangle \frac{\mu_i'}{\sqrt{V \langle \mu_i \rangle}} \underline{x}_i \, w \langle \underline{x}_i \rangle - \underline{\mathrm{fcc}} \langle \underline{\theta} \rangle \right) ,$$

where $\psi_c \langle \rangle$ is the Huber function and the vector constant $\underline{\mathrm{fcc}} \langle \underline{\theta} \rangle$ ensures the Fisher consistency of the estimator.

- The “weights” $w \langle \underline{x}_i \rangle$ can be used to down-weight leverage points.
- If $w \langle \underline{x}_i \rangle = 1$ for all observations i then the influence of position is not bounded (cf. regression GM-estimator).
- To bound the total influence, one may, e.g., use $w \langle \underline{x}_i \rangle = \sqrt{1 - H_{ii}}$.
  Since such an estimator will not yet have high breakdown point, we better use the inverse of the Mahalanobis distance for $\underline{x}_i$ which is based on a robust covariance estimator with high breakdown point (cf. Chap. 4).

Inference

The advantage of this approach is that inference results are available based

- on the asymptotic Gaussian distribution of the estimator (i.e., on $\mathcal{N}\langle\underline{\theta}, \mathbf{\Omega}\rangle$) and
- on robust quasi-deviances, i.e., on $\widehat{\Lambda} = D\left\langle \underline{y}, \underline{\widehat{\theta}}^{\,red}_{\text{Mqle}} \right\rangle - D\left\langle \underline{y}, \underline{\widehat{\theta}}^{\,full}_{\text{Mqle}} \right\rangle$.

- I do not dare to present the formulas for $\mathbf{\Omega}$ and $D\left\langle \underline{y}, \underline{\widehat{\theta}}_{\text{Mqle}} \right\rangle$, because they look frightening, but cf. LN 3.2.c,d.
- The test statistic $\widehat{\Lambda}$ is not just χ^2 distributed but it is rather distributed like a linear combination of χ^2_1 distributions.
- Because the calculation of $\widehat{\Lambda}$ is challenging, there is also an approximate version available, which is asymptotically χ^2 distributed.

Implementation

Theory and implementation in R cover responses which are **Poisson, binomial, gamma** or **Gaussian** (i.e., linear regression GM-estimator) distributed.

- Fitting in R is done by
```
glmrob(Y ~ ..., family=..., data=...,
      weights.on.x=c("none", "hat", "robCov", "covMcd"))
```
- Testing in R is done by
```
anova(Fit1, Fit2, test=c("Wald", "QD", "QDapprox"))
```

"QD"= robust quasi-deviance; "QDapprox"= approximate version of "QD"

Take Home Message Half-Day 2

- Least-squares estimation are unreliable if contaminated observations are present
- Better use a Regression MM- (or SMDM-) Estimator in the presence of potential leverage points
  - In the R packages **robustbase** you find:

`lmrob(...)`	Regression MM-Estimator
`lmrob(..., setting="KS2014")`	Regression SMDM-Estimator
`anova(...)`	Comparing models using robust procedures
`plot(‘‘lmrob object’’)`	graphics for a residual analysis

Remark: `lmrob(...)` is based on an improved algorithm (since `robustbase 0.9-2`) and can handle both numeric and factor variables as exploratory variables.

- Robust GM-estimators are also available for generalised linear models (GLMs)
  - In the R packages **robustbase** you find:

`glmrob(...)`	(Mallows type) Regression GM-Estimators
`anova(...)`	Comparing models using robust procedures