Half-Day 2: Robust Regression Estimation

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

WBL Statistik 2016 — Robust Fitting
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):

![Graphs showing comparison between LS and robust regression]

- or check the influence function:

\[ IF \left( \substack{ x, y; \hat{\theta}_M, \mathcal{N} } \right) = \psi \left( \frac{r}{\sigma} \right) \quad \text{M}x \text{ unbounded} \]

**Bound the total influence function!**
A First Workaround: GM-Estimation

An simple modification of the Huber’s \(\psi\)-function can remedy:

Either (Mallows)

\[
\sum_{i=1}^{n} \psi_c \left( \frac{r_i \langle \hat{\theta} \rangle}{\sigma} \right) x_i^{(k)} w \langle d \langle x_i \rangle \rangle = 0, \quad k = 1, \ldots, p,
\]

or (Schweppe)

\[
\sum_{i=1}^{n} \psi_c \left( \frac{r_i \langle \hat{\theta} \rangle}{\sigma \cdot w \langle d \langle x_i \rangle \rangle} \right) x_i^{(k)} w \langle d \langle x_i \rangle \rangle = \sum_{i=1}^{n} \psi_c \cdot w \langle d \langle x_i \rangle \rangle \left( \frac{r_i \langle \hat{\theta} \rangle}{\sigma} \right) x_i^{(k)} = 0,
\]

where \(w \langle \rangle\) is a suitable weight function and \(d \langle x_i \rangle\) is some measure of the “outlyingness” of \(x_i\).

Examples for \(w \langle \rangle\) and \(d \langle x_i \rangle\):

\[
d \langle x_i \rangle = \frac{(x_i - \text{median}(x_k))}{\text{MAD}(x_k)} \quad \text{or Mahalanobis distance} \quad \text{and} \quad w \langle d \langle x_i \rangle \rangle = \text{Huber's weight function}
\]

\[
w \langle x_i \rangle = 1 - H_{ii} \quad \text{or} \quad w \langle x_i \rangle = \sqrt{1 - H_{ii}} \quad \text{where} \ H_{ii} \ \text{is the leverage}
\]
Example Air Quality (modified)

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

![Graph showing log(Daily Mean) vs. log(Annual Mean) comparing LS, robust (M), and robust (GM).]
Example Air Quality (Initial Data)

- General Regression M-Estimation
- Robust Regression
- MM-estimation
- Robuste Inferenz

**Example Data:**

- Initial Data

**Graph:**

- Log(Annual Mean) vs. Log(Daily Mean)
- LS
- Robust (Mallows)
- Robust (GM tuned)
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!

The outliers are identified and analyse separately. The result of this 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)

That also means that it is not possible to detect clusters of leverage points with the projection matrix $H$ in residual analysis.

To see why, look at these “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 $\psi$

- 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 $\psi$ functions like Tukey’s biweight function ($\psi$ function (left) and corresponding weight function (right))

- But the equation defining the M-estimator has many solutions and only one identifies 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

$$\sum_{i=1}^{n} \left( \frac{r_i}{{\sigma}} \right)^2 = \min_{\theta} \frac{1}{n-p} \sum_{i=1}^{n} \rho \left( \frac{y_i - x_i^T \theta}{s} \right) = 0.5,$$

where $s$ must be as small as possible
(i.e. the equation should have a solution in $\theta$).

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

The function $\rho (\cdot)$ can be

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

(its derivative is Tukey's bisquare function). To get a breakdown point of 0.5
the tuning constant $b_o$ must be 1.548.
\( \rho \) 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-estimator which is **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** (modified M-estimator):

- An S-estimator with breakdown point $\varepsilon^* = 1/2$ is used as initial estimator
  Tukey’s bisquare function with $\rho_{b_0} = 1.548$: $\hat{\theta}^{(o)}$ and $s_o$
- The redescending regression M-estimator is applied
  using Tukey’s bisquare $\psi$-function $\psi_{b_1} = 4.687$ and fixed scale parameter $\sigma = s_o$ from the initial estimation; starting value is $\hat{\theta}^{(o)}$.

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

- $R_t = \text{return on the FoHF at time } t$
- $\alpha = \text{the excess return (a constant) of the FoHF}$
- $I_{k,t} = \text{the index return of sub-style } k (= \text{factor}) \text{ at time } t$
- $\beta_k = \text{the change in the return on the FoHF per unit change in factor } k$
- $p = \text{the number of used sub-indices}$
- $E_t = \text{residual (error) which cannot be explained by the factors}$
Robust MM-fit identifies clearly two different investment periods: one before April 2000 and one afterwards.
Residual Analysis with MM-Fit

```r
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) \( \theta \) and
- covariance matrix \( \sigma^2 \tau \mathbf{C}^{-1} \), where \( \mathbf{C} = (1/n) \sum_i x_i x_i^T \).

The covariance matrix of \( \hat{\theta} \) is estimated by

\[
\hat{\mathbf{V}} = \frac{s_o^2}{n} \hat{\tau} \hat{\mathbf{C}}^{-1}
\]

where

\[
\hat{\mathbf{C}} = \frac{1}{n} \sum_i w_i x_i x_i^T, \quad \hat{\tau} = \frac{1}{n} \sum_{i=1}^n \left( \frac{\psi_{b_1} \langle \tilde{r}_i \rangle}{\psi' \langle \tilde{r}_i \rangle} \right)^2
\]

with

\[
\tilde{r}_i := \frac{y_i - x_i^T \hat{\theta}^{(o)}}{s_o}, \quad w_i := \frac{\psi_{b_1} \langle \tilde{r}_i \rangle}{\tilde{r}_i}
\]

Note that \( \hat{\theta}^{(o)} \) and \( s_o \) come from the initial estimation.
A Further Modification to the MM-Estimator

- The estimated covariance matrix $\hat{V}$ depends on quantities $(\hat{\theta}^{(o)}$ 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="KS2011")
```

| Estimate | se   | Pr(>|t|) |
|----------|------|---------|
| (I)      | -0.0019 | 0.0017  | 0.2610 |
| RV       | 0.0062 | 0.3306  | 0.9850 |
| CA       | -0.0926 | 0.1658  | 0.5780 |
| FIA      | 0.0757 | 0.1472  | 0.6083 |
| EMN      | 0.1970 | 0.1558  | 0.2094 |
| ED       | -0.3010 | 0.1614  | 0.0655 |
| EDD      | 0.0687 | 0.1301  | 0.5986 |
| EDRA     | 0.0735 | 0.1882  | 0.6971 |
| LSE      | 0.4407 | 0.1521  | **0.0047** |
| GM       | 0.1723 | 0.0822  | **0.0390** |
| EM       | 0.1527 | 0.0667  | **0.0245** |
| SS       | 0.0282 | 0.0414  | 0.4973 |

Residual standard error: 0.009315

| Estimate | se   | Pr(>|t|) |
|----------|------|---------|
| -0.0030 | 0.0014 | 0.0381 |
| 0.3194  | 0.2803 | 0.2575 |
| -0.0671 | 0.1383 | 0.6288 |
| -0.0204 | 0.1279 | 0.8733 |
| 0.2721  | 0.1328 | **0.0434** |
| -0.4763 | 0.1389 | **0.0009** |
| 0.1019  | 0.1112 | 0.3621 |
| 0.0903  | 0.1583 | 0.5698 |
| 0.5813  | 0.1295 | **0.0000** |
| 0.0159  | 0.0747 | 0.8322 |
| 0.1968  | 0.0562 | **0.0007** |
| 0.0749  | 0.0356 | **0.0382** |

Residual standard error: 0.007723

The 95% confidence interval of $\beta_{SS}$ is

$$0.028 \pm 1.98 \cdot 0.041 = [-0.053, 0.109]$$

where 1.98 is the 0.975 Quantile of $t_{108}$

$$0.075 \pm 1.96 \cdot 0.036 = [0.004, 0.146]$$

where 1.98 is the 0.975 Quantile of $\mathcal{N}(0, 1)$
**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 $\theta$ 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 $H_0 : \theta_1 = 0$ where $\theta = (\theta_1^T, \theta_2^T)^T$.

- Further, let $\hat{V}_{11}$ be the quadratic submatrix containing the first $q$ rows and columns of $\hat{V}$.
The so-called **Wald-type test statistic** can now be expressed as

\[ W = \theta_1^T (\hat{V}_{11})^{-1} \theta_1. \]

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

This test statistic also provides the basis for confidence intervals of a single parameter \( \theta_k \):

\[ \hat{\theta}_k \pm q_{1-\alpha/2}^N \cdot \sqrt{\hat{V}_{kk}} \]

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

**Comments:**

- Since we have estimated the covariance matrix \( \text{Cov} (\hat{\theta}) \) by \( \hat{V} \), it seems obvious from classical regression inference that replacing \( \chi^2_q \) by \( F_{q, n-p} \) is a reasonable adjustment in the distribution of the test statistic for estimating the variance \( \sigma^2 \).

- However, to do so has no formal justification and it would be better to avoid too small sample sizes (thus \( \frac{1}{q} \cdot \chi^2_q \approx F_{q, n-p} \)) because all results are asymptotical in their nature anyway.
For an MM-estimator, we may define a robust deviance by

\[ D\left( y, \hat{\theta}_{\text{MM}} \right) = 2 \cdot s_o^2 \cdot \sum_{i=1}^{n} \rho \left( \frac{y_i - x_i^T \hat{\theta}_{\text{MM}}}{s_o} \right) \].

The robust generalisation of the F-test

\[ \left( \frac{SS_{\text{reduced}} - SS_{\text{full}}}{SS_{\text{full}}/(n-p)} \right) = \left( \frac{SS_{\text{reduced}} - SS_{\text{full}}}{\hat{\sigma}^2} \right) \]

is to replace the sum of squares by the robust deviance (similar to generalised linear models) so that we obtain the test statistic

\[ \Delta^* = \tau^* \cdot \frac{D\left( y, \hat{\theta}_{\text{MM}}^r \right) - D\left( y, \hat{\theta}_{\text{MM}}^f \right)}{s_o^2} \]

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

Then \( \Delta^* \sim \chi^2_q \) 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

<table>
<thead>
<tr>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt; F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0085024</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.0077231</td>
<td>7</td>
<td>0.00077937</td>
<td>1.2831</td>
<td>0.2679</td>
</tr>
</tbody>
</table>

# Robust with SMDM-estimator
> FoHF.rlm <- lmrob(FoHF ~ ., data=FoHF, setting="KS2011")
> anova(FoHF.rlm, 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

<table>
<thead>
<tr>
<th>pseudoDf</th>
<th>Test.Stat</th>
<th>Df</th>
<th>Pr(&gt; chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>96</td>
<td>7</td>
<td>0.0007727</td>
</tr>
</tbody>
</table>
Example from Finance - RBSA2 (cont.)

# Robust with SMDM-estimator and Wald test
> FoHF.rlm <- lmrob(FoHF ~ ., data=FoHF, setting="KS2011")
> anova(FoHF.rlm, 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

<table>
<thead>
<tr>
<th>pseudoDf</th>
<th>Test.Stat</th>
<th>Df</th>
<th>Pr(&gt;chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>96</td>
<td>7</td>
<td>0.0007727  ***</td>
</tr>
</tbody>
</table>

# Robust with SMDM-estimator and Deviance test
> anova(FoHF.rlm, 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

<table>
<thead>
<tr>
<th>pseudoDf</th>
<th>Test.Stat</th>
<th>Df</th>
<th>Pr(&gt;chisq)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>89</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>96</td>
<td>7</td>
<td>0.0007318  ***</td>
</tr>
</tbody>
</table>
## Example from Finance - RBSA2 (i.e., Slide 21 again)

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

<table>
<thead>
<tr>
<th></th>
<th>lm(FoHF ~ ., data=FoHF)</th>
<th>lmrob(FoHF ~ ., data=FoHF, setting=&quot;KS2011&quot;)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>se</td>
</tr>
<tr>
<td>(I)</td>
<td>-0.0019</td>
<td>0.0017</td>
</tr>
<tr>
<td>RV</td>
<td>0.0062</td>
<td>0.3306</td>
</tr>
<tr>
<td>CA</td>
<td>-0.0926</td>
<td>0.1658</td>
</tr>
<tr>
<td>FIA</td>
<td>0.0757</td>
<td>0.1472</td>
</tr>
<tr>
<td>EMN</td>
<td>0.1970</td>
<td>0.1558</td>
</tr>
<tr>
<td>ED</td>
<td>-0.3010</td>
<td>0.1614</td>
</tr>
<tr>
<td>EDD</td>
<td>0.0687</td>
<td>0.1301</td>
</tr>
<tr>
<td>EDRA</td>
<td>0.0735</td>
<td>0.1882</td>
</tr>
<tr>
<td>LSE</td>
<td>0.4407</td>
<td>0.1521</td>
</tr>
<tr>
<td>GM</td>
<td>0.1723</td>
<td>0.0822</td>
</tr>
<tr>
<td>EM</td>
<td>0.1527</td>
<td>0.0667</td>
</tr>
<tr>
<td>SS</td>
<td>0.0282</td>
<td>0.0414</td>
</tr>
<tr>
<td></td>
<td>Residual standard error: 0.009315</td>
<td>Residual standard error: 0.007723</td>
</tr>
</tbody>
</table>
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

- Standardized residuals vs. Robust Distances
- Normal Q–Q vs. Residuals
- Response vs. Fitted Values
- Residuals vs. Fitted Values
- Sqrt of abs(Residuals) vs. Fitted Values
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 reformulation of the linear regression model. Instead of

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

use

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

The expectation \( \mu_i \) 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

\[ g \langle \mu_i \rangle = \eta_i \]
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 \( \eta_i = x_i^T \theta = \sum_{j=1}^{p} x_i^{(j)} \cdot \theta_j \) its linear predictor. Then

**Binary / Binomial Regression**

\[ Y_i \text{ indep. } \sim B \left( \pi_i, m_i \right) \]

\[
\begin{align*}
E \left( \frac{Y_i}{m_i} \mid x_i \right) &= \pi_i \\
\text{Var} \left( \frac{Y_i}{m_i} \mid x_i \right) &= \frac{\pi_i \cdot (1 - \pi_i)}{m_i}
\end{align*}
\]

**Poisson Regression**

\[ Y_i \text{ indep. } \sim P \left( \lambda_i \right) \]

\[
\begin{align*}
E \left( Y_i \mid x_i \right) &= \lambda_i \\
\text{Var} \left( Y_i \mid x_i \right) &= \lambda_i
\end{align*}
\]

The mean response \( \pi_i \) and \( \lambda_i \), respectively, are related to the linear predictor \( \eta_i \) by the link function \( g \langle \cdot \rangle \):

\[
g \langle \pi_i \rangle = \log \langle \pi_i / (1 - \pi_i) \rangle \quad \text{Logit model}
\]

\[
g \langle \pi_i \rangle = \Phi^{-1} \langle \pi_i \rangle \quad \text{Probit model}
\]

\[
g \langle \pi_i \rangle = \log \langle - \log (1 - \pi_i) \rangle \quad \text{Complementary log-log-model}
\]

\[
g \langle \lambda_i \rangle = \log \langle \lambda_i \rangle \quad \text{log-linear model}
\]

\[
g \langle \lambda_i \rangle = \lambda_i \quad \text{identity}
\]

\[
g \langle \lambda_i \rangle = \sqrt{\lambda_i} \quad \text{square root}
\]


**Gamma Regression**

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

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

- $E \langle Y_i | x_i \rangle = \frac{\alpha_i}{\beta_i}$
- $\var \langle Y_i | x_i \rangle = \frac{\alpha_i}{\beta_i^2}$

Common link functions are

- $g \langle \mu_i \rangle = \frac{1}{\mu_i}$ inverse
- $g \langle \mu_i \rangle = \log \langle \mu_i \rangle$
- $g \langle \mu_i \rangle = \mu_i$ identity.

In GLM it is assumed that

- the response $Y_i$ is independently distributed according to a distribution from the exponential family with expectation $E \langle Y_i \rangle = \mu_i$.
- The expectation $\mu_i$ is linked by a function $g$ to the linear predictor $x_i^T \beta$:
  
  $g \langle \mu_i \rangle = x_i^T \beta$

- The variance of the response depends on $E \langle Y_i \rangle$: $\var \langle Y_i \rangle = \phi \, \var \langle \mu_i \rangle$.
  
  The so-called variance function $\var \langle \mu_i \rangle$ is determined by the distribution. $\phi$ is the dispersion parameter.
Estimating Equation

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

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

where $\mu'_i = \frac{\partial \mu}{\partial \eta_i}$ 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,

$$\operatorname{Var} \left( \frac{y_i - \mu_i}{\sqrt{V\langle \mu_i \rangle}} \right) \approx \phi \sqrt{1 - H_{ii}}.$$ 

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

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

\[ 0 = \sum_{i=1}^{n} \left( \psi_c \langle r_i \rangle \frac{\mu_i'}{\sqrt{V \langle \mu_i \rangle}} x_i w \langle x_i \rangle - fcc \langle \theta \rangle \right), \]

where \( \psi_c \langle \rangle \) is the Huber function and the vector constant \( fcc \langle \theta \rangle \) ensures the Fisher consistency of the estimator.

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

- The advantage of this approach is that standard inference based both on the asymptotic Gaussian distribution of the estimator and on robust quasi-deviances is available.

I do not dare to present the formulas here, because they look frightening. But they are implemented in R and you can run the inference procedure like `glm(...)` and `anova(...)`.

- The theory and the implementation cover Poisson, binomial, gamma and Gaussian (i.e., linear regression GM-estimator)

- 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"))`
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 R packages 'robustbase' you find:

- `lmrob(...)` — Regression MM-Estimator
- `lmrob(..., setting="KS2011")` — 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