mquantreg: An R package for estimating generalized linear M-quantile regression models

Felix Skarke[1], Timo Schmid[1],
Nicola Salvati[2]

[1] Freie Universität Berlin
[2] University of Pisa

The Use of R in official statistics
December 2, 2020

General information about the mquantreg package

- Package **mquantreg** is still under development
- Regression package for the estimation of linear and generalized linear M-quantile regression models
  - Continuous data
  - Binary data
  - Count data (poisson, quasi-poisson, negative binomial)
- Hypothesis testing via Wald- and Likelihood-Ratio Tests
- Estimation of q-scores (variable selection; further use in SAE-methods)
- Model diagnostics using diagnostic plots

Short excursion into theory (1)

What is M-estimation?

- Huber (1964) formalised a approach to robust estimation of location parameters called M-estimation:

$$\hat{\theta} = \arg\min_{\theta} \left(n^{-1} \sum_{i=1}^{n} \rho(x_i; \theta) \right)$$

or if $\rho(\cdot)$ is differentiable and convex a solution is easier found with **influence function** $\psi(x; \theta)$:

$$n^{-1} \sum_{i=1}^{n} \psi(x_i; \hat{\theta}) = 0, \text{where } \psi(x; \theta) = \frac{\partial}{\partial \theta} \rho(x; \theta)$$

- examples: $\rho(x; \bar{x}) = (x - \bar{x})^2$, $\rho(x; \widetilde{x}) = |x - \widetilde{x}|$

Short excursion into theory (2)

Huber Proposal 2 for $\psi(x;\theta)$:

$$\psi_k(x-\theta) = \begin{cases} x-\theta, & \text{if } |x-\theta| < k \\ k\,\text{sgn}(x-\theta), & \text{if } |x-\theta| \geq k \end{cases}$$

or alternatively $\psi_k(x-\theta) = \max(-k, \min(x-\theta, k))$ with **tuning constant** k, which has to be chosen beforehand.

- k determines the robustness of the estimator
- Huber estimator is equivalent to the mean for $k \to \infty$ and to the median for $k \to 0$

Short excursion into theory (3)

- Generalization of the concept to the regression case (Huber, 1973)
- Further development by Breckling and Chambers (1988) to M-quantile regression for continuous data:

Influence Function (depending on q and k):

$$\psi_{q,k}(y - MQ_{q,k}) = 2[(1-q)I_{y \leq MQ_{q,k}} + qI_{y > MQ_{q,k}}]\psi_k(y - MQ_{q,k})$$

Estimating equations (depending on q and k):

$$n^{-1}\sum_{i=1}^{n}\psi_{q,k}(y_i - \hat{MQ}_{q,k}(x_i))x_i = 0, \text{ where } \hat{MQ}_{q,k}(x_i) = x_i^T\hat{\beta}_{q,k}$$

Short excursion into theory (4)

- M-quantiles are not scale-equivariant: scale parameter $\sigma_{q,k}$ has to estimated at the same time as regression coefficients
- The parameters can be estimated via IRLS or ML using Asymmetric Least Informative (ALI) distribution (Bianchi et al. 2018)
- Categorical data
  - Generalizations of estimation approach to robust generalised linear models by Cantoni and Ronchetti (2001)
  - Binary: Chambers, Salvati and Tzavidis (2016)
  - Poisson: Tzavidis, Ranalli et al. (2015)
  - Negative Binomial: Chambers, Dreassi and Salvati (2014)

Short excursion into theory (5)

Q-scores:

- Application of M-quantiles by Kokic et al. (1997)
- Can be found by solving for q_i^*, which solves

$$MQ_{q_i^*,k}(x_i) = y_i$$

- Observations with high q-scores have values high values y_i conditional on x_i
- Q-scores can be used in a lot of different ways in a M-quantile regression context
  - Model diagnostics
  - Variable selection (Dawber, 2017)
  - Useful for example in Small Area Estimation for the estimation of pseudo-random effects (Chambers and Tzavidis (2006))

Estimation

Example for continuous data using a sample of size 1000 from the German EUSILC PUMD (Public Use Microdata) from 2013:

```
fit <- mquantreg(formula = income ~ income_self_empl +
    income_empl + old_age_benefits + sex,
    q = c(0.1, 0.5, 0.7),
    k = 1.345,
    data = dat,
    method = "continuous",
    scale.estimator = "Mad",
    cont = "lik",
    compute.qscores = FALSE)
```

Print and Summary functions (1)

```
> print(fit)
Call:
mquantreg(formula = income ~ income_self_empl + income_
    empl +old_age_benefits + sex, data = dat, q = c
    (0.1, 0.5, 0.7), method = "continuous", scale.
    estimator = "Mad", compute.qscores = FALSE, cont =
    "lik")

Coefficients:
     q   (Intercept) income_self_empl income_empl
[1,] 0.1    4.936694     0.0004245352   0.4365187
[2,] 0.5   12.418848     0.0005556530   0.5109797
[3,] 0.7   15.776415     0.0006699692   0.5382013

old_age_benefits sexFemale
0.4866095        1.523643
0.4279946        2.754370
0.4199769        3.209267
Degrees of freedom: 1000 total; 995 residual
```

Print and Summary functions (2)

```
> summary(fit)
Estimation for m-quantile  q = 0.5 with k =  1.345
n = 1000

Coefficients:
                coefficients   Std. Error
(Intercept)       12.418847514 5.827978e-01
income_self_empl   0.000555653 3.709961e-05
income_empl        0.510979679 1.179632e-02
old_age_benefits   0.427994581 3.231350e-02
sexFemale          2.754369615 6.575171e-01
Estimator for scale parameter sigma:  9.593204

Residuals:
-27.30647 -6.399448 -0.7536041 1.611043 6.858612
    102.6775

Proportion of residuals smaller than 0: 0.539
Residuals bounded at +- 12.90286
Proportion of Huberised residuals: 17.7 %
Pseudo R-squared (Bianchi et al, 2015):  0.1356437
```

Variable Selection

```
fit_sel <- stepmq(fit)
Starting model:
income ~ income_self_empl + income_empl + old_age_
    benefits + sex

                        Df      Corr.
<none>                  NA 1.0000000
- sex                    1 0.9862613
- income_self_empl       1 0.9757906
- old_age_benefits       1 0.9072897
- income_empl            1 0.7163004
                        Df      Corr.
<none>                  NA 1.0000000
- income_self_empl       1 0.9761506
- old_age_benefits       1 0.9162316
- income_empl            1 0.7169594
                        Df      Corr.
<none>                  NA 1.0000000
- old_age_benefits       1 0.9241252
- income_empl            1 0.7323867
```

Hypothesis Testing (1)

Wald tests for single coefficients:

```
> mqwald(fit, id.beta = c(3,4), each = TRUE, q = 0.5)
mqwald(fit, id.beta = c(3,4), each = TRUE, q = 0.5)
Call:
mqwald(object = fit, id.beta = c(3, 4), q = 0.5, each =
    TRUE)

Wald tests for single coefficients at q-value(s) 0.5
H0:[1] 0 0

test-statistic:
      income_empl old_age_benefits
[1,]    1876.353          175.4319

degrees of freedom
[1] 1 1

p-values
       income_empl old_age_benefits
[1,]             0                0
```

Hypothesis Testing (2)

Likelihood-Ratio Omnibus-Test:

```
> mqLRT(fit)
Call:
mqLRT(object = fit)

Estimated coefficients to be restricted to zero in
    reduced model:

    [...]

Result of Likelihood-Ratio test(s):
        q-value(s)  Chi^2  df   p-value(s)
[1,]           0.1  686.3380  4           0
[2,]           0.5  986.5277  4           0
[3,]           0.7  812.1397  4           0

The likelihood-ratio test gives significant results
    at a level alpha of 0.05 for the m-quantile(s)
    0.1, 0.5, 0.7 .
```

Graphical Diagnostics (1)

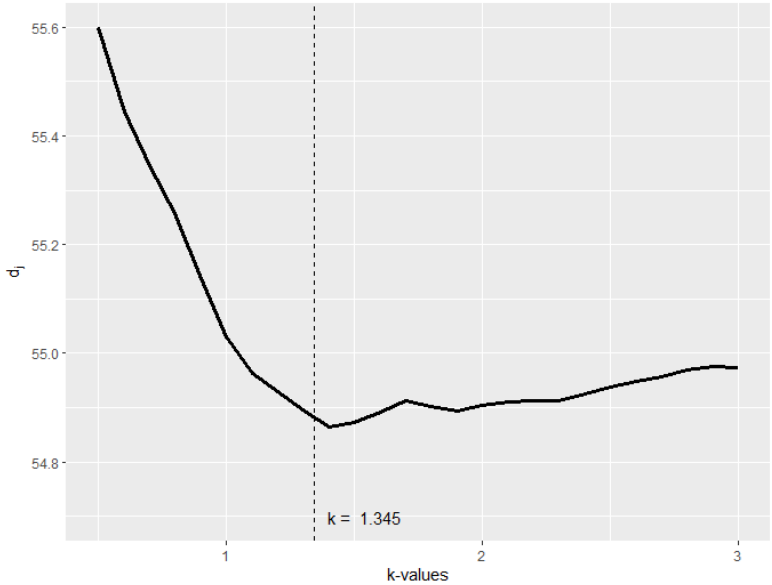
The packages includes the following ploting options (based on Dawber (2017)):

- "fit.res" - Fitted vs. Residuals Plot
- "prop.res" - Proportion of Residuals Smaller than 0 vs. Fitted Values
- "fitted.observed" - Fitted vs. Observerd Values
- "optk" - Comparing the Residual Distribution with Normal Distribution
- "coef.q" - Coefficients over q
- "prop.huber.k" - Proportion of Huberised residuals over k
- "prop.huber.q" - Proportion of Huberised residuals over q

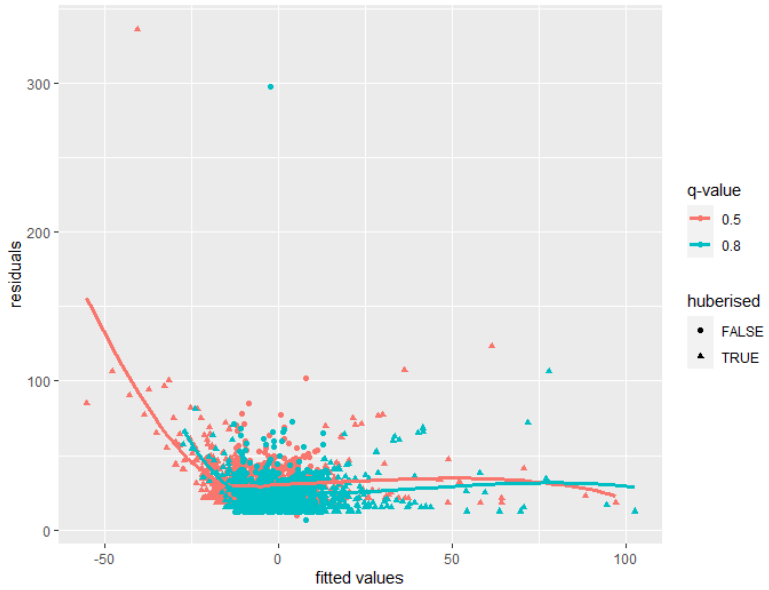
Graphical Diagnostics (2)

```
1    > plot(fit, plottype = "optk")
2    > plot(fit, plottype = "fit.res", add.qvals = c(0.8))
```

Optimal k



Fitted vs. Residuals



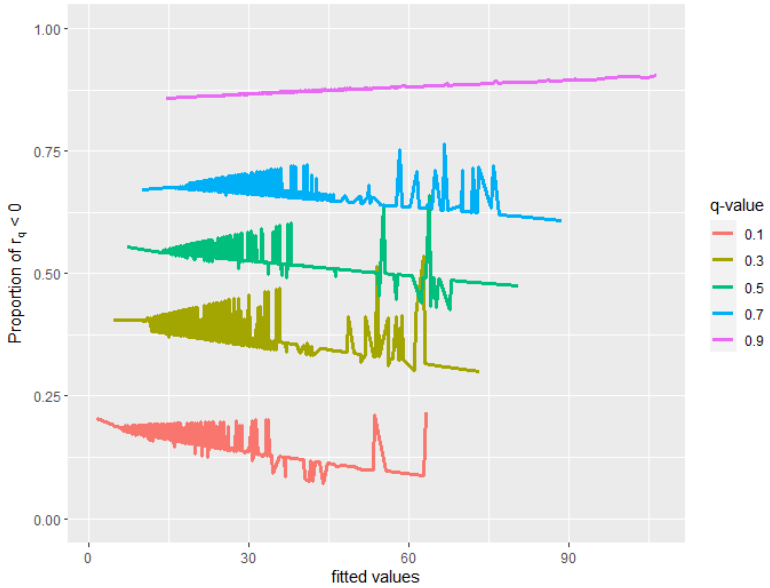
Freie Universität Berlin

UNIVERSITÀ DI PISA

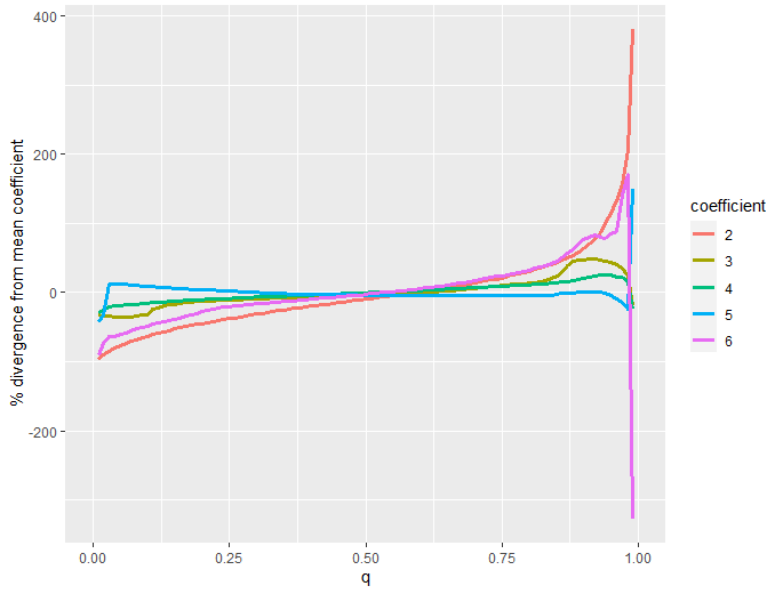
Graphical Diagnostics (3)

```
1    > plot(fit, plottype = "prop.res", add.qvals = c(0.1,0.3,0.7,
         0.9))
2    > plot(fit, plottype = "coef.q")
```

Fitted vs. Proportion negative Residuals



Coefficients over q-values



Using simulated data under contamination:

```
1       fit_binom <- mquantreg(formula = Y ~ x1 + x2,data =
            dat, q  = 0.5,
2           method = "binom")
```

Summary for simulated example

```
Call:
mquantreg(formula = "Y ~ x1 + x2", data = df, q =
    0.5, method = "binom")
Estimation for m-quantile  q = 0.5 with k =  1.6
n = 1000
Coefficients:
            coefficients Std. Error
(Intercept)   0.09966344 0.06557439
x1            0.36810515 0.06782330
x2            0.38303418 0.06855655
Estimator for scale parameter sigma:  0.4971013

Residuals:
-0.7280652 -0.4848358 0.2357902 0.001179976 0.4516844
     0.795725

Proportion of residuals smaller than 0: 0.475
Residuals bounded at +- 0.6686012
Proportion of Huberised residuals: 7.9 %
---
Estimation with Robust Quasi-Likelihood Estimation
```

Conclusion

- M-quantile regression is a field of continued research
- These methods are helpful not just as regression models, but can be applied in other contexts as for example SAE
- Package **mquantreg** can be the foundational package for further use of M-quantile regression
- In its current state package **mquantreg** already has the functionality users are accustomed to from other regression packages (i.e. summary(), print(), plot() based on classes)

References

Bianchi, A., Fabrizi, E., Salvati, N., and Tzavidis, N. (2018). Estimation and Testing in M-quantile Regression with application to small area estimation.

International Statistical Review, 85(3), 541 - 570.

Breckling, J. and Chambers, R. (1988). M-quantiles

Biometrika, 75(4), 761 - 771.

Cantoni, E. and Ronchetti, E. (2001). Robust Inference for Generalized Linear Models.

Journal of the American Statistical Association, 96(455), 1022 - 1030.

Chambers, R., Dreassi, E., and Salvati, N. (2014). Disease mapping via negative binomial regression M-quantiles.

Statistics in Medicine, 33(27), 4805 - 4824.

References

Chambers, R., Salvati, N., and Tzavidis, N. (2016). Semiparametric small area estimation for binary outcomes with application to unemployment estimation for local authorities in the UK.

Journal of the Royal Statistical Society: Series A (Statistics in Society), 179(2), 453 - 479.

Chambers, R. and Tzavidis, N. (2006). M-quantile Models for Small Area Estimation.

Biometrika, 93(2), 255 - 268.

Dawber, J. (2017). Advances in M-quantile estimation.

Doctor of Philosophy thesis, School of Mathematics and Applied Statistics, University of Wollongong. https://ro.uow.edu.au/theses1/188.

Huber, P. J. (1964). Robust Estimation of a Location Parameter.

The Annals of Mathematical Statistics, 35(1), 73 - 101

References

Huber, P. J. (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo.

The Annals of Statistics, 1(5), 799 - 821.

Kokic, P., Chambers, R., Breckling, J., and Beare, S. (1997). A Measure of Production Performance.

Journal of Business & Economic Statistics, 15(4),445 - 451.

Tzavidis, N., Ranalli, M. G., Salvati, N., Dreassi, E., and Chambers, R. (2015). Robust small area prediction for counts.

Statistical methods in medical research, 24(3), 373 - 395.

Thank you for your attention

Felix Skarke
f.skarke@fu-berlin.de