Lecture 10
Robust and Quantile Regression

1

Outliers

• Many definitions: Atypical observations, extreme values, conditional unusual values, observations outside the expected relation, etc.

• In general, we call an *outlier* an observation that is numerically different from the data. But, is this observation a "mistake," say a result of measurement error, or part of the (heavy-tailed) distribution?

• In the case of normally distributed data, roughly 1 in 370 data points will deviate from the mean by 3*SD. Suppose T=1,000 and we see **9** data points deviating from the mean by more than 3*SD indicates outliers... Which of the **9** observations can be classified as an outlier?

Problem with outliers: They can affect estimates. For example, with small data sets, one big outlier can seriously affect OLS estimates.

Outliers

• Many definitions: Atypical observations, extreme values, conditional unusual values, observations outside the expected relation, etc.

• In general, we call an *outlier* an observation that is numerically different from the data. But, is this observation a "mistake," say a result of measurement error, or part of the (heavy-tailed) distribution?

• In the case of normally distributed data, roughly 1 in 370 data points will deviate from the mean by 3*SD. Suppose T=1,000 and we see **9** data points deviating from the mean by more than 3*SD indicates outliers... Which of the **9** observations can be classified as an outlier?

Problem with outliers: They can affect estimates. For example, with small data sets, one big outlier can seriously affect OLS estimates.

Outliers

• Sometimes, a distinction is made between ***y*-outliers** & ***x*-outliers**. In a regression, we usually look at outliers in the residuals (y-outliers).

• There are many proposed measures to detect outliers. In general, these measures are evaluated informally, through ad-hoc rules ("**rules of thumb**"). Under some assumptions, usually assuming a normal distribution, there are some formal tests for outliers.

• In general, the outlier literature is more interested in the identification of outliers, not a lot of attention is devoted to miss-identification –i.e., Type I error (false positive).

Remark: It is common to evaluate results and, if they go with the intuition, ignore potential outliers.

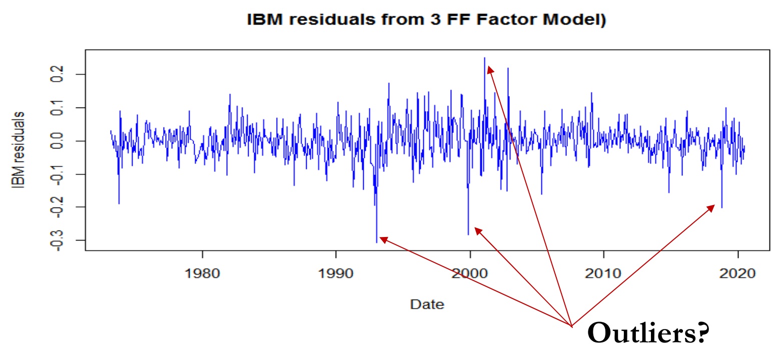
Outliers: Identification

• Informal identification method:

- *Eyeball*: Look at the observations away from a scatter plot.

Example: Plot residuals for the 3 FF factor model for IBM returns

```
x_resid <- residuals(fit_ibm_ff3)
plot(x_resid, typ ="l", col="blue", main ="IBM Residuals from 3 FF Factor Model",
xlab="Date", ylab="IBM residuals")
```



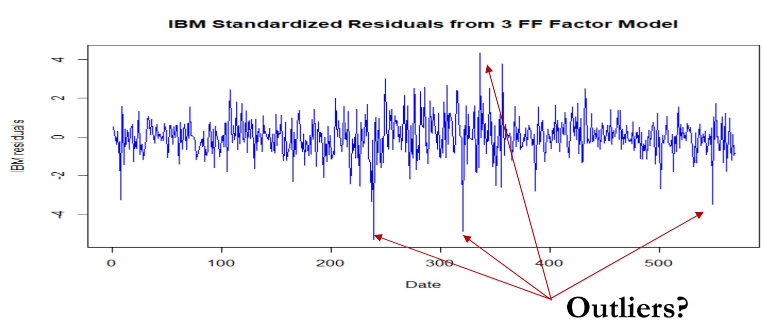
Outliers: Identification

• Formal identifications methods:

- **Standardized residuals**, $e_i/\mathrm{SD}(e_i)$: Check for standardized errors that are bigger than 2 (or 3).

Example: Plot standardized residuals for IBM residuals

```
x_stand_resid <- x_resid/sd(x_resid)      # standardized residuals
plot(x_stand_resid, typ ="l", col="blue", main ="IBM Standardized Residuals from 3 FF
Factor Model", xlab="Date", ylab="IBM residuals")
```



Outliers: Identification – Leverage & Influence

• Formal identifications methods:

- **Leverage statistics**: It measures the difference of an independent data point from its mean. High leverage observations can be potential outliers. Leverage is measured by the diagonal values of the **P** matrix:

$$h_{ii} = \ 1/T + (x_i - \bar{x})^2/(T-1)\ s_x^2]$$

Intuition: Recall $\hat{y} = Py$ $\Rightarrow \hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \cdots + h_{iT}y_T$

• h_{jj} quantifies the influence that the observed response y_i has on its predicted value $\hat{y}_i$. Large h_{ii}, y_i plays a large role in $\hat{y}_i$.

• It turns out $h_{ii} \in [0, 1]$ & the sum of the h_{ii} is equal to k.

• A standard cut-off point for h_{ii} is $(2k+2)/T$. But, other cut-off points are used, for example, $3k/T$.

Outliers: Identification – Leverage & Influence

• For multivariate sets, Mahalanobis distance (MD) is recommended.

$$\text{MD}(\boldsymbol{x}_i) = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})'\, \boldsymbol{S}_x^2\, (\boldsymbol{x}_i - \bar{\boldsymbol{x}})'$$

Suggested cut-off values for $\text{MD}(\boldsymbol{x}_i)/k$ are 3 or 4 for large T.

Note: An observation can have high leverage, but no *influence*.

- **Influence statistics**: **Dif beta**. It measures how much an observation influences a parameter estimate, say b_j. Dif beta is calculated by removing an observation, say i, recalculating b_j, say $b_j(-i)$, taking the difference in betas and standardizing it. Then,

$$Dif\ beta_{j(-i)} = \frac{\sum_{j=1}^{k}(b_j - b_j(-i))}{SE[b_j]}$$

• Usual threshold for declaring an observation "influential" is $2/\sqrt{T}$.

Outliers: Identification – Leverage & Influence

• A related popular influence statistic is **Distance D** *(as in Cook's D)*. It measures the effect of deleting an observation, say i, on the fitted values, say $\hat{y}_j$. Using the previous notation we have:

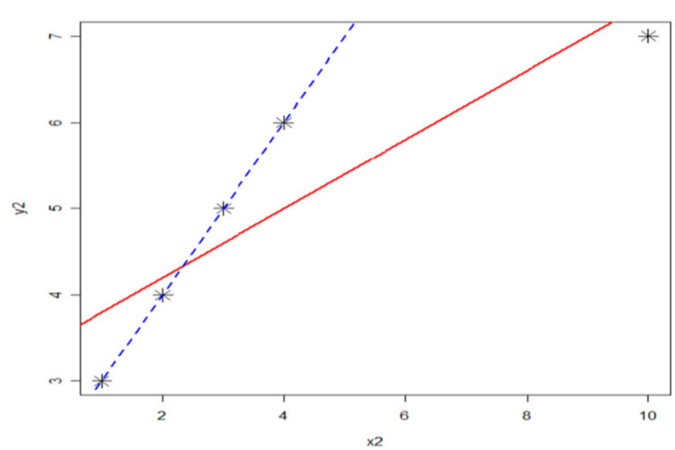
$$D_i = \frac{\sum_{i=1}^{T}(\hat{y}_j - \hat{y}_j(-i))}{k \, * MSE}$$

where k is the number of parameters in the model and MSE is mean square error of the regression model (MSE = RSS$/T$).

• Popular rule of thumb for Cook's D: If $D_i > 4T \Rightarrow$ observation i is considered a (potential) highly influential point.

• The textbook of Kutner et al. (2005), recommends comparing D_i to the $F_{k,T-k}$ distribution ⇒ greater than the 50% percentile signals an outlier.

Outliers: Leverage & Influence



• Deleting the observation in the upper right corner has a clear effect on the regression line. This observation has *leverage* and *influence*.

Outliers: Summary of Rules of Thumb

• Summary of popular rules of thumb used to identify outliers:

Measure	Value	
abs(stand resid)	> 2	(> 3 is another popular value)
leverage	$> (2\,k + 2)/T$	($> 3k/T$ is also used)
abs(*Dif Beta*)	$> 2/\sqrt{T}$	(If T is small, 1 can be used)
Cook's D	$> 4/T$	

Note: The analysis can also be carried out for groups of observations. In this case, we look for blocks of highly influential observations.

Outliers: Example

Example: We estimate the Fama-French 3-factor model for IBM and then we look for outliers in the residuals:

```
SFX_da <- read.csv("http://www.bauer.uh.edu/rsusmel/4397/Stocks_FX_1973.csv",
head=TRUE, sep=",")

## Extract variables from imported data
x_ibm <- SFX_da$IBM                    # extract IBM price data
x_Mkt_RF <- SFX_da$Mkt_RF              # extract Market excess returns (in %)
x_RF <- SFX_da$RF                      # extract Risk-free rate (in %)
x_SMB <- SFX_da$SMB
x_HML <- SFX_da$HML

# Define log returns & adjust size of variables accordingly
T <- length(x_ibm)                     # sample size
lr_ibm <- log(x_ibm[-1]/x_ibm[-T])     # create IBM log returns (in decimal returns)
Mkt_RF <- x_Mkt_RF[-1]/100             # Adjust sample size to (T-1) by removing 1st obs
RF <- x_RF[-1]/100                     # Adjust sample size and use decimal returns.
SMB <- x_SMB[-1]/100
HML <- x_HML[-1]/100
ibm_x <- lr_ibm - RF
```

Outliers: Example

Example (continuation):

```
fit_ibm_ff3 <- lm(ibm_x ~ Mkt_RF + SMB + HML)
> summary(fit_ibm_ff3)

Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)  -0.004947  0.002408    -2.054   0.04040 *
Mkt_RF       0.885706   0.054259    16.324   < 2e-16 ***
SMB          -0.228137  0.081180    -2.810   0.00511 **
HML          -0.058153  0.077791    -0.748   0.45502
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05857 on 607 degrees of freedom
Multiple R-squared:  0.3227,  Adjusted R-squared:  0.3193
F-statistic: 96.38 on 3 and 607 DF,  p-value: < 2.2e-16564
```

Note: Market and SMB are significant factors.

Outliers: Example

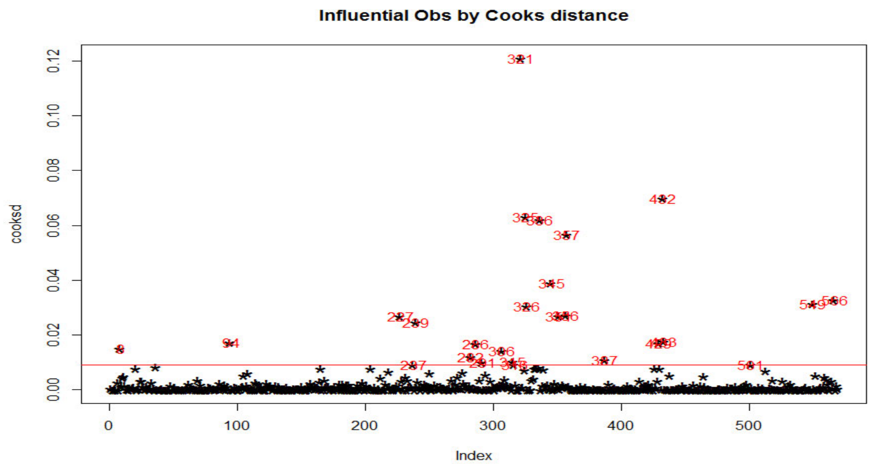
Example: Cook's D for IBM returns using the 3 FF Factor Model

```
mod <- lm(ibm_x ~  Mkt_RF + SML + HML)
cooksd <- cooks.distance(fit_ibm_ff3)
# plot cook's distance
plot(cooksd, pch="*", cex=2, main="Influential Obs by Cooks distance")
# add cutoff line
abline(h = 4*mean(cooksd, na.rm=T), col="red")  # add cutoff line
# add labels
text(x=1:length(cooksd)+1, y=cooksd, labels=ifelse(cooksd>4*mean(cooksd,
na.rm=T),names(cooksd),""), col="red")  # add labels

 # influential row numbers
influential <- as.numeric(names(cooksd)[(cooksd > 4*mean(cooksd, na.rm=T))])
 # print first 10 influential observations.
head(dat_xy[influential, ],n=10L)
```

Note: There are easier ways to plot Cook's D and identify the suspect outliers. The package *olsrr* can be used for this purpose too.

Outliers: Example

Example (continuation): Cook's D for IBM (3 FF Factor Model)



Outliers: Example

Example (continuation): Cook's D for IBM (3 Factor-Model)

```
> # print first 10 influential observations.
>head(dat_xy[influential, ],n=10L)
```

	y	V1	Mkt_RF	SMB	HML
8	-0.16095068	1	0.0475	0.0294	0.0219
94	0.01266444	1	0.0959	-0.0345	-0.0835
227	-0.04237227	1	0.1084	-0.0224	-0.0403
237	-0.19083575	1	0.0102	0.0205	-0.0210
239	-0.30648638	1	0.0153	0.0164	0.0252
282	0.07787100	1	-0.0597	-0.0383	0.0445
286	0.20734626	1	0.0625	-0.0389	0.0117
291	0.15218986	1	0.0404	-0.0565	-0.0006
306	0.13928315	1	-0.0246	-0.0512	-0.0096
315	0.16196934	1	0.0433	0.0400	0.0253

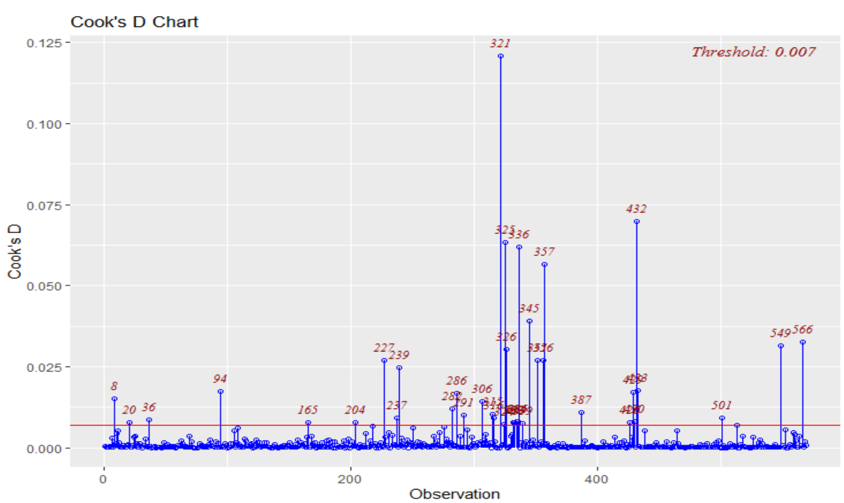
Outliers: Example

Example: Different tools to check for outliers for IBM residuals
We will use the package *olsrr*.

```
library(olsrr)                        # need to install package olsrr
x_resid <- residuals(fit_ibm_ff3)
x_stand_resid <- x_resid/sd(x_resid)  # standardized residuals
sum(x_stand_resid > 2)                # Rule of thumb count (5% count is OK)
x_lev <- ols_leverage(fit_ibm_ff3)    # leverage residuals
sum(x_lev >(2*k+2)/T)                 # Rule of thumb count (5% count is OK
ols_plot_resid_stand(fit_ibm_ff3)     # Plot standardized residuals
ols_plot_cooksd_bar(fit_ibm_ff3)      # Plot Cook's D measure
ols_plot_dffits(fit_ibm_ff3)          # Plot Difference in fits
ols_plot_dfbetas(fit_ibm_ff3)         # Plot Difference in betas

> sum(x_lev >(2*k+2)/T)
[1] 32                                 # 5%? = 32/569 = 0.0562
> sum(x_stand_resid > 2)
[1] 13                                 # 5%? = 13/569 = 0.0228
```

Outliers: Example

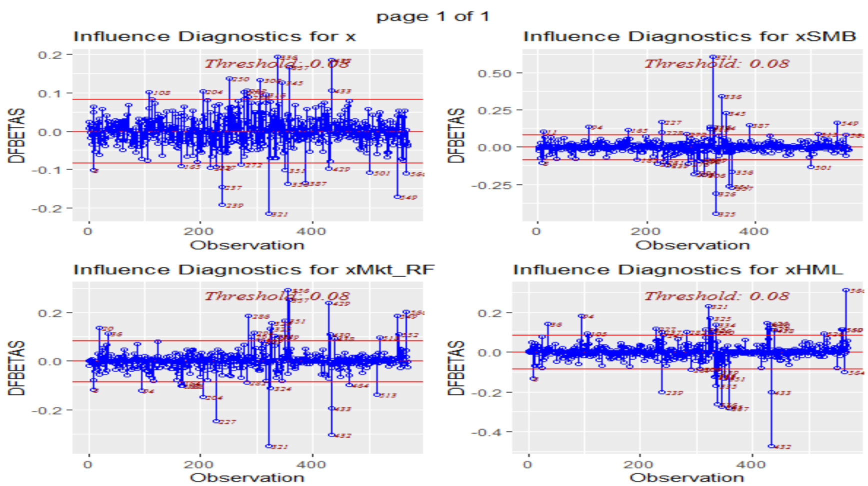
Example (continuation):
> ols_plot_cooksd_bar(fit_ibm_ff3)   # Plot Cook's D measure



Outliers: Example

Example (continuation):

> ols_plot_dfbetas(fit_ibm_ff3)



Outliers: Application – Rules of Thumb

• The histogram, Boxplot, and quantiles helps us see some potential outliers, but we cannot see which observations are potential outliers. For these, we can use Cook's D, $Dif\ beta$'s, standardized residuals and leverage statistics, which are estimated for each i.

Observation Type	Proportion	Cutoff	
Outlier	0.0356	2.00	(abs(standardized residuals) > 2)
Outlier	0.1474	$2/\sqrt{T}$	(diffit > 2/sqrt(1038)=0.0621)
Outlier	0.0501	$4/\ T$	(cookd > 4/1038=0.00385)
Leverage	0.0723	$(2\ k+2)/T$	(h = leverage > .00771)

Outliers: Example

Example (continuation): The FF3 model does not seem to suffer from outliers. Now, we corrupt the data, we **add 2 outliers** to IBM.

```
y <- ibm_x
y[10] <- -0.85                    # Corrupt observation (added outlier #1)
y[90] <- 0.95                     # Corrupt observation (added outlier #2)
fit_ibm_ff3_out <- lm(y ~ Mkt_RF + SMB + HML)
>summary(fit_ibm_ff3_out)
```

Coefficients:

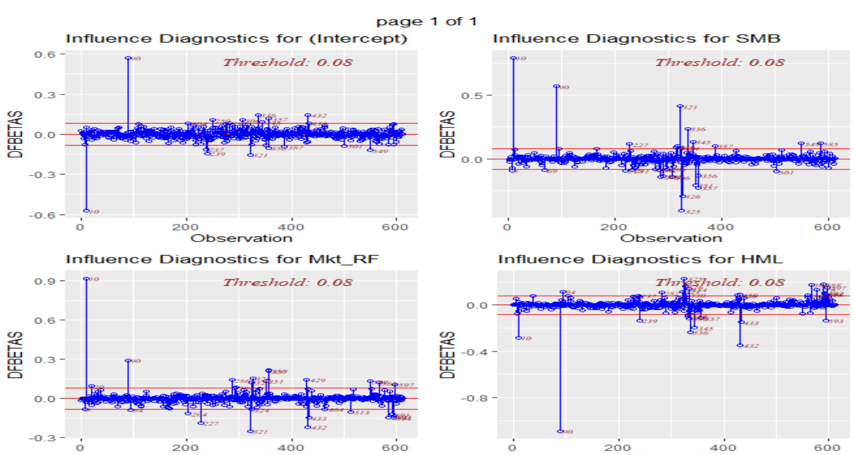
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.005215	0.003084	-1.691	0.0914 .
Mkt_RF	0.965445	0.069487	**13.894**	<2e-16 ***
SMB	-0.097430	0.103964	-0.937	0.3491
HML	-0.176166	0.099623	-1.768	0.0775 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Note: SMB and constant are not longer significant factors at the 5% level. HML is now significant at the 10%.

Outliers: Example

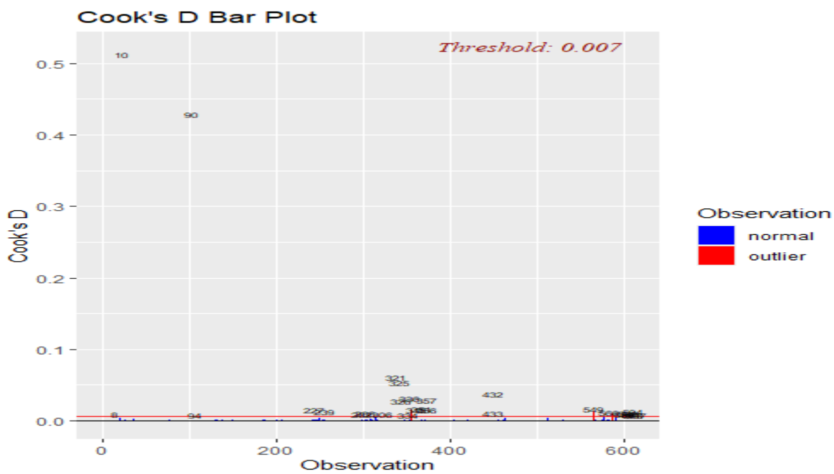
Example (continuation): The corrupted outliers are easily picked up by the standard influence measures
> ols_plot_dfbetas(fit_ibm_ff3_out)



Outliers: Example

Example (continuation):

> ols_plot_cooksd_bar(fit_ibm_ff3_out)   # Plot Cook's D measure



Outliers: What to do?

• Typical solutions:

- Use a non-linear formulation or apply a transformation (log, square root, etc.) to the data.

- Remove suspected observations. (Sometimes, there are theoretical reasons to remove suspect observations. Typical procedure in finance: remove public utilities or financial firms from the analysis.)

- Winsorization of the data (traditionally, the most common method).

- Use dummy variables.

- Use LAD (quantile) regressions, which are less sensitive to outliers.

- Weight observations by size of residuals or variance (robust estimation).

• General rule: Present results with or without outliers.

Robust Estimation

• Following Huber (1981), we will interpret **robustness** as insensitivity to small deviations from the assumptions the model imposes on the data.

• In particular, we are interested in *distributional robustness*, and the impact of skewed distributions and/or *outliers* on regression estimates.

– In this context, *robust* refers to the shape of a distribution –i.e., when the actual distribution differs from the theoretically assumed distribution.

– Although conceptually distinct, distributional robustness and outlier resistance are, for practical purposes, synonymous

– Robust can also be used to describe standard errors that are adjusted for non-constant error variance. But, we have already covered this topic.

25

Robust Estimation – Mean vs Median

• Intuition: Under normality, OLS has optimal properties. But, under non-normality, nonlinear estimators may be better than LS estimators.

Example: *i.i.d.* case

Let $\{y_t\} \sim F(\frac{y-\mu}{\sigma})$ where $F(0) = 0.5$.

where F is a symmetric distribution with scale parameter σ.

• Let the order statistics be $y_1 \leq \cdots \leq y_T$

• Sample median: $\tilde{\mu} = y_{(T+2)/2}$

• Laplace showed that

$$\sqrt{T}\,(\tilde{\mu} - \mu) \to N\left(0, \frac{1}{4\, f(\mu = 0)^2}\right)$$

26

Robust Estimation – Mean vs Median

• Using this result, one can show:

	$T\text{var}(\text{mean} = \mu)$	$T\text{var}(\text{median} = \tilde{\mu})$
Normal	1	1.57
Laplace	2	1
Average	1.5	1.28

• Intuitively, this occurs because Laplace is fat-tailed, and the median is much less sensitive to the information in the tails than the mean.

• The mean gives $1/T$ weight to all observations (close to the mean or in the tails). A large observation can seriously affect (*influence*) the mean, but not the median.

27

Robust Estimation – Mean vs Median

• Remark: The sample mean is the MLE under the Normal distribution; while the sample median is the MLE under the Laplace distribution.

• If we do not know which distribution is more likely, following Huber, we say the median is robust ("better"). But, if the data is normal, the median is not efficient (57% less efficient than mean).

• There are many types of robust estimators. Although they work in different ways, they all give less weight to observations that would otherwise influence the estimator.

• Ideally, we would like to design a weighting scheme that delivers a robust estimator with good properties (efficiency) under normality.

28

Robust Estimation – Mean vs Median

Examples: Robust estimators for central location parameter.

- The sample median, $\tilde{\mu}$.

- **Trimmed-Mean**, the mean of the sample after fraction a of the largest and smallest observations have been removed.

- The "**Winsorized Mean,**" $\hat{\mu}^W$:

$$\hat{\mu}^W = \frac{1}{T}\{(g+1) * y_{g+1} + y_{g+2} + \ldots + y_{T-g-1} + (g+1) * y_{T-g}\}$$

which is similar to the trimmed-mean, but instead of throwing out the extremes, we "accumulate" them at the truncation point.

• Q: All robust, which one is better? Trade-off: robustness-efficiency.

• The concept of robust estimation can be easily extended to the problem of estimating parameters in the regression framework.

29

Robust Regression

• There are many types of robust regression models. Although they work in different ways, they all give less weight to observations that would otherwise influence the regression line.

• Early methods:

– ***Least Absolute Deviation/Values*** (LAD/LAV) regression or least absolute deviation regression –i.e., minimizes $|e|$ instead of e^2.

• Modern methods:

- ***M-Estimation***
  - Huber estimates, Bi-square estimators
- ***Bounded Influence Regression***
  - Least Median of Squares, Least-Trimmed Squares

30

Review: M-Estimation

• An extremum estimator is one obtained as the optimizer of a criterion function, $q(\boldsymbol{z}, \mathbf{b})$.

Examples:

OLS: $\mathbf{b}$ = arg max $\{-\sum_{i=1}^{T} e_i^{\ 2} = -\boldsymbol{e'e}\ /T\}$

MLE: $\mathbf{b}_{MLE}$ = arg max $\{\ \ln L = \sum_{i=1}^{T} lnf(\boldsymbol{x}_i, y_i, \mathbf{b})\ \}$

GMM: $\mathbf{b}_{GMM}$ = arg max $\{-\ g(\boldsymbol{x}_i, y_i, \mathbf{b})'\ \mathbf{W}\ g(\boldsymbol{x}_i, y_i, \mathbf{b})\}$

• There are two classes of extremum estimators:

- M-estimators: The objective function is a sample average or a sum.

- Minimum distance estimators: The objective function is a measure of a **distance**.

• "**M**" stands for a maximum or minimum estimators –Huber (1967). 31

Review: M-Estimation

• The objective function is a sample average or a sum.

• We want to minimize a population (first) moment:

$$\min_{\mathbf{b}} \mathrm{E}[q(\boldsymbol{z}, \beta)]$$

– Using the LLN, we move from the population first moment to the sample average:

$$\sum_{i=1}^{T} q(\boldsymbol{z}_i, \mathbf{b})/T \xrightarrow{p} \mathrm{E}[q(\boldsymbol{z}, \beta)]$$

– We want to obtain: $\mathbf{b}$ = argmin $\sum_{i=1}^{T} q(\boldsymbol{z}_i, \mathbf{b})$ (or divided by T)

– In general, we solve the f.o.c. (or zero-score condition):

$$\text{Zero-Score:} \quad \sum_{i=1}^{T} \frac{\partial q(\boldsymbol{z}_i, \mathbf{b})}{\partial \mathbf{b}'} = \mathbf{0}$$

– To check the s.o.c., we define the (pd) Hessian:

$$\mathbf{H} = \sum_{i=1}^{T} \frac{\partial^2 q(\boldsymbol{z}^i, \mathbf{b})}{\partial \mathbf{b}\ \partial \mathbf{b}'}$$

32

Review: M-Estimation

- If $\mathbf{s}(\boldsymbol{z}, \mathbf{b}) = \frac{\partial q(\boldsymbol{z}_i, \mathbf{b})}{\partial \mathbf{b}'}$ exists (almost everywhere), we solve

$$\sum_{i=1}^{T} s(\boldsymbol{z}_i, \mathbf{b})/T = 0 \quad (*)$$

- If, in addition, $E_X[s(\boldsymbol{z}_i, \mathbf{b})] = \partial/\partial\mathbf{b}'\ E_X[q(\boldsymbol{z}, \boldsymbol{\beta})]$ –i.e., differentiation and integration are exchangeable–, then

$$E_X[\frac{\partial q(\boldsymbol{z}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}] = \mathbf{0}.$$

- Under these assumptions the M-estimator is said to be of *ψ-type* ($\psi = \mathbf{s}(\boldsymbol{z}, \mathbf{b})$ = score). Often, $\mathbf{b}_M$ is taken to be the solution of (*) without checking whether it is indeed a minimum).

- Otherwise, the M-estimator is of *ρ-type*. ($\rho = q(\boldsymbol{z}_i, \mathbf{b})$).

33

Review: M-Estimation

- Minimum L_p-estimators

$$
\begin{array}{lll}
- q(\boldsymbol{z}, \boldsymbol{\beta}) & = (1/p)|\boldsymbol{x} - \boldsymbol{\beta}|^p & \text{for } 1 \le p \le 2 \\
- \boldsymbol{s}(\boldsymbol{z}, \boldsymbol{\beta}) & = |\boldsymbol{x} - \boldsymbol{\beta}|^{p-1} & \boldsymbol{x} - \boldsymbol{\beta} < 0 \\
 & = -|\boldsymbol{x} - \boldsymbol{\beta}|^{p-1} & \boldsymbol{x} - \boldsymbol{\beta} > 0
\end{array}
$$

- Special cases:

– $p = 2$: We get the sample mean (LS estimator for $\boldsymbol{\beta}$).

$$\boldsymbol{s}(\boldsymbol{z}, \boldsymbol{\beta}) = \sum_{i=1}^{T}(x_i - \mathbf{b}_M) = 0 \Rightarrow \mathbf{b}_M = \sum_{i=1}^{T} x_i/T$$

– $p = 1$: We get the sample median as the estimator with the least absolute deviation (LAD) for the median $\boldsymbol{\beta}$. (There is no unique solution if T is even.)

Note: Unlike LS, LAD does not have an analytical solving method. Numerical optimization is not feasible. Linear programming is used. 34

M-Estimation: Asymptotic Normality

- Summary
  - $\boldsymbol{b}_M \xrightarrow{p} \boldsymbol{b}_0$
  - $\boldsymbol{b}_M \xrightarrow{a} N(\boldsymbol{b}_0, \text{Var}[\boldsymbol{b}_0])$
  - Var[$\boldsymbol{b}_M$] = $(1/T)\ \boldsymbol{H}_0^{-1}\ \boldsymbol{V}_0\ \boldsymbol{H}_0^{-1}$
  - If the model is correctly specified: -**H** = **V**.
    Then, Var[$\boldsymbol{b}$] = $\boldsymbol{V}_0$

  – **H** and **V** are evaluated at $\boldsymbol{b}_0$:
    - $\mathbf{H} = \sum_i [\partial^2 q\ (\mathbf{z}_i;\ \mathbf{b})/\partial\mathbf{b}\partial\mathbf{b}']$
    - $\mathbf{V} = \sum_i [\partial q(\mathbf{z}_i;\ \mathbf{b})/\partial\mathbf{b}][\partial q(\mathbf{z}_i;\ \mathbf{b})/\partial\mathbf{b}']$

35

M-Estimation: LAD Estimation

Example: We compute the CAPM for IBM, using LAD. We use the *quantreg* R package (default is tau=.50, the median).

```
rqfit_50 <- rq(ibm_x ~ Mkt_RF)
summary(rqfit)
> summary(capm_ibm)
Call: rq(formula = ibm_x ~ Mkt_RF)

tau: [1] 0.5

Coefficients:
            coefficients lower bd upper bd
(Intercept) -0.00667     -0.01033 -0.00269
Mkt_RF       0.87281      0.78756  0.95096
```

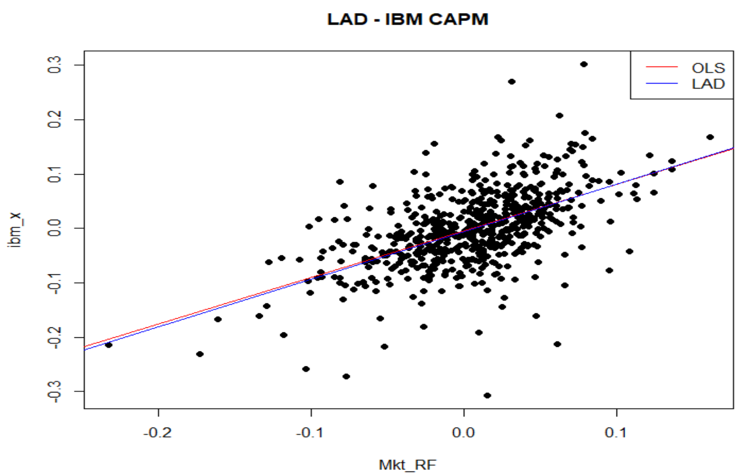
• CAPM

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.005361   0.002403  -2.231    0.026 *
Mkt_RF       0.856569   0.051438  16.653   <2e-16 ***
```

36

M-Estimation: LAD Example

Example (continuation): Below, we plot both lines:



Note: Overall, very similar results.

37

Breakdown Point: Intuition

• There are several measures of robustness of an estimator, attempting to quantify the change. One of the most commonly used is the **breakdown point**.

• Let $\boldsymbol{X}$ be a random sample and $\mathbf{T}(\boldsymbol{X})$ be an estimator. Informally, the breakdown point of the estimator is the proportion m/T of observations, which can be replaced by *bad observations* (outliers) without forcing $\mathbf{T}(\boldsymbol{X})$ to leave a bounded set –i.e., become infinity.

Example: The sample mean has a breakdown point equal to 0 (one observation can drive the sample mean, regardless of the other T-1 values). The median has a breakdown point 1/2 (it can tolerate 50% bad values) and α%-trimmed mean has a breakdown point α%.

38

Breakdown Point: Definition

• Assume a sample, $\boldsymbol{Z}$, with T observations, and let $\mathbf{T}$ be a regression estimator. That is, we apply $\mathbf{T}$ to $\boldsymbol{Z}$ we get the regression coefficients:

$$\mathbf{T}(\boldsymbol{Z}) = \mathbf{b}$$

• Imagine all possible "corrupted" samples $\boldsymbol{Z}^{\mathbf{0}}$ that replace any subset of observations, m, in the dataset with arbitrary values -*i.e.*, influential cases.

• The maximum bias that could arise from these substitutions is:

$$bias(m; \mathbf{T}, \boldsymbol{Z}) = \sup_{\mathbf{Z}'} \left\| \mathbf{T}(\boldsymbol{Z}') - \mathbf{T}(\boldsymbol{Z}) \right\|$$

• If the $bias(m; \mathbf{T}, \boldsymbol{Z})$ is infinite, the m outliers have an arbitrarily large effect on $\mathbf{T}$. In other words, the estimator *breaks down*.

39

Breakdown Point: Definition

• Then, the breakdown point for an estimator $\mathbf{T}$ for a finite sample $\boldsymbol{Z}$ is:

$$\varepsilon_n^*(\mathbf{T}, \boldsymbol{Z}) = \min\{\frac{m}{T}; bias(m, \mathbf{T}, \boldsymbol{Z}) \text{ is infinite}\}$$

• The breakdown point of an estimator is the smallest fraction of "*bad*" data (outliers or data grouped at the extreme of a tail) the estimator can tolerate without taking on values arbitrarily far from $\mathbf{T}(\boldsymbol{Z})$.

• For OLS regression one unusual case is enough to influence the coefficient estimates. Its breakdown point is then

$$\varepsilon_n^*(\mathbf{T}, \boldsymbol{Z}) = 1/T$$

• As T gets larger, $1/T$ tends towards 0, meaning that the breakdown point for OLS is 0%.

40

Robust Regression: Methods

• Robust regression methods attempt to limit the impact of unusual cases on the regression estimates:

– ***Least Absolute Values (LAV/LAD) regression*** is robust to outliers (unusual y_i values given x_i), but typically fares even worse than OLS for cases with high leverage.

- If a leverage point is very far away, the LAD line will pass through it. In other words, its breakdown point is also $1/T$.

– ***M-Estimators*** are also robust to outliers. More efficient than LAD estimators. They can have trouble handling cases with high leverage, meaning that the breakdown point is also $1/T$.

– ***Bounded influence methods*** have a much higher breakdown point (as high as 50%) because they effectively remove a large proportion of the cases. These methods can have trouble with small samples.

41

Estimating the Center of a Distribution

• In order to explain how robust regression works, we start with the simple case of robust estimation of the center of a distribution. Consider independent observations and the simple model:

$$y_i = \mu + \varepsilon_i$$

• If the underlying distribution is normal, the sample mean is the MLE.

• The mean minimizes the LS objective function:

$$q_{LS} = \boldsymbol{e}'\boldsymbol{e} = \sum_{i=1}^{T} e_i^2$$

• The derivative of the objective function with respect to e_i gives the influence function which determines the influence of observations: $\psi_{LS,i}(e) = 2 * e_i$. That is, influence is proportional to the residual e_i.

42

Estimating the Center of a Distribution

• As an alternative to the mean, we consider the median as an estimator of μ. The median minimizes the LAD objective function:

$$q_{LAD} = 1/T \ \sum_{i=1}^{T} |e_i|$$

• Taking the derivative of the objective function gives the shape of the influence function:

$$\psi_{LAD,i}(e) = 1 \quad \text{for } e_i > 0.$$
$$= 0 \quad \text{for } e_i = 0.$$
$$= -1 \quad \text{for } e_i < 0.$$

• Note that influence of e_i is bounded. The fact that the median is more resistant than the mean to outliers is a favorable characteristic.

43

Influence Function for Mean and Median

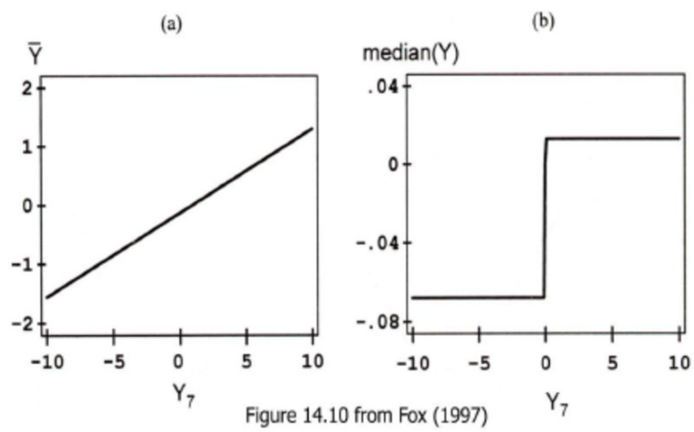


Figure 14.10 from Fox (1997)

44

M-Estimation: LAD Example with Outliers

Example: We compute the 3-factor F-F for IBM, using LAD, for the **Corrupted IBM data**. We use the *quantreg* R package (default is tau=.50, the median).

```
rqfit_50 <- rq(ibm_x ~ Mkt_RF)
summary(rqfit)
> summary(capm_ibm)
Call: rq(formula = ibm_x ~ Mkt_RF)

tau: [1] 0.5
Coefficients:
             Estimate   lower bd   upper bd
(Intercept)  -0.00549   -0.00874   -0.00296
Mkt_RF        0.95236    0.82872    1.02321
SMB          -0.20620   -0.34978   -0.07770
HML           0.02466   -0.09746    0.24427
```

Note: Again, SMB & constant are significant (5% level). HML is not.

45

M-Estimation: LAD Example with Outliers

Example (continuation): Below, we compare the LAD estimates and OLS estimates with the corrupted IBM data. Estimates and significance are affected for HML & SMB.

	LAD				OLS		
	Estimate	**lower bd**	**upper bd**		**Estimate**	**SE**	**t-value**
(Intercept)	-0.00549	**-0.0087**	**-0.00296**		-0.00522	0.003	-1.69
Mkt_RF	0.95236	**0.82872**	**1.02321**		0.96545	0.069	**13.89**
SMB	-0.2062	**-0.3498**	**-0.0777**		-0.09743	0.104	-0.94
HML	0.02466	-0.0975	0.24427		-0.17617	0.1	-1.77

46

M-Estimation: Huber Estimates

• But, the median is far less efficient, however. If y_i ~ N(μ, σ^2),

Var[mean] = σ^2/T

Var[median] = $\pi\ \sigma^2/2T$

⇒ The Var[median] is $\pi\ /2$ (≈ 1.57) times as large as Var[mean].

• A good compromise between the efficiency of LS and the robustness of LAD is the Huber (1964) objective function:

$q_{H,i}(e_i) = \frac{1}{2} * {e_i}^2$ for $|e_i| \leq k$. (k = tuning constant)

$= k|e_i| - \frac{1}{2}k^2$ for $|e_i| > k$.

with influence function:

$s_{H,i}(e_i)' = k$ for $e_i > k$.

$= e_i$ for $|e_i| \leq k$.

$= -k$ for $e_i > k$.

47

M-Estimation: Tuning constant, k

• k is called the *tuning constant*.

Note: For $k \rightarrow \infty$, the M-estimator turns into mean, for $k \rightarrow 0$, it becomes the median.

• Assuming $\sigma^2 = 1$, setting k = 1.345 produces 95% efficiency relative to the sample mean when the population is normal. It gives substantial resistance to outliers when it is not.

• In general, k is expressed as a multiple of the Y scale (the spread), S

⇒ $k = c\ S$

– We could use σ as a measure of scale, but it is more influenced by extreme observations than is the mean.

– Instead, we use the **median absolute deviation:**

MAD = median$|y_i - \hat{\mu}|$ = median$|e_i|$

48

M-Estimation: Tuning constant, k

• We use the median absolute deviation:

$$\text{MAD} = \text{median}|y_i - \hat{\mu}| = \text{median}|e_i|$$

• The median of Y serves as an initial estimate of $\hat{\mu}$, thus allowing us to define S = MAD/.6745, which ensures that S estimates σ when the population is normal –i.e., for the standard normal E[MAD] = 0.6745

• Using k = 1.345 S (1.345/.6745 is about 2 MAD) produces 95% efficiency relative to the sample mean when the population is normal and gives substantial resistance to outliers when it is not.

Note: A smaller k gives more resistance to outliers.

49

M-Estimation: Bi-weight Estimates

• Tukey's **bi-weight (bisquare) estimates** behave somewhat differently than Huber weights, but are calculated in a similar manner

• The **biweight objective function** is especially resistant to observations on the extreme tails:

$$q_{BW,i}(e_i) = \frac{k^2}{6} * \{1 - [1 - (\frac{e_i}{k})^2]^3\} \qquad \text{for } |e_i| \leq k.$$
$$= \frac{k^2}{6} \qquad \text{for } |e_i| > k.$$

with an influence function:

$$s_{BW,i}(e_i)' = \{e_i * [1 - (\frac{e_i}{k})^2]\} \qquad \text{for } |e_i| \leq k.$$
$$= 0 \qquad \text{for } |e_i| > k.$$

• For this function, k = 4.685 S (4.685/.6745 about 7 MADS) produces 95% efficiency when sampling from a normal population

50

M-Estimation and Regression

• Since regression is based on the mean, it is easy to extend the idea of M-estimation to regression. The linear model is:

$$y_i = x_i'b + \varepsilon_i$$

• The M-estimator then minimizes the objective function:

$$q = \sum_{i=1}^{T} q(y_i - x_i'b)$$

with f.o.c.'s:

$$\sum_{i=1}^{T} \psi\,(y_i - x_i'b)\, x_i' = \mathbf{0}$$

• We have a system of k equations. We replace $\psi(.)$ with the weight function, $w_i = \psi(.)/e_i$:

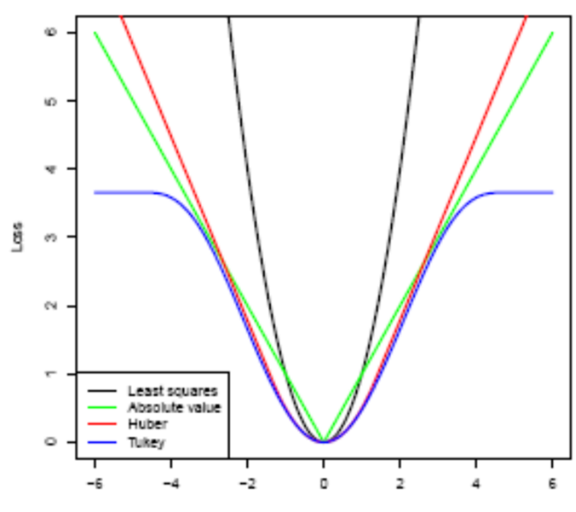
$$\sum_{i=1}^{T} w_i(y_i - x_i'b)\, x_i' = \mathbf{0}$$

Note: We assign a different weight to each i depending on the size of e_i; similar to WLS.

51

M-Estimation and Regression: Loss Functions

• Different loss functions:



52

M-Estimation and Regression: Weights

• The weight function: $w_i = w(e_i) = \psi(.)/e_i$:

Method	Objective Function	Weight Function
Least-Squares	$\rho_{LS}(e) = e^2$	$w_{LS}(e) = 1$
Huber	$\rho_H(e) = \begin{cases} \frac{1}{2}e^2 & \text{for } \lvert e\rvert \le k \\ k\lvert e\rvert - \frac{1}{2}k^2 & \text{for } \lvert e\rvert > k \end{cases}$	$w_H(e) = \begin{cases} 1 & \text{for } \lvert e\rvert \le k \\ k/\lvert e\rvert & \text{for } \lvert e\rvert > k \end{cases}$
Bisquare	$\rho_B(e) = \begin{cases} \frac{k^2}{6}\left\{1 - \left[1 - \left(\frac{e}{k}\right)^2\right]^3\right\} & \text{for } \lvert e\rvert \le k \\ k^2/6 & \text{for } \lvert e\rvert > k \end{cases}$	$w_B(e) = \begin{cases} \left[1 - \left(\frac{e}{k}\right)^2\right]^2 & \text{for } \lvert e\rvert \le k \\ 0 & \text{for } \lvert e\rvert > k \end{cases}$

53

Weight Functions for Various Estimators

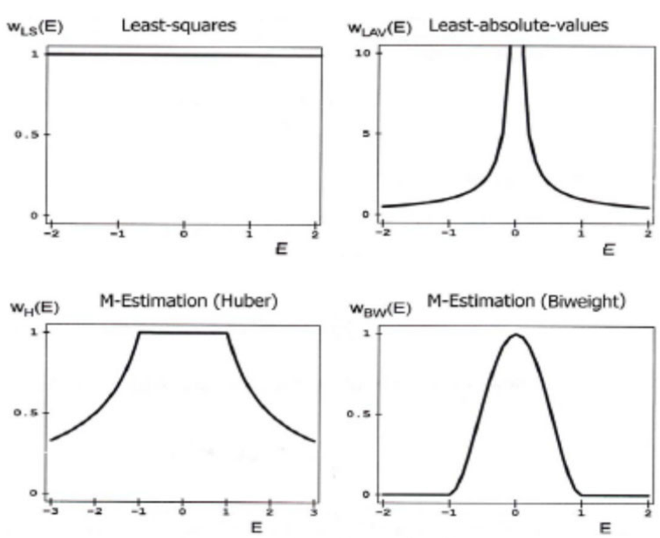


Figure 14.13 from Fox (1997)

54

M-Estimation and Regression: Algorithm

• The solution assigns a different weight to each case depending on the size of their residual, and thus minimizes the weighted sum of squares.

$$\sum_{i=1}^{T} w_i\, \varepsilon_i^2 = 0$$

• The w_i weights depend on the residuals in the model. An iterative solution (using *Iterative Re-weighted Least Squares*, IRLS) is needed.

• The solution to this problem is weighted LS:

(1) Set initial $\boldsymbol{b}^0$, say by using OLS. Get e_i^0.

(2) Estimate the scale of the residuals S^0 and the weights w_i^0.

(3) Estimate $\boldsymbol{b}^j$: $\quad j = 1, 2, ...$

$$\boldsymbol{b}^j = (\mathbf{X'WX})^{-1}\, \mathbf{X'Wy} \qquad \mathbf{W}=\text{diag}\{\, w_i^{j-1}\}$$

(4) With $\boldsymbol{b}^j$ go back to (1). Repeat steps (1)-(3) until convergence.

55

M-Estimation and Regression

• Usual weight functions: Huber and Biweight (bisquare) weights.

• M-Estimators are statistically equally efficient as OLS if the distribution is normal, while at the same time are more robust with respect to influential cases.

• However, M-estimation can still be influenced by a single very extreme X-value—*i.e.*, like OLS, it still has a breakdown point of 0

56

Bounded Influence Regression: LTS

• M-estimation can still be influenced by a single very extreme X-value—*i.e.*, like OLS, it still has a breakdown point of 0

• **Least-trimmed-squares** (**LTS**) estimators (Rousseeuw (1984)) can have a breakdown point up to 50% -i.e., half the data can be influential in the OLS sense before the LTS estimator is seriously affected.

– Least-trimmed-squares essentially proceeds with OLS after eliminating the most extreme positive or negative residuals.

• LTS orders the squared residuals from smallest to largest:

$$(e^2)_{(1)}, (e^2)_{(2)}, ..., (e^2)_{(T)}$$

• Then, LTS calculates **b** that *minimizes the sum of only the smaller half of the residuals*.

57

Bounded Influence Regression: LTS

• LTS calculates **b** that *minimizes the sum of only the smaller half of the residuals:*

$$\sum_{i=1}^{m} e_i^2$$

where $m = [T/2] + 1$; the square bracket indicates rounding down.

• By using only the 50% of the data that fits closest to the original OLS line, LTS completely ignores extreme outliers. The breakdown value for the LTS estimate is $(T - m)/T$.

• On the other hand, this method can misrepresent the trend in the data if it is characterized by clusters of extreme cases or if the data set is relatively small.

58

Bounded Influence Regression: LMS

• An alternative bounded influence method is ***Least Median Squares*** (**LMS**).

• Rather than minimize the sum of the least squares function, this model minimizes the median of the squared residuals, e_i^2.

• The breakdown value for the LTS estimate is also $(T - m)/T$.

• LMS is very robust with respect to outliers both in terms of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

• But, it performs poorly form the point of view of asymptotic efficiency. Also, relative to LMS, LTS's objective function is smoother, making the LTS estimate less jumpy -i.e., less sensitive to local effects.

Bounded Influence Regression: MM-estimator

• One application of bounded-influence estimators is to provide starting values for M-estimation.

• This procedure, along with using the bounded-influence estimate of the error variance, produces the so-called **MM-estimator**.

• The MM-estimator retains the high breakdown point of the bounded-influence estimator and shares the relatively high efficiency under normality of the traditional M-estimator.

• MM-estimators are especially attractive when paired with redescending -functions such as the bisquare, which can be sensitive to starting values.

M-Estimation: CAPM

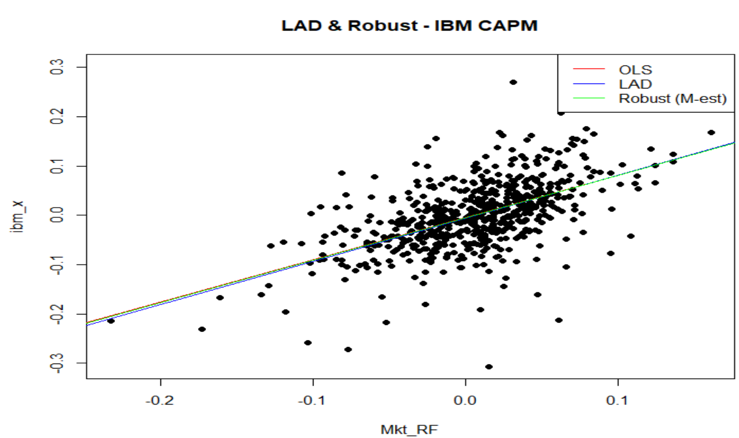
Example: We compute the CAPM for IBM using Huber's M-estimator. We use the *MASS* R package (default is M estimation).

```
library(MASS)
ibm_rob <- rlm(ibm_x ~ Mkt_RF)
> summary(ibm_rob)

Call: rlm(formula = ibm_x ~ Mkt_RF)
Residuals:
      Min         1Q     Median         3Q        Max
-0.3140923 -0.0323543 -0.0008752  0.0333639  0.2490740

Coefficients:
            Value   Std. Error t value
(Intercept) -0.0056  0.0021    -2.6126
Mkt_RF       0.8602  0.0455    18.9090

Residual standard error: 0.0487 on 609 degrees of freedom
```

61

M-Estimation: CAPM

Example (continuation): Below, we plot the 90th quantile fit, along the standard CAPM line.



Note: Again, very similar results to OLS. There is evidence of outliers affecting fit.

62

M-Estimation: F-F Model with Outliers

Example: We compute the 3-factor F-F model for the **Corrupted IBM data**, using Huber's M-estimator. We use the *MASS* R package (default is M-estimation).

```
library(MASS)
ibm_ff3_rob <- rlm(ibm_x ~ Mkt_RF + SMB+HML)
> summary(ibm__ff3_ rob)
Coefficients:
             Estimate    Std. Error   t value
(Intercept)  -0.0052     0.0021       -2.4525
Mkt_RF       0.9114      0.0479       19.0187
SMB          -0.2438     0.0717       -3.4002
HML          .0091       0.0687       0.1318
> summary(fit_ibm_ff3)
Coefficients:
             Estimate    Std. Error   t value
 (Intercept) -0.004947   0.002408     -2.054
Mkt_RF       0.885706    0.054259     16.324
SMB          -0.228137   0.081180     -2.810
HML          -0.058153   0.077791     -0.748
```

63

M-Estimation: F-F Model with Outliers

Example (continuation): Below, we compare the M-estimates and OLS estimates with the corrupted data. Again, estimates and significance are affected for HML & SMB. The M-estimates are very similar to the OLS estimates without the corrupting outliers.

	M-estimation				**OLS**		
	Estimate	**SE**	**t-value**		**Estimate**	**SE**	**t-value**
(Intercept)	-0.0052	0.0021	**-2.4525**		-0.00522	0.003	-1.69
Mkt_RF	0.9114	0.0479	**19.0187**		0.96545	0.069	**13.89**
SMB	-0.2438	0.0717	**-3.4002**		-0.09743	0.104	-0.94
HML	0.0091	0.0687	0.1318		-0.17617	0.1	-1.77

64

M-Estimation: Bi-square F-F Model (Outliers)

Example: We compute the 3-factor F-F model for the **Corrupted IBM data**, using Tukey's bi-square estimator. We use the *MASS* R package (default is M-estimation).

```
library(MASS)
ibm_ff3_out_bisq <- rlm(ibm_x ~ Mkt_RF + SMB+HML, method="MM")
> summary(ibm_ff3_out_bisq b)
Coefficients:
```

	Estimate	Std. Error	t value
(Intercept)	-0.0051	0.0021	**-2.3784**
Mkt_RF	0.9178	0.0480	**19.1386**
SMB	-0.2475	0.0717	**-3.4487**
HML	.00377	0.0688	0.5477

Note: Again, very similar results to the Huber's M-estimation.

Robust Regression: Application

De Long and Summers (1991) studied the national growth of 61 countries from 1960 to 1985 using OLS:

$$GDP_i = \beta_0 + \beta_1\ LFG_i + \beta_2\ GAP_i + \beta_3\ EQP_i + \beta_4\ NEQ_i + \varepsilon_i$$

where GDP growth per worker (GDP) and the regressors are labor force growth (LFG), relative GDP gap (GAP), equipment investment (EQP), and nonequipment investment (NEQ).

The REG Procedure
Model: MODEL1
Dependent Variable: GDP

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.01430	0.01028	-1.39	0.1697
LFG	1	-0.02981	0.19838	-0.15	0.8811
GAP	1	0.02026	0.00917	2.21	0.0313
EQP	1	0.26538	0.06529	4.06	0.0002
NEQ	1	0.06236	0.03482	1.79	0.0787

• The OLS analysis: GAP and EQP have a significant effect on GDP at the 5% level.

Robust Regression: Application

Zaman, Rousseeuw, and Orhan (2001) used robust techniques to estimate the same model (Zambia (observation #60) an outlier):

$$GDP_i = \beta_0 + \beta_1\ LFG_i + \beta_2\ GAP_i + \beta_3\ EQP_i + \beta_4\ NEQ_i + \varepsilon_i$$

The ROBUSTREG Procedure

Model Information

Data Set	MYLIB.GROWTH
Dependent Variable	GDP
Number of Covariates	4
Number of Observations	61
Name of Method	M-Estimation

Summary Statistics

Variable	Q1	Median	Q3	Mean	Standard Deviation
LFG	0.0118	0.0239	0.02805	0.02113	0.009794
GAP	0.57955	0.8015	0.88625	0.725777	0.21807
EQP	0.0265	0.0433	0.072	0.052325	0.029622
NEQ	0.09555	0.1356	0.1812	0.139856	0.056966
GDP	0.01205	0.0231	0.03095	0.022384	0.015516

Summary Statistics

Variable	MAD
LFG	0.009489
GAP	0.177764
EQP	0.032469
NEQ	0.062418
GDP	0.014974

The ROBUSTREG Procedure

Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square
Intercept	1	-0.0247	0.0097	-0.0437	-0.0058	6.53
LFG	1	0.1040	0.1867	-0.2619	0.4699	0.31
GAP	1	0.0250	0.0086	0.0080	0.0419	8.36
EQP	1	0.2968	0.0614	0.1764	0.4172	23.33
NEQ	1	0.0885	0.0328	0.0242	0.1527	7.29
Scale	1	0.0099				

Parameter Estimates

Parameter	Pr > ChiSq
Intercept	0.0106
LFG	0.5775
GAP	0.0038
EQP	<.0001
NEQ	0.0069
Scale	

• Huber M-estimates: Besides GAP and EQP, the robust analysis also show NEQ has significant effect on GDP.

67

Robust Regression: Diagnostics

• It is common to analyze the residuals for outliers (as usual) and leverage points. To check for leverage points, Rousseeuw (1984) proposes a robust version of the Mahalanobis distance by using a generalized minimum covariance determinant (MCD) method.

• Mahalanobis Distance is the square root of a standard Wald distance:

$$MD(\boldsymbol{x}_i) = \text{sqrt}\{(\boldsymbol{x}_i - \overline{\boldsymbol{x}})'\ \bar{C}(\boldsymbol{X})\ (\boldsymbol{x}_i - \overline{\boldsymbol{x}})\}$$

where $\bar{x}$ is the mean and $\bar{C}(\boldsymbol{X})$ is the variance (scale or scatter) of $\boldsymbol{X}$.

• Rousseeuw's Robust Distance is given by

$$RD(\boldsymbol{x}_i) = \text{sqrt}\{(\boldsymbol{x}_i - T(\boldsymbol{x}))'\ \bar{C}(\boldsymbol{X})\ (\boldsymbol{x}_i - T(\boldsymbol{x}))\}$$

where $T(\boldsymbol{x})$ & $C(\boldsymbol{x})$ are the robust multivariate location & scale, respectively, obtained by MCD.

68

Robust Regression: Diagnostics

• MD and RD are compared with thresholds to determine if an observation is an outlier.

• Thresholds tend to be data-specific, but, it is common to use thresholds based on Confidence intervals using the Chi-square distribution, with degrees of freedom are given by the number of parameters/variables in the model.

• Outlier detection can be also be done by looking at the standardized robust residuals.

• **Mass significance** issues appear in this context (we check every observation!), thus, many authors suggest using very small *p-values* (0.005 or 0.001). See Hair et all (2010) or Tabachnik and Fidell (2013).

Robust Regression: LTS - Application

Analysis of robust residuals. Lots of leverage observations, but only one outlier (Zambia, #60).

The ROBUSTREG Procedure

Diagnostics

Obs	Mahalanobis Distance	Robust MCD Distance	Leverage	Robust Residual	Outlier
1	2.6083	4.0639	*	-0.9424	
5	3.4351	6.7391	*	1.4200	
8	3.1876	4.6843	*	-0.1972	
9	3.6752	5.0599	*	-1.8784	
17	2.6024	3.8186	*	-1.7971	
23	2.1225	3.8238	*	1.7161	
27	2.6461	5.0336	*	0.0909	
31	2.9179	4.7140	*	0.0216	
53	2.2600	4.3193	*	-1.8082	
57	3.8701	5.4874	*	0.1448	
58	2.5953	3.9671	*	-0.0978	
59	2.9239	4.1663	*	0.3573	
60	1.8562	2.7135		-4.9798	*
61	1.9634	3.9128	*	-2.5959	

Diagnostics Profile

Name	Percentage	Cutoff
Outlier	0.0164	3.0000
Leverage	0.2131	3.3382

Robust Regression: LTS - Application

The analysis of robust residuals revealed Zambia (#60) as an outlier. Potentially, this can create problems for M-estimators. LTS estimation is has a better breakdown point.

The ROBUSTREG Procedure

LTS Profile

Total Number of Observations	61
Number of Squares Minimized	33
Number of Coefficients	5
Highest Possible Breakdown Value	0.4590

LTS Parameter Estimates

Parameter	DF	Estimate
Intercept	1	-0.0249
LFG	1	0.1123
GAP	1	0.0214
EQP	1	0.2669
NEQ	1	0.1110
Scale		0.0076
WScale		0.0109

The ROBUSTREG Procedure

Diagnostics

Obs	Mahalanobis Distance	Robust MCD Distance	Leverage	Robust Residual	Outlier
1	2.6083	4.0639	*	-1.0715	
5	3.4351	6.7391	*	1.6574	
8	3.1876	4.6843	*	-0.2324	
9	3.6752	5.0599	*	-2.0896	
17	2.6024	3.8186	*	-1.6367	
23	2.1225	3.8238	*	1.7570	
27	2.6461	5.0336	*	0.2334	
31	2.9179	4.7140	*	0.0971	
53	2.2600	4.3193	*	-1.2978	
57	3.8701	5.4874	*	0.0605	
58	2.5953	3.9671	*	-0.0857	
59	2.9239	4.1663	*	0.4113	
60	1.8562	2.7135		-4.4984	*
61	1.9634	3.9128	*	-2.1201	

Diagnostics Profile

Name	Percentage	Cutoff
Outlier	0.0164	3.0000
Leverage	0.2131	3.3382

Rsquare for LTS-estimation

Rsquare	0.7417678684

Robust Regression: LTS - Application

After removing the outlier (Zambia), we re-estimate model:

The ROBUSTREG Procedure

Parameter Estimates for Final Weighted LS

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square
Intercept	1	-0.0222	0.0093	-0.0403	-0.0041	5.75
LFG	1	0.0446	0.1755	-0.2995	0.3886	0.06
GAP	1	0.0245	0.0081	0.0085	0.0404	9.05
EQP	1	0.2824	0.0576	0.1695	0.3953	24.03
NEQ	1	0.0849	0.0311	0.0239	0.1460	7.43
Scale		0.0115				

Parameter Estimates for Final Weighted LS

Parameter	Pr > ChiSq
Intercept	0.0165
LFG	0.7995
GAP	0.0026
EQP	<.0001
NEQ	0.0064
Scale	

Figure 7. Final Weighted LS estimates

72

Robust Regression: Remarks

• Separated points can have a strong *influence* on statistical models

– Unusual cases can substantially influence the fit of the OLS model. Cases that are both *outliers* and *high leverage* exert *influence* on both the slopes and intercept of the model

– Outliers may also indicate that our model fails to capture important characteristics of the data

• Efforts should be made to remedy the problem of unusual cases before proceeding to robust regression

• If robust regression is used, careful attention must be paid to the model—different procedures can give completely different answers.

Robust Regression: Remarks

• No one robust regression technique is best for all data

• There are some considerations, but even these do not hold up all the time:

– LAD regression should generally be avoided because it is less efficient than other techniques and often not very resistant

– Bounded influence regression models, which can have a breaking point as high as 50%, often work very well with large datasets. But, they tend to perform poorly with small datasets.

• M-Estimation is typically better for small datasets, but its standard errors are not reliable for small samples. This can be overcome by using bootstrapping to obtain new estimates of the standard errors.

Quantile Regression

• Mosteller and Tukey (1977):

"What the regression curve does is a grand summary for the the averages of the distributions corresponding to the set of x's. We could go further and compute several different regression curves corresponding to the various percentage points of the distribution and thus get a more complete picture."

• One might be interested in behavior of say, lower tail of the conditional distribution rather than in its mean.

• For example, how does a 1% increase in market returns affect the returns of small size firms?

75

Quantiles: Characterizing a Distribution

• We are used to assume a distribution and describe it through its moments: mean, variance, skewness, etc. Some distributions are characterized by few parameters. For example, the normal is completely described by the mean and the variance.

• A different approach. Use quantiles instead. For example:

– Median

– Interquartile Range

– Interdecile Range

– Symmetry = $(\zeta_{.75} - \zeta_{.50})/(\zeta_{.50} - \zeta_{.25})$

– Tail Weight = $(\zeta_{.90} - \zeta_{.10})/(\zeta_{.75} - \zeta_{.25})$

Quantiles

Definition: We say that a firm is in the θ^{th} quantile if it is bigger than the proportion θ, of the reference group of firms, and smaller than the proportion $(1-\theta)$.

• The θ^{th} sample quantile is simply $y_{(k)}$, where k is the smallest integer such that $\frac{K}{T} < \theta$. (Note the relation between rank and quantile.)

77

Quantiles: Definition

Definition:

(1) Discrete RV. Given $\theta \in [0, 1]$. A θ^{th} quantile of a discrete RV Z is any number ζ_θ such that $P(Z < \zeta_\theta) \leq \theta \leq P(Z \geq \zeta_\theta)$.

Example: Suppose $Z = \{3, 4, 7, 9, 9, 11, 17, 21\}$ and $\theta = 0.5$ then $P(Z < 9) = 3/8 \leq 1/2 \leq P(Z \geq 9) = 5/8$.

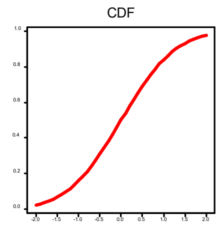
(2) Continuous RV. Let Z be a continuous r.v. with cdf F(.), then $P(Z < z) = P(Z \leq z) = F(z)$ for every z in the support and a θth quantile is any number ζ_θ such that $F(\zeta_\theta) = \theta$.

• If F is continuous and strictly increasing then the inverse exists and $\zeta_\theta = F^{-1}(\theta)$.

Quantiles: CDF and Quantile Function

- **Cumulative Distribution Function**

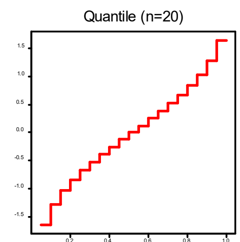
  $F(z) = P(Z \leq z)$



- **Quantile Function**

  $Q(\theta) = \min(z: F(z) \leq \theta)$

  $\Rightarrow$ Discrete step function



Quantiles

• It can be shown that quantile (θ) is the solution to

$$\min_{\zeta} \frac{1}{T}\{\sum_{y_i \geq \zeta} \theta\, |y_i - \zeta| + \sum_{y_i < \zeta} (1-\theta)\, |y_i - \zeta|\}$$

• If $\theta = 1/2$, then this becomes $\min_{\zeta} \frac{1}{T}\sum_{i=1}^{N} |y_i - \zeta|$, which yields a f.o.c.:

$$0 = \frac{1}{T}\sum_{i=1}^{N} \text{sgn}(y_i - \zeta)$$

where sng ("*signum*") function: $\text{sgn}(u) = 1 - 2 * \text{I}[u < 0]$, (defined to be right-continuous).

$\Rightarrow$ the sample median, $\zeta_{\theta=.50}$, solves this problem (easier to visualize with expectations).

80

Quantile Regression

• Basset and Koenker (1978, JASA) suggest simply replacing the ζ in the definition of the quantile estimator

$$\min_{\zeta} \frac{1}{T}\{\textstyle\sum_{y_i \geq \zeta} \theta\,|y_i - \zeta| + \sum_{y_i < \zeta} (1-\theta)\,|y_i - \zeta|\}$$

with $\boldsymbol{x}_i'\boldsymbol{\beta}$ to get the **quantile regression**:

$$\min_{\boldsymbol{\beta}} \{\textstyle\sum_{y_i \geq \boldsymbol{x}_i'\boldsymbol{\beta}} \theta\,|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\,| + \sum_{y_i < \boldsymbol{x}_i'\boldsymbol{\beta}} (1-\theta)\,|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\,|\}$$

or

$$\min_{\boldsymbol{\beta}} \{\textstyle\sum_{y_i \geq \boldsymbol{x}_i'\boldsymbol{\beta}} \theta\,|\varepsilon_i| + \sum_{y_i < \boldsymbol{x}_i'\boldsymbol{\beta}} (1-\theta)\,|\varepsilon_i|\,\}$$

• If $\theta = 1/2$, then this becomes LAD estimation. We have a symmetric weighting of observations with positive and negative residuals. But, if $\theta \neq 1/2$, the weighting is asymmetric.

81

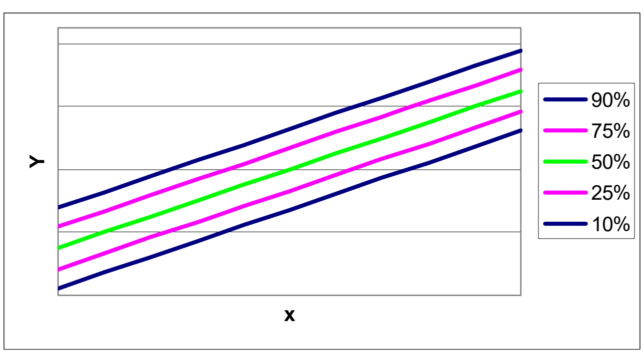
Quantile Regression

• We define a family of regressions:

$$\zeta_\theta = \mathrm{Q}(y_i \,|\, \boldsymbol{x}_i, \theta) = \boldsymbol{X}'\beta_\theta, \qquad \theta \in [0, 1]$$

- Median regression is obtained by setting θ = .50:

$$\zeta_{\theta=.50} = \mathrm{Q}(y_i \,|\, \boldsymbol{x}_i, .50) = \boldsymbol{X}'\beta_{\theta=.50}$$



82

Quantile Regression

Note: Median regression estimated by LAD. It estimates the same parameters as OLS if symmetric conditional distribution.

• We assume *correct specification* of the quantile, $Q(y_i \mid \boldsymbol{x}_i, \theta) = \boldsymbol{X}'\beta_\theta$. That is, $\boldsymbol{X}'\beta$ is a particular linear combination of the independent variables such that

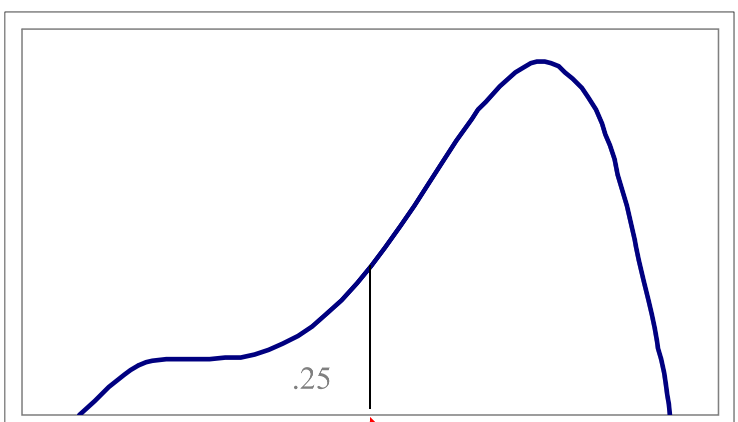
$$\theta = P(Y \leq \zeta_\theta(X)|X) = P(Y \leq \boldsymbol{X}'\beta) = F(\zeta_\theta(X)|X)$$

Q: Why use quantile (median) regression?

- Semiparametric
- Robust to some extensions (heteroscedasticity?)
- Complete characterization of conditional distribution.

83

Quantile Regression



$\zeta_{\theta=.25}(x) = \boldsymbol{X}'\beta_{\theta=.25}$

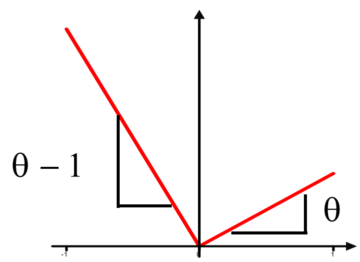
Quantile Regression: Loss Function

• Different from LS, now we minimize an asymmetric absolute loss function, given by

$$\min_{\boldsymbol{\beta}} \rho_\theta(y_i, \boldsymbol{x}_i'\boldsymbol{\beta}) = \min_{\boldsymbol{\beta}} \{\textstyle\sum_{y_i \geq \boldsymbol{x}_i'\boldsymbol{\beta}} \theta\,|\varepsilon_i| + \sum_{y_i < \boldsymbol{x}_i'\boldsymbol{\beta}} (1-\theta)\,|\varepsilon_i|\}$$

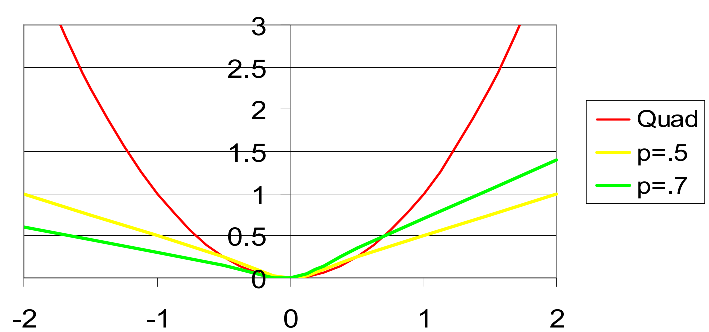
where $\varepsilon_i = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$, for some θ.

• We call ρ_θ the **tilted absolute value function**. It is convex. The local minimum is a global one, which assures uniqueness (and identification).



Quantile Regression: Loss Function

Absolute Loss vs. Quadratic Loss over errors



A quadratic loss penalizes large errors very heavily. When p=.5 our best predictor is the median; it does not give as much weight to outliers. When p=.7 the loss is asymmetric; large positive errors are more heavily penalized then negative errors.

Quantile Regression: Estimation

• Optimization problem:

$$\min_{\boldsymbol{\beta}} \{\textstyle\sum_{\varepsilon_i \geq 0} \theta\, |\varepsilon_i| + \sum_{\varepsilon_i < 0} (1-\theta)|\varepsilon_i| = \sum_{i=1}^{T} (\theta - I[y_i < 0])\varepsilon_i\}$$

where $\varepsilon_i = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$, for some θ.

• <u>Simple intuition</u>: number of negative residuals $\leq T\,\theta \leq$ number of negative residuals + number of zero residuals.

• The loss function is piecewise linear ⇒ A linear programming problem. <u>Trick</u>: replace absolute values by positivity constraints. Thus,

$$\min_{\beta} \{\sum_{t=1}^{T} \theta\varepsilon_t^{\,+} + (1-\theta)\varepsilon_t^{\,-} = \theta\iota'\varepsilon^{+} + (1-\theta)\iota'\varepsilon^{-}\}$$

$$s.t. \quad y = X\beta + \varepsilon^{+} - \varepsilon^{-} \qquad (\varepsilon_t^{\,-} \leq y_t - X_t\beta \leq \varepsilon_t^{\,+})$$

$$\varepsilon_t^{\,+} \geq 0, \quad \varepsilon_t^{\,-} \geq 0$$

87

Quantile Regression: Estimation

• The usual software packages will use the Barrodale and Roberts (1974) simplex algorithm or a Frisch-Newton (FN) algorithm.

• For large data sets, the FN method is used. It combines a log-barrier Lagrangian (Frisch part) with steepest descent steps (Newton part). For very large data sets, FN algorithm is combined with a preprocessing step, which makes the computations faster.

• Solution at vertex of feasible region. The solution need not be unique (along the edge). The fitted line will go through k data points.

• Well known program in R, written by Koenker and described in Koenker's Vignetee article (2005).

88

M-Estimation: LAD Estimation

Example: We compute the CAPM for IBM for the 90th quantile, using LAD. We use the *quantreg* R package (default is tau=.50, LAD).

```
rqfit_90<- rq(ibm_x ~ Mkt_RF, tau=.90)
summary(rqfit_90)
> summary(c rqfit_90)
Call: rq(formula = ibm_x ~ Mkt_RF)

tau: [1] 0.9

Coefficients:
            coefficients lower bd upper bd
(Intercept) 0.06707      0.05898  0.07734
Mkt_RF      0.90736      0.64632  1.06355
```

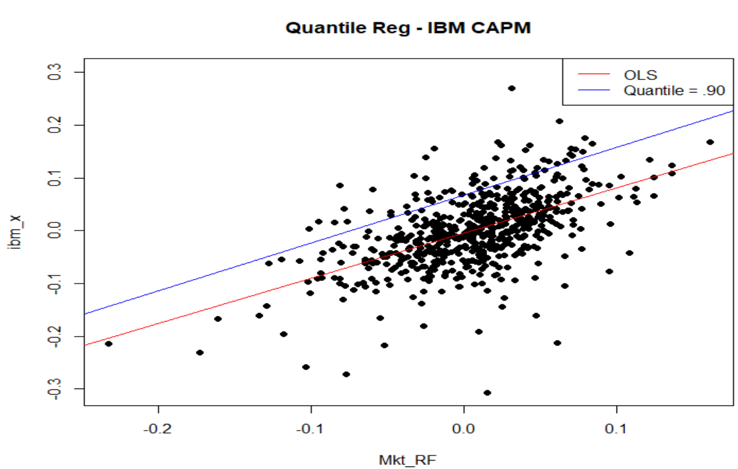
• **CAPM**

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.005361  0.002403  -2.231   0.026 *
Mkt_RF       0.856569  0.051438  16.653   <2e-16 ***
```

89

M-Estimation: LAD Estimation

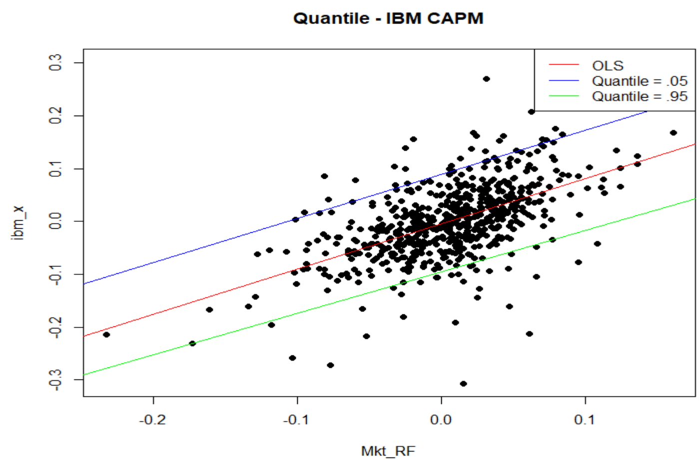
Example (continuation): Below, we plot the 90th quantile fit, along the standard CAPM line.



90

M-Estimation: LAD Estimation

Example (continuation): Below, we plot the 95th and 5% quantile fits, along the standard CAPM line.



91

Quantile Regression: Optimality

• Proposition

Under the asymmetric absolute loss function ρ_θ a best predictor of Y given X = x is the θ^{th} conditional quantile, ζ_θ.

Example: Let θ = .5. Then, the best predictor is the median fitted value.

• That is, under asymmetric absolute loss, the quantile regression estimator is more efficient than OLS.

• We offer this without proof. The proof would be similar in construction to the Gauss-Markov Theorem, which states that the conditional mean is best linear unbiased.

Properties of the Estimator

• Consistency

Consistency of $\hat{\beta}_\theta$ is easy. The minimand $S_n(.)$ is continuous in β with probability 1. In fact, $S_n(.)$ is convex in β; then, consistency follows if S_n can be shown to converge *pointwise* to a function that is uniquely minimized at the true value β_0.

• To prove consistency, we impose conditions on the model:

1. The data $(x_i; y_i)$ are *i.i.d.* across i.
2. The regressors have bounded second moment.
3. $\varepsilon_i \,|\, x_i$ is continuously distributed; with conditional density $f_\varepsilon(\varepsilon_i|x_i)$ satisfying the conditional quantile restriction.
4. The regressors and error density satisfy a .local identification condition: C E[$f_\varepsilon(0)\ \boldsymbol{x}\boldsymbol{x}'$] is a pd matrix.

Properties of the Estimator

• Asymptotic Normality (under *i.i.d* assumption)

The lack of continuously differentiable $S_n(\beta)$ complicates the usual derivation of asymptotic normality (through Taylor's expansion).

• But, an approximate f.o.c. can be used -through sgn(.). Additional conditions (*stochastic equicontinuity*) need to be established before using the Lindeberg-Levy CLT, which establishes:

$$\sqrt{T}\left(\hat{\beta}_\theta - \beta_\theta\right) \xrightarrow{L} N\left(0, \Lambda_\theta\right)$$

where

$$\Lambda_\theta = \theta(1-\theta)\left(E\left[f_\varepsilon(0 \mid x_i) x_i x_i'\right]\right)^{-1} E\left[x_i x_i'\right]\left(E\left[f_\varepsilon(0 \mid x_i) x_i x_i'\right]\right)^{-1}$$

• We have a sandwich estimator. The variance matrix depends on the unknown $f_\varepsilon(.|x_i)$ and the x_i, at which the covariance is being evaluated.

Properties of the Estimator

• We need to estimate E[$f_\varepsilon(0)\ \boldsymbol{x}\boldsymbol{x}'$], complicated without knowing $f_\varepsilon(.\,|x_i)$! It can be done through non-parametric kernel estimation.

• When the error is independent of x –i.e., $f_\varepsilon(\varepsilon_i|x_i) = f_\varepsilon(\varepsilon_i)$– then the coefficient covariance reduces to

$$\Lambda_\theta = \frac{\theta\,(1-\theta)}{f_\varepsilon"(\varepsilon_i)}\,(\hat{E}[\boldsymbol{x}\boldsymbol{x}'])^{-1}$$

where

$$\hat{E}[\boldsymbol{x}\boldsymbol{x}'] = \frac{1}{T}\sum_{i=1} \boldsymbol{x}_i\boldsymbol{x}_i{}'$$

• The variance is related to a Bernoulli variance [$\theta\ (1-\theta)$] –divided by the square density of y at the quantile, analogous to a sample size.

Properties of the Estimator

• The previous results can be extended to multivariate cases –i.e., joint estimates of several quantiles. We obtain convergence to a multivariate normal distribution.

• In general, the quantile regression estimator is more efficient than OLS. But, efficiency requires knowledge of the true error's pdf.

• Robust to outliers. As long as the sign of the residual does not change, any y_i can be arbitrarily changed without shifting the conditional quantile line.

• The regression quantiles are correlated.

Partial Effects and Prediction

• The marginal change in the Θ*th* conditional quantile due to a marginal change in the $\boldsymbol{j}$-th element of $\boldsymbol{x}$:

$$\frac{\partial\, Q_\theta(y_i|x_i)}{\partial x_{i,j}}$$

• Under linearity, the effect will be β_j. But, if non-linearities are included, the partial effect will be a function of $\boldsymbol{x}$.

Note: There is no guarantee that the $\boldsymbol{i}$-*th* observation will remain in the same quantile after $x_{i,j}$ changes.

• Using $\hat{\beta}_\theta$ and X values, predicted vales of $\hat{\boldsymbol{y}}_\theta$ can be computed. Suppose we have $\boldsymbol{X} = \boldsymbol{x_0}'$, the predicted 90*th* quantile is $\boldsymbol{x}_0'\, \hat{\beta}_\theta$.

Hypothesis Testing: Standard Errors

• Given asymptotic normality, one can construct asymptotic t-statistics for the coefficients. But which standard errors should be used?

• We can use the asymptotic estimator, but in non-*i.i.d.* situations is complicated. Inversion of a rank test --Koenker (1994, 1996)-- can be used to construct C.I.'s in a non-*i.i.d.* error context.

• Bootstrapping works well. Parzen, Wei, and Ying (1994) have suggested that rather than bootstrapping (y_i, $\boldsymbol{x}_i$) pairs, instead bootstrap the quantile regression gradient condition. It produces a pivotal approach.

Hypothesis Testing

• Alternatively, confidence regions for the quantile regression parameters can be computed from the empirical distribution of the sample of bootstrapped $b_j(\theta)$'s, the so-called percentile method.

• These procedures can be extended to deal with the joint distribution of several quantile regression estimators $\{\boldsymbol{b}_j(\theta_k), k = 1, 2, \ldots, K\}$. This would be needed to test equality of slope parameters across quantiles.

• The error term may be heteroscedastic. Efficiency issue. There are many tests for heteroscedasticity in this context.

• A test for symmetry, resembling a Wald Test, can be constructed which could not be done under Least Squares estimation.

Crossings

• Since quantile regressions are typically estimated individually, the quantile curves can cross, leading to strange (an invalid) results.

• Crossings problems increase with the number of regressors..

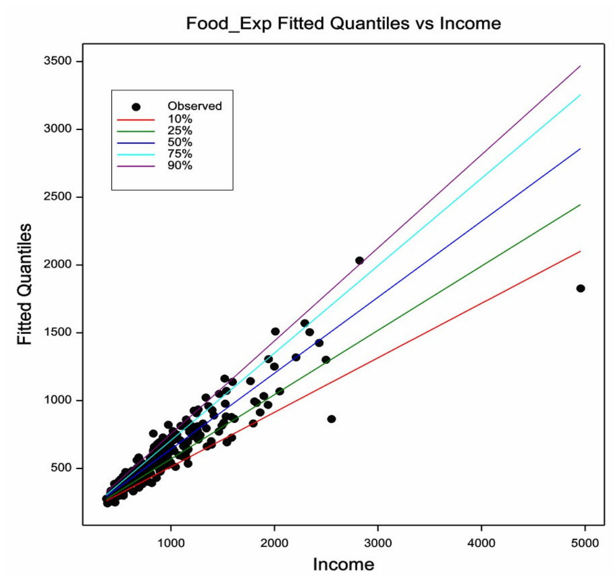
• Simultaenous estimation, with constraints are one solution.

• Individual specification of each quantile also works. For example:

$$y = X\beta_0 + \varepsilon^0, \qquad P[\varepsilon^0 < 0 \mid X] = \theta_0 \quad (\text{say}, \theta_0 = .5)$$
$$y = X\beta_0 - \exp(X\beta_1) + \varepsilon^1, \qquad P[\varepsilon^1 < 0 \mid X] = \theta_1 \quad (\text{say}, \theta_0 = .25)$$
$$y = X\beta_0 + \exp(X\beta_2) + \varepsilon^2, \qquad P[\varepsilon^2 < 0 \mid X] = \theta_2 \quad (\text{say}, \theta_0 = .75)$$

Note: Since exp(.) is positive, the quantiles by design never cross.

Quantile Linear Regression: Application 1

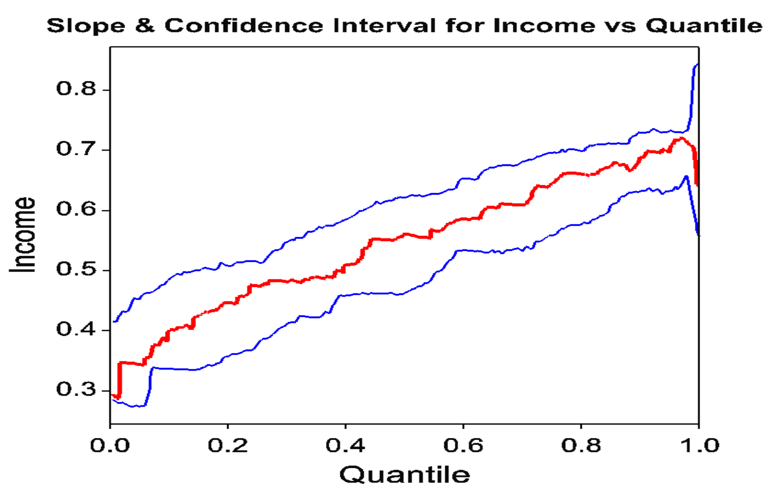


Food Expenditure vs Income

Engel 1857 survey of 235 Belgian households

Q: Change of slope at different quantiles?

Quantile Linear Regression: Application 1



Note: Variation of Parameter with Quantiles.

Quantile Linear Regression: Application 2

```
----------------------------------------------------------------------------
Quantile Regression Model. Quantile =       .250000
Linear Programming estimation method
LHS=HHNINC    Mean                  =        .44583
              Standard deviation    =        .21650
              Number of observs.    =          3377
              Minimum               =        .04000
              t= .25000 quantile    =        .30000
              Maximum               =       3.00000
Model size    Parameters            =             5
              Degrees of freedom    =          3372
Residuals     Sum of squares        =     193.75951
              Standard error of e   =        .20226
Fit           R-squared             =        .12721
              PseudoR2=1-F(0)/F(b)  =        .11046
Not using OLS or no constant. Rsquared may be <= 0
Functions F= Sum r(t)[y(i)-x(i)b]   =     164.31749
          F0=Sum r(t)[y(i)-Qy(t)]   =     184.72281
              r(t)[u]=t*u-u*[u<0].t=        .250000
--------+-------------------------------------------------------------------
        |                 Standard             Prob.       95% Confidence
  HHNINC|  Coefficient      Error        z    |z|>Z*          Interval
--------+-------------------------------------------------------------------
Constant|    -.07580***     .01839    -4.12   .0000     -.11185    -.03975
     AGE|    -.00036**      .00016    -2.25   .0244     -.00068    -.00005
    EDUC|     .02393***     .00137    17.51   .0000      .02125     .02661      α = .25
 MARRIED|     .11459***     .00547    20.96   .0000      .10388     .12531
    HSAT|     .00773***     .00122     6.31   .0000      .00533     .01013
--------+-------------------------------------------------------------------
Constant|    -.01504        .03479     -.43   .6656     -.08323     .05315
     AGE|    -.00035        .00039     -.90   .3669     -.00112     .00041
    EDUC|     .02707***     .00167    16.19   .0000      .02379     .03035      α = .50
 MARRIED|     .11361***     .01115    10.19   .0000      .09175     .13547
    HSAT|     .00777***     .00195     3.99   .0001      .00396     .01158
--------+-------------------------------------------------------------------
Constant|     .03738        .03246     1.15   .2495     -.02624     .10099
     AGE|     .00020        .00039      .51   .6100     -.00057     .00097
    EDUC|     .03240***     .00237    13.68   .0000      .02776     .03704      α = .75
 MARRIED|     .08042***     .01112     7.23   .0000      .05863     .10222
    HSAT|     .00693***     .00231     3.00   .0027      .00240     .01145
--------+-------------------------------------------------------------------
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
----------------------------------------------------------------------------
```

Quantile Linear Regression: SAS - Application 3

```
proc quantreg data=ab ;
model y1 = xm SMB HML /quantile=0.25 0.5 0.75
run;
```

```
                The QUANTREG Procedure
           Quantile and Objective Function
Quantile                        0.25
Parameter DF Estimate     95% Confidence  Limits
Intercept  1  -1.6310  -1.7793  -1.5164
xm         1   0.9855   0.9477   1.0069
SMB        1   1.2018   1.1505   1.3219
HML        1   0.5071   0.4250   0.5615

Quantile                        0.75
Parameter DF Estimate     95% Confidence  Limits
Intercept  1   0.9056   0.6957   1.1344
xm         1   0.9919   0.9626   1.0535
SMB        1   1.4267   1.3340   1.5025
HML        1   0.5435   0.4593   0.6213
```

Heteroscedasticity

- Model: $y_i = \boldsymbol{x}_i{}'\beta + \varepsilon_i$, with *i.i.d.* errors.
  - The quantiles are a vertical shift of one another.

- Model: $y_i = \boldsymbol{x}_i{}'\beta + \sigma(\boldsymbol{x}_i)\,\varepsilon_i$, errors are now heteroscedastic.
  - The quantiles now exhibit a location shift as well as a scale shift.

- Khmaladze-Koenker Test Statistic

Quantile Regression: Bibliography

• Buchinsky, M. (1994), "Changes in the u.s. wage structure 1963-1987: Application of quantile regression," Econometrica, 62, 405-458.

• Koenker and Hullock (2001), "Quantile Regression," Journal of Economic Perspectives, Vol. 15, Pps. 143-156.

• Koenker (2005), **Quantile Regression,** Cambridge University Press.

Quantile Regression

- S+ Programs - Lib.stat.cmu.edu/s
- www.econ.uiuc.edu/~roger
- http://Lib.stat.cmu.edu/R/CRAN
- SAS
- Limdep