

# Regression III

## Lecture 10: Robust Regression

Dave Armstrong

University of Wisconsin – Milwaukee  
Department of Political Science

e: armstrod@uwm.edu  
w: www.quantoid.net/ICPSR.php

1 / 64

### Goals of the lecture

- Introducing the idea of “robustness”, in particular distributional robustness
  - Discuss various measures of the robustness of an estimator, such as the breakdown point and influence function
- Discuss *robust* and *resistant* regression methods that can be used when there are unusual observations or skewed distributions
  - Particular emphasis will be placed on M-estimation and some extensions (in particular MM-estimation)
  - Revisit diagnostics for outliers, showing how robust regression can be used as a diagnostic tool
  - Extend M-estimation to GLMs
- As usual, we’ll see how to do these things in  $\mathbf{R}$ .

2 / 64

### Defining “Robust”

- Statistical inferences are based both on observations and on prior assumptions about the underlying distributions and relationships between variables
  - Although the assumptions are never *exactly* true, some statistical models are more sensitive to small deviations from these assumptions than others
- Following Huber (1981) *robustness* signifies insensitivity to deviations from the assumptions the model imposes
  - A model is robust then, if it is (1) reasonably efficient and unbiased, (2) small deviations from model assumptions will not substantially impair the performance of the model and (3) somewhat larger deviations will not invalidate the model completely
- Robust regression is concerned with distributional robustness and outlier resistance
  - Although conceptually distinct, these are for practical purposes synonymous

3 / 64

### Breakdown Point (1)

- Assume a sample,  $\mathbf{Z}$ , with  $n$  observations, and let  $T$  be a regression estimator.
- Applying  $T$  to  $\mathbf{Z}$  gives us the vector of regression coefficients:

$$T(\mathbf{Z}) = \hat{\beta}$$

- Imagine all possible “corrupted” samples  $\mathbf{Z}'$  that replace any observations  $m$  of the dataset with arbitrary values (i.e., influential cases)
- The maximum bias that could arise from these substitutions is:

$$\text{effect}(m; T, \mathbf{Z}) = \sup_{\mathbf{Z}'} \| T(\mathbf{Z}') - T(\mathbf{Z}) \parallel$$

where the supremum is over all possible  $\mathbf{Z}'$

4 / 64

## Breakdown Point (2)

- if the bias( $m; T, Z$ ) is infinite, the  $m$  outliers have an arbitrarily large effect on  $T$
- The breakdown point for an estimator  $T$  for a finite sample  $Z$  is:

$$BDP(T, Z) = \min \left\{ \frac{m}{n}; \text{effect}(m; T, Z) \text{ is infinite} \right\}$$

- In other words, the breakdown point is the smallest fraction of “bad” data (outliers or data grouped in the extreme tail of the distribution) the estimator can tolerate without taking on values arbitrarily far from  $T(Z)$ 
  - For OLS regression, one unusual case is enough to influence the coefficient estimates. Its breakdown point then is:

$$BDP = \frac{1}{n}$$

- As  $n$  gets larger,  $\frac{1}{n}$  tends toward 0, meaning that the breakdown point for OLS is 0%

5 / 64

## Influence Function (or Influence Curve)

- While the breakdown point measures global robustness, the influence function (IF) measures local robustness
- More specifically, the IF measures the impact of a single observation  $Y$  that contaminates the theoretically assumed distribution  $F$  on an estimator  $T$

$$IF(Y, F, T) = \lim_{\lambda \rightarrow 0} \frac{T\{(1 - \lambda) + \lambda\delta_Y\} - T\{F\}}{\lambda}$$

where  $\delta_Y$  is the probability distribution that puts its mass at the point  $Y$  (i.e.,  $\delta_Y=1$  at  $Y$  and 0 otherwise), and  $\lambda$  is the proportion of contamination at  $Y$

- Simply put, the IF indicates the bias caused by adding outliers at the point  $Y$ , standardized by the proportion of contamination
- The IF can be calculated from the first derivative of the estimator

6 / 64

## Influential Cases and OLS

- OLS is not robust to outliers. It can produce misleading results if unusual cases go undetected - even a single case can have a significant impact on the fit of the regression surface
- Moreover, the efficiency of the OLS regression can be hindered by heavy-tailed distributions and outliers
- Diagnostics should be used to detect heavy-tails or influential cases, but once they are found we are left with a decision as to what to do
  - Investigate whether the deviations are a symptom of model failure that can be repaired by deleting cases, transformations or adding more terms to the model
  - In cases when the unusual data cannot be remedied, robust regression can provide an alternative to OLS

7 / 64

## Estimating Location

- In order to explain how robust regression works, it is helpful to start with the simple case of robust estimation of the center of a distribution
- Consider independent observations and the simple model:

$$Y_i = \mu + e_i$$

- If the underlying distribution is normal, the sample mean is the maximally efficient estimator of  $\mu$ , producing the fitted model:

$$Y_i = \bar{Y} + E_i$$

- The mean minimizes the least squares objective function:

$$\sum_{i=1}^n \rho_{LS}(E_i) = \sum_{i=1}^n \rho_{LS}(Y_i - \hat{\mu}) \equiv \sum_{i=1}^n (Y_i - \hat{\mu})^2$$

8 / 64

### Estimating Location (2)

- The derivative of the objective function with respect to  $E$  gives the influence function which determines the influence of observations:  $\Psi_{LS}(E) \equiv \rho'_{LS}(E) = 2E$ . In other words, influence is proportional to the residual  $E$
- Compared to the median, the mean is sensitive to extreme cases. As an alternative, then, we now consider the median as an estimator of  $\mu$
- The median minimizes the least-absolute-values (LAV) objective function:

$$\sum_{i=1}^n \rho_{LAV}(E_i) = \sum_{i=1}^n \rho_{LAV}(Y_i - \hat{\mu}) \equiv \sum_{i=1}^n |Y_i - \hat{\mu}|$$

- This method is more resistant to outliers because, in contrast to the mean, the influence of an unusual observation on the median is *bounded*

9 / 64

### Estimating Location (3)

- Again, taking the derivative of the function gives the shape of the influence function

$$\Psi_{LAV}(E) \equiv \rho'_{LAV}(E) = \begin{cases} 1 & \text{for } E > 0 \\ 0 & \text{for } E = 0 \\ -1 & \text{for } E < 0 \end{cases}$$

- the fact that the median is more resistant than the mean to outliers is a favorable characteristic
- It is far less efficient, however. If  $Y \sim \mathcal{N}(\mu, \sigma^2)$ , the sampling variance is  $\frac{\sigma^2}{n}$ ; the sampling variance for the median is  $\frac{\pi\sigma^2}{2n}$ 
  - In other words, the sampling variance for the median is  $\frac{\pi}{2} \approx 1.57$  times as large as for the mean
- The goal, then, is to find an estimator that is more resistant than the mean, but more efficient than the median

10 / 64

### M-estimation of Location

- A large class of estimators that generalize the idea of maximum likelihood to robust measures of scale and location, and also extend to robust regression
- M-estimates are very robust for estimation location and relatively efficient compared to other robust measures for large samples ( $n > 40$ )
- Let  $T_n(y_1, \dots, y_n)$  be an estimate of an unknown parameter that characterizes the distribution  $F(Y; \theta)$

$$lF(\theta; y_1, \dots, y_n) = \prod_{i=1}^n f(Y, \theta)$$

where  $f(Y; \theta)$  is the density corresponding to  $F(Y; \theta)$

- The ML estimator is the value of  $\theta$  that maximizes the likelihood function, or equivalently minimizes:

$$-\log l = \sum_{i=1}^n \rho(Y; \theta) = \sum_{i=1}^n \rho\left(\frac{y_i - \hat{\mu}}{cS}\right)$$

11 / 64

### M-estimation of Location (2)

- Restricting the objective function  $\rho(Y; \theta)$  to any function that is differentiable with an absolutely continuous derivative  $\Psi(\cdot)$  results in the ML estimator  $T_n$

$$\sum_{i=1}^n \Psi(Y; \theta) = 0$$

where

$$\Psi(Y; \theta) = -\left(\frac{\partial}{\partial \theta}\right)\rho(Y; \theta) = \left(\frac{\partial}{\partial \theta}\right)\log f(Y; \theta)$$

- M-estimation relies on the least squares objective function:  $\rho(y; \theta) = \frac{1}{2}(y - \hat{\mu})^2$  whose derivative shows the influence function:  $\Psi(y; \theta) = (y - \hat{\mu})$  is proportional to the value of  $y$ .
- If  $\rho(Y; \theta)$  is symmetric around 0, the breakdown point of the estimator is  $\varepsilon_n^* = \lim_{n \rightarrow \infty} \varepsilon_n^* = .5$ 
  - The commonly used Huber and Bisquare estimates fit these criteria

12 / 64

### M-Estimation of Location Huber Estimates (1)

- A good compromise between the efficiency of the least-squares and the robustness of the least-absolute values estimators is the *Huber Objective Function*
- At the center of the distribution, the Huber function behaves like the OLS function, but at the extremes, it behaves like the LAV function

$$\rho_H(y; \theta) = \begin{cases} \frac{1}{2}y^2 & \text{for } |y| \leq c \\ c|y| - \frac{1}{2}c^2 & \text{for } |y| > c \end{cases}$$

- The Influence function is determined by taking the derivative:

$$\Psi_H(y; \theta) = \begin{cases} c & \text{for } y > c \\ y & \text{for } |y| \leq c \\ -c & \text{for } y < -c \end{cases}$$

- The tuning constant,  $c$ , defines the center and tails

13 / 64

### M-estimation, the Tuning Constant

- The tuning constant can be expressed as a multiple of the *scale* (the spread) of  $Y$ ,  $k = cS$  where  $S$  is the measure of the scale of  $Y$  (i.e., the spread)
  - Since the standard deviation is influenced by extreme observations, we instead use the median absolute deviation to measure spread:

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

- The median of  $Y$  serves as an initial estimate of  $\hat{\mu}$ , thus allowing us to define  $S = \text{MAD}/.6745$  which ensures that  $S$  estimates  $\sigma$  when the population is normal
  - Using  $k = 1.345$  ( $\frac{1.345}{.6745} \approx 2$ ) produces a 95% efficiency relative to the sample mean when the population is normal and gives substantial resistance to outliers when it is not
- A smaller  $k$  gives more resistance

14 / 64

### M-Estimation of Location: Biweight Estimates

- Biweight 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:

$$\rho_{BW}(y) = \begin{cases} \left\{ \frac{c^2}{6} \left[ 1 - \left[ 1 - \left( \frac{y}{c} \right)^2 \right]^3 \right] \right\}, & \text{if } |y| \leq c \\ \frac{c^2}{6}, & \text{if } |y| > c \end{cases}$$

- The influence function, then, tends toward zero:

$$\Psi_{BW}(y) = \begin{cases} y \left[ 1 - \left( \frac{y}{c} \right)^2 \right]^2, & \text{if } |y| \leq c \\ 0, & \text{if } |y| > c \end{cases}$$

- For this function a tuning constant of  $c = 4.685 \times S \approx 7$  MADs, produces 95% efficiency when sampling from a normal population

15 / 64

### Weights in M-Estimation

Taking the derivative of  $\Psi_H(y; \theta)$  gives the weights applied to each observation.

$$w_{H_i}(y) = \begin{cases} 1 & \text{if } y \leq c; \\ \frac{c}{|y|} & \text{if } y > c \end{cases}$$

$$w_{BW_i}(y) = \begin{cases} \left[ 1 - \left( \frac{y}{c} \right)^2 \right]^2 & \text{if } |y| \leq c; \\ 0 & \text{if } |y| > c \end{cases}$$

The difference between the two weight functions happens more in the extremes. The Bi-weight function is a bit more resistant to outliers

16 / 64

## Weight Functions for Various Estimators

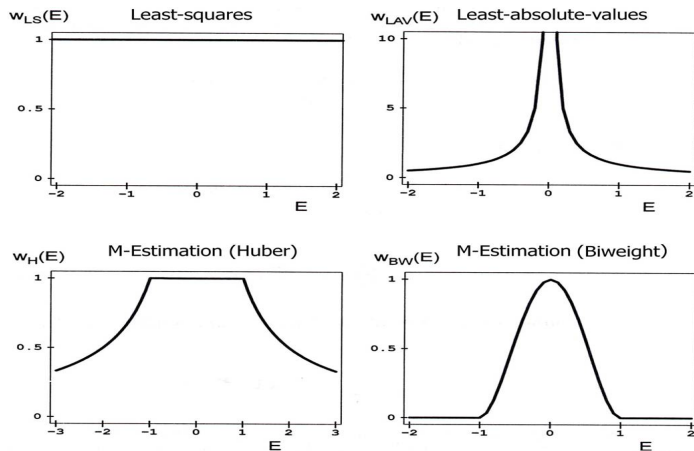


Figure 14.13 from Fox (1997)

## M-Estimation for Regression

- OLS minimizes the sum of squares function

$$\min \sum_{i=1}^n (E_i)^2$$

- Following from M-estimation of location, a robust regression M-estimate minimizes the sum of a less rapidly increasing function  $\rho$  of the residuals

$$\min \sum_{i=1}^n \rho(E_i)$$

- Since the solution is not scale invariant. the residuals must be standardized by a robust estimate of their scale,  $\sigma_\epsilon$ , which is estimated simultaneously. Usually, the median absolute deviation is used:

$$\min \sum_{i=1}^n \rho\left(\frac{E_i}{\hat{\sigma}_E}\right)$$

## M-estimation for Regression (2)

- Taking the derivative and solving produces the shape of the influence function:

$$\sum_{i=1}^n \Psi\left(\frac{E_i}{\hat{\sigma}_E}\right) x_i, \text{ where } \psi = \rho'$$

- We then substitute  $\Psi$  with an appropriate weight function

$$\sum_{i=1}^n w_i \left(\frac{E_i}{\hat{\sigma}_E}\right) x_i$$

- Typically the Huber or bisquare weight is employed. In other words, the solution assigns a different weight to each case depending on the size of its residual and thus minimizes the weighted sum of squares

$$\sum_{i=1}^n w_i E_i^2$$

## M-Estimation and Regression (3)

- Since the weights can be estimated before fitting the model and estimates can't be found without the weights, an iterative procedure is needed to find estimates
- Initial estimates of  $\mathbf{b}$  are selected using weighted least squares
- The residuals from this model are used to calculate an estimate of the scale of the residuals  $\sigma_\epsilon^{(0)}$  and the weights  $w_i^{(0)}$
- The model is then refit with several iterations minimizing the weighted sum of squares to obtain new estimates of  $\mathbf{b}$

$$\mathbf{b}^{(l)} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \mathbf{X}'\mathbf{W}\mathbf{y}$$

- where  $l$  is the iteration counter; in the  $i^{\text{th}}$  row of the model matrix are  $x_i^l$  and  $\mathbf{W} \equiv \text{diag}\{w_i^{l-1}\}$
- This process continues until the model converges ( $\mathbf{b}^{(l)} \approx \mathbf{b}^{(l-1)}$ )

## Asymptotic Standard Errors

- For all M-estimators (including the MM-estimator), asymptotic standard errors are given by the square root of the diagonal entries of the estimated asymptotic covariance matrix  $(X'WX)^{-1}\sigma_E^2$  from the final IWLS fit
- The ASE for a particular coefficient, is then given by:

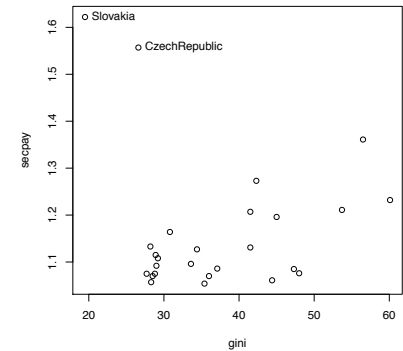
$$SE\hat{\beta} = \sqrt{\frac{\sum [W(E_i)]^2}{[\sum W'(E_i)/n]^2} (X'X)^{-1}}$$

- The ASEs are reliable if the sample size  $n$  is sufficiently large relative to the number of parameters estimated
  - Other evidence suggests that their reliability also decreases as the proportion of influential cases increases
- As a result, if  $n < 50$  bootstrapping should be considered

21 / 64

## Inequality Model Revisited (1)

- The scatterplot clearly shows Slovakia and the Czech Republic as unusual cases - they are outliers in  $Y$  and have high leverage according to  $X$ , a combination that results in high influence
- The OLS line fit to the data indicates that they significantly pull the line toward them



22 / 64

## R-Script for plot with case names

```
> library(boot)
> weakliem <- read.table("Weakliem2.txt")
> plot(secpay ~ gini, data=weakliem)
> outs <- which(rownames(weakliem) %in%
+ c("Slovakia", "CzechRepublic"))
> with(weakliem, text(gini[outs], secpay[outs],
+ rownames(weakliem)[outs], pos=4))
```

23 / 64

## Inequality Model Revisited (2)

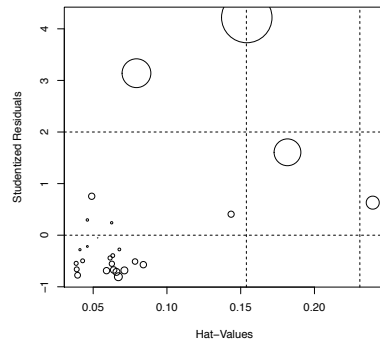
(Intercept)	1.1948*
	(0.1107)
gini	-0.0008
	(0.0029)
$N$	26
$R^2$	0.0029
adj. $R^2$	-0.0387
Resid. sd	0.1485

Standard errors in parentheses

\* indicates significance at  $p < 0.05$

24 / 64

### Inequality Model Revisited (3)



We see the Czech Republic and Slovakia standing out, but Chile also has relatively high influence

25 / 64

### Inequality Model Revisited (4)

(Intercept)	0.9408*
	(0.0526)
gini	0.0050*
	(0.0013)
$N$	24
$R^2$	0.3887
adj. $R^2$	0.3609
Resid. sd	0.0627

Standard errors in parentheses  
\* indicates significance at  $p < 0.05$

26 / 64

### M-Estimation in R (1)

```
> library(MASS)
> mod3 <- rlm(secpay ~ gini, data=weakliem)
> summary(mod3)
```

Call: rlm(formula = secpay ~ gini, data = weakliem)

Residuals:

Min	1Q	Median	3Q	Max
-0.091680	-0.041786	-0.005509	0.038944	0.545452

Coefficients:

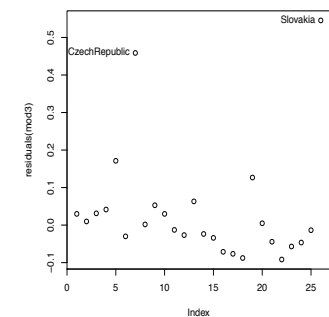
	Value	Std. Error	t value
(Intercept)	1.0169	0.0563	18.0643
gini	0.0031	0.0015	2.0837

Residual standard error: 0.06363 on 24 degrees of freedom

27 / 64

### M-Estimation in R (2)

- As we would expect, the Czech Republic and Slovakia have the largest residuals meaning that they would have the largest influence on the OLS regression line

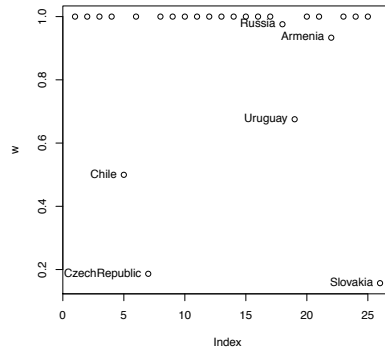


```
> plot(residuals(mod3))
> text(outs, residuals(mod3)[outs],
+ rownames(weakliem)[outs], pos=2)
```

28 / 64

### M-Estimation in R (3)

- The plot on the right shows the weight given to each case
- In line with the previous graph, the Czech Republic and Slovakia received the least weight of all the observations
- Note: this technique can also be used to assess influential cases



```
> loww <- with(mod3, which(w < 1))
> with(mod3, plot(w))
> text(loww, mod3[["w"]][loww],
+ rownames(weakliem)[loww], pos=2)
```

29 / 64

### M-estimation in R (4)

```
> mod4 <- rlm(secpay ~ gini, data=weakliem,
+ psi=psi.bisquare)
> summary(mod4)
```

Call: rlm(formula = secpay ~ gini, data = weakliem, psi = psi.bisquare)

Residuals:

	Min	1Q	Median	3Q	Max
	-0.095520	-0.023978	0.005806	0.046531	0.577087

Coefficients:

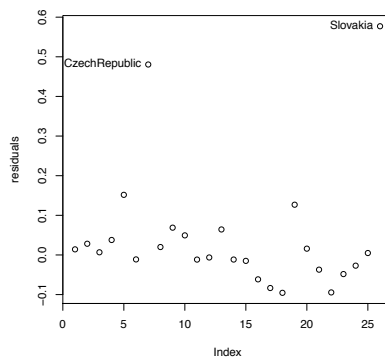
	Value	Std. Error	t value
(Intercept)	0.9583	0.0491	19.5245
gini	0.0044	0.0013	3.4724

Residual standard error: 0.05552 on 24 degrees of freedom

30 / 64

### M-estimation in R (5)

- Again, the Czech Republic and Slovakia have large residuals

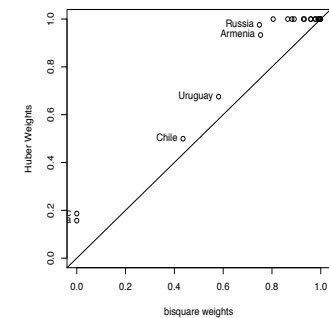


```
> with(mod4, plot(residuals))
> text(outs, mod4[["residuals"]][outs],
+ rownames(weakliem)[outs], pos=2)
```

31 / 64

### Comparing Huber and Bisquare Weights

- Notice that although the relative weights are similar, the Huber method gives a weight of 1 to far more cases than does the bisquare weight



```
> plot(mod4[["w"]], mod3[["w"]], xlim=c(0,1),
+ ylim=c(0,1), xlab="bisquare weights",
+ ylab="Huber Weights")
> abline(0,1)
> text(mod4[["w"]][loww], mod3[["w"]][loww],
+ rownames(weakliem)[loww], pos=2)
```

32 / 64

### Limitation of M-estimation

- M-estimation for regression is designed to be robust against heavy-tailed error distributions
  - They are more resistant to vertical outliers than are OLS estimators
  - They are also more efficient than OLS estimators when there are vertical outliers
- M-estimates are not immune to unusual observations however
  - They have a breakdown point of 0, and an unbounded influence function because they pay no attention to leverage
  - As a result, “bad” leverage points that have small residuals can have as much influence on M-estimates as they do on OLS estimates

33 / 64

### Resistant Regression Methods

- Also referred to as *bounded influence methods*
- Have the highest possible breakdown point ( $\epsilon_i^* = .5$ ). Some of these are:
  - S-estimation
  - Least Trimmed Squares (LTS)
  - Least Median Squares (LMS)
- Although very resistant in the presence of unusual observations, these methods have the serious limitation that they are very inefficient relative to OLS when the errors *are* normally distributed
  - For all of these, relative efficiency is less than 40%
- These methods have utility, however, in that they all can play an important role in MM-estimation, which produces estimates that are both highly resistant and efficient

34 / 64

### Bounded Influence Regression: LTS

- Least-Trimmed Squares orders the squared residuals from smallest to largest:  $(E^2)_{(1)}, (E^2)_{(2)}, \dots, (E^2)_{(n)}$
- It then calculates  $\mathbf{b}$  that minimizes the sum of only the smaller half of the residuals

$$\min \sum_{i=1}^m (E^2)_{(i)}$$

where  $m = \lfloor \frac{n+k+2}{2} \rfloor$ ; the floor brackets indicate rounding *down* to the nearest integer

- By using only the 50% of the data that fits closest to the original OLS line, LTS completely ignores extreme outliers (i.e., it eliminates the largest residuals)
- Although highly resistant, LTS is very inefficient and can also misrepresent the trend in the data if it is characterized by clusters of extreme cases or if the dataset is relatively small

35 / 64

### Bounded Influence Regression: LMS

- An alternative bounded influence method is Least Median Squares
- Rather than minimize the sum of the least squares function, LMS minimizes the median of the squared residuals  $E_i^2$

$$\min \text{MED}(E_i^2)$$

- LMS is very robust with respect to outliers both in terms of  $X$  and  $Y$  values
- Again, however, like LTS it performs poorly from the point of view of asymptotic efficiency (at best, it is 37%)

36 / 64

## Bounded Influence Regression: S-estimation

- Attempt to find the solution with the smallest possible dispersion of the residuals

$$\min \hat{\sigma} (E_1(\hat{\beta}), \dots, E_n(\hat{\beta}))$$

- Of course, OLS does this too, by minimizing the variance
- S-estimation, however, minimizes a robust M-estimate of the residual scale  $\hat{\sigma}$

$$\sum_{i=1}^n \rho \left( \frac{E_i}{\hat{\sigma}_E} \right) = (n - p)K$$

$K = 0.5$  ensures consistency at the normal distribution of errors

- Although it has a high breakdown point, S-estimation isn't attractive on its own because it has very low efficiency (approximately 30% of OLS estimators under the condition of normally distributed errors)

37 / 64

## Resistant Regression in R

- LTS, LMS and S-estimation can all be done using the MASS library:
  - LTS regression: `ltsreg(y ~ x)`
  - LMS regression: `lmsreg(y ~ x)`
  - S-estimation: `lqs(y ~ x, method = "S")`
- Although the `fitted`, `residuals`, and `coefficients` functions work for these objects, many other functions available for `lm` objects are not available (including the `summary` function)
- In any event, these methods really have limited utility on their own

38 / 64

## Combining Resistance and Efficiency: MM-estimation (1)

- MM estimation is perhaps the most commonly employed method today
- "MM" in the name refers to the fact that more than one M-estimation procedure is used to calculate the final estimates
- Combine a high breakdown point (50%), bounded influence function and high efficiency under normal errors ( $\approx 95\%$ )

39 / 64

## Steps to MM-estimation (1)

1. Initial estimates of the coefficients  $\mathbf{B}^{(1)}$  and corresponding residuals  $e_i^{(1)}$  are taken from a highly resistant robust regression (i.e., a regression with a breakdown point of 50%)
  - Although the estimator must be consistent, it is not necessary that it is efficient. As a result, S-estimation with Huber or Bisquare weights is typically employed here
2. The Residuals  $E_i^{(1)}$  from the S-estimation stage 1 are used to compute an M-estimation of the scale of the residuals
3. Finally, the initial estimates of the residuals  $e_i^{(1)}$  from stage 1 and of the residual scale  $\sigma_E$  from stage 2 are used to compute a single-step M-estimate

$$\sum_{i=1}^n w_i \left( \frac{E_i^{(1)}}{\hat{\sigma}_E} \right) x_i$$

where the  $w_i$  are typically Huber or bisquare weights. In other words, the M-estimation procedure at this stage needs only a single iteration of weighted least squares

40 / 64

## MM-estimation in R

```
> mod5 <- rlm(secpay ~ gini, data=weakliem,
+ method="MM")
> summary(mod5)

Call: rlm(formula = secpay ~ gini, data = weakliem, method = "MM")
Residuals:
    Min       1Q   Median       3Q      Max
-0.097533 -0.023654  0.004203  0.046373  0.578463

Coefficients:
            Value Std. Error t value
(Intercept)  0.9546   0.0489   19.5070
gini          0.0046   0.0013    3.5759

Residual standard error: 0.06068 on 24 degrees of freedom
```

41 / 64

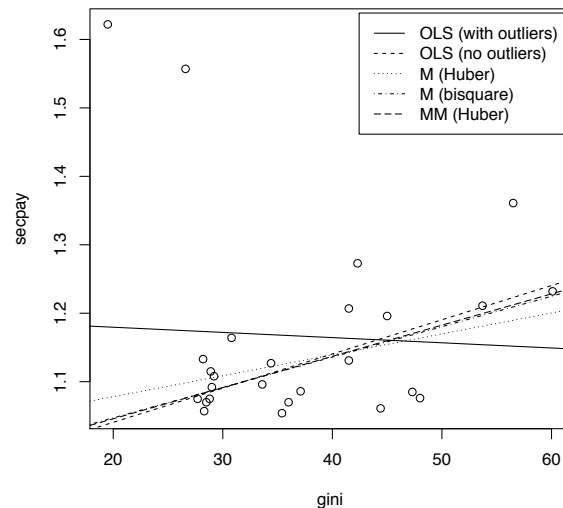
## Comparing the models

	Estimate	SE	t
OLS (with outliers)	-0.0008	0.0029	-0.2630
OLS (no outliers)	0.0050	0.0013	3.7400
M (Huber)	0.0031	0.0015	2.0837
M (bisquare)	0.0044	0.0013	3.4724
MM (Huber)	0.0046	0.0013	3.5759

- the M-estimation techniques give very similar estimates to the OLS with the two influential cases deleted
  - The MM-estimation is closest
- Don't just fit these models blindly - look at the pattern in the data and make sure you pick the method that works best

42 / 64

## Comparison Plot of Different Models



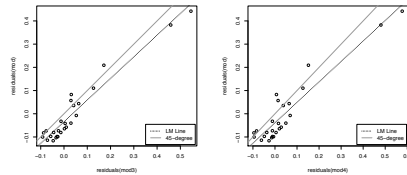
43 / 64

## Diagnostics Revisited RR-Plots

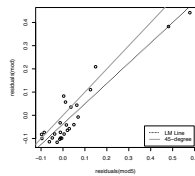
- A criticism of common measures of influence, such as Cooks' D is that they are not robust
  - Since their calculation is based on the sample mean and covariance matrix, they can often miss outliers because of a "masking effect" where a group of influential points can mask each other
  - Partial regression plots can overcome this effect with respect to individual coefficients, but don't help detect overall influence
- RR-plots ("residual-residual" plots), which plot the residuals from an OLS fit against the residuals from many different robust regressions, are a solution
  - If the OLS assumptions hold perfectly, there will be a perfect positive relationship, with a slope = 1 (the identity line), between OLS and the residuals from any robust regression
  - If there are outliers, however, the slope will equal a value other than 1 since the OLS regression will not resist them while the robust regression will

44 / 64

## RR-Plots



(a) M (Huber)      (b) M (Bisquare)



(c) MM (Huber)

45 / 64

## Robust GLMs

- Robust GLMs are useful especially for improving efficiency when the errors are widely spread, but can also limit the impact of outliers
- Cantoni and Ronchetti's (2001) robust estimator is a recent development that uses quasi-likelihood estimation
- The estimator follows from M-estimation for a regression, solving:

$$\Psi(Y; \mu) = v(Y; \mu)w(X)\mu' - a(\beta)$$

where

$$a(\beta) = \frac{1}{2} \sum_{i=1}^n E[v(y_i, \mu_i)]w(X)\mu'$$

- And the  $v_i$  and  $w_i$  are weight functions, with the  $v_i$  being based on Huber weights for the residuals (i.e., outlyingness) and the  $w_i$  being based on robust distances from the centroid point of the  $X$ 's, (i.e., on leverage)

46 / 64

## Robust GLM: A Poisson Regression Example

- Outcome variable: number of voluntary associations to which respondents belong
- Explanatory variables:
  - Gender (with women as the reference category)
  - AGE (in years)
  - SES *Upper*, Middle, Skilled, Unskilled
- The goal is to assess the impact on membership of SES
- I'll fit both a regular GLM and after some diagnostics, a robust GLM

47 / 64

## Robust GLM Example (2)

(Intercept)	0.8433*
	(0.0859)
SEXMen	-0.1056*
	(0.0494)
AGE	-0.0001
	(0.0015)
SESmiddle	-0.1282*
	(0.0618)
SESskilled	-0.5883*
	(0.0697)
SESunskilled	-0.6719*
	(0.1031)
<i>N</i>	1000
AIC	3766.8528
BIC	3884.6389
log <i>L</i>	-1859.4264

Standard errors in parentheses

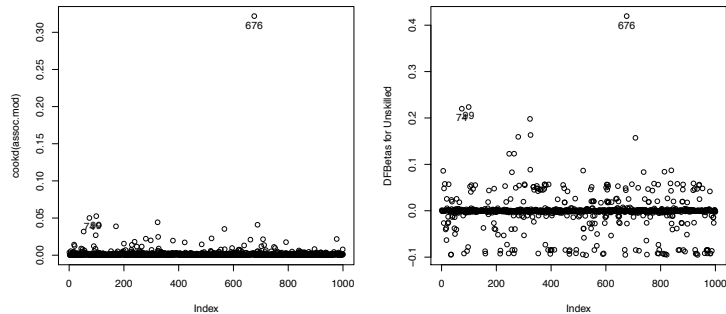
\* indicates significance at  $p < 0.05$

- SES seems to be a strong predictor of Membership
- Now, let's do some diagnostics

48 / 64

### Robust GLM Example (3)

```
> assoc[bigd, ]
  ASSOC COUNTRY  SEX AGE      SES
74      7  Canada Women 36 unskilled
99      7  Canada Women 57 unskilled
676     15  Canada Women 18 unskilled
```



49 / 64

### Robust GLM Example (4)

- The Cook's Ds and DFBetas suggest that there is one observation that has unusually high influence (676)
- This case may be affecting the estimates and/or standard errors, so we try fitting a robust GLM

50 / 64

### Robust GLM Example (5)

- Using the `glmrob` function of the `robustbase` package, we now come to the conclusion

```
> library(robustbase)
> glm2 <- glmrob(ASSOC ~ SEX + AGE + SES, data=assoc, family=poisson(link=log))
> summary(glm2)
```

Call: `glmrob(formula = ASSOC ~ SEX + AGE + SES, family = poisson(link = log),`

Coefficients:

	Estimate	Std. Error	z-value	Pr(> z )
(Intercept)	0.785625	0.095518	8.225	< 2e-16 ***
SEXMen	-0.058482	0.055002	-1.063	0.2877
AGE	-0.002138	0.001669	-1.281	0.2002
SESmiddle	-0.146813	0.068356	-2.148	0.0317 *
SESskilled	-0.556185	0.076265	-7.293	3.04e-13 ***
SESunskilled	-0.985988	0.129939	-7.588	3.25e-14 ***

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

Robustness weights w.r \* w.x:

758 weights are ~ = 1. The remaining 242 ones are summarized as

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.08398	0.49370	0.73350	0.67540	0.90170	0.99800

Number of observations: 1000  
Fitted by method `âMqleâ` (in 8 iterations)

51 / 64

### Robust GLM Example (6)

- Taking the exponent of the coefficients, we see that for the regular GLM, upper class workers belong to about twice as many organizations as unskilled workers
- In the robust GLM, we can see that people in the upper class belong to almost 3 times as many organizations as unskilled workers
- The large difference suggests that using the robust GLM is a good idea

	GLM	Robust GLM
(Intercept)	2.32	2.19
SEXMen	0.90	0.94
AGE	1.00	1.00
SESmiddle	0.88	0.86
SESskilled	0.56	0.57
SESunskilled	0.51	0.37

52 / 64

## Bootstrapping Robust Regression

- Robust Regression (MM-estimation) gave a significantly better fit to the data than OLS
  - The slope for the Gini coefficient didn't only reach statistical significance, but it changed direction
  - The residual standard error also decreased

```
> mod2 <- rlm(secpay ~ gini, data=weakliem, method="MM")
> summary(mod2)
```

```
Call: rlm(formula = secpay ~ gini, data = weakliem, method = "MM")
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-0.097533 -0.023654  0.004203  0.046373  0.578463
```

```
Coefficients:
```

```
              Value Std. Error t value
(Intercept)  0.9546  0.0489    19.5070
gini         0.0046  0.0013     3.5759
```

```
Residual standard error: 0.06068 on 24 degrees of freedom
```

53 / 64

## Bootstrapping Regression: Example (5)

- Despite that the slope coefficient is now statistically significant, the standard errors are not reliable because of the small sample size
  - Standard errors produced by the `rlm` function rely on asymptotic approximations, and thus are not trustworthy for samples of size 26
- The bootstrap provides an alternative way to get standard errors
- I could proceed to bootstrap the regression in 2 ways:
  - Random-X resampling or observation resampling
  - Fixed-X resampling

54 / 64

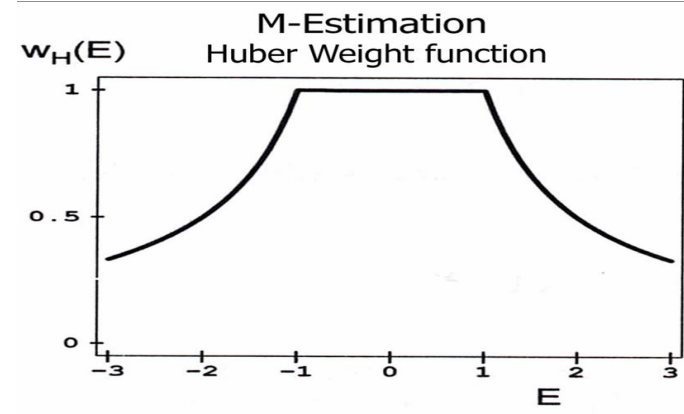
## Random-X Resampling

- I start by creating a function that extracts the Huber weights
- Recall that the Huber weight behaves like OLS in the center of the distribution, but similar to absolute values regression on the tails, giving low weight to observations with extreme residuals

```
> boot.huber <- function(data, inds, maxit=20){
+   assign(".inds", inds, envir=.GlobalEnv)
+   mod <- rlm(secpay ~ gini, data=data[.inds,],
+     maxit=maxit)
+   remove(".inds", envir=.GlobalEnv)
+   coefficients(mod)
+ }
```

55 / 64

## Huber Weight Function



56 / 64

## Random-X Resampling (3)

- The `boot` function returns the original robust coefficients (from the `rlm` model), bootstrap standard errors and the bias of the bootstrap estimate

```
> boot.ranx <- boot(data=weakliem, statistic=boot.huber,  
+ R=1500, maxit=100)  
> boot.ranx
```

- `t1*` represents the first coefficient in the model frame (the intercept); `t2*` represents the second (gini)

57 / 64

## Bootstrap Confidence Interval

- According to all three measures, the coefficient is not statistically significant (the default for `boot.ci` is a 95% CI) - still, the  $BC_a$  is slightly larger and centered differently than the normal

```
> boot.ci(boot.ranx, type=c("normal", "perc", "bca"), index=2)
```

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS  
Based on 1500 bootstrap replicates

CALL :

```
boot.ci(boot.out = boot.ranx, type = c("normal", "perc", "bca"),  
index = 2)
```

Intervals :

Level	Percentile	BCa
95%	(-0.0071, 0.0071 )	(-0.0082, 0.0067 )

Calculations and Intervals on Original Scale

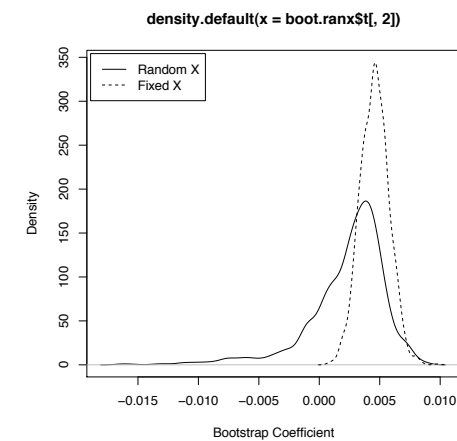
58 / 64

## Fixed-X Resampling

```
> fit <- fitted(mod2)  
> e <- residuals(mod2)  
> boot.huber.fixed <- function(data, inds, maxit=20){  
+ assign(".inds", inds, envir=.GlobalEnv)  
+ y.star <-< fit + e[.inds]  
+ mod <- update(mod2, y.star ~ .)  
+ remove(".inds", envir=.GlobalEnv)  
+ coefficients(mod)  
+ }  
> boot.fixx <- boot(weakliem, boot.huber.fixed,  
+ R=1500, maxit=100)  
> boot.fixx
```

59 / 64

## Bootstrap Distributions



60 / 64

## Make Confidence Intervals

```
> ranx.ci <- boot.ci(boot.ranx, index=2,
+ type=c("perc", "bca"))
> fixx.ci <- boot.ci(boot.fixx, index=2,
+ type=c("perc", "bca"))
> analytical.ci <- coef(mod2)[2] + qt(.975,
+ nrow(weakliem)-mod2$rank)*
+ sqrt(diag(vcov(mod2)))[2] * c(-1,1)
> cis <- rbind(ranx.ci$bca[4:5],
+ fixx.ci$bca[4:5],
+ analytical.ci)
> rownames(cis) <- c("randomX", "fixedX", "analytical")
> colnames(cis) <- c("lower", "upper")
> round(cis, 3)
```

	lower	upper
randomX	-0.008	0.007
fixedX	0.002	0.007
analytical	0.002	0.007

61 / 64

## Comparison

- The standard error in the original data was 0.0015, essentially the same as the fixed-X bootstrap. This is considerably less than the standard error of 0.0033 found in the random-X bootstrap
- This result is also confirmed in the distribution of the bootstrap - although there is still a skew, it is less than in the random-X case

62 / 64

## Summary and Conclusions (1)

- Separate points can have 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 can also cause problems for efficiency
- Efforts should first be made to remedy the problem of unusual cases before proceeding to robust regression
- If robust regression is used, careful attention should be paid to the model - different procedures can give different answers
  - Plots of OLS residuals against robust residuals can be helpful for detecting problems with the tails of the error distribution
  - If there are no problems, the residuals should all be close to perfectly correlated

63 / 64

## Summary and Conclusions (2)

- MM-estimation is typically the best robust method for the linear model
  - Inference from these models is better than OLS in the presence of "bad" cases, but its standard errors are not reliable in small samples
  - This can be overcome by using bootstrapping to obtain new estimates of the standard errors
- Robust regression can also be extended to the GLM

64 / 64