SOS3003 **Applied data analysis for social science** Lecture note 09-2009

Erling Berge
Department of sociology and political science
NTNU

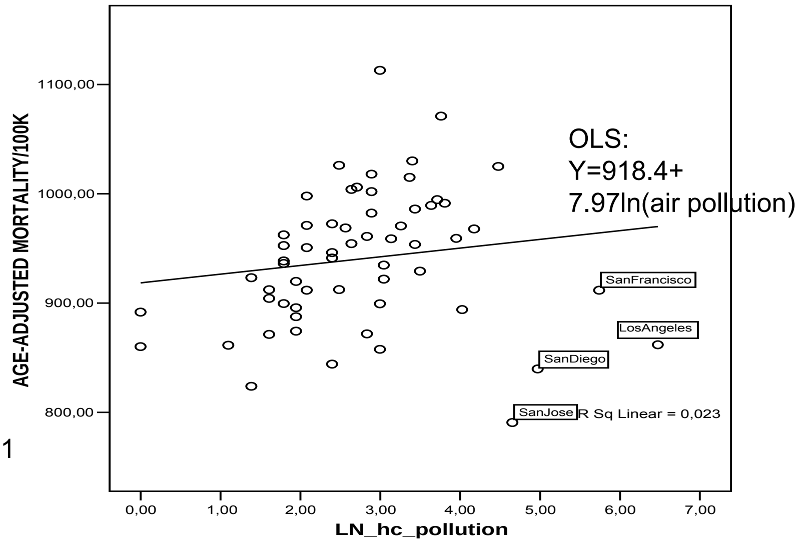
Literature

- Robust Regression
  Hamilton Ch 6 p183-212

Robust Regression

- Has been developed to work well in situations where OLS breaks down. Where the OLS assumptions are satisfied robust regression are not as good as OLS, but not by very much
- Even if robust regression is better suited for those who do not want to put much effort into testing the assumptions, it is so far difficult to use
- Robust regression has focused on residuals with heavy tails (many cases with high influence on the regression)

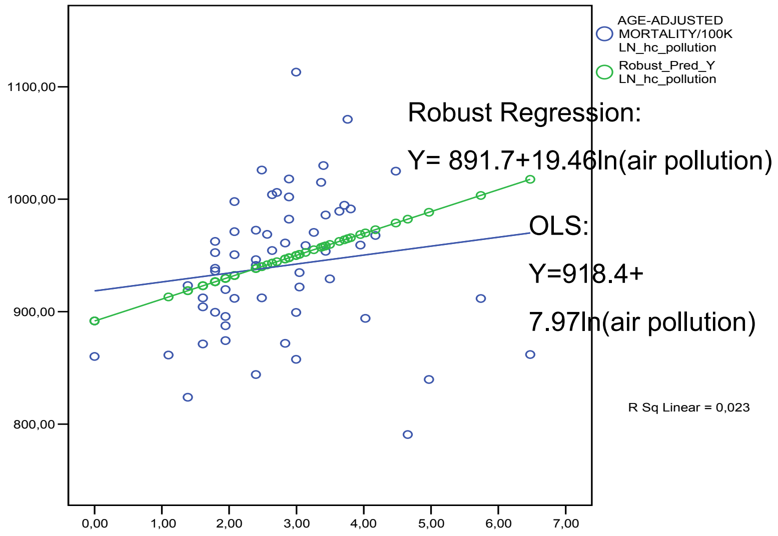
Regression of mortality on air pollution



OLS:
Y=918.4+
7.97ln(air pollution)

Figure 6.1
Hamilton

Robust regression of mortality on air pollution



Robust Regression:

Y= 891.7+19.46ln(air pollution)

OLS:

Y=918.4+

7.97ln(air pollution)

Figure 6.2
Hamilton

Robust regression and SPSS

- SPSS do not have a particular routine that performs robust regression
- It can possibly be done within the Generalized linear models procedure <but I have not tested it our>
- It can be done by weighted OLS regression, but then it is required that we make the weight functions and go through the iterations one by one including computation of weights every time
- This procedure will be outlined below

ROBUST AND RESISTANT

- RESISTANT methods are not affected by small errors or changes in the sample data
- ROBUST methods are not affected by small deviations from the assumptions of the model
- Most resistant estimators are also robust in relation to the assumption about normally distributed residuals
-
- **OLS is neither ROBUST nor RESISTANT**

Outliers is a problem for OLS

Outliers affect the estimates of

- Parameters
- Standard errors (standard deviation of parameters)
- Coefficient of determination
- Test statistics
- And many other statistics

Robust regression tries to protect against this by giving

less weight to such cases,

not by excluding them

Protection against NON-NORMALE residuals

Robust methods can help when

- the tails in the distribution of the residuals are heavy, i.e. when it is too many outliers compared to the normal distribution
- Unusual X-values have leverage and may cause problems

But for other causes of non-normality

robust methods will not help

Estimation methods for robust regression

- M-estimation (maximum likelihood) minimizes a weighted sum of the residuals. This can be approximated by the weighted least squares method (WLS)
- R-estimation (based on rank) minimizes a sum where a weighted rank is included. The method is more difficult to use than M-estimation
- L-estimation (based on quantiles) uses linear functions of the sample order statistics (quantiles)

IRLS- Iterated Reweighted Least Squares

M-estimation by means of IRLS needs

1. Start values from OLS. Save the residuals
2. Use OLS residuals to find weights. Larger residuals gives less weight
3. Find new parameter values and residuals with WLS
4. Go to step 2 and find new weights from the new residuals, go on to step 3 and 4, until changes in the parameters become small

Iteration: to repeat a sequence of operations

IRLS

- IRLS is in theory equivalent to M-estimation
- To use the method we need to compute
- Scaled residuals, u_i , and a
- Weight function, w_i ,that gives least weight to the largest residuals

Scaling of residuals I

- Scaled residual u_i
  - s is the scale factor and e_i residual
- The scale factor in OLS is the estimate of the standard error of the residual: <u>nb! s_e is not resistant</u>
- A resistant alternative is based on MAD, "median absolute deviation"

$$u_i = \frac{e_i}{s}$$

$$s_e = \sqrt{\frac{RSS}{n-K}}$$

$$MAD = median\,|\,e_i - median\left(e_i\right)|$$

Scaling of residuals II

$$MAD = median\,|\,e_i - median\left(e_i\right)|$$

The scale factor (standard error of the distribution) Using a resistant estimate will be

- s = MAD/ 0.6745 = 1.483MAD

and the scaled residual

- u_i = [e_i / s] = (0.6745*e_i)/MAD

In a normal distribution s= MAD/ 0.6745 will estimate the standard error correctly like s_e

In case of non-normal errors s= MAD/ 0.6745 will be better.

This is a resistant estimate, s_e is not resistant

Weight functions I

- Properties is measured in relation to OLS on normally distributed errors.
- The method should be “almost as good” as OLS on normally distributed errors and much better when the errors are non-normal
- Properties are determined by a “calibration constant” (c in the formulas)

Weight functions II

- **OLS-weights**: w_i = 1 for all i
- **Huber-weights**: weights down when the scaled residual is larger than c, c=1,345 gives 95% of the efficiency of OLS on normally distributed errors
- **Tukey’s bi-weighted** estimates get 95% of the efficiency of OLS on normally distributed errors by gradually weighting down scaled errors until $|u_i| \leq c = 4.685$ and by dropping cases where the residual is larger

Huber-weights

$$w_i = 1 \ \forall \ |\mathrm{u_i}| \leq \mathrm{c}$$

$$w_i = \frac{c}{u_i} \forall \ |\mathrm{u_i}| > \mathrm{c}$$

$$\forall = \text{for alle}$$

Tukey weights

$$w_i = \left[1 - \left(\frac{u_i}{c} \right)^2 \right]^2 \forall \ |\mathrm{u_i}| \leq c$$

$$w_i = 0 \ \forall \ |\mathrm{u_i}| > c$$

$$\forall = \ \mathit{for\ alle}$$

- Tukey weighting in IRLS is sensitive for start values of the parameters (one may end up at local minima)

Standard errors and tests in IRLS

- The WLS program cannot estimate standard errors and test statistics correctly by IRLS
- A procedure that works is described by Hamilton on page 198-1999

Use of Robust Estimation

- If OLS and Robust estimates are different it means that outliers have influence on the OLS results making them unreliable. Results cannot be trusted
- Robust predicted values will better portray the bulk of the data
- Robust residuals will better at discovering which cases are unusual
- Weights from the robust regression will show which cases are outliers
- OLS and RR can support each other

Fig 6.9 Hamilton: OLS and RR on untransformed data

Mortality regressed on air pollution

Effect of high leverage

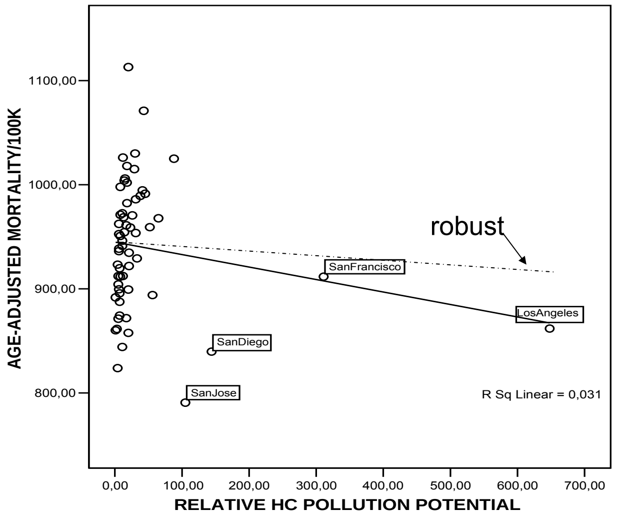
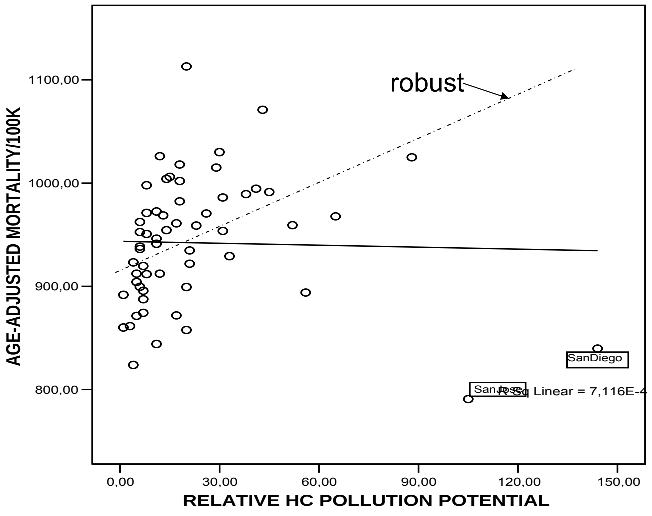


Fig 6.10 Hamilton: OLS and RR on untransformed data when two outliers are removed

Mortality regressed on air pollution



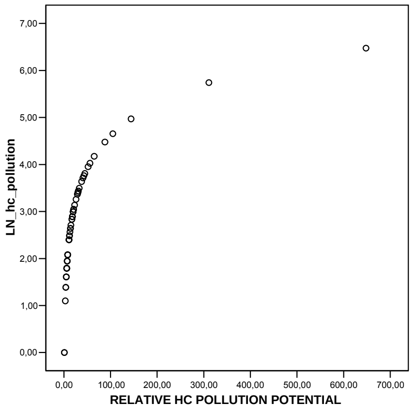
RR do not protect against leverage

- RR with M-estimation protects against unusual y-values (outliers) but not necessarily against unusual x-values (leverage)
- Efforts to test and diagnose are still needed (heteroscedasticity is still a problem for IRLS)
- Studies of the data and transformation to symmetry will reduce the risk of problems appearing
- No method is “safe” if it is used without forethought and diagnostic studies of data

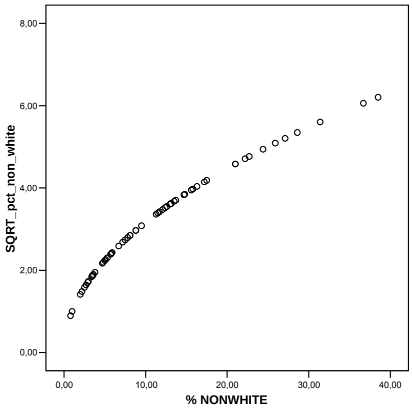
Robust Multippel Regresjon

X_1	RELATIVE HC POLLUTION POTENTIAL	(natural log)
X_2	AVG. YEARLY PRECIP. INCHES	
X_3	AVG. JANUARY TEMPERATURE, F	
X_4	MEDIAN EDUCATION OF POP 25+	
X_5	% NON-WHITE	(square root)
X_6	POPULATION PER HOUSEHOLD	
X_7	% 65 AND OVER	
X_8	% SOUND HOUSING UNITS	
X_9	PEOPLE PER SQUARE MILE	(natural log)
X_{10}	AVG. JULY TEMPERATURE, F	
X_{11}	% WHITE COLLAR EMPLOYMENT	
X_{12}	% FAMILIES WITH INCOME<$3000	(negative reciprocal root)
X_{13}	AVG RELATIVE HUMIDITY, %	

Multiple OLS regression with transformed variables:
effect of transformation



ln of air pollution



Square root of % non-white

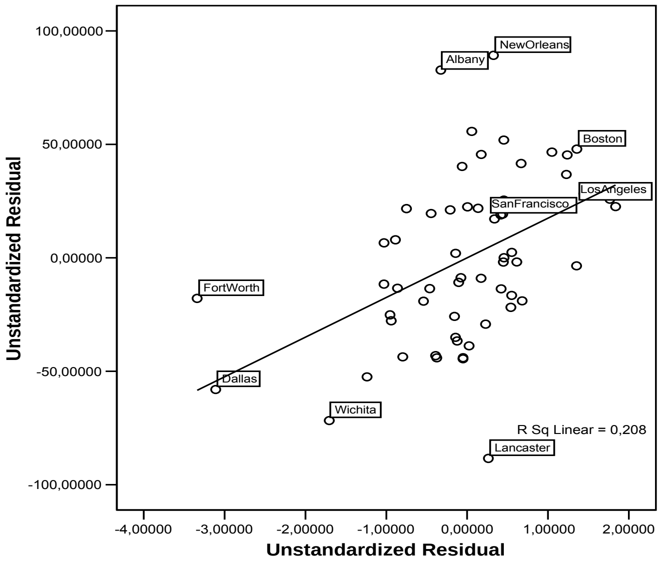
OLS with backward elimination gives

Dependent Variable: AGE-ADJUSTED MORTALITY/100K	B	Std. Error	t	Sig.
(Constant)	986,261	82,674	11,929	,000
LN_hc_pollution	17,469	4,636	3,768	,000
AVG. YEARLY PRECIP. INCHES	2,352	,640	3,677	,001
AVG. JANUARY TEMPERATURE, F	-2,132	,504	-4,228	,000
MEDIAN EDUCATION OF POP 25+	-17,958	6,204	-2,895	,005
SQRT_pct_non_white	27,335	4,398	6,215	,000

- Robust regression gives predicted y:
- $Y= 1001.8+17.77x_{1i}+2.32x_{2i}-2.11x_{3i}-19.1x_{4i}+26.2x_{5i}$

Multiple OLS regression with transformed variables

Leverage plot of residual from mortality (y) and residual of ln_air_pollution (x)



Los Angeles and San Francisco are no longer outliers

Four estimates of the relationship mortality – air pollution

Effect of air pollution

	OLS	Robust
1 variable	7.97	19.46
5 variables	17.47	17.77

- Note that in RR the bivariate regression comes pretty close to the result of the multivariate regression

- In the five-variable model there are new cases with influence on the line of regression
- Removing the 5 cases that have the highest leverage parameter (h_i) do not give substantial changes in the coefficients

Robust Regression vs Bounded Influence Regression

- Robust Regression protect against the effect of outliers (unusual y-values) if these do not go together with unusual x-values

- Bounded Influence Regression is designed to protect against influence from unusual combinations of x-values

BI - Bounded Influence Regression

- BI-methods are made to limit the influence of high leverage cases (large h_i = high leverage)
- The simplest way of doing this is to modify the Huber-weights or Tukey-weights in the IRLS procedure for RR (robust regression) with a factor based on the leverage statistic

Bounded influence: modification of weights

- Expand the weight function with a weight based on the leverage statistic h_i
- $w^H_i = 1$ if $h_i \leq c^H$
- $w^H_i = (c^H / h_i)$ if $h_i > c^H$
- c^H is often set to the 90% percentile in the distribution of h_i
- Then the IRSL weight becomes $w_i \, w^H_i$ where w_i is either the Tukey- or Huber-weight that changes from iteration to iteration while w^H_i is constant

Bounded influence as a diagnostic tool

- Estimation of standard errors and test statistics becomes even more complicated than for the M-estimators mentioned above
- We can use BI estimates as a descriptive tool to check up on other estimates
- One (somewhat) extreme example: PCB pollution in river mouths in 1984 and 1984 (Hamilton table 6.4)

Fig 6.15 and 6.16 Hamilton

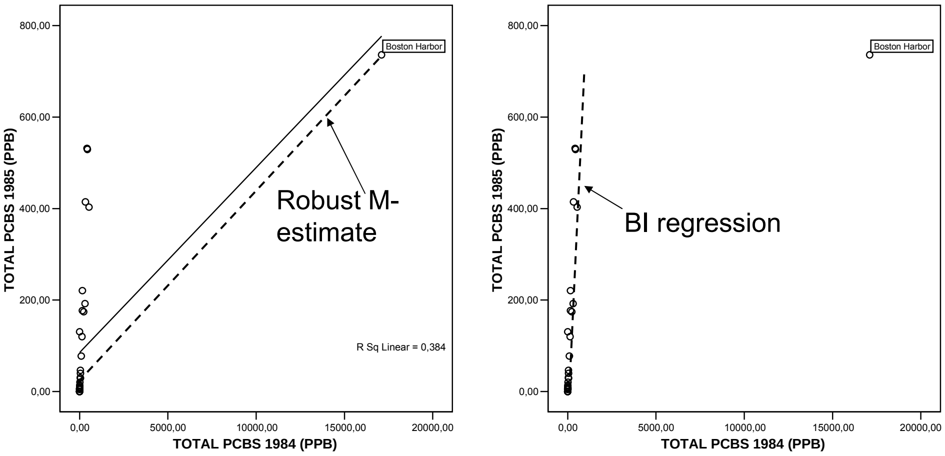
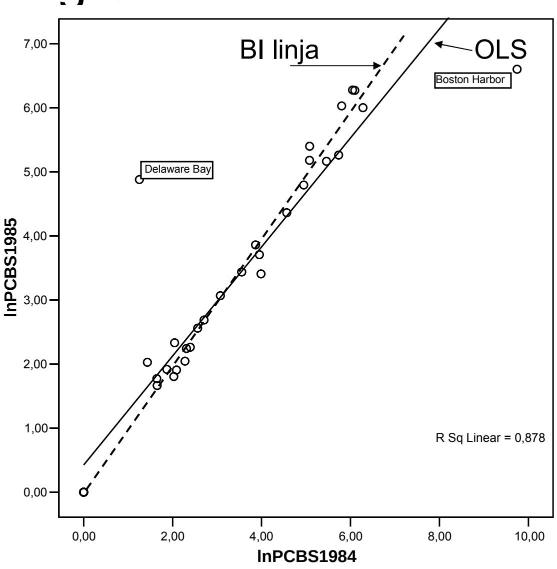
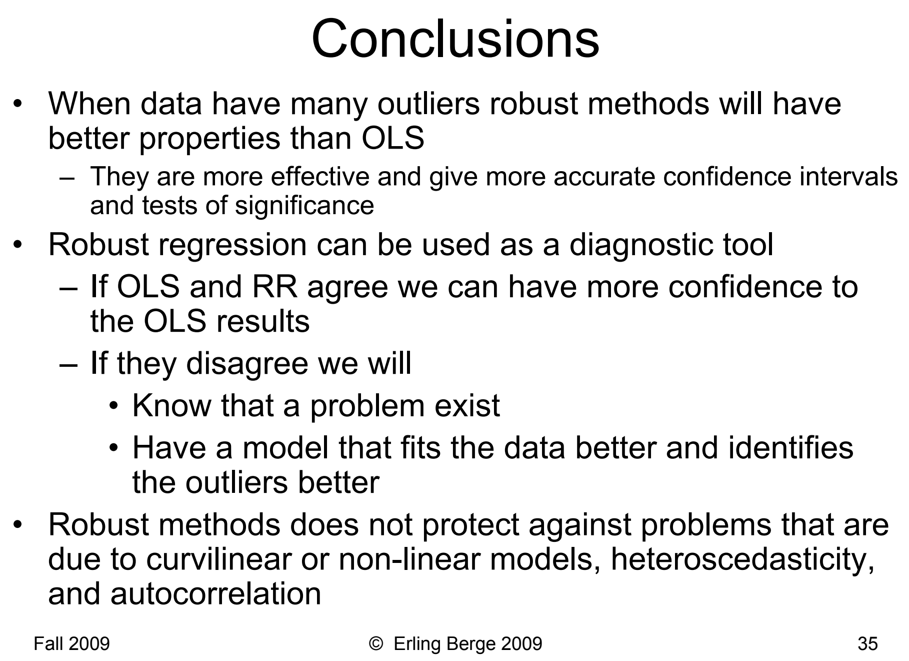


Fig 6.17 Hamilton

OLS and BI estimates with transformed variables give about the same result





Conclusions

- When data have many outliers robust methods will have better properties than OLS
  - They are more effective and give more accurate confidence intervals and tests of significance
- Robust regression can be used as a diagnostic tool
  - If OLS and RR agree we can have more confidence to the OLS results
  - If they disagree we will
    - Know that a problem exist
    - Have a model that fits the data better and identifies the outliers better
- Robust methods does not protect against problems that are due to curvilinear or non-linear models, heteroscedasticity, and autocorrelation