Advanced Topics in Applied Regression

Day 5: Robust regression

Constantin Manuel Bosancianu

Doctoral School of Political Science
Central European University, Budapest
bosancianu@icloud.com

August 4, 2017

Perils of outliers

MARITAL COITAL FREQUENCY AND THE PASSAGE OF TIME: ESTIMATING THE SEPARATE EFFECTS OF SPOUSES' AGES AND MARITAL DURATION, BIRTH AND MARRIAGE COHORTS, AND PERIOD INFLUENCES*

GUILLERMINA JASSO
University of Minnesota

To the extent that marital coital frequency is linked to both within-couple and societal fertility and sex ratio, it may be implicated in a wide range of behavioral and social phenomena. Variation in marital coital frequency over the life course as well as across cohorts may thus affect many aspects of the social life. This paper reports estimates of the ceteris paribus cohort-free effects of spouses' ages, marital duration, and contemporaneous period influences on coital frequency, as well as of the correlations between coital frequency and birth- and marriage-cohort factors. These estimates are obtained by (a) using the properties of the fixed-effects statistical model in order to separate the effects of cohort influences from the age/duration and period effects and to control for the operation of couple-specific unobservables, and (b) using strictly monotonic nonlinear transformations in order to separate the effects of wife's age, husband's age and marital duration.

American Sociological Review: frequency decreases with marital duration and age; some period effects are also found, related to risks of contraceptive use. $N = 2062$.

Perils of outliers

MARITAL COITAL FREQUENCY:
UNNOTICED OUTLIERS AND
UNSPECIFIED INTERACTIONS
LEAD TO ERRONEOUS CONCLUSIONS*

JOAN R. KAHN J. RICHARD UDRY
Carolina Population Center, The University of North Carolina at Chapel Hill

In a recent application of fixed-effects modeling, Jasso estimates age and period effects on marital coital frequency that take into account cohort effects. Although the application is an innovation for researchers interested in identifying net age and period effects, the results are in some respects problematic. Jasso claims that over the period

American Sociological Review: removing 8 cases makes age effects disappear; estimating model on marriages > 2 *yrs* (88 % of sample) makes age, duration and period effects disappear.

Perils of outliers

IS IT OUTLIER DELETION OR IS IT SAMPLE TRUNCATION? NOTES ON SCIENCE AND SEXUALITY*

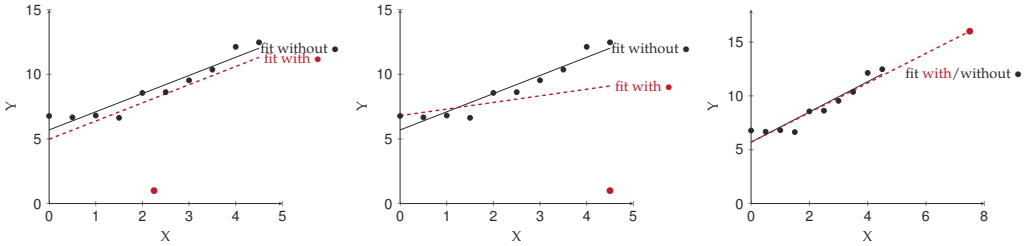
(Reply to Kahn and Udry, *ASR*, October, 1986)

GUILLERMINA JASSO
University of Minnesota

I argue that the estimates of the effects of spouses' ages, marital duration, and contemporaneous period influences on martial coital frequency reported in Jasso (1985) are superior to the outlier-deleting estimates reported in Kahn and Udry (1986)—for chiefly one reason: If the

American Sociological Review: removing cases leads to sample truncation and an unwarranted reduction in variance in the outcome.

Outliers and high leverage cases



Left panel: Outlier, but with low leverage. **Center panel**: Outlier, with high leverage. **Right panel**: High leverage, but not an outlier.

Influence on coefficients

$$\textit{Influence} = \textit{Leverage} \times \textit{Discrepancy} \tag{1}$$

The case in the second panel has high influence (on the regression slope).

The case in the third panel is nevertheless problematic.

$$V(b) = \frac{\sigma_\epsilon^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2}$$

The sampling variance is "artificially" reduced in such cases.

Assessing leverage

"Hat-values" are used.

It's possible to express every $\hat{Y}_i$ as a weighted sum of Y_i.

$$\hat{Y}_j = h_{1j} Y_1 + h_{2j} Y_2 + \cdots + h_{jj} Y_j + \cdots + h_{nj} Y_n \tag{3}$$

Any observation that has a h_{ij} larger than $2 \times \bar{h}$ or $3 \times \bar{h}$, should be considered a high leverage case.

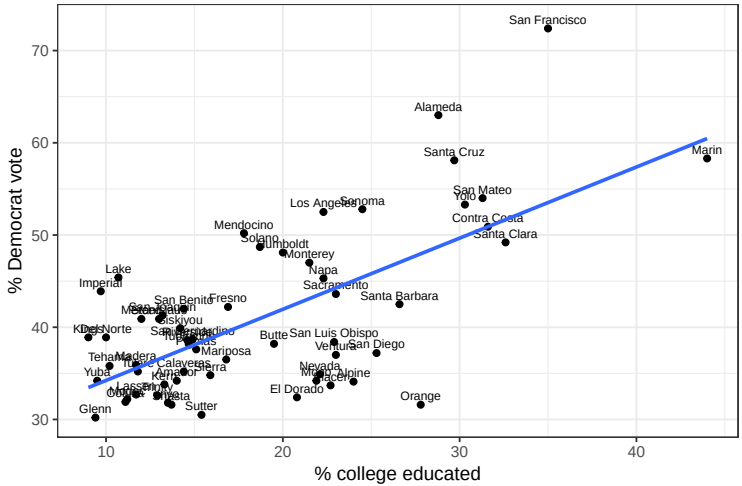
Example: California 1992

Califonia counties in 1992

	DF: perc vote for Democrat
(Intercept)	41.000***
	(0.887)
Share college educated	0.771***
	(0.116)
R^2	0.440
Adj. R^2	0.430
Num. obs.	58

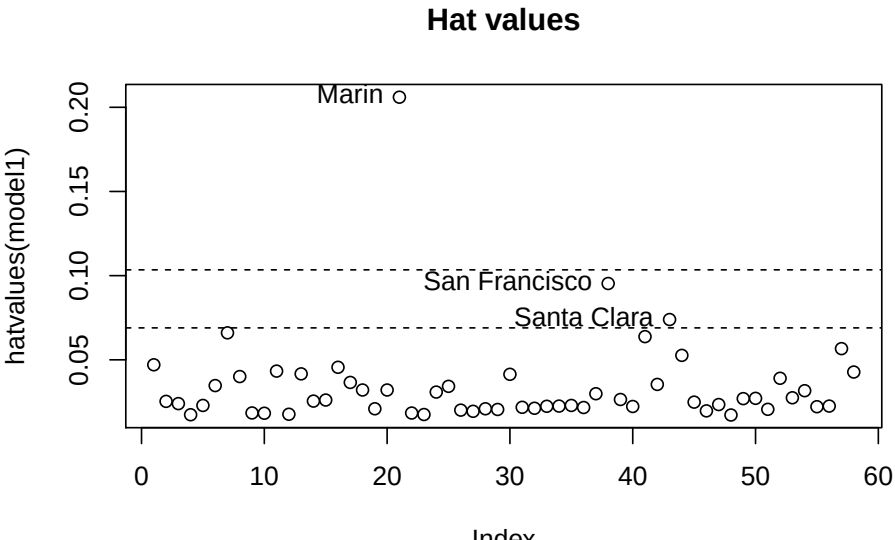
$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Example: California 1992



Bivariate relationship between education and vote

Example: California 1992



Detecting outliers

Studentized residuals:

$$E_i^* = \frac{e_i}{S_{E(-i)}\sqrt{1-h_i}} \tag{4}$$

Computed from:

- ✓ OLS residuals, e_i;
- ✓ standard error of the regression, $S_E = \sqrt{\frac{\sum_{i=1}^{n} \epsilon_i^2}{n-k-1}}$;
- ✓ hat values, h_i.

Detecting outliers

Instead of the regression SE, S_E, we use the regression S_E without the i observation, $S_{E(-1)}$.

This makes the top and bottom part of Equation 4 independent of each other $\Rightarrow E_i^*$ has a t distribution ($n - k - 2$ degrees of freedom).

We're interested in the maximum value of the studentized residual, E_{max}^*.[1]

[1] The procedure is a bit more complicated, involving a *Bonferroni adjustment* to the p value of this E_{max}^*. Specific details can be found in Fox (2008, p. 248).

Example: California 1992

```
outlierTest(model1)

No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
   rstudent unadjusted p-value Bonferroni p
38 3.166318           0.002518      0.14605
```

Observation 38 = San Francisco.

The *Bonferroni-adjusted* p-value suggests that it's not unusual to see a residual of such magnitude in a sample of 58 observations.

Assessing influence

A large number of quantities have been proposed.

$DFBETA_{ij}$: a distance between the OLS estimate with and without a particular observation i in the sample.

A derivate measure for influence is $DFBETAS_{ij}$, which simply standardizes the D_{ij}.

The problem with both is that each measure can be computed for each observation and each predictor.

Cook (1977) introduces a distance measure based on the *standardized* residuals, which applies only to observations: Cook's D.

Cook's D

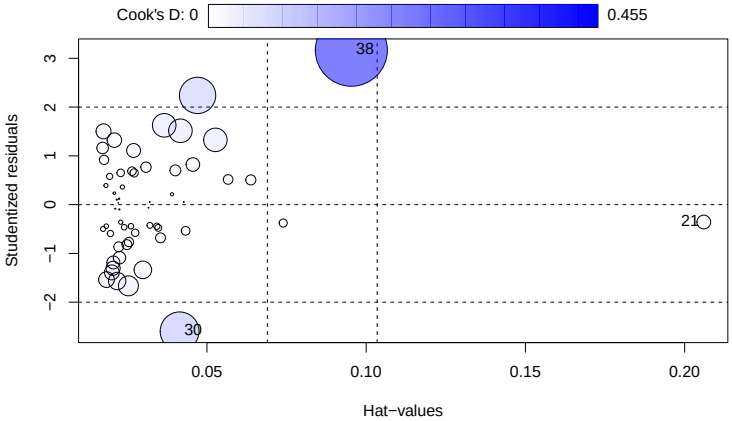
$$D_i = \underbrace{\frac{E_i'^2}{k+1}}_{\text{discrepancy}} \times \underbrace{\frac{h_i}{1-h_i}}_{\text{leverage}} \tag{5}$$

k is the number of predictors in the model, and h_i is the hat-value for observation i. $E_i'^2$ represent the squared standardized residuals[2], where

$$E_i' = \frac{\epsilon_i}{S_E\sqrt{1-h_i}} \tag{6}$$

[2]Belsley, Kuh, and Welsch (2004, p. 15) propose a similar distance measure, called $DFFITS_i$, but computed based on *studentized* residuals.

Bubble plot



"Bubble plot" of hat-values versus studentized residuals (the area of the circle is proportional to Cook's D).

Robust estimation

“Ideal” robust estimator

Ideally, we would like a robust estimator to have 2 characteristics (Mosteller & Tukey, 1977):

- ✓ presence of outliers will not result in change in estimate;
- ✓ presence of outliers will not result in change in efficiency;

ROBUSTNESS: ability of an estimation procedure to discount unusual observations, i.e. to use *all* available information, but weight it differentially.

Breakdown point (BDP)

We have previously talked about bias, consistency, or efficiency of an estimator.

In the context of robustness, there is also the *breakdown point*: smallest fraction of outliers or influential data that an estimator can handle without producing a vastly different result.

Golden standard for BDP = 0.5 (otherwise the estimator is based on a minority of the data).

Sensitivity of the mean

Country	Without China	With China
Seychelles	94677	94677
Greenland	56186	56186
Tajikistan	8734951	8734951
United States	323127513	323127513
Bolivia	10887882	10887882
Suriname	558368	558368
Belize	366954	366954
Guinea-Bissau	1815698	1815698
Mauritius	1263473	1263473
Lithuania	2872298	
China		1378665000
MEAN	34,977,800	172,557,070.2

Population mean with/without China

Sensitivity of the mean & σ

We know that the mean is very sensitive to outliers $\Rightarrow$ it's not a robust estimate for the location of a distribution.

$BDP_{mean} = \frac{1}{n}$, which for $n \to \infty$ is essentially ≈ 0.

The same problem is encountered with σ, as a measure of the scale of a distribution.

$$\sigma_Y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{Y})^2}{n}} \tag{7}$$

Alternatives to mean: α-trimmed mean

Order observations by size:

$$y_{(1)} \leq y_{(2)} \leq \dots y_{(n)} \tag{8}$$

α is a trimming factor. If $\alpha = 0.1$, trim bottom and top 10% of sample, and compute the mean.

Alternatives to mean: median

Has been known to be very resistant to outliers, e.g. its use as an indicator of income.

In more formal terms, it has the best *breakdown point* (BDP=0.5).

However, it is less efficient than the mean.

Alternatives to σ: MD and MDM

Mean deviation from the mean, or $MD = \frac{\sum_{i=1}^{n} |y_i - \bar{Y}|}{n}$.

More efficient than σ for outliers, but $BDP \approx 0$.

Alternatively, there is the *mean deviation from the median*, or $MDM = \frac{\sum_{i=1}^{n} |y_i - M_y|}{n}$.

This roughly exhibits the same flaws as the MD.

Alternatives to σ: q-quantile range & MAD

The difference between specific quantiles of the distribution.

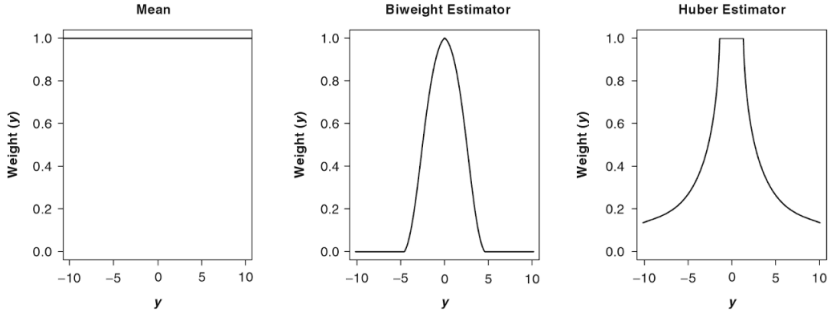
If $q = 0.25$, then we have the middle 50% of observations: the *inter-quartile range*, IQR.

Alternatively, we have the *median absolute deviation*, MAD = $median(|y_i - M_y|)$.

$BDP_{MAD} = 0.5$, which puts it way ahead of the IQR, with a BDP of 0.25.

M-estimation

M-estimation: weight functions



Compared to the mean, the goal of the robust estimators is to downweight the outlying observations.

M-estimation for regression

It borrows some of the same ideas from M-estimation for location, particularly the weighting.

It attempts to minimize a function of the residuals: $\sum_{i=1}^{n} \psi(e_i)$.

Catch-22: we need the model to obtain residuals, but we also need residuals to estimate the model.

M-estimation: IWLS

IWLS: *iteratively weighted least squares*.

- ✓ Fit an OLS and, based on the estimated bs, obtain e_i;
- ✓ Select a weighting function, and construct initial weights w_i based on the e_i;
- ✓ Use WLS to minimize $\sum_{i=1}^{n} w_i(e_i^2)$, and produce a new set of bs.

With these bs obtain a new set of e_i, and the cycle begins anew.

M-estimation: IWLS

The process stops when the difference between successive rounds of bs is lower than a particular relative threshold, e.g. 0.01%.

M-estimators are about 95% as efficient as OLS estimators, and are very robust ($BDP = 0.5$) if tweaked a bit (see below the *MM* type).

However, they are not completely robust to observations with high leverage (as these can produce small residuals).

SEs for *M*-estimators

There are analytical (formula-based) solutions for this, but they tend to be very complex, and biased for small sample sizes.

The preferred solution is to obtain empirical standard errors, through *bootstrapping*.

- ✓ Take a sample size of n 1,000 times from the Xs and Ys (this is naturally done with replacement);
- ✓ For each of these, re-estimate the model, and save the values of bs;
- ✓ The σ of these distributions of bs are the SEs.

MM-estimator

It tackles the difficulty that an *M*-estimator of regression has with *high-leverage* outliers.

It combines the efficiency of a standard estimator (*M*), with the robustness that comes from a *least-trimmed-squares* approach.

$$e^2_{(1)},\ e^2_{(2)},\ \ldots,\ e^2_{(n)} \tag{9}$$

The LTS removes $\frac{1}{2}$ of the largest residuals, and minimizes the sum of the rest.

MM-estimator

Mechanism is the same as the IWLS before.

At the first stage, though, initial estimates for b and e_i are obtained through the LTS approach, instead of an OLS.

In this stage a measure of the scale of the residuals is obtained as well: the MAD of the distribution.

Then the IWLS proceeds as usual, by obtaining regression coefficients that minimize a weighted function of the scaled residuals, $\sum_{i=1}^{n} w_i(\frac{e_i}{MAD_e})$, where Huber weights or biweights are used.

Quantile regression

Examining the whole distribution

Instead of modelling the mean, we can try using the median, as a more robust measure of central location of a distribution.

At the same time, the median is just a special case of quantile.

Quantile regression adopts the same reasoning for predicting a number of quantiles of the distribution, e.g. 0.10, 0.20, . . . 0.90.

Examining the whole distribution

This can show both how the conditional median of the distribution changes with the predictors, as well as how its shape changes.

On the one hand, it can give us a good understanding of how particular populations are impacted by the predictors, e.g. the poorest in society.[3]

On the other, it can also directly tell us how the shape of the distribution changes, maybe toward more inequality between income groups, for example.

[3] Interactions can also do this, so it's not a strong argument.

Quantile regression model (QRM)

The QRM is similar to the linear regression model (LRM), in that it models a continuous outcome with a linear function of predictors.

QRM is different, though, in terms of what it models (conditional quantiles instead of conditional means).

$$Y_i = a^{(p)} + b_1^{(p)} X1_i + b_2^{(p)} X2_i + e_i^{(p)} \tag{10}$$

p is the particular percentile we're trying to obtain estimates for.

Quantile regression model (QRM)

In LRM, the sum of the squares of vertical distances between fit line and observations is minimized ($\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$).

In QRM, it's a weighted sum of absolute vertical distances, where the distances are not squared, and the weights are $1 - p$ for points above the line, and p for points below it.

$$p \sum_{y_i \geq b_0^{(p)} + b_1^{(p)} X_i} |y_i - b_0^{(p)} - b_1^{(p)} X_i| + (1-p) \sum_{y_i < b_0^{(p)} + b_1^{(p)} X_i} |y_i - b_0^{(p)} - b_1^{(p)} X_i| \qquad (11)$$

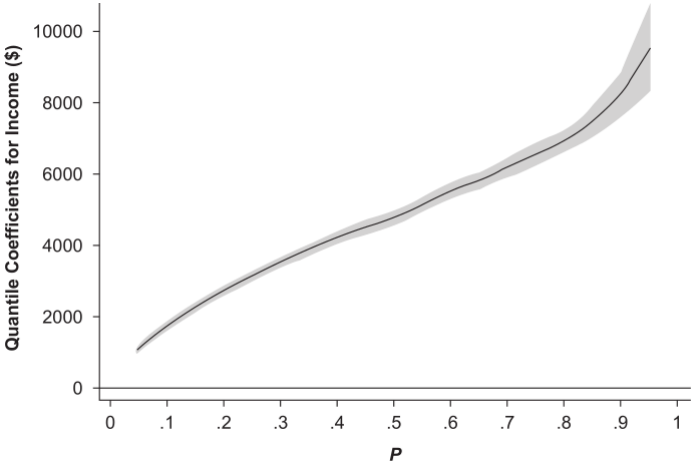
Quantile regression model (QRM)

Even though the estimation takes place for each separate quantile, it uses the entire sample.

Coefficients are interpreted in the same way as for a standard OLS, only that they don't refer to the expectation of Y, rather to the specific quantile of Y.

Measures of uncertainty are obtained through *bootstrapping*.

Presenting results



Predicting quintiles of income using education and race—results for education.

Presenting results

If the line were flat, then we would expect that only the location of the distribution changes, not the scale as well.

With an ascending/descending line, we have evidence that the distribution is spreading out / compressing.

The model estimation gives us even more precise measures of how this shifts occurs, though.

Presenting results

Scale Shifts of Income Distribution
From 11-Year to 12-Year Education

Quantile and Quantile Range	*Sample-Based*			*Model-Based*
	Education = 11 (1)	*Education = 12* (2)	*Difference* (2) − (1)	
$Q_{.025}$	3387	5229	1842	665
$Q_{.05}$	5352	7195	1843	1130
$Q_{.10}$	6792	10460	3668	1782
$Q_{.25}$	12098	18694	6596	3172
$Q_{.75}$	38524	53120	14596	6598
$Q_{.90}$	58332	77422	19090	8279
$Q_{.95}$	74225	95804	21579	9575
$Q_{.975}$	87996	117890	29894	11567
$Q_{.75} - Q_{.25}$	26426	34426	8000	
$\hat{\beta}^*_{.75} - \hat{\beta}^*_{.25}$				3426
$Q_{.90} - Q_{.10}$	51540	66962	15422	
$\hat{\beta}^*_{.90} - \hat{\beta}^*_{.10}$				6497
$Q_{.95} - Q_{.05}$	68873	88609	19736	
$\hat{\beta}^*_{.95} - \hat{\beta}^*_{.05}$				8445
$Q_{.975} - Q_{.025}$	84609	112661	28052	
$\hat{\beta}^*_{.975} - \hat{\beta}^*_{.025}$				10902

Predicting spread of income based on education.

Thank you for the kind attention!

References I

Belsley, D. A., Kuh, E., & Welsch, R. E. (2004). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Hoboken, NJ: Wiley-Interscience.

Cook, R. D. (1977). Detection of Influential Observation in Linear Regression. *Technometrics*, *19*(1), 15–18.

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models* (2nd ed.). Thousand Oaks, CA: Sage.

Mosteller, F., & Tukey, J. W. (1977). *Data Analysis and Regression. A Second Course in Statistics*. Reading, MA: Addison–Wesley.