An Exploration of Influential Observations

In the Panel Study of Income Dynamics

Statistics 499 Honors Research

Xuanzhong Wang

Instructor: Professor Ed Rothman

1. Introduction

In the least square analysis of data based on a full rank linear regression model, an observation may be judged influential if important features of the analysis—i.e. estimated regression coefficients are altered substantially when the observation is isolated from the analysis. In microeconomics researches, influential observations usually can be attributed to various factors. This paper will explore the causes and ways to deal with the influential observations using the Panel Study of Income Dynamics (PSID) data as an example.

The Panel Study of Income Dynamics is a nationally representative longitudinal study of nearly 8,000 US families. Following the same families and individuals since 1968, the PSID collects data on economic health and social behavior. There are many possible factors that can be attributed to the influential observations in the PSID. Information obtained in an interview can always be noisy. Respondents may not know or remember the exact answer to certain types of questions—for example, the amount of money spent on gasoline over the past one year. It is also possible that respondents tried to hide the truth to certain sensitive questions from the interviewer—tax evasion for example. Improperly recorded data can be a cause of influential observations too—either at the stage when people were interviewed or in the process that the data were coded into the data base.

Identifying an influential observation, as well as controlling the effect of such observation is an important step in the regression diagnostics. This project examined the characteristics of an influential observation in least square analysis and used data from the PSID as an example illustrating some ways of eliminating the influence of influential observations.

2. Discussion

In a full rank linear regression model, an observation should be considered as influential if there is substantial change in the estimated parameters of the regression model when the observation is excluded from the analysis. Cook (1979) presented in his paper the following relationship: $\hat{\beta} - \hat{\beta}_{s(i)} = (X^T X)^{-1} x_i^T \frac{(y_i - \hat{y}_i)}{1 - H_{ii}}$ where x_i the ith row of the data matrix X, and H_{ii} is the ith diagonal entry in the hat matrix $H = X\left(X^T X\right)^{-1} X^T$.

The influence of an observation is defined as the amount change in the parameters estimated from a regression model with the observation excluded. Cook's finding in 1979 shows that the influence of an observation is a combination effect of the residual from a full rank regression and the value of H_{ii}.

Holding H_{ii} constant, when the residual from a full rank regression of a specific observation is small, the influence of that observation is small; the opposite can be concluded if the residual is large. Similarly, apart from the effect in influence from the size of residuals, as value of H_{ii} approaches 1, the change in the estimated parameters will approach infinity. In order to identify influential observations, it is important to understand the mechanism behind factors that affect the influence of an observation. In this project, the H_{ii} value was of great interest.

2.1-a) Simple Linear Regression Case (Derivation)

First, consider a simple linear regression shown below,

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, this can be written as $y = X\beta + \varepsilon$ where

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},\ X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix},\ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \text{ and } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

The hat matrix H is:

$$H = X\left(X^T X\right)^{-1} X^T = \frac{1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix} \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix}$$

And the i[th] diagonal entry in the hat matrix is: $H_{ii} = \dfrac{\sum_{i=1}^{n} x_i^2 - 2x_i \sum_{i=1}^{n} x_i + n x_i^2}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$

Let $x_i = \bar{X} + d_i$, where d_i is the deviation of x_i from the mean of X and it can be shown that:

$$H_{ii} = \frac{\sum_{i=1}^{n} x_i^2 - 2(\bar{X} + d_i)\sum_{i=1}^{n} x_i + n(\bar{X} + d_i)^2}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$= \frac{\sum_{i=1}^{n} x_i^2 - 2\bar{X}\sum_{i=1}^{n} x_i - 2d_i\sum_{i=1}^{n} x_i + n(\bar{X})^2 + n(d_i)^2 + 2nd_i\bar{X}}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$= \frac{\sum_{i=1}^{n} x_i^2 - 2n(\bar{X})^2 - 2d_i\sum_{i=1}^{n} x_i + n(\bar{X})^2 + n(d_i)^2 + 2d_i\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$= \frac{\sum_{i=1}^{n} x_i^2 - n(\bar{X})^2 + n(d_i)^2}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{1}{n} + \frac{n(d_i)^2}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{1}{n} + \frac{(d_i)^2}{n(Var(X))}$$

The above identity shows that the H_{ii} value of a simple linear regression model depends on three factors listed below:

1) Total number of observations in a full rank regression

   As n goes top infinity, H_{ii} value of each observation goes to 0. This is because $\sum_{i=1}^{n} H_{ii} = p$ and p is a constant pre-determined by the number of independent variables in the analysis.

2) The deviation of the X value of this observation from the mean of X

   If x_i is close to the mean of X, then H_{ii} will be small and vice versa.

3) The variance in X

When the variance in X is large, H_{ii} will be small as opposite to the situation when variance in X is small.

Similarly, the identity can be written as:

$$H_{ii} = \frac{1}{n} + \frac{\text{variance of X contributed by } x_i}{\text{Var(X)}}$$

$$= \frac{1}{n} + \text{proportion of variance in X contributed by } x_i$$

2.1-b) Simple Linear Regression Case (Simulation Example)

To varify what was obtained above, a simulation study was conducted. In the simulation study below, eight different cases will be considered. Table 1 summarized the condition of each case and the corresponding estimates in each case.

Table 1 Summary of the result from a simulation study

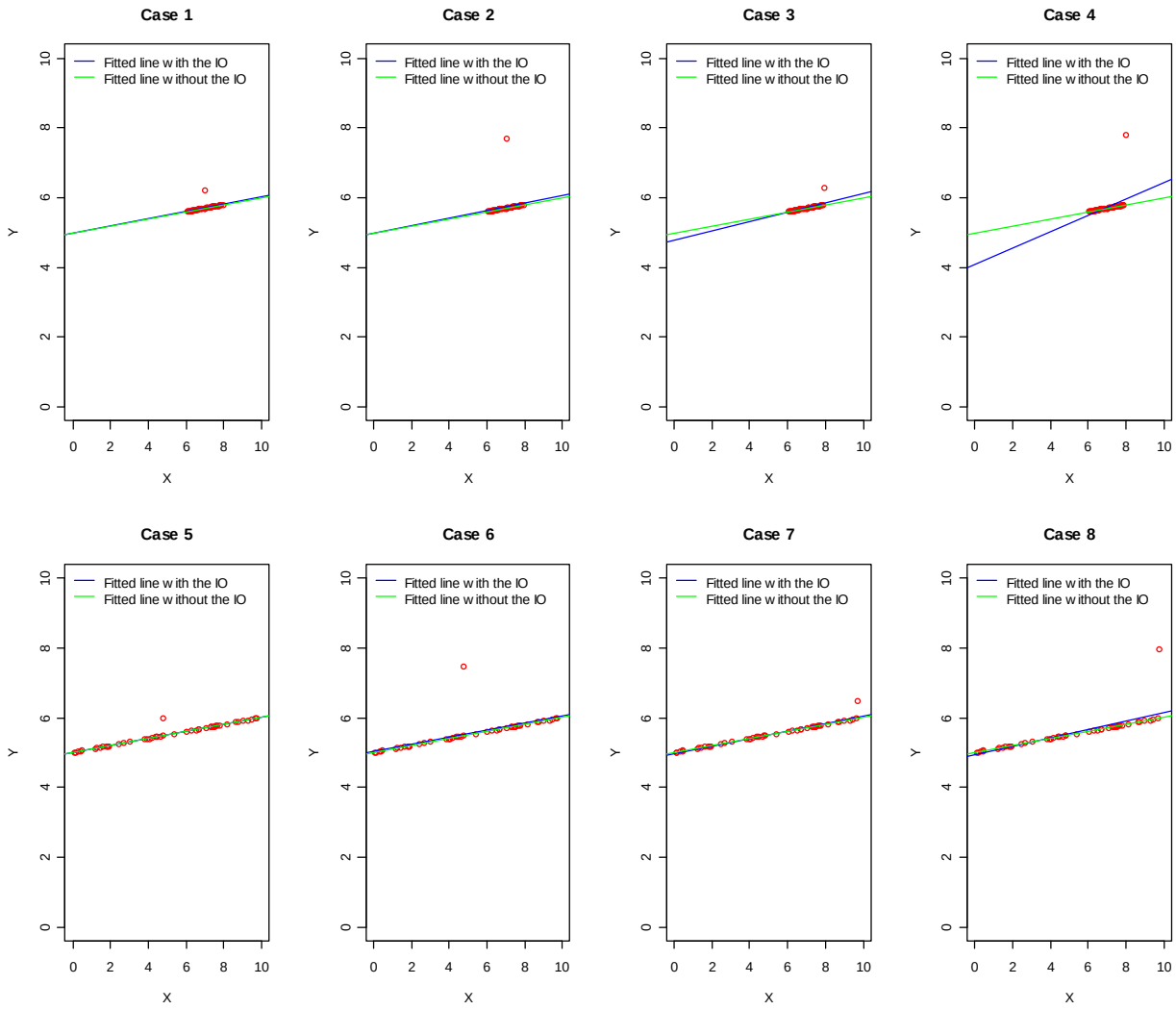
	Var(X)	x_i from $\bar{X}$	$y_i - \hat{y}_i$	Estimate of β_0 with the observation removed	Estimate of β_0 from a full rank regression	Estimate of β_1 with the observation removed	Estimate of β_1 from a full rank regression
Case 1	small	close	small	5	4.997409	0.1	0.1018085
Case 2			large		4.98632		0.1077104
Case 3		far	small		4.773075		0.1340312
Case 4			large		***4.088985***		***0.2366008***
Case 5	large	close	small		5.010946		0.0998395
Case 6			large		5.044802		0.09907133
Case 7		far	small		4.982216		0.1055619
Case 8			large		4.929882		0.1219611

From Table 1, it can be seen that when x_i is close to $\bar{X}$, meaning that x_i only contribute a minute amount of variation to the total variation in X, the influence of the observation is small. The influence of an observation is most significant when a large proportion of variance in X is contributed by X and the residual from a full rank regression is large.

Chart 1 presents a graphic representation of each case. The simulated observations are in represented by red circles. Most (49/50) of the red circles are located along the straight line and there is only one point that deviates from the trend of the simulated points. The fitted line with a full rank regression is plotted in blue and the fitted line from the regression with the influential observation removed is plotted in green.

The deviation of the green line from the blue one indicates the influence of the observation. As Table 1 summarized, Case 4 has the most influential point as the discrepancy between the blue line and the green line is most significant. Thus the criteria for an influential observation in a simple linear regression model were verified.

Chart 1 A graphic representation of the simulation study



2.2-a) Multiple Linear Regression with Two Predictors (Derivation)

Next, consider a simple case of multiple linear regressions. Suppose there is a regression model $y_i' = \beta_0 + \beta_1 x_i' + \beta_2 z_i' + \varepsilon_i'$, i.e.

$$Y' = X'\beta + \varepsilon' \text{ Where } Y' = \begin{pmatrix} y_1' \\ \vdots \\ y_n' \end{pmatrix}, X' = \begin{pmatrix} 1 & x_1' & z_1' \\ \vdots & \vdots & \vdots \\ 1 & x_n' & z_n' \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \text{ and } \varepsilon' = \begin{pmatrix} \varepsilon_1' \\ \vdots \\ \varepsilon_n' \end{pmatrix}$$

If the following transformations are taken,

$$y_i = \frac{y_i' - \bar{y}'}{sd(y_i')}, x_i = \frac{x_i' - \bar{x}'}{sd(x_i')}, z_i = \frac{z_i' - \bar{z}'}{sd(z_i')} \text{ and } \varepsilon_i = \frac{\varepsilon_i' - 0}{sd(\varepsilon_i')}$$

The model can be re-written as $y_i = \beta_1 x_i + \beta_2 z_i + \varepsilon_i$, i.e. $Y = X\beta + \varepsilon$

$$\text{Where } Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} x_1 & z_1 \\ \vdots & \vdots \\ x_n & z_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \text{ and } \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

The following relationships are observed:

$$\sum_i^n x_i^2 = \sum_i^n \left(\frac{x_i' - \bar{x}'}{sd(x_i')} \right)^2 = \frac{1}{\text{var}(x_i')} \sum_i^n \left(x_i' - \bar{x}' \right)^2 = n - 1$$

$$\text{Similarly } \sum_i^n y_i^2 = \sum_i^n z_i^2 = n - 1, \sum_i^n x_i z_i = \sum_i^n \left(\frac{x_i' - \bar{x}'}{sd(x_i')} \right) \left(\frac{z_i' - \bar{z}'}{sd(z_i')} \right) = r(n - 1)$$

Where r is the Pearson Correlation Coefficient

$$\left(X^T X\right)^{-1} = \frac{1}{\sum_i^n x_i^2 \sum_i^n z_i^2 - \left(\sum_i^n x_i z_i\right)^2} \begin{pmatrix} \sum_i^n z_i^2 & -\sum_i^n x_i z_i \\ -\sum_i^n x_i z_i & \sum_i^n x_i^2 \end{pmatrix}$$

$$= \frac{1}{(n-1)^2 - r^r (n-1)^2} \begin{pmatrix} n-1 & -r(n-1) \\ -r(n-1) & n-1 \end{pmatrix}$$

$$= \frac{1}{(n-1)(1-r^2)} \begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix}$$

$$H = X\left(X^T X\right)^{-1} X = \frac{1}{(n-1)(1-r^2)} \begin{pmatrix} x_1 & z_1 \\ \vdots & \vdots \\ x_n & z_n \end{pmatrix} \begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix} \begin{pmatrix} x_1 & \cdots & x_n \\ z_1 & \cdots & z_n \end{pmatrix}$$

The diagonal entries of the Hat matrix H are:

$$H_{ii} = \frac{x_i^2 + z_i^2 - 2r x_i z_i}{(n-1)(1-r^2)}$$

The above result shows that holding residuals constant, the following factors will have an effect in the influence of an observation:

1) The larger the deviation that the observation lies from the mean of all its predictors, the larger the H_{ii} value, i.e. the larger the influence of an observation

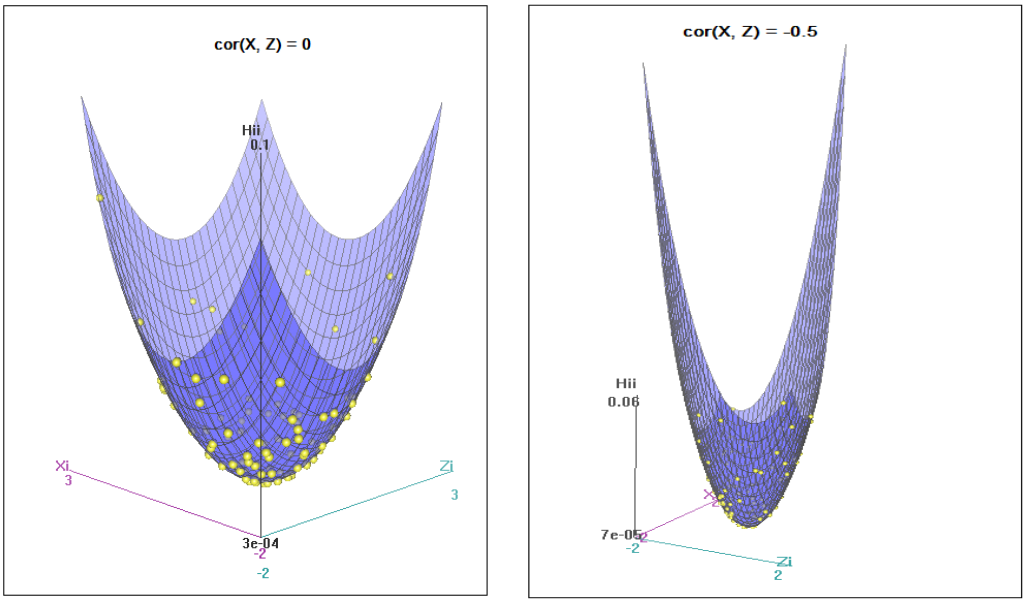
2) The correlation coefficient between the predictors of this observation, the larger the correlation coefficient, the larger the influence will be. The influence will be amplified especially when the observation has two predictors opposite the trend of correlation between them. For example, the two predictors are positively correlated overall, then observations with predictor pairs that appear to be negatively correlated will have larger H_{ii} value.

3) The larger the sample size, the smaller the H_{ii} value, thus the influence.

2.2-b) Multiple Linear Regression with Two Predictors (Simulation Study)

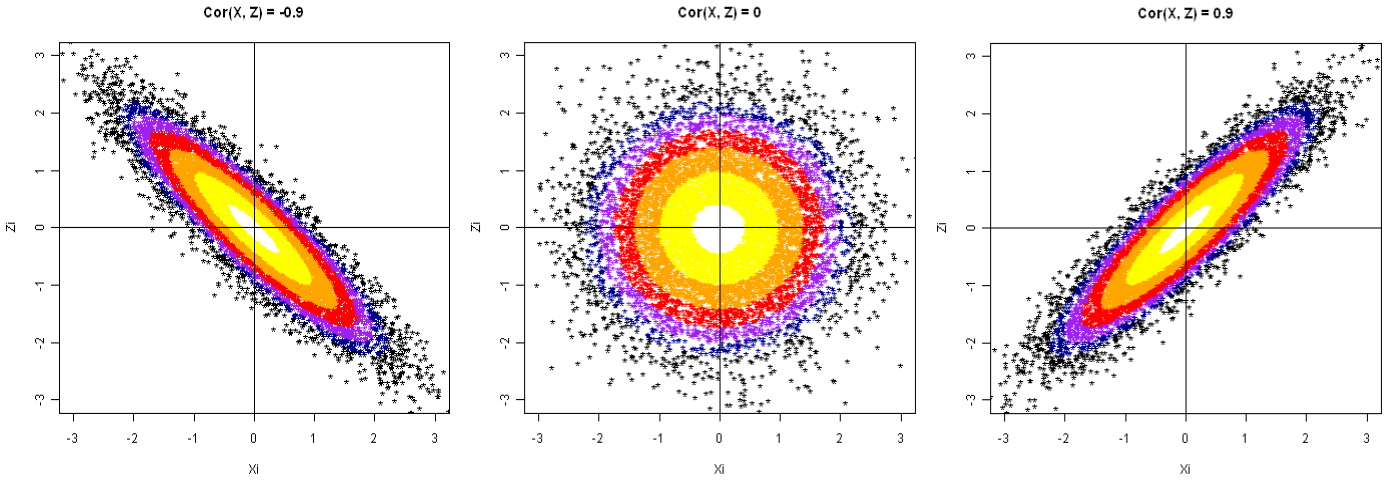
A simulation study was conducted to verify the first two factors. 100 observations were randomly generated from a population where X and Z have correlation coefficient r. Xi and Zi are the corresponding standardized scores of X and Z, the corresponding H_{ii} values are calculated and plotted in the 3-D space. The blue surface indicates the H_{ii} value corresponding to each combination of the Xi and Zi pair.

Chart 2 Simulated H_{ii} Surfaces in 3-D Space



Both Xi and Zi have mean 0 and standard deviation of 1. When Xi and Zi are close to 0, the H_{ii} value will be small. If the H_{ii} values are plotted against corresponding Xi, Zi pairs in a 3 dimensional space as shown in Chart 2, a paraboloid can be observed. The minimum point of the paraboloid locates at the mean of (X_i, Z_i), i.e. when both Xi and Zi equals to 0. In the plot when the correlation coefficient between X and Z is -0.5, a moderate negative correlation, it is obvious that along the line of X=Z, the H_{ii} value increase the fastest when Xi and Zi start to deviate from their means. This shows the influence has been amplified when the observations fail to follow the correlation pattern between predictors.

Chart 3 (X_i, Z_i) Pairs Color Coded By H_{ii}



Next, 10,000 pairs of Xi and Zi values are simulated for different correlation coefficients r. The corresponding H_{ii} value to each Xi Zi pair was then calculated. In Chart 3, when Zi is plotted against Xi, the points are color coded by H_{ii} values, it is obvious that when the correlation coefficient between X and Z is 0, the equal H_{ii} contours are circles centered at $(X_i, Z_i) = (0,0)$, and radius $\sqrt{H_{ii}(n-1)}$. This shows that when X and Z are not correlated, the distance of an observation from the means of both predictors will determine the H_{ii} value of that observation, and in turn, the influence of the observation. It also can be observed that when X and Z are strongly correlated, either negatively or positively, the H_{ii} value increase more rapidly in the direction opposite the general trend of all observations.

2.3) Generalizations

In order to generalize the result above, the following model will be examined:

$Y = X\beta + \varepsilon$ where $X = \begin{pmatrix} x_{11} & x_{21} & x_{31} \\ \vdots & \vdots & \vdots \\ x_{1n} & x_{2n} & x_{3n} \end{pmatrix}$, there are three predictors in this model and all of them are standardized in order to simplify the analysis. It can be shown that

$$X^TX = (n-1)\times\begin{pmatrix}1 & \rho_{12} & \rho_{13}\\ \rho_{12} & 1 & \rho_{23}\\ \rho_{13} & \rho_{23} & 1\end{pmatrix}$$ where ρ_{ij} indicates the correlation between predictor i and j.

$$(X^TX)^{-1} = \frac{1}{(n-1)(1-\rho_{12}^2-\rho_{13}^2-\rho_{23}^2+2\rho_{12}\rho_{13}\rho_{23})}\times\begin{pmatrix}1-\rho_{23}^2 & \rho_{13}\rho_{23}-\rho_{12} & \rho_{12}\rho_{23}-\rho_{13}\\ \rho_{13}\rho_{23}-\rho_{12} & 1-\rho_{13}^2 & \rho_{12}\rho_{13}-\rho_{23}\\ \rho_{12}\rho_{23}-\rho_{13} & \rho_{12}\rho_{13}-\rho_{23} & 1-\rho_{12}^2\end{pmatrix}$$

And the corresponding value H_{ii} is

$$H_{ii} = \frac{(1-\rho_{23}^2)X_{1i}^2+(1-\rho_{13}^2)X_{2i}^2+(1-\rho_{12}^2)X_{3i}^2+2X_{1i}X_{2i}(\rho_{13}\rho_{23}-\rho_{12})+2X_{1i}X_{3i}(\rho_{12}\rho_{23}-\rho_{13})+2X_{2i}X_{3i}(\rho_{12}\rho_{13}-\rho_{23})}{(n-1)(1-\rho_{12}^2-\rho_{13}^2-\rho_{23}^2+2\rho_{12}\rho_{13}\rho_{23})}$$

From the above result, similar conclusion can be made—the sample size, variation in predictors, correlation between predictors, as well as observations that has opposite trend with the correlation will inflate the influence.

On a separate note:

$$X^TX = \begin{pmatrix}\sum_{i=1}^n X_{1i}^2 & \sum_{i=1}^n X_{1i}X_{2i} & \cdots & \sum_{i=1}^n X_{1i}X_{(p-1)i} & \sum_{i=1}^n X_{1i}X_{pi}\\ \sum_{i=1}^n X_{1i}X_{2i} & \sum_{i=1}^n X_{2i}^2 & \cdots & \sum_{i=1}^n X_{2i}X_{(p-1)i} & \sum_{i=1}^n X_{2i}X_{pi}\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ \sum_{i=1}^n X_{1i}X_{(p-1)i} & \sum_{i=1}^n X_{2i}X_{(p-1)i} & \cdots & \sum_{i=1}^n X_{(p-1)i}^2 & \sum_{i=1}^n X_{(p-1)i}X_{pi}\\ \sum_{i=1}^n X_{1i}X_{pi} & \sum_{i=1}^n X_{2i}X_{pi} & \cdots & \sum_{i=1}^n X_{(p-1)i}X_{pi} & \sum_{i=1}^n X_{pi}^2\end{pmatrix}$$

$$= (n-1)\times\begin{pmatrix}1 & \rho_{12} & \cdots & \rho_{1(p-1)} & \rho_{1p}\\ \rho_{12} & 1 & \cdots & \rho_{2(p-1)} & \rho_{2p}\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ \rho_{1(p-1)} & \rho_{2(p-1)} & \cdots & 1 & \rho_{p(p-1)}\\ \rho_{1p} & \rho_{2p} & \cdots & \rho_{p(p-1)} & 1\end{pmatrix}$$

When all predictors in X are standardized, the matrix X^TX gives the correlation matrix.

When all predictors in X are orthogonal, i.e. all entries in X^TX except diagonal entries equal to 0,

$$X^TX = (n-1) \times \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

$H_{ii} = \frac{1}{(n-1)} \sum_{j=1}^{p} X_{ji}^2$ which means that H_{ii} only depends on total variation that the observation contributes to the variation in each predictor.

This shows that to control the influence of influential observations, several methods can be taken: increasing the sample size—this is the most straightforward approach, since $E(H_{ii}) = \frac{p}{n}$, larger sample size reduces overall influence of an observation. However, this approach seldom works in reality for increasing sample size cannot be realized without increasing cost which study institutions try to avoid.

Increasing the variability in predictors also helps to reduce influence. Say age is considered an independent variable in a certain analysis, in the data collection process, researchers should be aware that a larger age range should be covered as to increase variability in predictors. However, this is not possible with a give data set. Besides that, avoiding highly correlated predictors is important in controlling for influence: high correlation between predictors inflates the H_{ii} value and instability in the inverse of the matrix X^TX. In the next section, an example using the PSID data will be presented as an illustration of possible factors that affect the influence of an observation and ways of dealing with influential observations.

3. Example from the PSID

In a complex research project—like the PSID, rarely is the case that various predictors are uncorrelated; instead, many predictors are strongly correlated and highly influential points are likely to occur in those cases. In this section, an example will be drawn from the PSID to figure out the various causes of influential observations and a discussion of methods to deal with those observations will be discussed.

Suppose the interest of study is to examine the relationship between amount of investment in child's education and parent's income and variation in family income as well as other characteristics in children. The data used are extracted from the PSID Child Development Supplement (CDS) wave II interview in 2002. The data set used for this analysis consists of two parts. Part 1 is the Child file that includes variables of child's age, sex, type of school attended, parents' expectation in the child and total school related expenditures (school cost, school supplies and extra lessons) reported by the primary care giver (PCG) of the child being interviewed. Part 2 is the family file that contains the household information extracted from the PSID core data including the family money income—the sum of all taxable income, transfer income and social security income of members in the family in 2002; the variation in family income estimated by the standard deviation in family money income from year 1984 to 2002 (in 2002 dollars); and the annual property tax paid in 2002. The Child file and the Family File were merged using the Family Identification Mapping System (FIMS). A list of variables is shown in Table 2. Case wise deletion was applied to observations with missing values, resulting a sample size of 2772.

Table 2 List of Variables Used (Total Number of Observations n = 2772)

	Variable Name	Remarks		Variable Name	Remarks
Child File	Sex	1: Male; 0: Female	**Family File**	Total Family Income in 2002	Ranges from -99260 to 2069000 dollars in 2002
	Age	5-18 Years Old			
	Type of School Attended	0: Not in School; 1: Public School; 2: Private School; 3: Attend School at Home		Variation in Income	Ranges from 270.7 to 760700 dollars (in 2002 dollars)
	Expectation	0-8 where 0 is lowest and 8 is highest		Total Property Tax	Ranges from 0 to 19800 dollars in 2002
	School Related Expenditure	Ranges from 0 to 21340 dollars in 2002			

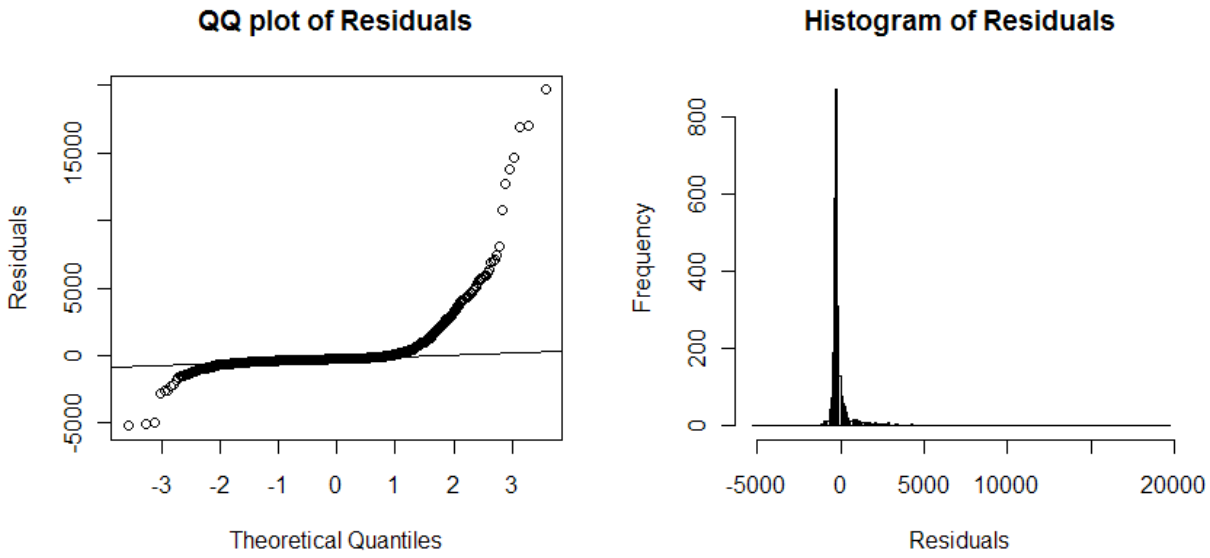
A general approach of this type of research question is usually regression analysis. Suppose in this case, a multiple regression model with two predictors was constructed as shown below:

Model 1:

$$\text{School Related Expenditure}_i = \beta_0 + \beta_1 \text{Family Income}_i + \beta_2 \text{Variation in Family Income}_i + \varepsilon_i$$

The coefficients of the model estimated using ordinary least squares are presented in the Table 3, a summary of regression models. Chart 4 shows that the normality assumption of residuals was violated and box-cox transformation was applied to the model. A log transformation was made to the response variable yielding Model 2 shown below.

Chart 4 Diagnostic Plots of Model 1



Model 2:

$$\log(\text{School Related Expenditure}_i) = \beta_0 + \beta_1 \textit{Family Income}_i + \beta_2 \textit{Variation in Family Income}_i + \varepsilon_i$$

Both Model 1 and Model 2 are naïve approaches to the question; such approaches were subject to the influence of influential data points. As discussed in the previous section, the influence of an observation depends on both the estimated residuals in a full rank linear regression and the H_{ii} value as well. In Chart 5, a scatter plot of Family Income against Variation in Family Income was

shown and the color of the points indicates the corresponding H_{ii} value—the darker the color, the larger the H_{ii} value. Similarly a scatter plot of Family Income against Variation in Family Income color coded by residual size is presented in Chart 6.

Chart 5 A Scatter Plot of Family Income vs Variation in Family Income (Color Coded by H_{ii})

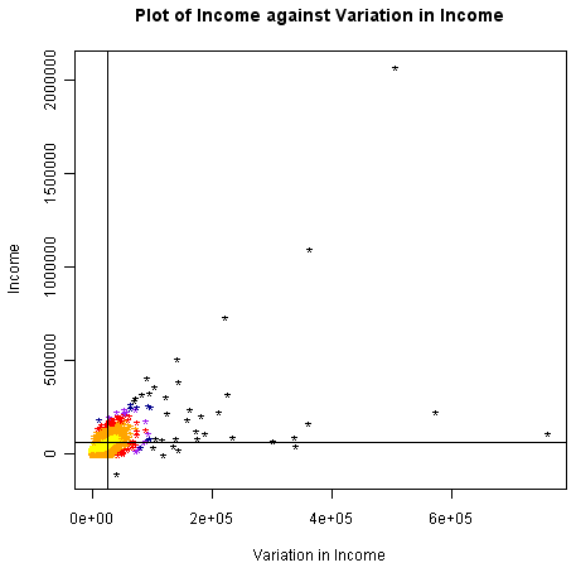
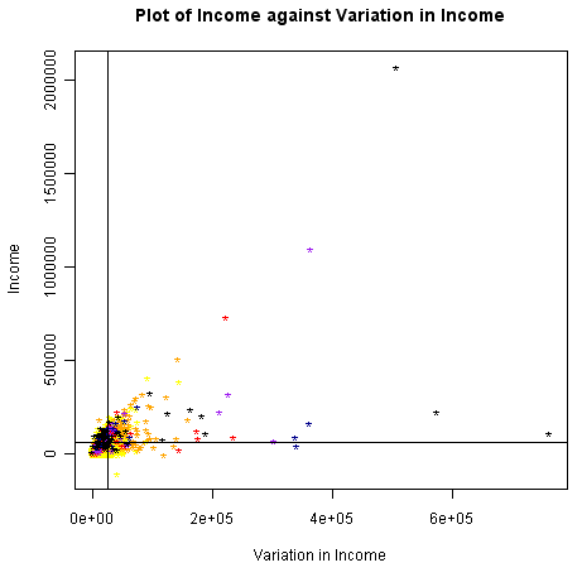
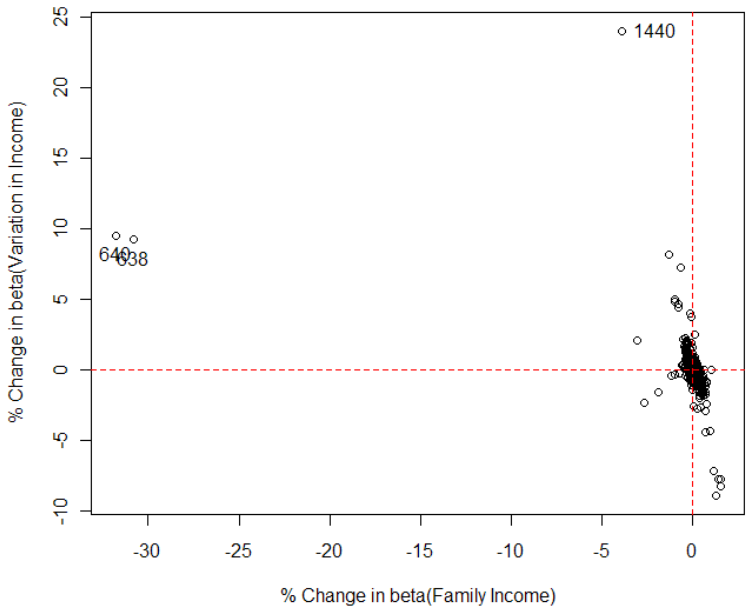


Chart 6 A Scatter Plot of Family Income vs Variation in Family Income (Color Coded by Residuals)



From Chart 5 and Chart 6, it can be seen that there are several observations that have both large H_{ii} values and large residuals. Also, there is a strong positive correlation between family income and variation in family income—the two independent variables. Thus, we should anticipate influential observations in the regression analysis. Next, 2772 regression models were estimated, each with only one observation isolated from the model. This produced 2772 sets of estimated coefficients. The newly estimated coefficients were then compared to the full rank linear regression estimates. In Chart 7, the percent change in estimated coefficients of variation in family income is plotted against the percent change in estimated coefficients of family income.

Chart 7 A Scatter Plot of Percent Change in Estimated Coefficients of Variation in Family Income vs. Percent Change in Estimated Coefficients of Family Income



From Chart 7, it can be seen that most of the observations resulted only less than 10% change in estimated coefficients when they were isolated from the regression analysis. However, at the same time, three observations labeled as 638, 640 and 1440 resulted almost 30% change in one of the estimated coefficients. Despite a large sample size of 2772, influential observations affected the analysis substantially. This can be a great concern especially in cases when such analysis was used to aid policy making,

In the case of the example, most of the observations clustered close to the means of the two predictors; there are very few observations lying far away from the means. In another sense, when the more influential data points located further from the means are removed from the regression analysis, the coefficient of a predictor can be almost anything. Thus, it is the more influential data points that determine the estimated coefficient of the regression model. However, the result obtained may not make any economic sense. In this case, the influential observations are those children in a family unit with exceptionally high family income and therefore, the pattern of spending in child's education may be different from those in moderately rich households. It is

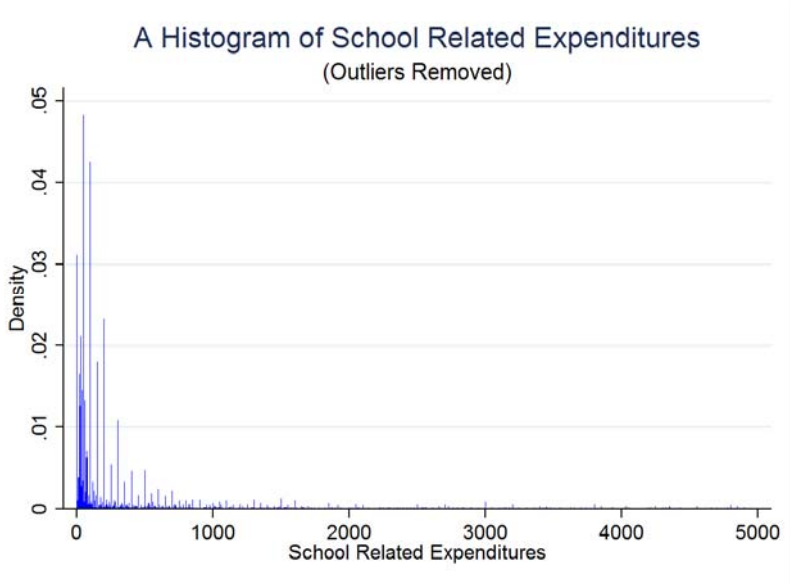
inaccurate to let the extremely wealthy families to influence the regression results without taking any additional measure.

To reduce the influence of those influential observations, it is necessary to figure out the source of influence of the observations. First, as mentioned before, an association between predictors will inflate the influence. From Chart 5, it is obvious that there is a strong positive correlation between family income in 2002 and variation in family income—usually wealthy families are those whose wealth build up rapidly and thus, there is a larger variation in family income over the years comparing to those households whose earning remained relatively unchanged—in terms of purchasing power of earning. This strong association is a source of highly influential observations.

Also, large residuals from a full rank regression are the source of influential observations. Large residuals are sometimes due to improperly recorded data, it can also be attributed to the fact sometimes people did not report the right amount. Using the same example, in this case, parents have to report the amount of money that they spend on the very specific child being interviewed. Several explanations can be offered for mis-reporting the amount:

The interview of the PSID CDS is in a retrospective manner—i.e. they interview is carried out in 2003 for the 2002 wave, so parents would have to recall the amount and this may cause some discrepancies between the real amount and reported amount. Chart 8 shows a histogram of reported school related spending (with outliers eliminated). It reflects a pattern that parents report at whole intervals of hundreds or thousands instead of specific amount simply because the parents may not be able to remember how much they spend and they started guessing or making up numbers in the interview. It is also possible that parents overstates the amount of money that is allocated to the child's education because spending more on education is something that's both socially and morally desirable. The incentive for parents to overstate their spending on child's education is beyond the study of this project and will not be further discussed. Such pattern is possibly a cause of large residuals in a full rank regression and thus large influence.

Chart 8 Histogram of School Related Expenditures



With the causes of influential observations—strong correlation between family income and variation in family income, and large residuals from full rank regression due to mis-reporting by parents; it is possible to work out several measure that can reduce the influence of influential observations and reach a conclusion that makes more statistical and economic sense.

Method 1

One cause of highly influential observations in the previous example is extreme value in family income. With a progressive tax system in the United State, taxation helps to reduce the disparity in income by taxing more heavily on high income families. By using disposable income, that is income after tax as a predictor instead of taxable income will move extreme family income observations close to the mean family income standard and can possibly reduce the influence of those influential data points. The model will be modified as

Model 3:

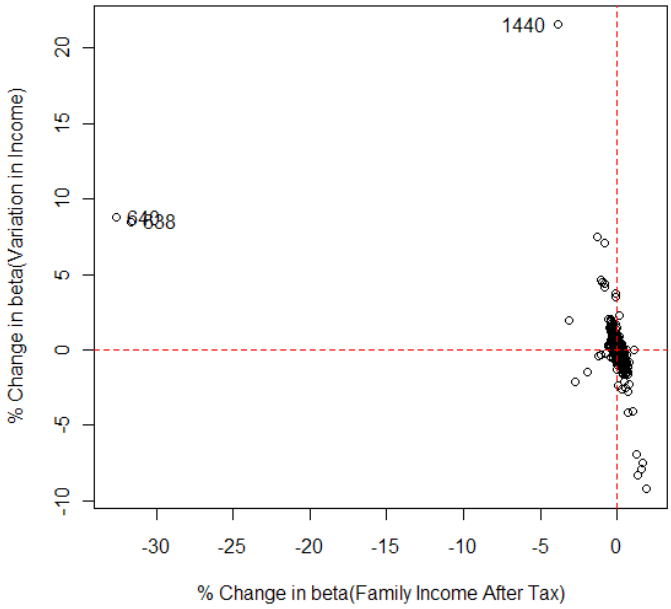
$$\log(School\ \ Related\ Expenditure_i) = \beta_0 + \beta_1 Family\ Disposable\ Income_i + \beta_2 Variation\ In\ Family\ Income_i + \varepsilon_i$$

Where $\text{Family Disposable Income}_i = \text{Family Income}_i - \text{Annual Tax}_i$

A log transformation of the response variable was used to avoid violation of normality assumption in residuals.

Comparing the summary of Model 3 and Model 2 at the end of this paper, it can be seen that there is not much difference in terms of goodness of fit between the two models. Approximately 5.8% of the variations in the response variables were explained by the predictors in Model 2 and 5.6% of the variations in same response variables were explained by the predictors in Model 3. Therefore, any improvement in influential observations will make Model 3 a more desirable choice in analysis than Model 2. The percent change in estimated coefficients of variation in family income is plotted against the percent change in estimated coefficients of family income after tax in Chart 9.

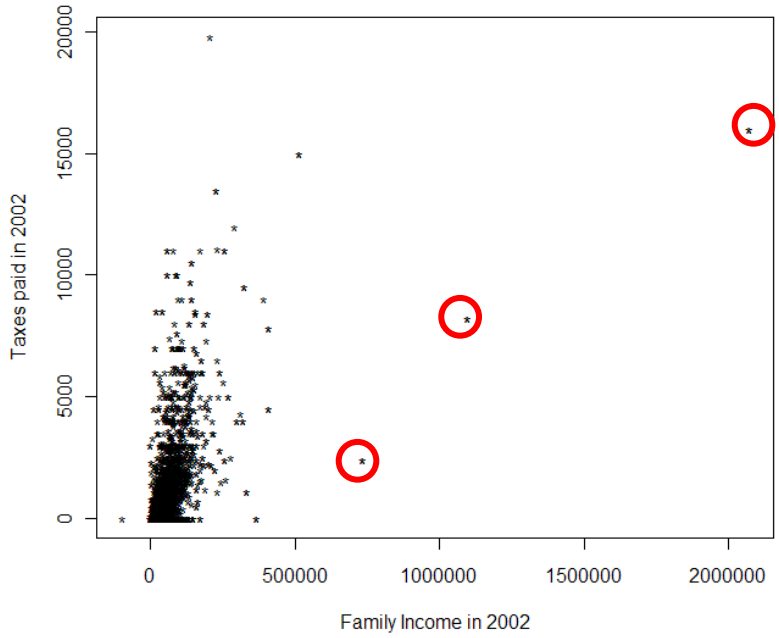
Chart 9 A Scatter Plot of Percent Change in Estimated Coefficients of Variation in Family Income vs. Percent Change in Estimated Coefficients of Disposable Family Income (After Tax)



Comparing Chart 7 with Chart 9, it can be seen that the layout of the effect of influential observations did not improve significantly. In Chart 10, taxes paid in 2002 were plotted against total family income. The observations circled by red circles indicate observations that had very high income but disproportionally low amount of tax paid. This shows that tax evasion existed in such

families and therefore, tax did not always work well in order to bring extremely high income closer to the average incomes. Thus, Method 1 that tried to reduce the variation of predictors contributed by highly influential observations did not work well.

Chart 10 A Scatter Plot of Taxes Paid in 2002 vs. Total Family Income in 2002



Method 2

In Method 1, it was found that although taxation serves as an instrument whose purpose was to bring down extremely high incomes to be close to their means, it was not an effective one—thus the influence reduction was not effective. Consider the strong positive correlation between family income and variation in income; it is possible to predict family income given variation in family income over the past years. By using predicted family income as a single predictor **(Model 4)**, i.e. the use of an instrumental variable, the influence of an extreme data points can be reduced.

$$\underset{}{\textit{Estimated Family Income}_i} = \underset{(1470^*)}{27,885} + \underset{(0.035^*)}{1.42} \times \textit{Variation In Family Income}_i$$

R-Squared=0.3791

The above regression results shows that approximately 38% of the variation in family income can be explained by the variation in the predictor—variation in family income over the past years. When family income is predicted using the model above, it can be used as a predictor for school related expenditure on child.

Model 4:

$$School\ \ Related\ Expenditure_i = \beta_0 + \beta_1 Estimated\ Family\ Income_i + \varepsilon_i$$

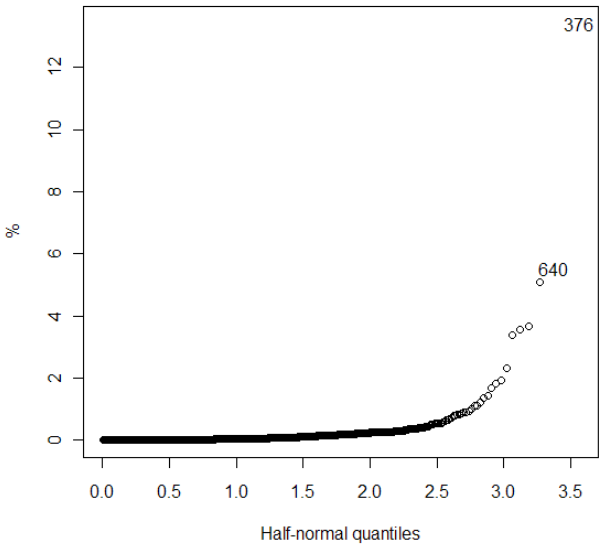
Diagnostics again shows that a log transformation of the response variable was necessary and resulted Model 5.

Model 5:

$$\log(School\ \ Related\ Expenditure_i) = \beta_0 + \beta_1 Estimated\ Family\ Income_i + \varepsilon_i$$

When the instrumental variable is used, there is a slight compromise in R-square comparing to the example before. Only 3.4% of the variation in the response variable was explained by the predictors in Model 5 as opposed to 5.8% in Model 2.

Chart 11 Half Normal Plot of Percent Change in Estimated Coefficients of Estimated Family Income



However, in Chart 11 it can be seen that the percent change in estimated coefficients has decreased to around 12% for the maximum. More interestingly, the previously three most influential observations-638, 640 and 1440 were nowhere to be seen among the most influential ones in Model 5. The use of instrumental variables is then an effective way of reducing influences of extreme data points. The reduction of influence by the use of instrumental variables should be attributed to the regression to the mean effect. In this case, each predicted family income is the mean of population family income given the variation in family income. Uncertainties and deviations from the means are removed in the predicted value—this on one hand, sacrifice some of the information contained in the original data points, causing a compromise in R-square; on another hand, has successfully reduce the influence.

In order to verify the influence reduction effect of using instrumental variables, the following models were estimated and compared.

Model 6:

$$\begin{aligned}&School\ \text{Re}lated\ Expenditure_i = \\ &\beta_0 + \beta_1 Family\ Income_i + \beta_2 Variation\ In\ Family\ Income_i + \beta_3 Sex\ Of\ Child_i + \beta_4 Age\ Of\ Child_i + \beta_5 Expectation_i \\ &+\beta_6 Public\ School_i + \beta_7\ \text{Pr}ivate\ School_i + \beta_8 Home\ School_i + \beta_9 Total\ Number\ Of\ Children\ In\ Family_i + \varepsilon_i\end{aligned}$$

Model 7:

$$\begin{aligned}&\log(School\ \text{Re}lated\ Expenditure_i) = \\ &\beta_0 + \beta_1 Family\ Income_i + \beta_2 Variation\ In\ Family\ Income_i + \beta_3 Sex\ Of\ Child_i + \beta_4 Age\ Of\ Child_i + \beta_5 Expectation_i \\ &+\beta_6 Public\ School_i + \beta_7\ \text{Pr}ivate\ School_i + \beta_8 Home\ School_i + \beta_9 Total\ Number\ Of\ Children\ In\ Family_i + \varepsilon_i\end{aligned}$$

Model 8:

$$\begin{aligned}&School\ \text{Re}lated\ Expenditure_i = \\ &\beta_0 + \beta_1 Estimated\ Family\ Income_i + \beta_2 Sex\ Of\ Child_i + \beta_3 Age\ Of\ Child_i + \beta_4 Expectation_i \\ &+\beta_5 Public\ School_i + \beta_6\ \text{Pr}ivate\ School_i + \beta_7 Home\ School_i + \beta_8 Total\ Number\ Of\ Children\ In\ Family_i + \varepsilon_i\end{aligned}$$

Model 9:

$$\begin{aligned}&\log(School\ \text{Re}lated\ Expenditure_i) = \\ &\beta_0 + \beta_1 Estimated\ Family\ Income_i + \beta_2 Sex\ Of\ Child_i + \beta_3 Age\ Of\ Child_i + \beta_4 Expectation_i \\ &+\beta_5 Public\ School_i + \beta_6\ \text{Pr}ivate\ School_i + \beta_7 Home\ School_i + \beta_8 Total\ Number\ Of\ Children\ In\ Family_i + \varepsilon_i\end{aligned}$$

Please note that both Model 6 and Model 8 have violated the normality assumption and thus, log transformation in Model 7 and Model 9 was necessary.

In Model 7, approximately 28% of the variation was explained and in Model 9, approximately 26.4% of the same variation was explained—the use of instrumental variable sacrificed the goodness of fit by 1.6%. However, Model 9 reduced the influence of influential observations in Model 7. Chart 12 shows that in Model 7, there are two highly influential observations. Using the estimated coefficients of family income as an example, the coefficient will change by almost 30% if any of the observations was isolated from the regression analysis. However, the same observations will only change the estimated coefficient of estimated family income by about 5%. Another interesting feature is there was one observation (in red circle) which was not considered influential in Model 7—the change in estimated coefficients when the observation was isolated from the analysis was almost negligible; in Model 9, the same observation was much more influential than before—the change in estimated coefficients when the observation was isolated from the analysis was more than 10%.

Chart 12 A Scatter Plot of Percent Change in Estimated Coefficients of Estimated Family Income in Model 9 vs. Percent Change in Estimated Coefficients of Family Income in Model 7

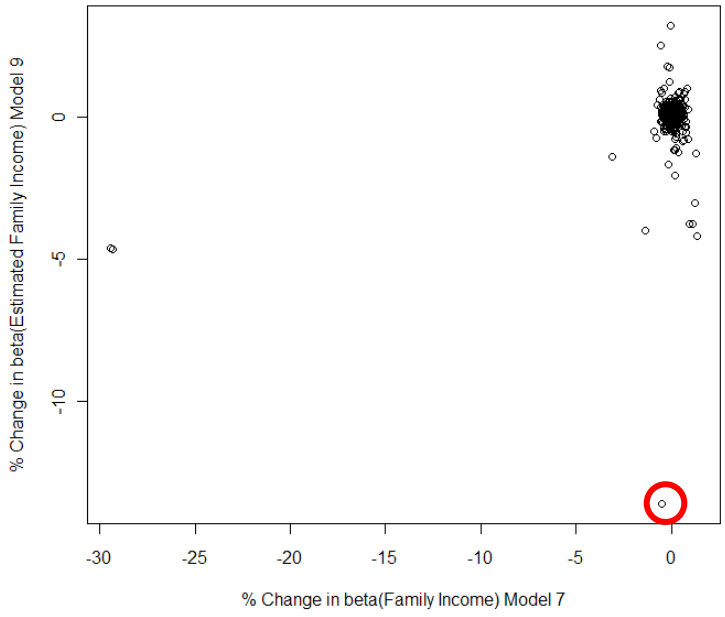


Table 3 Regression Results Summary

		Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
Response Variable (y = School Related Expenditure on Child)		**y**	**log(y)**	**log(y)**	**y**	**log(y)**	**y**	**log(y)**	**y**	**log(y)**
					Estimated Beta					
					Standard Error					
n					2772					
Intercept		2.07E+02	4.18E+00	4.189E+00	6.842E+01	4.120E+00	-1.435E+02	2.937E+00	-2.767E+02	2.735E+00
		*2.91E+01****	*4.42E-02 ****	*4.419E-02 ****	*3.608E+01 ^*	*5.553E-02 ****	*1.302E+02*	*2.191E-01 ****	*1.310E+02 **	*2.228E-01 ****
Family Income in 2002		8.35E-04	4.54E-06	---	---	---	3.015E-04	3.709E-06	---	---
		*3.54E-04 **	*5.38E-07 ****				*2.838E-04*	*4.775E-07 ****		
Variation in Family Income		7.06E-03	3.25E-06	3.486E-06	---	---	6.117E-03	1.460E-06	---	---
		*8.15E-04 ****	*1.24E-06 ***	*1.237E-06 ***			*6.476E-04 ****	*1.090E-06*		
Family Disposable Income (After Tax)		---	---	4.450E-06	---	---	---	---	---	---
				*5.436E-07 ****						
Estimated Family Income in 2002		---	---	---	5.806E-03	6.830E-06	---	---	4.604E-03	4.651E-06
					*4.528E-04 ****	*6.969E-07 ****			*3.624E-04 ****	*6.163E-07 ****
Sex of Child	**Male**	---	---	---	---	---	-9.790E+01	-3.039E-01	-9.667E+01	-2.888E-01
							*3.582E+01 ***	*6.026E-02 ****	*3.580E+01 ***	*6.087E-02 ****
	Female	---	---	---	---	---	---	---	---	---
Age of Child		---	---	---	---	---	1.415E+01	-1.290E-02	1.466E+01	-6.617E-03
							*4.974E+00 **	*8.369E-03*	*4.951E+00 **	*8.418E-03*
Parents' Expectation on Child		---	---	---	---	---	4.532E+01	1.757E-01	4.671E+01	1.927E-01
							*9.717E+00 ****	*1.635E-02 ****	*9.629E+00 ****	*1.637E-02 ****
Type of	**Not in**	---	---	---	---	---	---	---	---	---

School Attended	**School**									
	Public School	---	---	---	---	---	-2.047E+00 *8.945E+01*	9.302E-01 *1.505E-01 *****	-1.295E+00 *8.945E+01*	9.394E-01 *1.521E-01 ****
	Private School	---	---	---	---	---	2.880E+03 *1.128E+02 ****	3.781E+0 *1.897E-01 ****	2.882E+03 *1.127E+02 ****	3.810E+00 *1.917E-01 ****
	Home School	---	---	---	---	---	3.181E+02 *1.707E+02 ^*	1.581E+00 *2.872E-01 ****	3.195E+02 *1.707E+02 ^*	1.598E+00 *2.903E-01 ****
Total Number of Children in Family		---	---	---	---	---	-4.824E+01 *1.686E+01 **	-1.310E-01 *2.837E-02 ****	-4.798E+01 *1.686E+01 **	-1.278E-01 *2.867E-02 ****
R-Squared		0.058	0.058	0.056	0.056	0.034	0.411	0.280	0.411	0.264
Adjusted R-Squared		0.057	0.057	0.056	0.056	0.033	0.410	0.277	0.409	0.262

Signif. Codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '^' 0.1 ' ' 1**

4. Conclusion

As the example shown above, there is no cure-all approach that solves all the problems introduced by influential observations. The occurrence of influential observations—most of the time is an combination effect from both the predictors and the response variables. When there is no sufficient background knowledge about the specific observation—i.e. whether the influential observation contains useful information or was is simply an error in data input. With such information, it is easier to make decision on methods to deal with influential observations. However, most of the time in practice, such knowledge are not available. In such cases, depends on researcher's specific goal, improvements can be made in terms of influential observations. The use of instrumental variable is a method that shrinks predictors to its mean—a.k.a. reduces variation in a predictor contributed by a specific observation so as to reduce the H_{ii} value. This can be particularly helpful in when there is strong correlation between predictors. However, using instrumental variable sacrifice the goodness of fit of the model—information is lost when predicted value is used as a predictor.

5. Reference

Detection of Influential Observation in Linear Regression

Author(s): R. Dennis Cook

Source: *Technometrics*, Vol. 19, No. 1 (Feb., 1977), pp. 15-18

Published by: American Statistical Association and American Society for Quality

Influential Observations in Linear Regression

Author(s): R. Dennis Cook

Source: *Journal of the American Statistical Association*, Vol. 74, No. 365 (Mar., 1979), pp. 169-174

Published by: American Statistical Association