Sample Size and Outliers, Leverage, and Influential Points, and Cooks Distance Formula

Morteza Marzjarani

Saginaw Valley State University (Retired)

Abstract

In this article, a method for determining the sample size based on the confidence level selected by the user is developed. Outliers leverage, and influential data points are presented. Also, an alternative form of Cook's Distance formula involving the user selected standard error and confidence level is also presented.

Introduction

Sample size plays a significant role is data analysis. A small sample size may results in an unreliable estimate. A large sample size on the other hand might be impossible or too costly to collect. Therefore, it is very important to have at least a rough idea about the sample size needed in any situation where there is a need to estimate a parameter.

Determining the proper sample size:

Formulas have been developed to determine the required sample size for estimating parameters. To estimate a population mean with Sampling Error (SE) and 100(1- α) % confidence the required sample size can be estimated using the following formula:

$n= ((Z_{\alpha/2})^2 \sigma^2 N)/ ((N-1)*SE^2+ (Z_{\alpha/2})^2 \sigma^2)$ (1)

where N is population size. If N is not known but large enough, then (1) is reduced to

$n= (Z_{\alpha/2})^2 \sigma^2/ (SE^2)$. (2)

In the absence of σ, the value of R/4 where, R=Range=largest data point-smallest data point can be used. This is a very conservative estimate for the standard deviation and generally produces a high sample size. The answer in (1) or (2) is always rounded up to the nearest whole number. We use the latter formula here to get a rough idea about how large of a sample size we need depending on the desired level of confidence and the standard error selected by the user. The following table represents the results of the sample size needed based randomly generated 1000 data points.

Table 1: Partial list of sample size for different SE and $Z_{\alpha/2}$ values

100(1-α)% confidence	$Z_{\alpha/2}$	SE*	σ known**	σ unknown	Required sample size(σ known)	Required sample size(σ unknown)
99%	2.576	0.025	0.151626	0.370713	244	1459
95%	1.96	0.03	0.151626	0.370713	98	587
90%	1.645	0.04	0.151626	0.370713	39	232
85%	1.44	0.05	0.151626	0.370713	19	114
80%	1.282	0.06	0.151626	0.370713	10	63

(*)Formal polls such as political polls usually use 3% or 4% in determining sample size they need

(**)Data set used here consists of 1000 records which was large enough to estimate standard deviation and assumed that it was the population standard deviation.

For the purpose of further analysis, a few values for the sample size were selected and the following table was generated.

Table 2: ANOVA performed using different sample sizes

Sample size	F-Value	Pr>F	R-Square	Intercept			X		
				Value	t-value	Pr >|t|	Value	t-value	Pr>|t|
20	2.12	0.1620	0.1002	0.1727	1.00	0.3285	0.22542	1.45	0.1620
100	68.14	<.0001	0.4077	-.55452	-4.63	<.0001	0.82572	8.25	<.0001
300	671.93	<.0001	0.6920	-1.21076	-18.53	<.0001	1.37178	25.92	<.0001
500	1893.43	<.0001	0.7914	-1.45217	-32.22	<.0001	1.56832	43.51	<.0001
700	3616.92	<.0001	0.8380	-1.57122	-45.24	<.0001	1.66314	60.14	<.0001
1000	7453.61	<.0001	0.8819	-1.68556	-65.87	<.0001	1.7536	86.33	<.0001

Going back to Table (1), the sample size 100 is roughly equivalent to 95% confidence (σ known) and standard error of 0.03. In case of σ unknown, the sample size 100 is equivalent to 85% confidence and standard error of 0.05.

Outliers, leverage and influential data points

In general, unusual data points will impact the model and need to be identified. We want the model to be a representative of the whole population. Therefore it is important to identify the data points which impact the model significantly. Some outliers are valid data points whereas others might be due to errors such as reading or entering incorrect values. Regardless, unusual data points need to be addressed. For demonstration purposes, in the following sections, without loss of generality, it is assumed that the model consists of only one continuous predictor. That is:

$$y=f(x) \qquad (2)$$

The generated data set used above does not have any unusual data point. We add one outlier, one leverage, and one influential point to this data set. We consider three cases:

a. Outlier: An observation with a large(positive or negative) residual
b. Leverage: An observation with a large predictor
c. Influential: An observation with a large (positive or negative) predictor and residual (some call it the product of a and b.

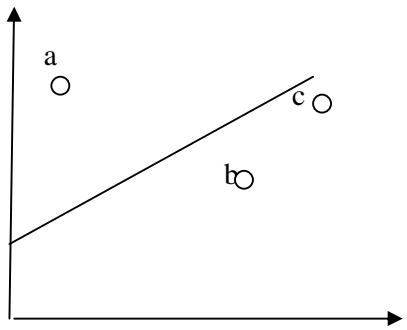

Figure 2: Outlier, Leverage and influential points

Figure 2 shows these points. To identify each unusual data point, methods have been proposed and we present each briefly below:

a. Outlier:

There are several methods one can use to locate outliers. The following method appears in most introductory books in statistics. For a random variable X with probability distribution function F(X) =Pr (X≤ x) the τ-th quartile of X is defined as the inverse function (τ)=inf{x :F(x)≥τ}where 0< τ <1. In particular, the first (Q1), the second (median, Q2), and the third (Q3) quartiles are Q(1/4), Q (1/2), and Q(3/4) respectively. Data points outside the interval (Q1-k *IQR, Q3+k*IQR) are considered outliers, where k is a chosen any value between 1.5 and 3, and IQR=Q3-Q1 is the abbreviation for Inter Quartile Range. Here, the value k=3 for locating outliers in the data set is used. The only outlier is the one we intentionally added to the data set (0,15).

b: Leverage:

Leverage points are unusually large predictors. They do not depend on the response variable. To locate a leverage point, the so called 'Hat Matrix' H=X(X'X) $^{-1}$X' is used. If X is of dimension n x k, then H is square matrix of dimension n x n. Consider the j-th row of this matrix $\mathbf{X}_j=(X_{j1}, X_{j2}, X_{j3},\ldots, X_{jn})$. The i-the diagonal element in this matrix $h_{ii}= 1/n+(X_{ii}- \bar{X})/\Sigma(X_i^2)$ represents a leverage if it is unusually large. That is, if h_{ii} is unusually large, the i-th predictor has a large predictor $(X_{j1}, X_{j2}, X_{j3},\ldots, X_{jn})$. The cutoff value we use here for a large data set and k predictors is: $2(k+1)/n=2(k+1)SE^2/ (Z_{\alpha/2})^2 \sigma^2$. That is, any predictor greater than this value is considered a leverage point and needs to be reviewed. For the given data set again point (12,1) is a leverage point.

c: Influential:

Again, several methods have been proposed to deal with influential points. Here, we use the Cook's formula and attempt to identify the influential points (if any) in the data set.

$D_i=\sum_{j=1}^{j=n}(\hat{y}j - \hat{y}j(i))/kMSE$, (3)

Or equivalently,

$D_i =((n-p)/k)\ \sum_{j=1}^{j=n}(\hat{y}j - \hat{y}j(i))^2 / \sum_{j=1}^{j=n}(yj - \hat{y}j)^2$ (4)

where inside parenthesis$(\hat{y}j\ and \hat{y}j(i))$is the difference between the predicted values for the response variable with and without the i-th observation, k is the number of predictors and MSE is the mean square error from the ANOVA table. Clearly, the point (6, 9) is an influential point.

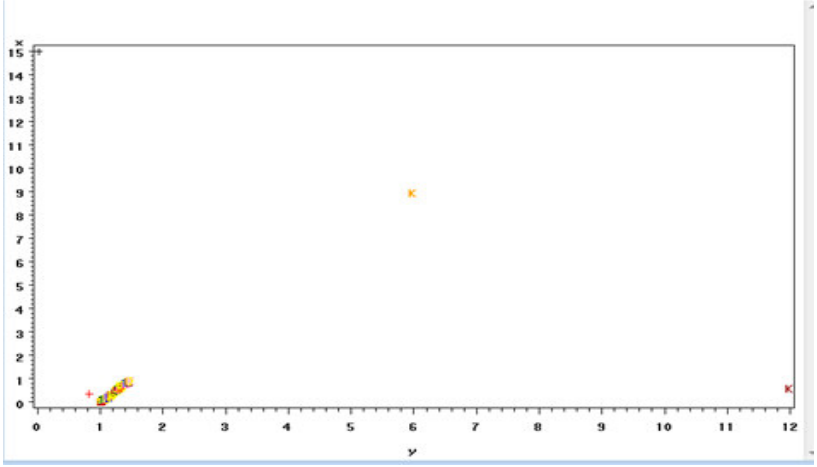

Figure 3: Leverage, influential and outliers

Without loss of generality, assume that there is only one influential point in the given data set and it is the first data point. Formula (3) can be written as:

$D_i=[(\hat{y}1-\hat{y}1(i))^2+\sum_{j=2}^{j=n}(\hat{y}j-\hat{y}j(i))^2]/kMSE$ (5)

The sum in (5) consists of some values close to each other. We could assume that they are roughly equal to a common value, say, a. Also, for simplicity, assume the first quantity is b. The (5) is reduced to:

$D_i=[(b+(n-1)a]/kMSE$ (6)

We represented n in terms of confidence level and standard error in (2). Substituting for n will result in the following:

$D_i=[(b-a+a(Z_{\alpha/2})^2\ \sigma^2/\ (SE^2)]/kMSE$ (7)

The formula in (7) represents an approximate value for the Cook's distance in terms of confidence level and standard error. The average values of residuals squared is a possible value for b in (7).

Conclusion:

In this article, the impact of confidence level on the sample size was presented. Unusual data points and an alternative form of Cook's distance formula involving the confidence level was also presented.

References:

(1) Breslow, N.E.; Clayton, D.G**.** (1993). "*Approximate Inference in Generalized Linear Mixed Models*". Journal of the American Statistical Association88 (421): 9–25.

(2) Clayton, D. (1996). *Generalized linear mixed models, in Markov Chain Monte Carlo Methods in Practice,* W.R. Gilks, S. Richardson & D.J. Spiegelhalter, eds, Chapman & Hall, New York, pp. 275–303.

(3) Cook, R. Dennis (February 1977). "Detection of Influential Observations in Linear Regression". *Technometrics*, 19 (1): 15–18.

(4) Cook, R. Dennis (March 1979). "Influential Observations in Linear Regression". *Journal of the American Statistical Association*, 74 (365): 169–174.

(5) Fitzmaurice, Garrett M.; Laird, Nan M.; Ware, James H. (2004). *Applied longitudinal analysis*. Hoboken, NJ: Wiley-Interscience. ISBN 0-471-21487-6.

(6) McCullagh, P, &Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman & Hall/CRC Press..