Statistica Sinica **5**(1995), 399-403

UPDATING SIMPLE LINEAR REGRESSION

Jerome H. Klotz

University of Wisconsin at Madison

Abstract. Closed form expressions are given for updating slope, intercept, and error estimates in the simple linear regression model. The formulae, which have good accuracy, involve only previous estimates and new observations and eliminate the need to store previous observations.

Key words and phrases: Linear regression; update, deletion formula; chisquare distribution.

1. Introduction

We consider the simple linear regression model

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where ε_i are independent $\mathcal{N}(0, \sigma^2)$. The usual unbiased estimates for n observations are

$$\begin{aligned} \hat{\beta}_n &= S_{XY}(n)/S_X^2(n), \\ \hat{\alpha}_n &= \bar{Y}_n - \hat{\beta}_n \bar{X}_n, \\ \hat{\sigma}_n^2 &= S_e^2(n)/(n-2), \end{aligned} \tag{1}$$

where $\bar{Y}_n = \sum_{i=1}^n Y_i/n$, $S_e^2(n) = \sum_{i=1}^n (Y_i - \bar{Y}_n - \hat{\beta}_n(X_i - \bar{X}_n))^2$ and

$$\begin{aligned} S_X^2(n) = \sum_{i=1}^n (X_i - \bar{X}_n)^2, \quad S_Y^2(n) = \sum_{i=1}^n (Y_i - \bar{Y}_n)^2, \\ S_{XY}(n) = \sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n). \end{aligned} \tag{2}$$

When additional observations and covariates $(Y_i, X_i), i = n+1, n+2, \ldots, N$, become available, we desire updates $\hat{\beta}_N$, $\hat{\alpha}_N$, $\hat{\sigma}_N^2$. If updates can be calculated using only new data and prior estimates, then storage of previous individual observations is unnecessary. A small computing device can then be programmed to handle large sample sizes.

Despite loss of accuracy from cancellation in floating point subtraction, many calculators update $\sum X_i$, $\sum Y_i$, $\sum X_i^2$, $\sum Y_i^2$, $\sum X_i Y_i$ in one pass and then compute estimates from

$$\begin{aligned} S_e^2(n) &= S_Y^2(n) - \hat{\beta}_n^2 S_X^2(n), && (3)\\ \hat{\beta}_n &= S_{XY}(n)/S_X^2(n), && (4)\\ \hat{\alpha}_n &= \bar{Y}_n - \hat{\beta}_n \bar{X}_n, \end{aligned}$$

$$\begin{aligned} S_X^2(n) = (\sum_{i=1}^{n} X_i^2) - n\bar{X}_n^2, \quad S_Y^2(n) = (\sum_{i=1}^{n} Y_i^2) - n\bar{Y}_n^2,\\ S_{XY}(n) = (\sum_{i=1}^{n} X_i Y_i) - n\bar{X}_n\bar{Y}_n. \end{aligned} \qquad (5)$$

For the problem of calculating $S_X^2(n)$ for estimating the sample variance, Chan, Golub and LeVeque (1983) give a thorough discussion of a variety of algorithms. An additional formula is given by

$$S_X^2(N) = S_X^2(n) + S_X^2(*), \qquad (6)$$

where

$$S_X^2(*) = \sum_{i=n+1}^{N} (X_i - \bar{X}_*)^2, \quad \bar{X}_* = (\bar{X}_n\sqrt{n} + \bar{X}_N\sqrt{N})/(\sqrt{n} + \sqrt{N}).$$

We similarly have

$$S_{XY}(N) = S_{XY}(n) + S_{XY}(*), \qquad (7)$$

where

$$S_{XY}(*) = \sum_{i=n+1}^{N} (X_i - \bar{X}_*)(Y_i - \bar{Y}_*), \quad \bar{Y}_* = (\bar{Y}_n\sqrt{n} + \bar{Y}_N\sqrt{N})/(\sqrt{n} + \sqrt{N}).$$

Identity (7) can be derived by expanding the right side to obtain $\sum_{i=1}^{N} X_i Y_i - N\bar{X}_N\bar{Y}_N$ after substituting $N\bar{X}_N - n\bar{X}_n$ for $\sum_{i=n+1}^{N} X_i$, $N\bar{Y}_N - n\bar{Y}_n$ for $\sum_{i=n+1}^{N} Y_i$, and noting $(N-n)/(\sqrt{N} + \sqrt{n}) = \sqrt{N} - \sqrt{n}$ when collecting terms.

2. Updating Regression

Theorem. *Let $\hat{\beta}_n$, $\hat{\alpha}_n$, $\hat{\sigma}_n^2$ given in* (1) *be regression estimates based on the data $\{(Y_i, X_i),\ i = 1, 2, \ldots, n\}$. Then, when additional observations $\{(Y_i, X_i), i = n+1, n+2, \ldots, N\}$ are available, for the combined data*

$$\begin{aligned}\hat{\beta}_N &= (S_{XY}(n) + S_{XY}(*))/(S_X^2(n) + S_X^2(*)),\\ S_e^2(N) &= S_e^2(n) + S_e^2(*),\end{aligned} \tag{8}$$

where

$$S_e^2(*) = \sum_{i=n+1}^{N} (Y_i - \bar{Y}_* - \bar{\beta}_*(X_i - \bar{X}_*))^2, \tag{9}$$

$$\bar{\beta}_* = (\hat{\beta}_n S_X(n) + \hat{\beta}_N S_X(N))/(S_X(n) + S_X(N)), \tag{10}$$

and $\bar{X}_*$, $\bar{Y}_*$ *are given after* (6), (7).

Proof. Consider the equation $S_e^2(N) = S_e^2(n) + S_e^2(*)$ where we hope to find a value of β for which

$$S_e^2(*) = \sum_{i=n+1}^{N} (Y_i - \bar{Y}_* - \beta(X_i - \bar{X}_*))^2 = \beta^2 S_X^2(*) - 2\beta S_{XY}^2(*) + S_Y^2(*).$$

We also have

$$\begin{aligned}S_e^2(*) &= S_e^2(N) - S_e^2(n) = (S_Y^2(N) - S_{XY}^2(N)/S_X^2(N)) - (S_Y^2(n) - S_{XY}^2(n)/S_X^2(n))\\ &= S_Y^2(*) - [S_{XY}^2(N)/S_X^2(N) + S_{XY}^2(n)/S_X^2(n)].\end{aligned}$$

Equating the right sides of these two equations, and canceling $S_Y^2(*)$, we can solve for β in the quadratic equation

$$\beta^2 S_X^2(*) - 2\beta S_{XY}(*) + [(S_{XY}^2(N)/S_X^2(N) + S_{XY}^2(n)/S_X^2(n)] = 0.$$

A stable root is $\beta = \bar{\beta}_*$ as given by Equation (10).

Alternatively, Equation (9) could be directly verified by substituting $\bar{\beta}_*$ and using identities (6) and (7).

Formula (8) was discovered by applying the more general updating procedure of Golub and Styan (1973) (Sections 3 and 4) to the special case of simple linear regression. Householder transformations as discussed in Golub (1965) were used.

There is a considerable literature on update formulae for multiple linear regression but Equation (8) for simple linear regression is believed to be new. For a comprehensive exposition of updating in the general case see Bingham (1977).

3. Special Cases

For the special case of $N = n + 1$ Equation (8) can be written as

$$S_e^2(n+1) = S_e^2(n) + \frac{n}{(n+1)}(Y_{n+1} - \bar{Y}_n - \hat{\beta}_n(X_{n+1} - \bar{X}_n))^2 \frac{S_X^2(n)}{S_X^2(n+1)} \tag{11}$$

with $\hat{\beta}_n = S_{XY}(n)/S_X^2(n)$ obtained by updating $S_{XY}(n)$ by

$$S_{XY}(n+1) = S_{XY}(n) + \frac{n}{(n+1)}(X_{n+1} - \bar{X}_n)(Y_{n+1} - \bar{Y}_n)$$

and similarly for $S_X^2(n)$. Starting values of $\hat{\beta}_1 = 0$, and $S_e^2(1) = 0$, $S_e^2(2) = 0$, $S_X^2(1) = 0$, and $S_{XY}(1) = 0$ can be used.

To compare the accuracy for a small data set with commonly used computing equations for simple linear regression, define the one pass formula $S_{e(1)}^2(n)$ using Equations (3), (4) and (5).

Similarly define the two pass formula

$$S_{e(2)}^2(n) = S_{Y(2)}^2(n) - \hat{\beta}_{n(2)}^2 S_{X(2)}^2(n),$$

where $\hat{\beta}_{n(2)}$, and $S_{X(2)}^2(n)$, $S_{Y(2)}^2(n)$, $S_{XY(2)}$ are calculated from Equations (1) and (2).

The three pass formula is

$$S_{e(3)}^2(n) = \sum_{i=1}^{n}(Y_i - \bar{Y}_n - \hat{\beta}_{n(2)}(X_i - \bar{X}_n))^2.$$

For the following data for which $\hat{\alpha} = 20$ and $\hat{\beta} = 10$

X_i	10.1	20.1	30.1	40.1
Y_i	121.1	220.7	321.3	420.9

we compare the accuracy of $S_{e(0)}^2(n)$ computed from Equation (11) with the one, two, and three pass formula and the exact value for $S_e^2(n)$.

Calculation was done in standard IEEE single precision floating point in which nonzero numbers are represented by 32 binary bits in the form $x = (-1)^s 2^{e-127}(1+f)$. The single sign bit $s = 0, 1$, the 8 binary bit biased exponent $0 \leq e \leq 255$, and the 23 bit binary fraction $0 \leq f \leq 1 - 2^{-23}$ gives approximately $\log_{10}(2^{24}) \doteq 7.22$ digit accuracy.

(k)	$S_{X(k)}^2$	$S_{XY(k)}$	$S_{Y(k)}^2$	$S_{e(k)}^2$
0	499.999939	4999.99951	50000.19922	0.1999945
1	499.999725	4999.99805	50000.18750	0.1958923
2	499.999939	4999.99951	50000.19531	0.2014160
3	″	″	″	0.1999916
exact	500	5000	50000.2	0.2

In other examples computed, Equation (11) for $S_{e(0)}^2(n)$ (which also requires only one pass) gave accuracy comparable to the three pass formula. It is well known that the one pass equation (3) for $S^2_{e(1)}$ can lose accuracy from subtraction of terms with common leading digits. The two pass formula can similarly lose accuracy when S^2_Y and $\hat{\beta}^2 S^2_X$ have common leading digits.

For $N = n+1$ we can also derive a deletion formula from Equation (8) given by

$$\begin{aligned} S^2_e(n) =& S^2_e(n+1) \\ &- \frac{(n+1)}{n}(Y_{n+1} - \bar{Y}_{n+1} - \hat{\beta}_{n+1}(X_{n+1} - \bar{X}_{n+1}))^2 \frac{S^2_X(n+1)}{S^2_X(n)}. \end{aligned} \tag{12}$$

For a calculator, Formula (11) is convenient for adding a single new data pair and Equation (12) is useful for deleting a data pair that is in error.

When $n = 2$, $S^2_e(n) = 0$ since a straight line can fit 2 points exactly. It is well known that $S^2_e(N)$ has a $\sigma^2\chi^2_{N-2}(0)$ distribution. By Equation (8), $S^2_e(*)$ for this case of $n = 2$ forms an explicit representation of $S^2_e(N)$ as a sum of squares of $N-2$ independent $\mathcal{N}(0,\sigma^2)$ random variables using Cochran's theorem for quadratic forms. The $\sigma^2\chi^2_{N-2}(0)$ distribution could be proved directly by showing that $Y_i - \bar{Y}_* - \bar{\beta}_*(X_i - \bar{X}_*)$ are independent $\mathcal{N}(0,\sigma^2)$ random variables for $i = 3, 4, \ldots, N$.

Acknowledgements

Thanks go to the referee for a careful review and suggestions for improvement. Thanks also go to the editors for helpful recommendations and references.

References

Bingham, C. (1977). Some identities useful in the analysis of residuals from linear regression. Technical Report 300, School of Statistics, University of Minnesota.

Chan, T. F., Golub, G. H. and LeVeque, R. J. (1983). Algorithms for computing the sample variance: Analysis and recommendations. Amer. Statist. **37**, 242-247.

Golub, G. H. (1965). Numerical methods for solving least squares problems. *NumerischMathematik* **7**, 206-216.

Golub, G. H. and Styan, G. P. H. (1973). Numerical computations for univariate linear models. J. Statist. Comput. Simulation **2**, 256-274.

Department of Statistics, University of Wisconsin, Madison, WI 53706, U.S.A.

(Received June 1993; accepted July 1994)