Open Journal of Statistics, 2014, 4, 173-177
Published Online April 2014 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2014.43016

On Minimizing the Standard Error of the Slope in Simple Linear Regression

Steven M. Crunk, Xinh Huynh

Department of Mathematics and Statistics, San Jose State University, San Jose, USA
Email: steven.crunk@sjsu.edu, xinh.huynh@gmail.com

Received 8 February 2014; revised 8 March 2014; accepted 15 March 2014

Abstract

A common homework problem in texts covering calculus-based simple linear regression is to find a set of values of the independent variable which minimize the standard error of the estimated slope. All discussions the authors have heard regarding this problem, as well as all texts with which the authors of this paper are familiar and which include this problem, provide no solution, a partial solution, or an outline of a solution without theoretical proof and the provided solution is incorrect. Going back to first principles we provide the complete correct solution to this problem.

Keywords

Minimization; Variance; Coefficient; Beliefs about Statistics; Statistical Literacy

1. Introduction

A homework question, occurring in several oft cited best-selling introductory texts covering calculus-based simple linear regression, goes something like this:

Suppose we are to collect data and fit a straight-line simple linear regression, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. The errors are assumed to have mean zero, unknown variance σ^2 and to be uncorrelated with one another. Further suppose that in this designed experiment, the region of interest for x is $A \le x \le B$, $A < B$, and that the primary goal is to make the standard error of the estimate of the slope as small as possible. For a given sample size n, at what values of the independent variable should the observations be taken? That is, how should $x_1, x_2, \cdots, x_n$ be chosen so as to minimize the standard error of the estimate of β_1.

From [1], which does not include the above noted problem, and virtually any other text covering simple linear regression, we know the following: the estimate of the slope is

How to cite this paper: Crunk, S.M. and Huynh, X. (2014) On Minimizing the Standard Error of the Slope in Simple Linear Regression. *Open Journal of Statistics*, **4**, 173-177. http://dx.doi.org/10.4236/ojs.2014.43016

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

which has standard deviation

$$\sqrt{V(\hat{\beta}_1)} = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

The estimated standard deviation, or standard error, is found by replacing σ^2 by its estimate

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}.$$

The error variance σ^2 is an unknown constant and its estimator cannot be formed until data are collected. Thus in the case of either the theoretical standard deviation or the estimated standard error, the numerator under the radical is unknown and not under the control of the experimenter in the question. Consequently the minimization of the standard deviation or the standard error is achieved by maximizing the quantity

$$\mathrm{SXX} = \sum_{i=1}^{n}(x_i - \bar{x})^2,$$

the corrected sum of squares of the *x*'s.

Many texts which include this problem provide no solution. Every discussion that the authors have heard discussed or seen in a solutions manual suggests, without proof, that in order to maximize SXX if *n* is even, half of the observations should be taken at *A* and half at *B*. Many texts that include a solution ignore the possibility that *n* is odd, even though no condition on *n* was provided in the question. When a solution is provided for *n* odd, every solution we have seen suggested without proof that $(n-1)/2$ observations should be taken at each of *A* and *B* with the remaining single observation being taken half way between these values, at $(A+B)/2$. That this solution is incorrect which can be seen with a simple example where $n=3$. The result using the "usual" solution outlined above is to take $x_1 = A,\ x_2 = B,$ and $x_3 = (A+B)/2$ from whence $\bar{x} = (A+B)/2$ and $\mathrm{SXX} = (B-A)^2/2$. Alternatively, if we take $x_1 = x_2 = A$ and $x_3 = B$, we have $\bar{x} = (2A+B)/3$ and $\mathrm{SXX} = 2(B-A)^2/3$ which are larger than the value obtained using the "usual" solution, showing that the usual solution is not correct. We suppose that the desire for symmetry led to the belief in the incorrect solution; however symmetry has not been neither mentioned nor required for the problem under discussion.

In the sequel we show that for *n* even, the "usual" solution of choosing half of the observations to be taken at *A* and the other half to be taken at *B* is correct. For *n* odd we show that in order to minimize the standard error, $(n+1)/2$ observations should be taken at one end of the interval (either at *A* or at *B*) and the remaining $(n-1)/2$ observations should be taken at the other end of the interval. An example of this result is given in **Figure 1**. Throughout we will assume that the sample size *n* is a given constant.

2. The Objective Function; Sum of Squares

Our goal is to find the set of x_i which maximize

$$\mathrm{SXX} = \sum_{j=1}^{n}(x_j - \bar{x})^2 = \sum_{j=1}^{n} x_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)^2.$$

Since the x_i are continuous variables (not in the statistical sense but rather in the algebraic sense) on the interval $[A, B]$, we may use techniques of calculus in order to find the values that maximize this function (see, e.g., [2]). We have

$$\frac{\partial \mathrm{SXX}}{\partial x_i} = 2x_i - \frac{2}{n}\left(\sum_{j=1}^{n} x_j\right) = 2x_i - 2\bar{x} \quad \forall i.$$

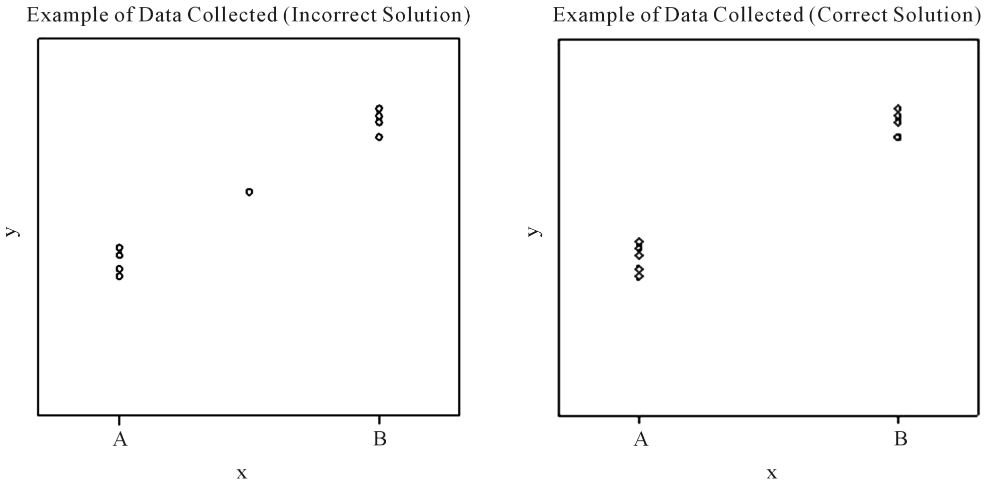

Figure 1. The graphs above represent *n* (an odd number) data points collected according to two plans for minimizing the standard error of the slope in simple linear regression. The figure on the left represents the common but incorrect solution whereby one observation is taken in the middle of the interval. In the graph to the right, the number of observations taken at either end of the interval differ by one. Although lacking symmetry, this is the correct solution for minimizing the standard error of the slope.

Setting this equal to zero we have $x_i = \bar{x}\ \forall i$ being stationary points. Of course our variables exist on a closed interval so we must also investigate the endpoints. As a result it must be true that $x_i \in \{A, \bar{x}, B\}\ \forall i$.

If $x_i = \bar{x}\ \forall i$ then SXX = 0, which is the smallest possible value of SXX, *i.e.*, choosing $x_i = \bar{x}\ \forall i$ leads to a minimum rather than a maximum. The same is true if observations are taken either all at *A* or all at *B*. We would then say it is obvious that at least one observation must be taken at *A* and at least one observation must be taken at *B*, but authors saying "it is obvious that..." is what led to this note in the first place. Consider the case where some observations are taken at $x = A$ and the rest at $x = \bar{x}$ distinct from *A*; this is a contradiction as the mean would then not be at $\bar{x}$. Similarly, it is impossible to have some observations at *B* and the rest at $\bar{x}$. Accordingly it must be true that at least one observation must be taken at each of *A* and *B*.

Let n_1 be the number of observations taken at *A*, n_2 be the number of observations taken at $\bar{x}$, and n_3 be the number of observations taken at *B*. From the argument in the previous paragraph we have $n_i \ge 1$ for $i = 1, 3$, $n_2 \ge 0$, all integers, and $n_1 + n_2 + n_3 = n$, a given constant. Then $\bar{x} = (n_1 A + n_2 \bar{x} + n_3 B)/(n_1 + n_2 + n_3)$, the simplification of which leads to $\bar{x} = (n_1 A + n_3 B)/(n_1 + n_3)$. Consequently, substituting these values, we have

$$\begin{aligned} \text{SXX} &= \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} \left(x_i - \frac{n_1 A + n_3 B}{n_1 + n_3} \right)^2 \\ &= n_1 \left(A - \frac{n_1 A + n_3 B}{n_1 + n_3} \right)^2 + n_2 \left(\bar{x} - \frac{n_1 A + n_3 B}{n_1 + n_3} \right)^2 + n_3 \left(B - \frac{n_1 A + n_3 B}{n_1 + n_3} \right)^2 \\ &= \frac{n_1 n_3 (B - A)^2}{n_1 + n_3}. \end{aligned}$$

The quantity $(B - A)$ is an arbitrary non-negative constant. Some texts give as their example $A = -1$ and $B = 1$, some give $A = 0$ and $B = 1$, and still other books use other choices for these given constants. The choice of *A* and *B*, as seen in the final formula for SXX, have no bearing on the solutions for n_1, n_2 and n_3 which maximize SXX. Thus we shall simply attempt to find parameters n_1, n_2 and n_3 that maximize $f(n_1, n_2, n_3) = n_1 n_3/(n_1 + n_3)$ with the constraints imposed previously that n_1, n_2 and n_3 are non-negative integers, n_1, $n_3 \ge 1$, and $n_1 + n_2 + n_3 = n$, a known/given constant.

3. Optimization

The function with constraints given in the previous paragraph may be maximized in any number of ways. Possibilities considered by the authors include the following: taking the variables of interest to be continuous and maximizing the function through the use of calculus, hoping for integer values which would then be the optimal solution [3]; using integer programming [4]; and other possibilities. However, it seems that a simple algebraic manipulation may be the most elegant solution.

Let

$$g(n_1,n_2,n_3) = 1/f(n_1,n_2,n_3) = (n_1+n_3)/(n_1n_3) = 1/n_3 + 1/n_1.$$

Thus maximizing $f(n_1,n_2,n_3)$ is equivalent to minimizing $g(n_1,n_2,n_3)$. We now show that n_2 must be zero. Assume that $(n_{1,0},n_{2,0},n_{3,0})$ is an ordered triple which meets the constraints and which minimizes $g(n_1,n_2,n_3)$ with $n_{2,0} > 0$. Let $n_{1,1} = n_{1,0} + n_{2,0}$, $n_{2,1} = 0$, and $n_{3,1} = n_{3,0}$. Then the ordered triple $(n_{1,1},n_{2,1},n_{3,1})$ also satisfies the constraints, and furthermore

$$g(n_{1,1},n_{2,1},n_{3,1}) = 1/n_{3,1} + 1/n_{1,1} < 1/n_{3,0} + 1/n_{1,0} = g(n_{1,0},n_{2,0},n_{3,0}),$$

which is a contradiction to the assumption that $g(n_{1,0},n_{2,0},n_{3,0})$ minimizes $g(n_1,n_2,n_3)$, hence $n_2 = 0$.

Now one of our constraints reduces to $n_1 + n_3 = n$, and maximizing $f(n_1,n_2,n_3) = n_1n_3/(n_1+n_3)$ reduces to maximizing

$$h(n_1,n_3) = n_1n_3/(n_1+n_3) = n_1n_3/n \propto n_1n_3 = n_1(n-n_1).$$

This last is simply a parabola which we need to maximize over $n_1 \in \{1,2,\cdots,n-1\}$. To find the maximum, treat the parabola as a function of a continuous variable *z*. The maximum occurs when $\partial z(n-z)/\partial z = n - 2z = 0$, that is, when $z = n/2$. As *n* is integer valued, for *n* even this implies $n_1 = n/2$ gives the maximum value, while for *n* odd either of the two points surrounding $n/2$, $(n-1)/2$ or $(n+1)/2$, gives the same maximum value. **Figure 2** graphically demonstrates this result. The contradiction in the previous paragraph gives $n_2 = 0$ and this with the original constraint that $n_1 + n_2 + n_3 = n$, a known/given constant, gives the value of n_3.

4. Conclusions

For the common homework problem appearing in approximately half of the texts covering calculus-based simple linear regression with which the authors are familiar, and which was posed at the beginning of this paper, we have shown that if *n* is even, the oft given solution to choose half of the points at which to take observations at either end of the interval is correct. However, for odd *n* we have shown that the only previously given solution to place one point in the center of the interval and half of the remaining points at each end of the interval is incorrect, and that the correct solution is to choose nearly half, either $(n-1)/2$ or $(n+1)/2$, at one end of the interval and the remaining points at the opposite end of the interval.

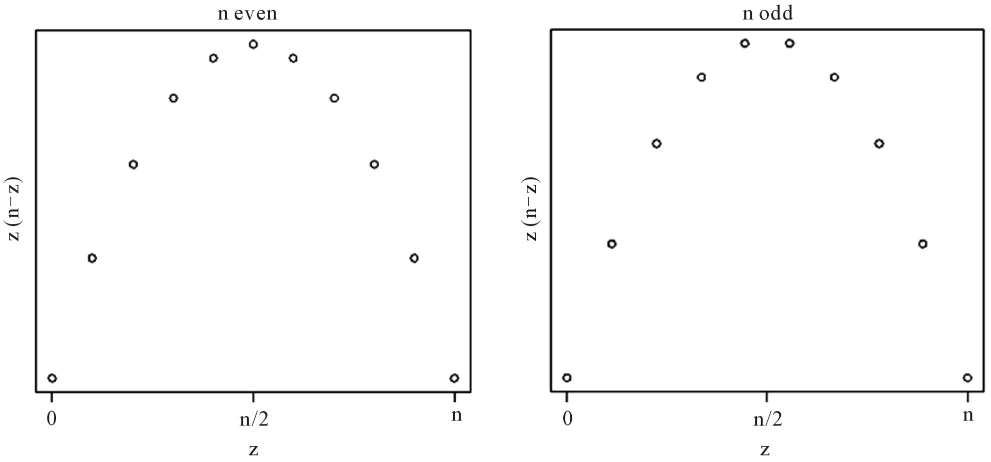

Figure 2. When *n* is even, the maximum of the objective function occurs at $n/2$. When *n* is odd, the maximum value occurs at $(n-1)/2$ and $(n+1)/2$.

We part with the common caveat that this oft given textbook problem is of little use in most realistic applications unless it is known that the true relationship among the data is linear, as the solution affords us no opportunity to check this assumption with the observed data. However, the authors would submit that there is a difference between being "useless in practical situations" and "understanding something fundamental about simple linear regression". We believe that it is important for a student to understand the theory underlying simple linear regression, and this importance is supported by the inclusion of the problem in a large number of highly cited and best-selling texts. Unfortunately, many of these texts provide no solution, some provide a partial solution and others provide an incorrect solution. No texts with which we are familiar, nor their solutions manuals, provide a complete and correct solution. This common textbook problem affords the student the opportunity to understand what drives the variance of the parameter estimate, and as such deserves a correct solution.

Acknowledgements

The authors wish to thank Dr. Ho Kuen Ng for a useful discussion on optimization.

References

[1] Weisberg, S. (2005) Applied Linear Regression. 3rd Edition, John Wiley & Sons, Inc., Hoboken. http://dx.doi.org/10.1002/0471704091

[2] Stewart, J. (2011) Calculus. 7th Edition, Thomson Brooks/Cole, Belmont.

[3] Greenberg, H. (1971) Integer Programming. Academic Press, New York.

[4] Li, D. and Sun, X. (2006) Nonlinear Integer Programming. Springer, New York.