OLS in one dimension

Steven Brawer
11/2017

introduction

In this document, we treat OLS (Ordinary Least Squares) with one predictor variable (one dimension), with a focus on the variance of several quantities. Both theoretical and numerical results are presented.

This analysis digs into some details of OLS results. While the OLS equations result from an average over all samples, the results show how different predictors lead to different variances.

We consider the following:

- Variance of the estimated slope, denoted as β (the hat is suppressed).
-
- $D\widehat{y_x} \equiv \widehat{y_x} - E[\widehat{y_x}]$, where $\widehat{y_x}$ is the estimated output value at predictor x - <x>. Here, x may be a sample point or an out-of-sample (ie, test) point. We show that $E\left[D\widehat{y_x}^2\right]$ increases with increasing |x-<x>|.
-
- $\Delta\widehat{y_x} \equiv \widehat{y_x} - y_x$, where y_x is the true output for predictor x. It is shown that $E\left[\Delta\widehat{y_x}^2\right]$ decreases with |x-<x>| for x a point in the training set, and increases with |x-<x>| for x a test point (out-of-sample). The former is analogous to **overfitting**.

notation and averages

The following equation generates both training and test data.

$$y_x = b\, g(x)\, x + \varepsilon_x \qquad (1)$$

The function g(x) is not random. g(.) = 1 means that the generated data is linear in x, with slope b. The intercept is 0. We refer to x as the **independent variable,** and y as the **dependent variable**. The variable x is also referred to as a **feature** or **predictor**, and y is also referred to as an **output**. In this note, x and y are continuous.

The residual ε_x in eq (1) is a random variable, which has mean 0 and variance σ^2. The value of σ is known only in ensemble averages (see below). The variance is taken here to be independent of x. The variables ε_x and $\varepsilon_{x'}$ are independent if $x \neq x'$. Because of (1), y is also a random variable. Each ε_x is iid.

There are **N** values of x and of y in each training set. These are **samples**. The quantities x_i, y_i are the values of x, y for the ith sample. Each training set has **N** different and independent values of residuals.

The collection of residuals, for many different training sets, form an **ensemble**. The residuals are generated from the same distribution in all ensembles. Each ensemble has in general different numerical values for the residuals.

We will assume that the **values of** x_i **in each ensemble are the same.** So the x values are not averaged over in ensemble averages.

The average E[...] represents an average over the ensemble of residuals. The theoretical averages assume that the size of the ensemble goes to infinity. We have

$$E[\varepsilon_x] = 0$$
$$E[\varepsilon_x \varepsilon_{x'}] = \begin{cases} 0 & if\ x \neq x' \\ \sigma^2 & otherwise \end{cases} \tag{3}$$

The samples of the training set are labeled with subscripts i,j,k...., integers, and which take on values 1,2,...,N for the N samples. Residuals in the training set are $\varepsilon_i, \varepsilon_j$, etc. In each member of the ensemble, ε_i for given i has in general a different value.

The second of eqs (3) states that, no matter how close x and x' are, the second of eqs (3) holds. This has implications which may seem strange in some circumstances, making the second of eqs (3) a very strong assumption.

In general, x represents any value of the independent variable, whether in or out of sample, and x_i represents a **sample** - a member of the training set. If x is continuous, then $x_i \neq x$ almost everywhere for all i. We shall see that statistics (that is, ensemble averages) depending on x and of x_i are in principle (that is, theoretically) sometimes significantly different if the value of x is not in the training set, no matter how close in value they are.

The **regression function** is assumed to have the linear form

$$\widehat{y_x} = f(x) = \alpha + \beta x \tag{4}$$

which is supposed to hold for both training and test predictors x. For estimated values of α, β , we should write them with a hat (^) above the symbols. However the hat is implicit for these parameters. All such values are estimated in this note.

It is convenient to define

$$< x > \equiv \frac{1}{N} \sum_{i=1}^{N} x_i$$
$$< y > \equiv \frac{1}{N} \sum_{i=1}^{N} y_i \tag{5}$$

$$q_i \equiv x_i - < x >$$
$$z_i \equiv y_i - < y > \tag{6}$$

Note that

$$\sum_{i=1}^{N} q_i = 0, \quad \sum_{i=1}^{N} z_i = 0 \tag{7}$$

a relation that is used extensively. Note also that <...> does not refer to an ensemble average, but rather an average over all samples in a given training set (given realization). It is convenient to define the **variance W** of the q values

$$W \equiv \frac{1}{N} \sum_{i=1}^{N} {q_i}^2 \tag{8}$$

The solution to the OLS equations are:

$$\alpha = < y > - \beta < x > \tag{9}$$

is the estimated intercept, and the *estimated* slope is

$$\beta \equiv \frac{1}{NW} \sum_{i=1}^{N} q_i \, z_i \tag{10}$$

The *observed* residuals are $y_i - \widehat{y_i}$, and this in general is not the same as ε_i We note that

$$\sum_i x_i \left(y_i - \widehat{y_i} \right) = \sum_i x_i \left(y_i - \alpha - \beta x_i \right) = 0 \tag{11}$$

Eq (11) is true regardless of g(x) and regardless of the statistics of the residuals. In fact, the residuals need not be statistically independent (so eq (3) does not hold). The relation (11) is simply the consequence of eqs (9) and (10).

several convenient relations

From (6) and (1) we find

$$z_i = b g(x_i) x_i + \varepsilon_i - < y > \tag{12}$$

Putting this into (10), and using (7) and the fact that <y> is independent of i, the estimate of β in a given realization is

$$\beta = \frac{1}{WN} \sum_i q_i \left(b g(x_i) x_i + \varepsilon_i \right) \tag{13}$$

Another relation that will be used is the following:

$$< y >= \frac{b}{N}\sum_i x_i\ g(x_i) + \frac{1}{N}\sum_i \varepsilon_i \qquad (14)$$

which is obtained directly from eqs (1) and (5).

mean and variance of the estimated slope

From (13) and (3), we see that,

$$E[\beta] = \frac{1}{WN}\sum_i q_i\ bg(x_i)x_i \qquad (15)$$

If g(.) = 1, this becomes

$$E[\beta] = \frac{1}{N}\sum_i \frac{bq_i^2}{W} = b \qquad \text{(g(.)=1)} \qquad (16)$$

The average $E[\beta]$ in (16) is a sum over $|x - < x >|^2$. The larger this quantity, the larger the contribution to the average slope.

As a simple example, suppose W=1 and <x>=0. Consider a continuum analog of eq (16)

$$A = \int_0^{1/2} dq\ q^2 = 1/24$$

$$B = \int_0^{1} dq\ q^2 = 1/3$$

so that

B / A = 8

This indicates that the larger q values make a much greater contribution to the average slope, eq (16), than the smaller ones. Then, since

$$\widehat{y_x} = \alpha + \beta\ x$$

for estimated y, we expect that the variance of $\widehat{y_x}$ is larger for large q (= x - <x>) than for smaller ones. This will turn out to be correct.

For the variance of β, define

$$\Delta\beta \equiv \beta - E[\beta] \tag{17}$$

Now we assume any g(x). From (13) and (15)

$$\Delta\beta = \frac{1}{WN}\sum_i q_i\,\varepsilon_i \tag{18}$$

and from (3) it follows that

$$E[\Delta\beta^2] = \frac{1}{W^2N^2}E\left[\sum_{i,j} q_i\,\varepsilon_i\,q_j\,\varepsilon_j\right] = \frac{1}{W^2N^2}\sum_i {q_i}^2$$

Finally from (8)

$$E[\Delta\beta^2] = \frac{\sigma^2}{WN} \tag{19}$$

The first of eqs (19) shows that for the variance, as for the average, larger values of $|x - < x >|^2$ make a larger contribution to the variance than smaller values.

discussion

Note that the variance (19) is independent of the slope bg(x)x, so eq (19) would be valid even if b were 0. It is the statistical error in the estimate of the slope.

The variance decreases with increasing N, as more data means in general a more certain fit.

> Note that, *when eq (19) shows that the variance in the slope is very small, it means ONLY that the fit to the data by a linear function has little uncertainty, and nothing else. It does not mean that it makes real-world sense to fit the data with the particular function* (4).

The unit of β is y/x. The quantity W is the variance of x (relative to <x>). So the smaller is W, relative to σ^2, the greater the variance of β. This is because a larger $\frac{\sigma^2}{W}$ means that the x values are clustered relative to the spread σ of y values. That is, there is a large statistical variation in y relative to the spread of x, so the uncertainty in slope is greater because of the spread of y in the different ensembles. On the other hand, if the spread in x values is much greater than the spread in y values, the variance is small as the slope is estimated much more certainly.

variance of yhat - E[yhat]

Here we consider the fluctuation in $\hat{y}$ relative to the mean of $\hat{y}$. This does not describe how well the data is fit but rather the internal uncertainty in fitting a line to the data. For example, it is possible that this variance is very small and yet the actual data is severely non-linear and the linear model makes no real-world sense. The reason is that, in this case, the mean values cancel, so that the variance is independent of true slope. (In other words, this quantity does ot reflect *bias*.)

By definition of the regression function (4), in a given ensemble

$$\widehat{y_x} = \alpha + \beta x \tag{20}$$

where the hat means the estimated value (both α, β are implicitly hatted, as discussed above).

From (1), (5), (6), (9), we find

$$\widehat{y_x} = \beta q_x + < y > \tag{21}$$

where <y> is given by eq (14) and β by eq (13). We want to compute $E[\widehat{y_x}]$ and the variance of

$$D\widehat{y_x} \equiv \widehat{y_x} - E[\widehat{y_x}] \tag{22}$$

Since

$$E[\widehat{y_x}] = q_x E[\beta] + E[< y >] \tag{23}$$

then from (14) and (13) and (3) we find

$$E[\widehat{y_x}] = \frac{b}{N}\sum_i x_i g(x_i)\left(1 + \frac{q_x q_i}{W}\right) \tag{24}$$

The first term in parens in (24) comes from E[<y>] and the second term from the average of β.If g(.) = 1, then (24) becomes

$$E[\widehat{y_x}] = bx \qquad \text{(g(.)=1)}$$

To calculate the variance of (22), the non-random terms in $\widehat{y_x}, E[\widehat{y_x}]$ are the same (obviously) and therefore cancel, so

$$D\widehat{y_x} = \frac{1}{N}\sum_i \varepsilon_i + \frac{q_x}{NW}\sum_i q_i \varepsilon_i \tag{25}$$

The variance is easily seen to be

$$E[D\widehat{y_x}^2] = \frac{\sigma^2}{N}\left(1 + \frac{q_x^2}{W}\right) \tag{26}$$

Eq (26) is valid whether x is in the training set or not. If x is in the training set, then $D\widehat{y_x}$ is labeled $D\widehat{y_i}$ and

$$\frac{1}{N}\sum_i D\widehat{y_i}^2 = \frac{2\sigma^2}{N} \tag{26a}$$

The first term of (26) is the contribution from the variance of <y> (the random part of which is just a sum of random variates), the second the contribution from the variance of β. There is no cross term because of (7).

We see that the larger is q_x, the larger is the variance. This is the same effect as for the variance of the slope.

We can get insight by considering the correlation

$$E[D\widehat{y_x}\ D\widehat{y_{x'}}] = \frac{\sigma^2}{N}\left(1 + \frac{q_x q_{x'}}{W}\right) \tag{27}$$

The correlation is positive if the qs have the same sign, and negative if they have different signs.

In going from ensemble to ensemble, we can picture the change of the estimated slope as being due to two factors:

(1) A translation of the whole (fitted) line up and down (ie, along the y axis), and
(2) A rotation of the fitted line about <x>.

In Eq (27), the first term is due to the translation (it is due only to the residuals), while the second term is due to the rotation. Since in a rotation, one end of a given fitted line goes up (down) if the other end goes down (up), we expect the estimate of y from one end to be the opposite of an estimate of y from the other end, assuming both estimates are made from the same ensemble member. When averaging over ensembles, we get the correlation (27).
Since for a given rotation amount, the amount of change in y and $\widehat{y}$ is greater the farther from <x>, the variance should increase with distance from <x>.

variance of yhat - ytrue

Finally, we calculate the variance of the difference between the estimated value of y and the true value, at a given x. This is really what we want to know. Define

$$\Delta\widehat{y_x} \equiv \widehat{y_x} - y_x \tag{28}$$

where $\widehat{y_x}$ is given by (21) and y_x is given by (1). We will see that the variance of (28) depends on whether x is one of the sample values x_i or not.

It is important to be clear about the whether x in (20) is or is not in the training set. In both cases, the value of x itself is the same in all ensembles. However, the residuals are different. If x is in the training set, so $x = x_i$ for some i, then eq (1) has $\varepsilon_x = \varepsilon_i$. However, if x is not in the training set, then the value of ε_x in any ensemble is different from any of the ε_i in that ensemble. This leads to different averages when computing the variance.

Substituting as usual, we find

$$\Delta\widehat{y_x} = A_x + \frac{1}{N}\sum_i(\varepsilon_i - \varepsilon_x) + \frac{q_x}{WN}\sum_i q_i\varepsilon_i \tag{29}$$

where

$$A_x = \frac{bq_x}{NW}\sum_i q_i x_i g(x_i) + \frac{b}{N}\sum_i x_i g(x_i) - bxg(x) \tag{30}$$

The quantity A is the **bias**. Note the presence of ε_x in eq (29). It is straightforward to show that

$$A_x = 0 \quad if\ g(.) = 1 \tag{31}$$

So A_x (bias) is non-zero only if the true slope is non-linear. We find:

$$E[\Delta\widehat{y_x}^2] = A_x^2 + \begin{cases} \sigma^2\left(1 - \frac{1}{N}\right) - \frac{q_k^2\sigma^2}{NW} & if\ x = x_k \\ \sigma^2\left(1 + \frac{1}{N}\right) + \frac{q_x^2\sigma^2}{NW} & if\ x \neq any\ x_k \end{cases} \tag{32}$$

For the top equation (32), there is a non-zero cross term for the two terms with residuals in (30), and that is responsible for the subtraction. In the bottom equation, that cross term is 0. In a sense, the first of eqs (32) might be a kind of overfitting, in that in-sample predictors are better than out-of-sample ones.

Let g(.) = 1. In this case, the top equation of (32) gives

$$\frac{1}{N}\sum_i E[\Delta\hat{y}_i^{\,2}] = \sigma^2\left(1-\frac{2}{N}\right) \tag{33}$$

When N = 2, the total variance is 0. This means each term, (32), is also zero for N=2 and *x a sample variable.* (This can be shown directly.) This result is due to the fact that, for only two samples, a line is an exact fit.

The difference between the top and bottom of eqs (32) is due to the independence of the residuals, as expressed in the second of eqs (3). Even if x_k and x are different by a tiny amount (for some k), the variances (32) could be very different. This is a direct consequence of eq (3) and of the iid generation of residuals. This independence assumption may or may not make sense in a given situation.

If x = x_k for some k, then the values of $\hat{y}_k$, y_k are correlated since their deviation from the fitted line in a given ensemble member is related to the same ε_k . This is not true if x is not a sample point, so the values are uncorrelated and the variance is larger.

a numerical study

This section presents a numerical study of the equations of this note. All data is computer generated. We have the following:

N = 20 (number of samples)
b = 0.5 (true slope)
true intercept = 0 (consistent with eq (1))
σ = 0.3 (standard deviation of variates)
Number of ensemble realizations = 5000

The x sample values are generated from a square distribution, and normalized to have mean 0 and W = 1 (eq. (8)).

The y values are generated from eq (1), for g(.) = 1 - that is, constant (linear) slope.

The histogram of variates ε_i, accumulated over the entire ensemble, is shown in Fig 1 below. The variates are generated from a triangular distribution.

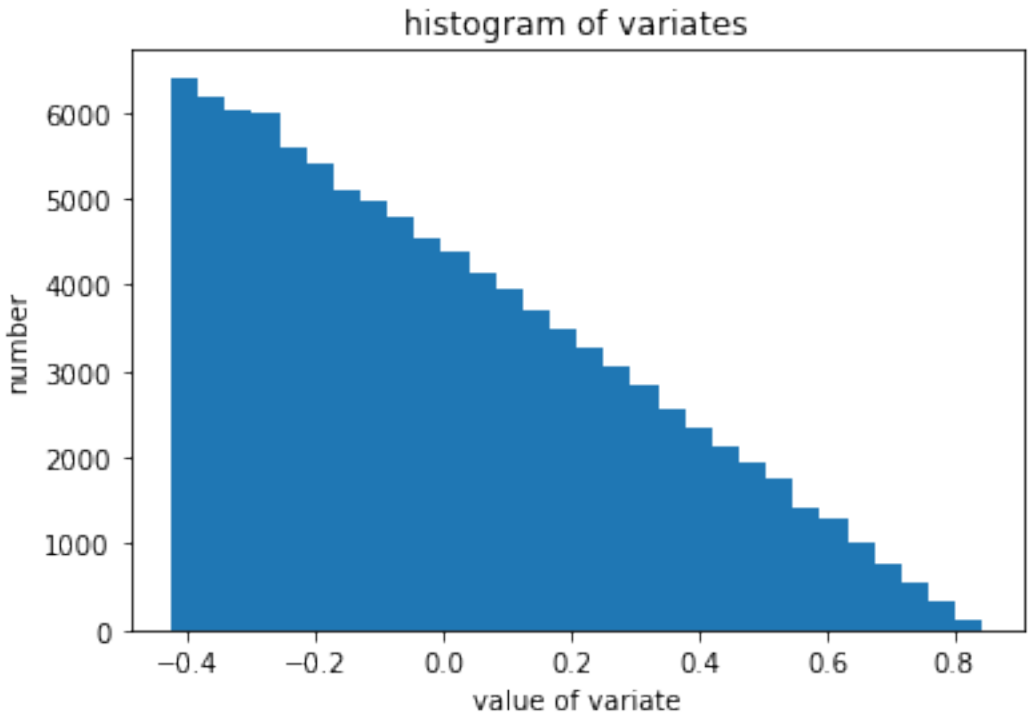

Figure 1

Fig 2 below shows y_i values and the fitted line vs x-<x> for one particular realization, but with N = 200. This is just for illustration purposes. The non-symmetric nature of the distribution of variates is clear. All other figures have N = 20. The line in Fig 2 is the OLS fit.

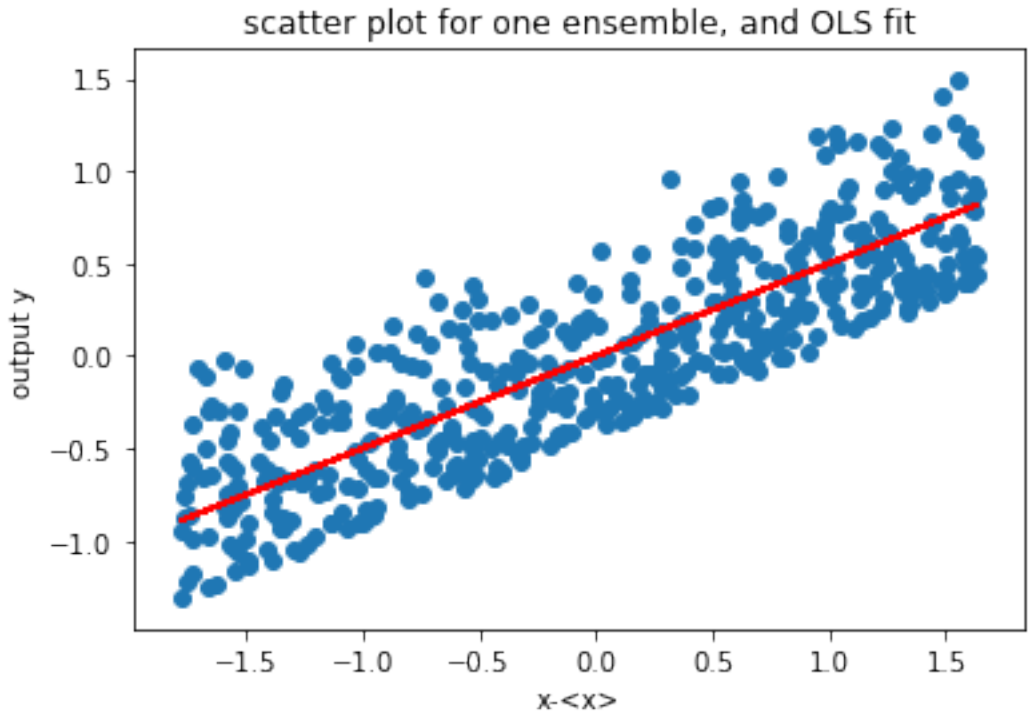

Figure 2

`

Back to N = 20. Figs 3 and 4 below show the histograms (values aggregated over the ensemble) of estimated slopes and estimated intercepts.

The value for the variance of the estimated slope, computed from the ensemble, is 0.00464. The theoretical value from eq (19) is 0.0045.

The average slope, computed from the ensemble, is 0.499 The exact value is 0.5.

The correlation coefficient between slopes and intercepts is -0.00295.

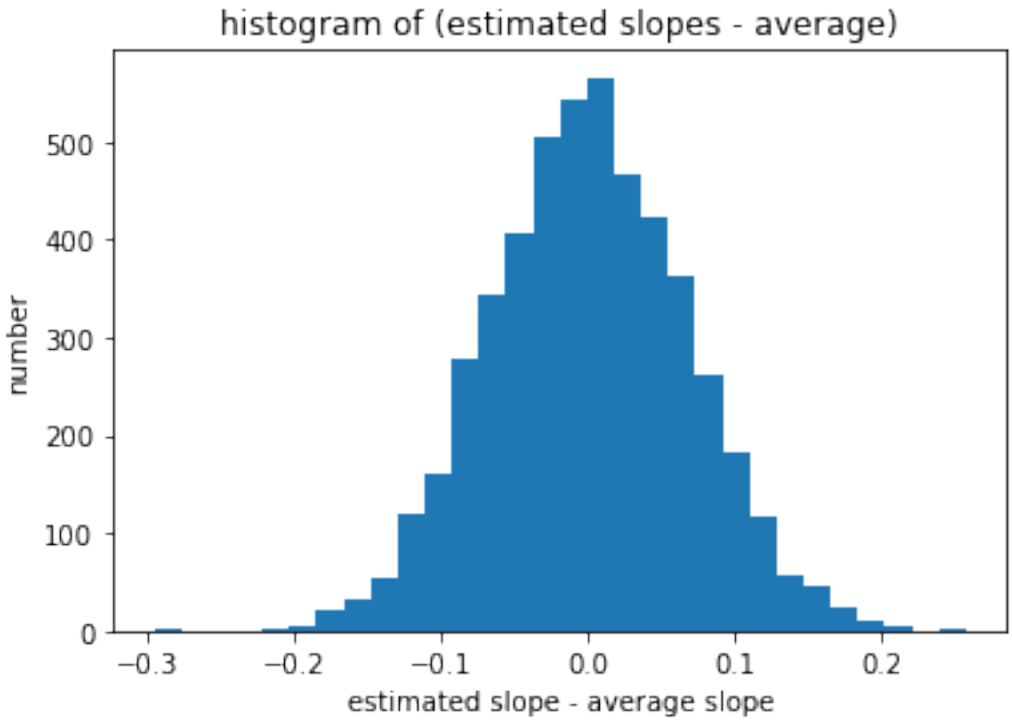

Figure 3

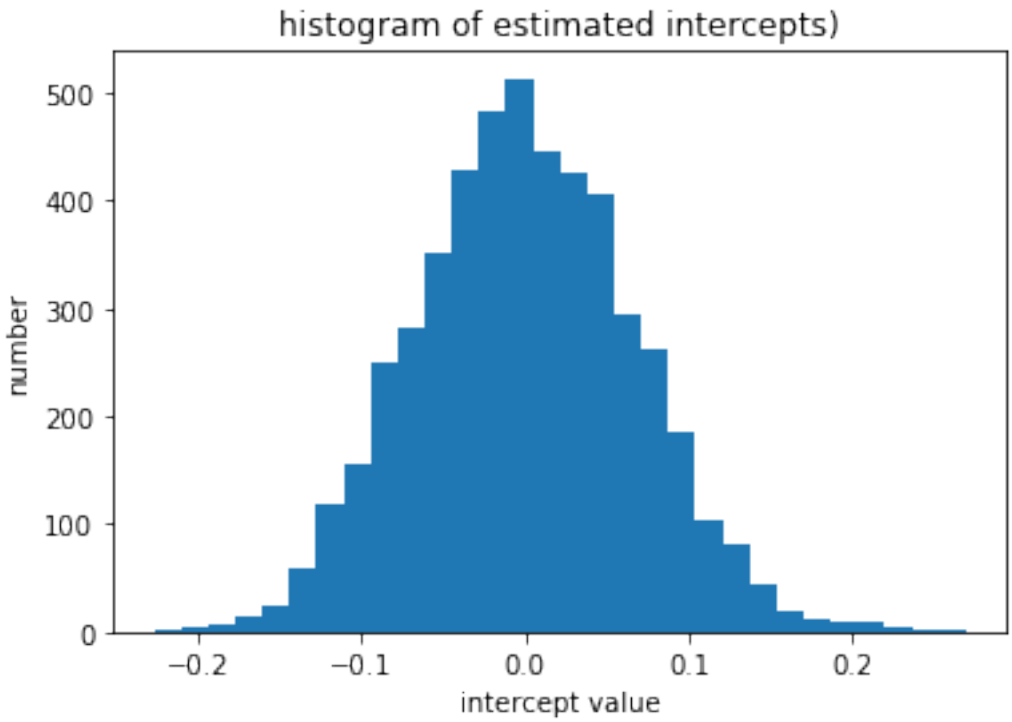
histogram of estimated intercepts)
number
500
400
300
200
100
0
−0.2
−0.1
0.0
0.1
0.2
intercept value

Figure 4

Fig. 5 shows $E[D\widehat{y}_x^{\,2}]$, defined in eq (22), as a function x-<x>. The filled circles "o" are computed from the ensemble. The "x" are values computed from eq (26).

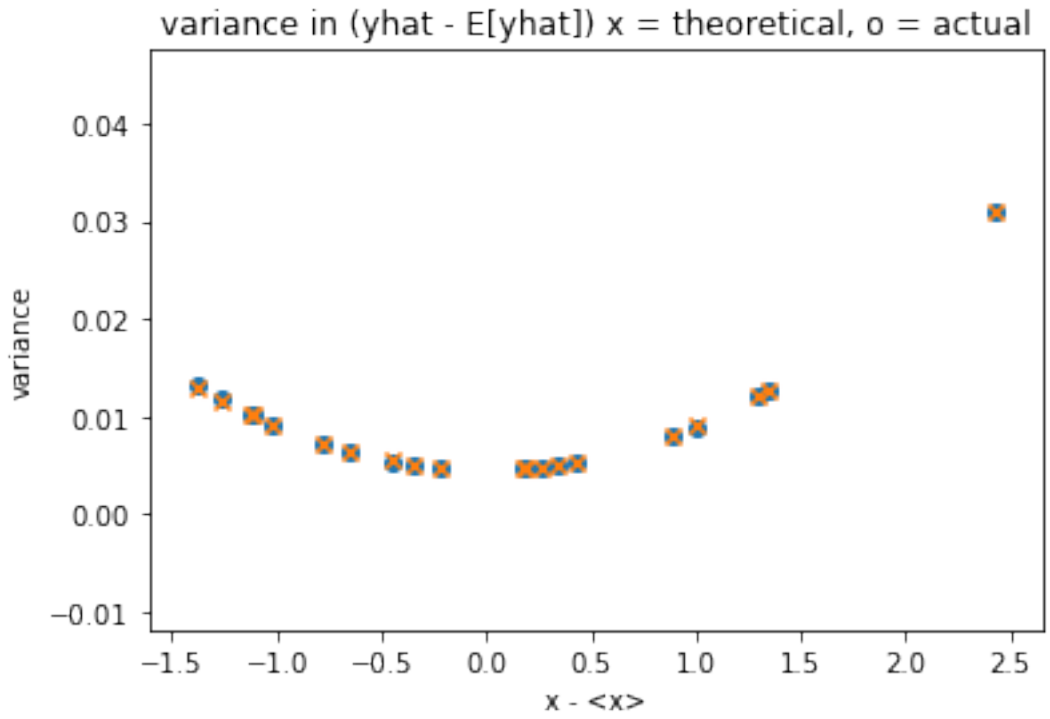

Figure 5

Figures 6 and 7 show values of $E[\Delta\widehat{y}_x^{\,2}]$, defined in eq (28), as a function of x-<x>. The filled circles are computed from the ensemble and the "x" points are from eqs (32). The ensemble size ids 5000. Figure 8 is the same as Fig 7, except that the ensemble size is 500.

In Fig 6, the values of x are the training-set values, and so the top eq (32) is used to compute the theoretical values (marker "x").

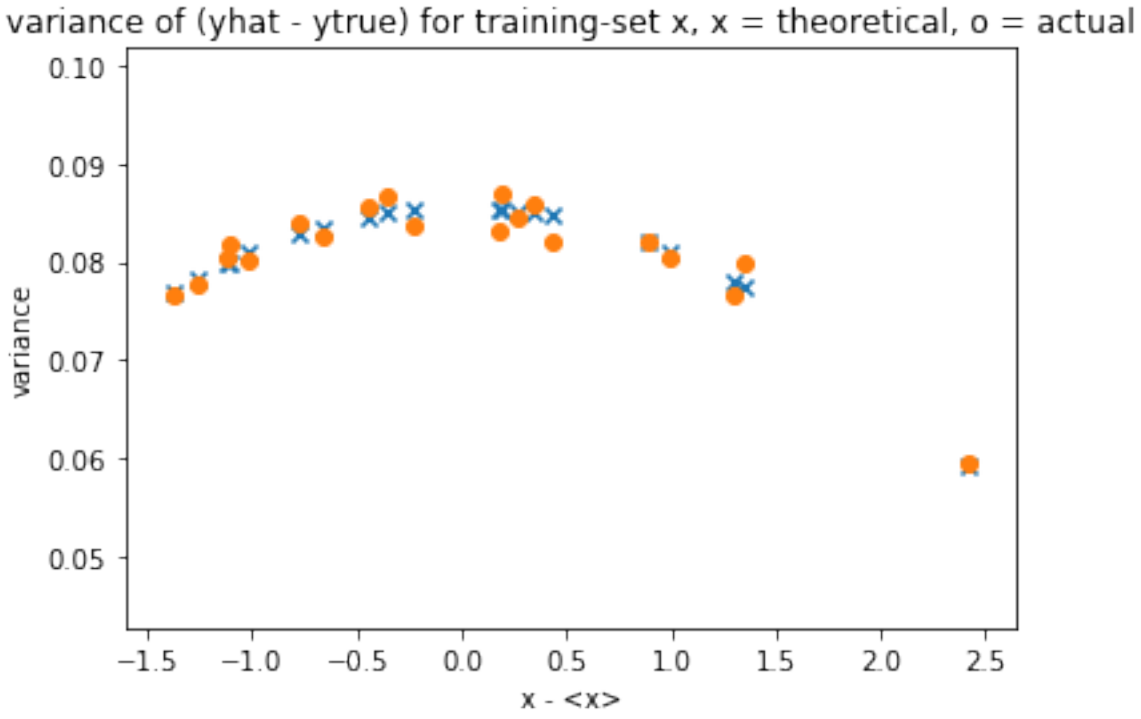

Figure 6

In Fig 7, the values of x are test values (not in the training set), so the bottom of eqs (32) is used to compute theoretical values.

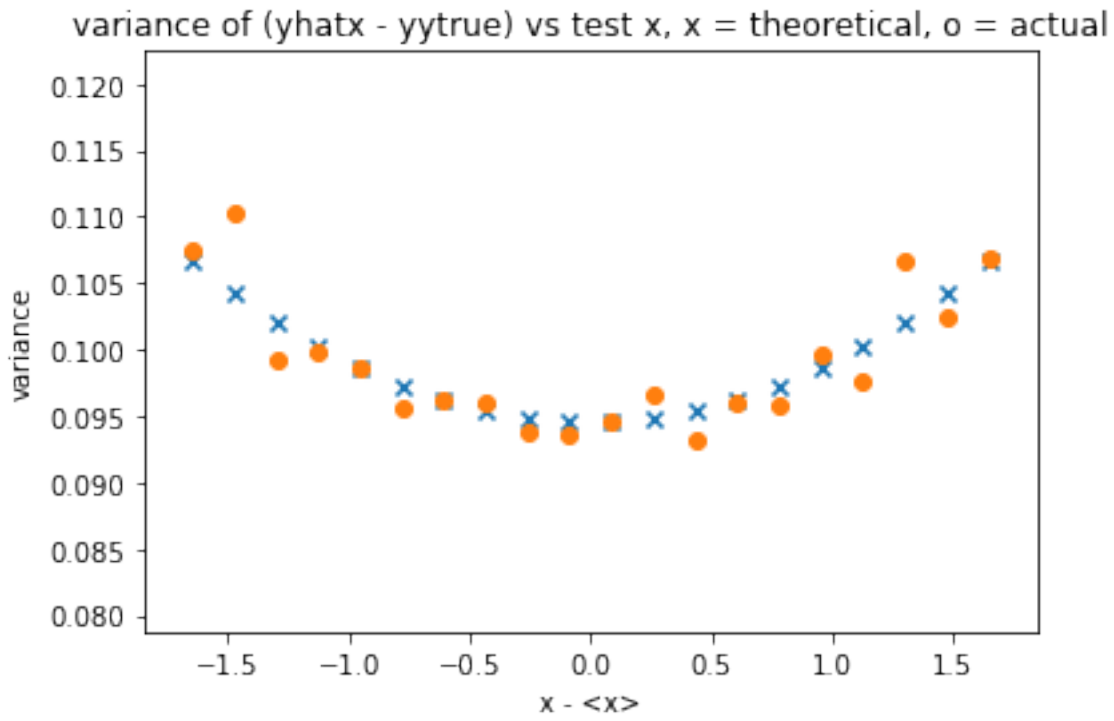

Figure 7

The fluctuation of simulated values (filled circles) in Figs 6, 7 are statistical fluctuations due to the finite size of the ensemble. Fig 8 shows the same graph as Fig 7, but with an ensemble size of only 500. The statistical fluctuations are seen to be much greater.

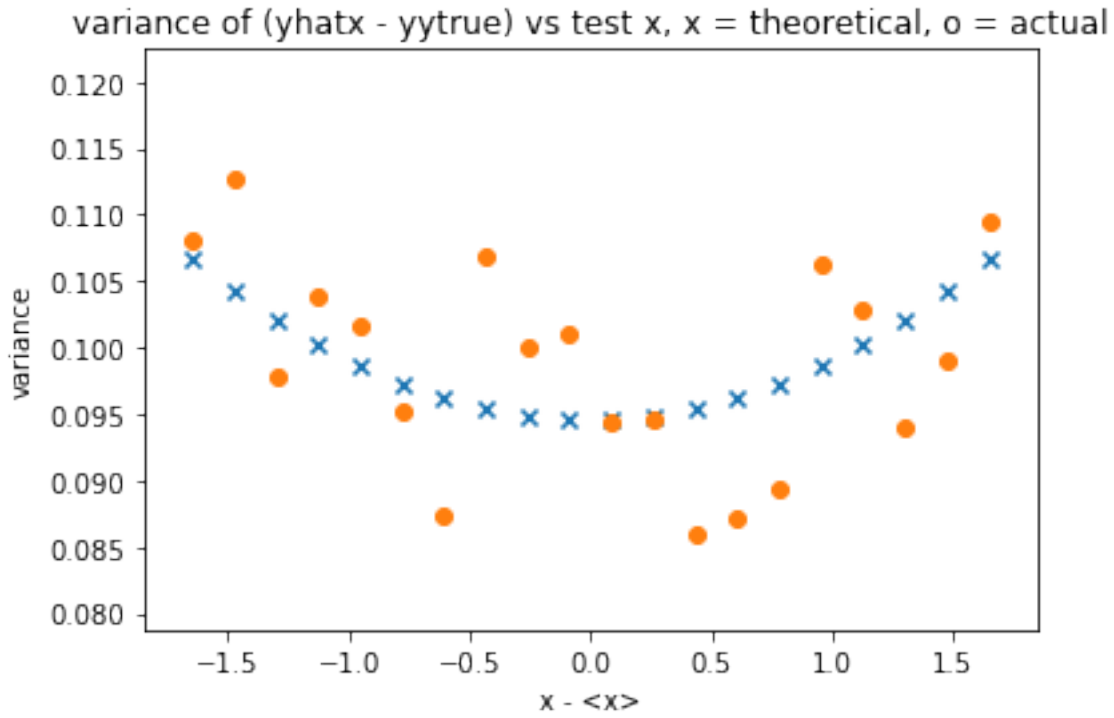

Figure 8