Measurement re-calibration the right and fair way

Bertrand Haas*

Abstract

Measurements of a physical quantity by measuring devices are usually noisy enough that we need to correct, or at least mitigate, the effects of noise. For this purpose, it's important to distinguish between **systematic** and **random** noise since they are of a different nature and independent from each other (when defined properly), so should be dealt with differently. For example, random noise can be significantly mitigated by averaging repeated measurements... when possible (and optimally in different controlled conditions). In contrast, systematic noise (or **bias**) can be corrected using regression with some ground truth (or accurate proxy of ground truth) as independent variable. However, regression is often used in a predictive way with ground truth as a response variable, with the effect of minimizing total errors (random plus systematic). Although this might sound reasonable, it turns out that besides blurring the distinction between random and systematic noise, it achieves nothing for the former and does not completely correct the latter. Moreover, the residual bias might be small for the majority of the population, but usually grows larger further away from the mean. In many cases this yields to **unfair** outcomes, in particular when the measurement is on human biological quantities. This is especially important when the biological quantities are used to determine if and how much a subject is sick or unhealthy, usually by how much its measurements deviate from the mean. We argue here that re-calibration done with this in mind should be bound to eliminate systematic noise, that random noise should be mitigated by other means (like improved device engineering), and that KPIs[1] based on total errors should be revised.

1. Introduction

Calibrating or re-calibrating a device for optimizing accuracy and precision of measurements is an important part of device engineering. We will first define "accuracy" and "precision", the former being an inverse measure of systematic noise (or "bias", a synonym), and the latter being an inverse measure of random noise. We usually write X for the ground truth variable, and Y for the measurement variable. Bias correction can be done in an "explanatory" way by regressing Y as a response from X (we write $Y \sim X$), or a "predictive" way, by $X \sim Y$. We will see that the former way makes more sense then the latter, despite yielding slightly worse Mean Square Errors (*MSE*).

We will demonstrate our points with algebra and illustrate it with simulations. For simplicity we will usually assume the data is normally distributed and relations between variables are linear but in section 5 at the end we will mention techniques to deal with more general distributions and non-linear relations.

We tried to make the exposition a little more engaging through a dialog between a statistician, Proserpina Statopoulos (PS) and a machine learning scientist, Michael Learner (ML), who might embody different viewpoints on the problem.

*Senior Data Scientist at Amazon

[1] Key Performance Index

2. Basic background

ML: Hi Proserpina, I am working on some re-calibration project and I would like to know your statistician's opinion about a couple of things. I did try to find some simple perspective in the literature, for example (Brown 1994), but I haven't been convinced yet.

PS: Sure. You might also look at (Fuller 1987), but these are mostly academic books. In the meantime it would certainly be nice to share and compare our own practical views and techniques on the subject. But let's start with your project.

ML: Thanks! I suppose my project is quite typical. We have a device that measures a continuous variable and we want to make it as accurate as possible.

2.1 Accuracy vs Precision: One-shot ground truth, many measurements

PS: Accurate? How about precision.

ML: Well, yes, both accurate and precise. Just to be sure we are on the same page about these terms, can you remind me the difference?

PS: Of course. A quick search for "Precision vs Accuracy" on the internet ("reliability" and "validity" are also used) will yield self-explanatory visual explanations in the context of target shooting. In our context, measurements of a physical quantity, the bull's eye is the ground truth and the "shots" are the measurements. But measurements are usually 1-dimensional, so let's draw them as distributions (for the population) and histograms (the sampling version of a distribution) into a 2x2 matrix (high/Low Accuracy/Precision) with respect to a single ground truth value (figure 1).

ML: Right. A distribution, univariate or multivariate, is firstly described by its location (or "center") and its spread. Statistics for the location are usually the mean, the median (or geometric median in more than one dimension), or the mode (assuming it is a nice unimodal distribution). The typical statistic for the spread of a multivariate distribution is the variance-covariance matrix. For univariate distributions, that corresponds simply to the variance, or its square-root, the standard deviation, and sometimes other statistics, such as the *MAD* (Mean Absolute Deviation), or the MAPE (Mean Absolute Percentage Error) are preferred if they make more sense for the problem at hand.

For simplicity, suppose our measurement or ground-truth distributions are unimodal, symmetric, and at least their 2 first moments are finite, like a normal distribution, so mean, median, and mode exist and are the same, and we'll use the variance as the statistic for the spread.

PS: Great set up. So accuracy is a measure of how close is the distribution of measurements located from the ground truth, like the distance between the bull's eye and the center of the shots, or between the mean of the measurements and the actual ground truth.

ML: Right, and precision a measure of how little the distribution spreads around its center.

PS: So we can define *random noise* as what causes the lack of precision and *systematic noise*, or *bias*, as what causes the lack of accuracy.

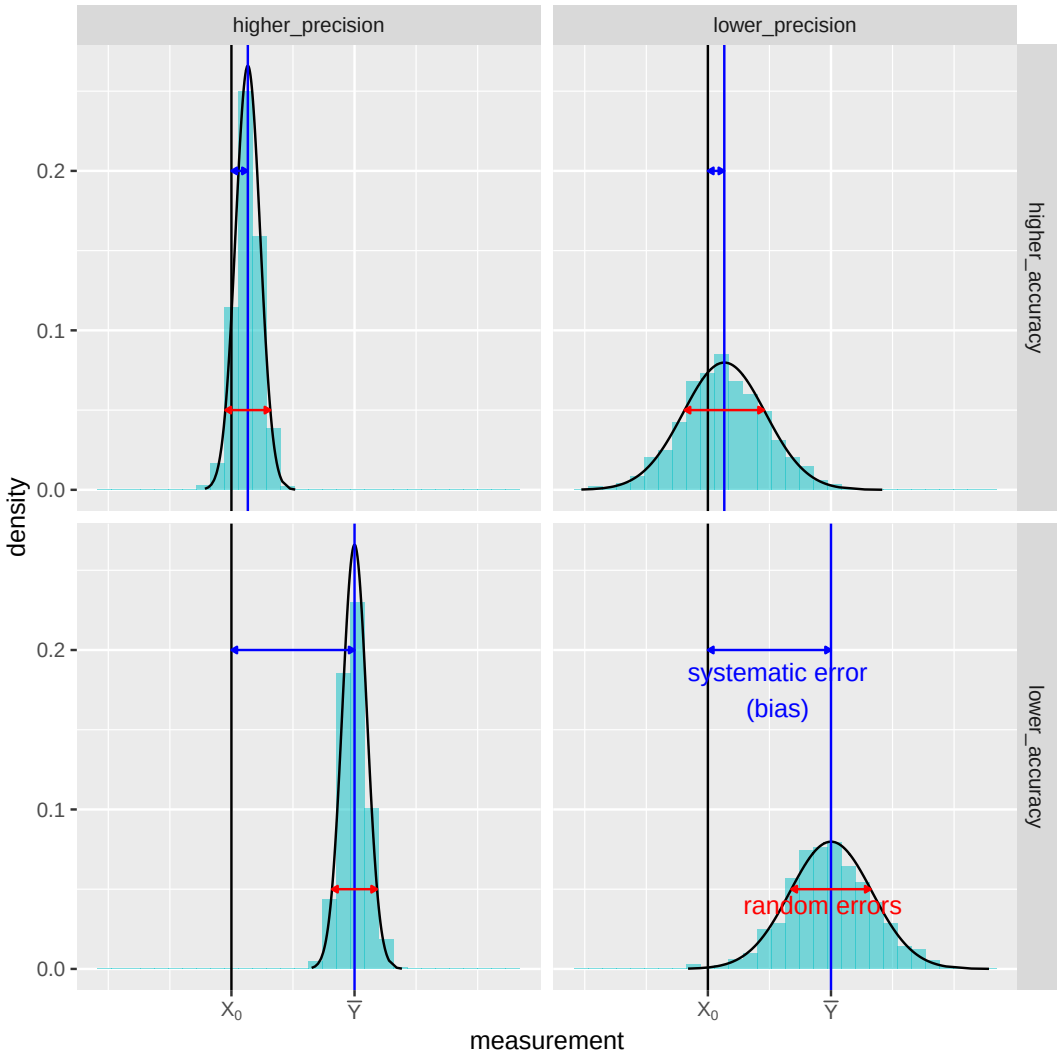


Figure 1: Accuracy/Precision High/Low for many measurements Y and one ground truth value X0.

ML: Exactly, so it looks relatively easy to improve accuracy (as opposed to precision) by *bias correction*: For the shots, just fix the cross-hair, or aim to a point opposite to the distribution center from the bull's eye. For measurements, just subtract the difference between the mean of the distribution and the ground truth X_0 to your measurements.

PS: Well, it makes sense for target shooting, but for measurements, we usually don't know the ground truth for all the measurements, otherwise we would not need measurements. Let's assume we have a training set, a sample of n pairs (ground-truths, measurements), say (X_i, Y_i), and both are continuous variables.

2.2 Measurements on continuous ground truth: Regression

ML: Yes, here we have sampled many ground truth values, each with one measurement. So it is indeed different, but with some assumption, we have a similar setting.

PS: Let's clarify these assumptions, since they are often unspoken. I think the main assumption is that noise (both random and systematic) is not too wild. More precisely, it is not very different from one ground truth value to one very close to it. In other words, the noise distribution varies continuously with ground truth values.

ML: Right. And we can usually verify that on a scatter plot with ground truth as one variable and measurement as the other.

PS: So despite having only one measurement per ground truth sample, we can "borrow" data from neighboring ground truths to determine the distribution of measurements and its location and spread, that is, determine the random and systematic noise.

ML: Yes, and that is done by *regression* analysis. Often, like in my project, the scatter plot looks rather elliptical, so we can make further assumptions, like the random error distribution depends on the ground truth only in location, and linearly so, and it has constant spread. This is the right set-up for *linear regression*. If we write Y for the measurements' random variable and X for the ground truths' random variable, we have a linear model $Y \sim X$, in other words:

$$Y = a_0 + a_1 X + \epsilon \tag{1}$$

Where ϵ is a random error variable with mean 0 and variance η^2. Often random errors are assumed to be normally distributed, but they don't have to be, as long as the distribution is "well-behaved" enough, like unimodal, symmetric, and with at least the two first moments finite.

PS: Let's plot a similar 2x2 matrix (low/high accuracy/precision), so we can see clearly what we mean by random and systematic noise in this regression set-up (figure 2).

PS: Here I drew the 95% confidence ellipse on top of the scatter plot, so its vertical width gives an idea of the amount of random errors. And the more the regression line departs from the diagonal, the more the bias. Also, if you take vertical sections of the plot (borrowing some points nearby the section if necessary), you should see 1-dimensional plots like in figure 1.

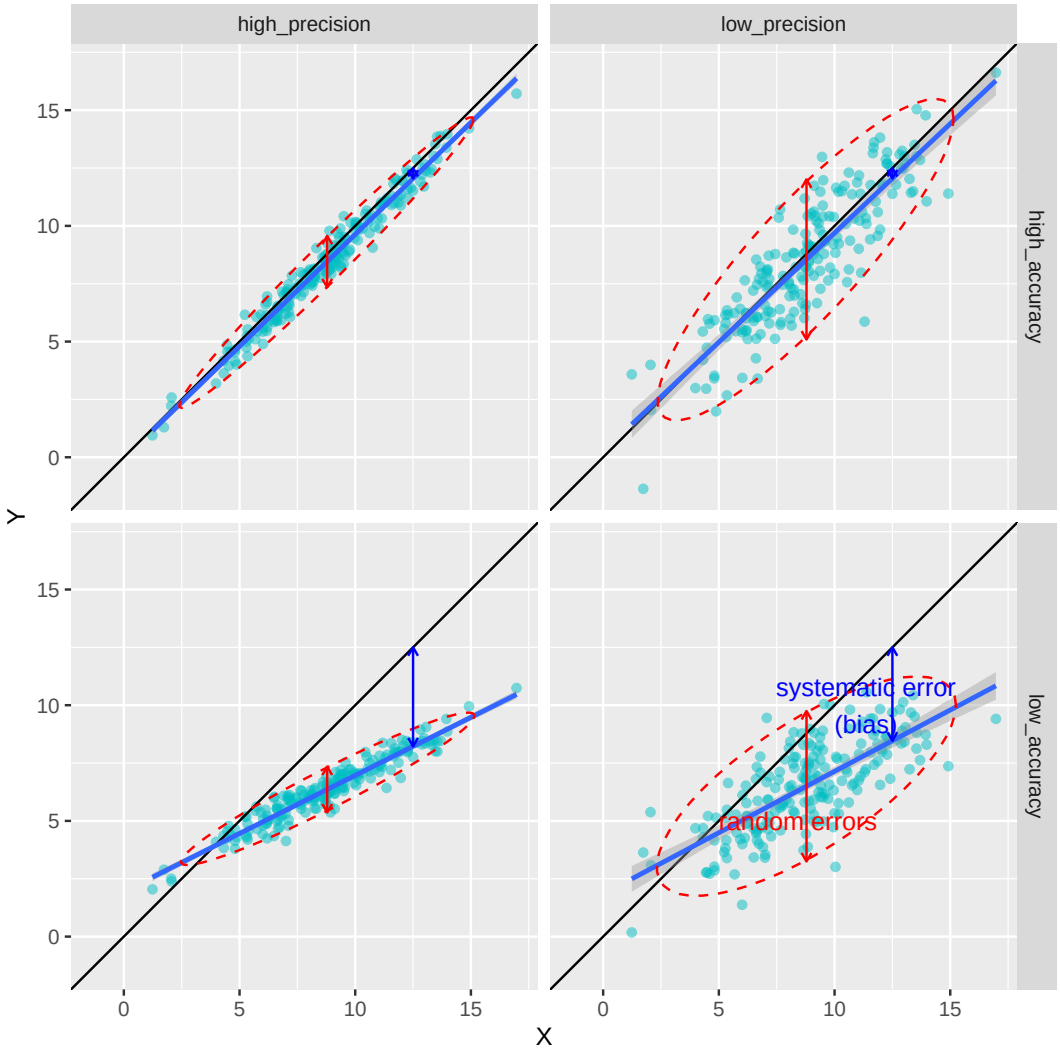


Figure 2: Accuracy/Precision High/Low for measurements Y on continuous ground truth X.

ML: So, here there are 2 ways the regression line can depart from the diagonal. It stems from the 2 parameters of a line, slope and intercept.

PS: Correct. So we can split our systematic errors further in to 2 types:

- On the one hand, it could be parallel to the diagonal (slope $a_1 = 1$) but vertically offset from it (large absolute intercept a_0). Let's call that "additive bias" since a_0 is just added to X, or "independent bias", since the bias is constant with X (it does not depend on it).
- On the other hand, it could have a large absolute slope a_1, but the center of the scatter-plot $(\overline{X}, \overline{Y})$ could be right on the diagonal. If we recenter the values by subtracting $\overline{X}$ and $\overline{Y}$ respectively to X and Y, the center of the plot would be $(0, 0)$ and the intercept $a_0 = 0$. We can call that "mutiplicative bias", since X is multiplied by a_1, or "dependent bias", since the bias as depicted in figure 1 depends on X.
- "Linear bias" is usually a combination of both but we can also imagine more complicated bias (imagine a "curvy" regression line).

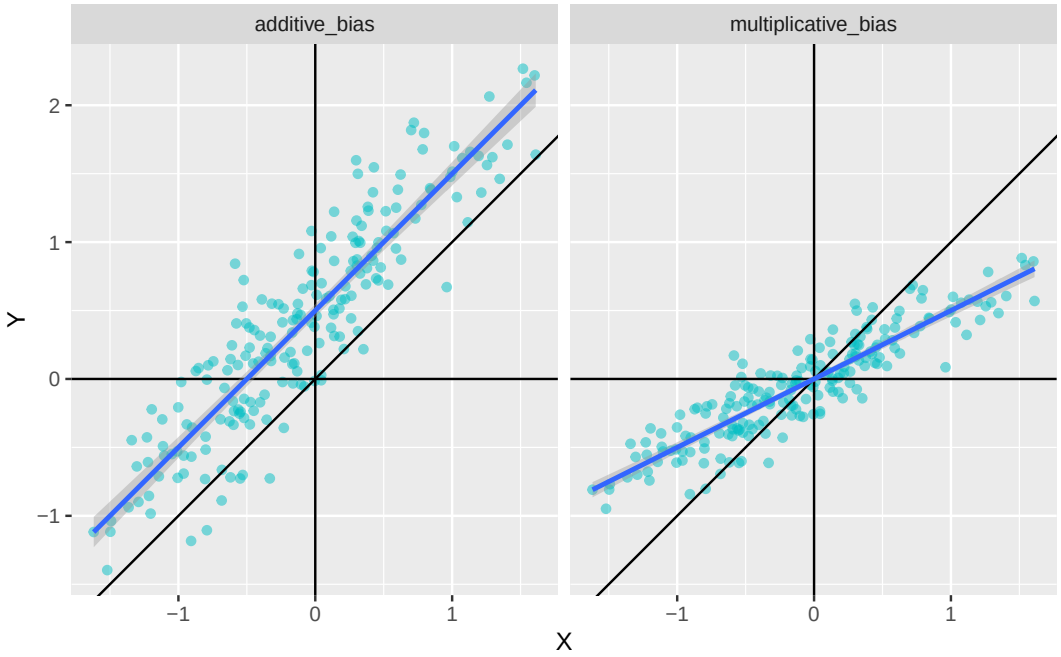


Figure 3: Additive bias (left) vs multiplicative bias (right).

3. Bias correction

3.1 Additive-bias correction

ML: It sounds relatively easy to correct for the additive bias, then:

1. Compute the sample means $\overline{X}$ and $\overline{Y}$ and let $a_0 = \overline{Y} - \overline{X}$ for the additive systematic error.
2. Then set $Y' = Y - a_0$, so now $\overline{Y'} = \overline{X}$.

We actually don't even need a 2-D plot for that (figure 4).

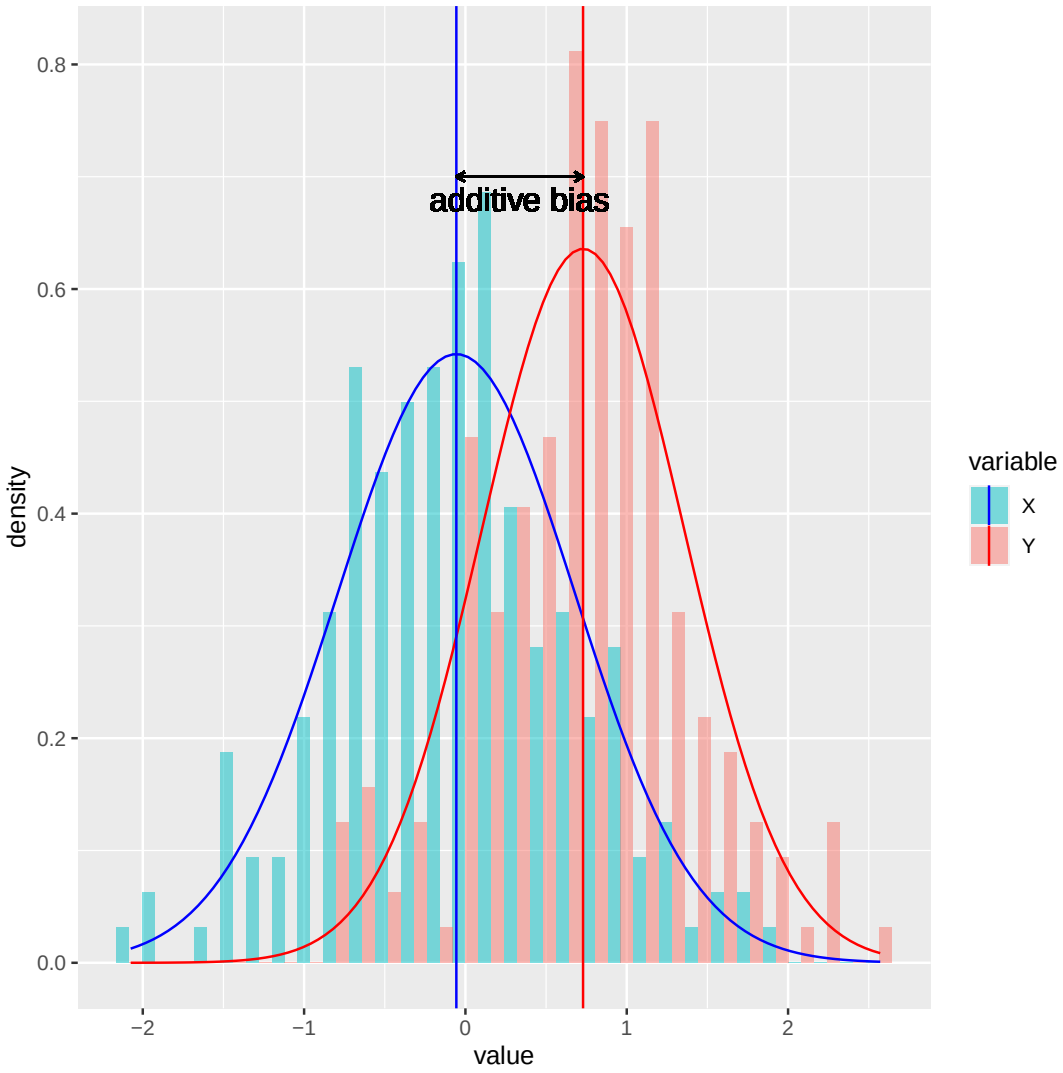


Figure 4: Additive bias is the difference between the centers of the X and Y distributions.

3.2 Multiplicative-bias correction and the independence of random noise and systematic noise

ML: Ok, so suppose we have corrected the additive bias a_0, so now we have only a multiplicative bias:

$$Y = a_1 * X + \epsilon$$

Where ϵ is the random noise variable (with mean zero). So, it is also quite straightforward to correct for that. Just set $Y' = Y/a_1$.

PS: That's correct, but you are touching another important point here. By doing that we get now

$$Y' = X + \epsilon/a_1$$

So it looks like we modified our random noise too. However, it should be quite intuitive that systematic noise and random noise, being physically different types of noise, should be independent from each other.

- Systematic noise stems from a poor calibration of the device and should easily be corrected by re-calibration.
- In contrast, random noise comes from variables out of our control so far:
  - Internal from the device, where it could be mitigated by hardware or software engineering improvement.
  - From external conditions (temperature, movement, atmospheric pressure, etc), where it can be mitigated, when possible, by controlling for the corresponding external variable.
  - Or from a combination of both... Or even from quantum noise, if you want to push it that far :-).

So, if you apply or correct for some systematic noise, that should not affect the random noise.

ML: Does this mean that we should define random noise differently, that is, after bias correction?

3.2.1 Modeling the generative process: First generate random noise then systematic noise

PS: Yes, that would be more consistent. Another way to view it is by "reverse engineering the noise", or by a "generative process", that is, start with a noise-less signal

$$Y^{(0)} = X$$

Then add noise in this order:

- Random noise ϵ:

$$\begin{aligned} Y^{(1)} &= Y^{(0)} + \epsilon \\ &= X + \epsilon \end{aligned}$$

- Systematic noise (both additive a_0 and multiplicative a_1):

$$\begin{aligned} Y = Y^{(2)} &= a_0 + a_1 Y^{(1)} \\ &= a_0 + a_1(X + \epsilon) \end{aligned} \tag{2}$$

ML: I see. So what I called random noise before was a scaled version of what we should call random noise.

3.2.2 Random noise versus random errors

PS: Actually, this is important, so let's be precise and distinguish between **random errors** and **random noise**.

Definition 3.1. The **random error** variable is ϵ' in the setting $Y = a_0 + a_1 X + \epsilon'$.

This is the usual definition in the context of linear regression. So, if you scale the measurements Y, then you scale the random errors as well.

Definition 3.2. The **random noise** variable is ϵ in the setting $Y = a_0 + a_1(X + \epsilon)$ as in equation (2).

When defined like that, random noise is indeed independent of measurement re-scaling.

In contrast, systematic noise is not defined as a random variable. While we are at it, let's also define it formally:

Definition 3.3. Systematic noise is the discrepancy between the regression line (straight line if the regression is linear) and the diagonal when the measurement is the response variable and the ground truth is the explanatory variable.

Random errors and noise are centered on the regression lines (respectively before and after bias correction, so the regression line is the diagonal $Y = X$ for random noise). So in simpler cases, like homoskedastic normal noise they are defined by a single number, the variance. In the case of a linear model, systematic noise, as we saw earlier, is defined by two numbers one related to the additive part and one to the multiplicative part (respectively a_0 and a_1 in (2)). We can also define one number, a **systematic errors** metric, like MSE.

3.2.3 The analytic process: First correct systematic noise then estimate random noise

ML: Alright, I think it's an important point, but before delving into that, let's come back to the more natural, non-reverse, non-generative, "observative" way. Here we are given a large enough sample set of pairs of the variables X and Y. So the way to compute the systematic and the random noise is:

1. Regress Y on X, that is, find the coefficients a_0 (corresponding to the additive bias) and a_1 (corresponding to the multiplicative bias) in:

$$Y = a_0 + a_1 X + \epsilon'$$

2. Re-calibrate your measurement by:

$$Y' = (Y - a_0)/a_1$$

   Which corrects the bias.
3. Estimate and report the random noise as the variance (or other measure of spread) of the residuals $X - Y'$. Or, in other words, the variance η^2 of the variable $\epsilon = \epsilon'/a_1$.

PS: Sounds right! By the way, the **error residuals** from our regressions are defined by the variable $R' = Y - \widehat{Y}$ (where $\widehat{Y} = a_0 + a_1 X$ corresponds to the fitted values), whereas the **noise residuals** are defined by $R = Y' - X = (Y - a_0) / a_1 - X =$

R'/a_1 $. The latter relates to the random noise, and therefore is the variable of interest for us.

ML: So plotting R versus X could be useful to detect fitting problems. Isn't that related to Bland-Altman plots?

PS: Well, it's in the same spirit, but Bland-Altman plots are usually for 2 methods of measurements, and not for one measurement versus ground truth, so they would actually be relevant for a different discussion where we do have 2 method of measurements but no ground truth.

3.3 SSTE = SSSE + SSRE

PS: For now, let's go back to random versus systematic errors. We can decompose the **S**um of **S**quared **T**otal **E**rrors (**SSTE**) into **S**um of **S**quared **S**ystematic **E**rrors (**SSSE**) and **S**um of **S**quared **R**andom **E**rrors (**SSRE**). Let's show that:

$$\text{SSTE} = \sum_i (Y_i - X_i)^2 = \sum_i (Y_i - \widehat{Y}_i)^2 + \sum_i (\widehat{Y}_i - X_i)^2 \tag{3}$$

- We call the first term on the rightmost side of the equalities the **SSRE**. Geometrically, the vertical difference between the points and the regression line are error residuals we just mentioned and the SSRE is the sum of the squares of these distances (so we can think of the **R** in SS**R**E as **Random** or **Residual**).
- We call the second term the **SSSE**. Geometrically, it is the sum of the squared vertical distances between the regression line and the diagonal (where $Y = X$) at each measurement.

The principle is the same as in textbooks' sum of square decomposition: Split a variable into two independent ones so their covariance (or cross-term when expanding the square) vanishes.

$$\begin{aligned}
\sum_i (Y_i - X_i)^2 &= \sum_i ((Y_i - \widehat{Y}_i) + (\widehat{Y}_i - X_i))^2 \\
&= \sum_i ((a_1\epsilon) + (a_0 + a_1 X_i) - X_i)^2 \\
&= \sum_i (a_1\epsilon)^2 + \sum_i (a_0 + (a_1 - 1)X_i)^2
\end{aligned}$$

Where the last line is due to the independence of the random noise variable ϵ with X.

Of course by dividing all the terms by $\frac{1}{n}$ you get the equivalent decomposition MSTE = MSSE + MSRE, where **S**um is replaced by **M**ean. Just don't confuse MSRE with RMSE (**R**oot **M**ean **S**quare **E**rror); All these terms look quite alike :-).

ML: This looks like the "bias-variance" decomposition.

PS: It is exactly that, but beware that the "bias" in that sense, that is MSSE = $\frac{1}{n}\sum_i (\widehat{Y}_i - X_i)^2 = \frac{1}{n}\sum_i (a_0 + (a_1 - 1)X_i)^2$ is only a measure of our more general notion of "bias" (or less ambiguously "systematic noise"). And the "variance" in that sense, that is MSRE = $\frac{1}{n}\sum_i (Y_i - \widehat{Y}_i)^2 = \frac{a_1^2}{n}\sum_i \epsilon^2$ is the variance of the random

errors, whereas we are interested in the variance of the random noise, that is $\frac{1}{a_1^2}\text{MSRE} = \sum_i \epsilon^2$.

4. Differences in perspectives: Explanatory versus predictive

ML: Actually, This decomposition might be related to a crucial point. Some of my colleagues have a different view on re-calibration. The view is that we are trying to predict ground truth values from measurement values. And in statistics or machine learning textbooks, we are taught in the same linear regression context that with a predictor (the measurement, Y) and a predicted variable (the ground truth X) we should set up a linear regression model $X \sim Y$, and not $Y \sim X$.

ML: To be specific, let's call that $Y \sim X$ model the **predictive model** and your $Y \sim X$ model the **explanatory model**[2]. So, the predictive model is:

$$X = b_0 + b_1 Y + e \tag{4}$$

Where e is the noise. It sounds like it makes sense and it's a simpler method than regressing Y on X and then inverting the linear relation. There are also precedents in literature and in practice. For example in pulse oximetry, SpO2 is computed from the measurements of a red light and a infra-red (*ired*) light signals. Their AC and DC components are measured and the perfusion indexes $P = AC/DC$ are computed. Finally the ratio $R = P_{\text{red}}/P_{\text{ired}}$ is supposed to be in linear relation with the ground truth SpO2, SpO2_{ref}:

$$\text{SpO2}_{\text{ref}} = c_0 + c_1 R$$

A large proportion of research papers (if not most) that mention this "calibration" process use the predictive model to infer those coefficients c_0 and c_1. Is there anything wrong with that way of doing calibration or re-calibration?

PS: Well, as the famous statistician George E.P. Box said, "All models are wrong, but some are useful". Let's add "Some are more wrong than others". Remember that in a linear regression $V \sim U$ (to avoid confusion with the letters used so far), the noisy variable is the one on the left side, V, and the one on the right side, U, is assumed noise-less. In our case the noisy variable is definitely the measurement variable, Y, not the ground-truth variable X, which by assumption is noise-less. So the "more right" model is definitely $Y \sim X$ and not $X \sim Y$. The latter might still be useful, but the former will be more useful, in the sense that the model will be closer to reality and the re-calibration should yield better results.

4.1 Simulation: Troubles for the explanatory model

ML: This is an important point. Let's do a quick simulation to make sure I get it right.

PS: Good idea.

ML: Assume the systematic and random noise for the measurements Y over the ground truth X is given by the linear relation (1), with the following parameters (they are close to the project I am working on):

[2] For a different perspective on Explanatory versus Predictive models see (Shmueli 2010)

Variable	Symbol	value
Ground Truth mean	μ_X	9
Ground Truth variance	σ_X^2	6
Mean Systematic noise	a_0	1.8
Systematic noise slope	a_1	0.7
Random noise variance	η^2	1.429
Number of observations	n	2000

So here is the data:

```
# A tibble: 2,000 x 3
       X      EY     Y
   <dbl>   <dbl> <dbl>
 1  8.37 -1.15    6.85
 2  5.67 -0.555   5.38
 3  9.97  0.585   9.19
 4 10.9   0.0755  9.45
 5 10.6   1.23   10.1
 6 12.8  -0.0347 10.7
 7  8.58  0.140   7.90
 8 10.2   0.751   9.49
 9 12.1  -1.28    9.35
10  8.07  0.912   8.08
# ... with 1,990 more rows
```

ML: Of course we already have the coefficients (a_0, a_1) since we just defined them, but let's be fair, pretend we do not know that, and infer those coefficients from the data by regressing $Y \sim X$.

$$\widehat{a_0} = 1.78$$
$$\widehat{a_1} = 0.702$$

Then, we invert them, that is determine $c_0 = 1/\widehat{a_0}$ and $c_1 = -\widehat{a_0}/\widehat{a_1}$ in the relation $X = c_0 + c_1\widehat{Y}$.

$$c_0 = -2.536$$
$$c_1 = 1.425$$

Besides that, we get directly the coefficients b_0 and b_1 from the linear regression $X \sim Y$. They correspond to the relation $\widehat{X} = b_0 + b_1 Y$.

$$b_0 = -0.426$$
$$b_1 = 1.166$$

ML: Now from your perspective, the correctly re-calibrated variable is $Y' = c_0 + c_1 X$ (exploratory model), whereas the incorrectly re-calibrated one is $Y'' = b_0 + b_1 X$ (predictive model).

```
# A tibble: 2,000 x 5
       X      EY     Y Y_recal_explan Y_recal_predic
   <dbl>   <dbl> <dbl>          <dbl>          <dbl>
 1  8.37 -1.15    6.85           7.23           7.56
```

```
 2  5.67 -0.555   5.38          5.13          5.85
 3  9.97  0.585   9.19         10.6          10.3
 4 10.9   0.0755  9.45         10.9          10.6
 5 10.6   1.23   10.1          11.8          11.3
 6 12.8  -0.0347 10.7          12.7          12.1
 7  8.58  0.140   7.90          8.72          8.79
 8 10.2   0.751   9.49         11.0          10.6
 9 12.1  -1.28    9.35         10.8          10.5
10  8.07  0.912   8.08          8.98          9.00
# ... with 1,990 more rows
```

To visualize that let's plot the data as a scatter-plot with the two lines of regression, and how to use them to re-calibrate a value Y into Y' using the explanatory model or into Y'' using the predictive model (figure 5).

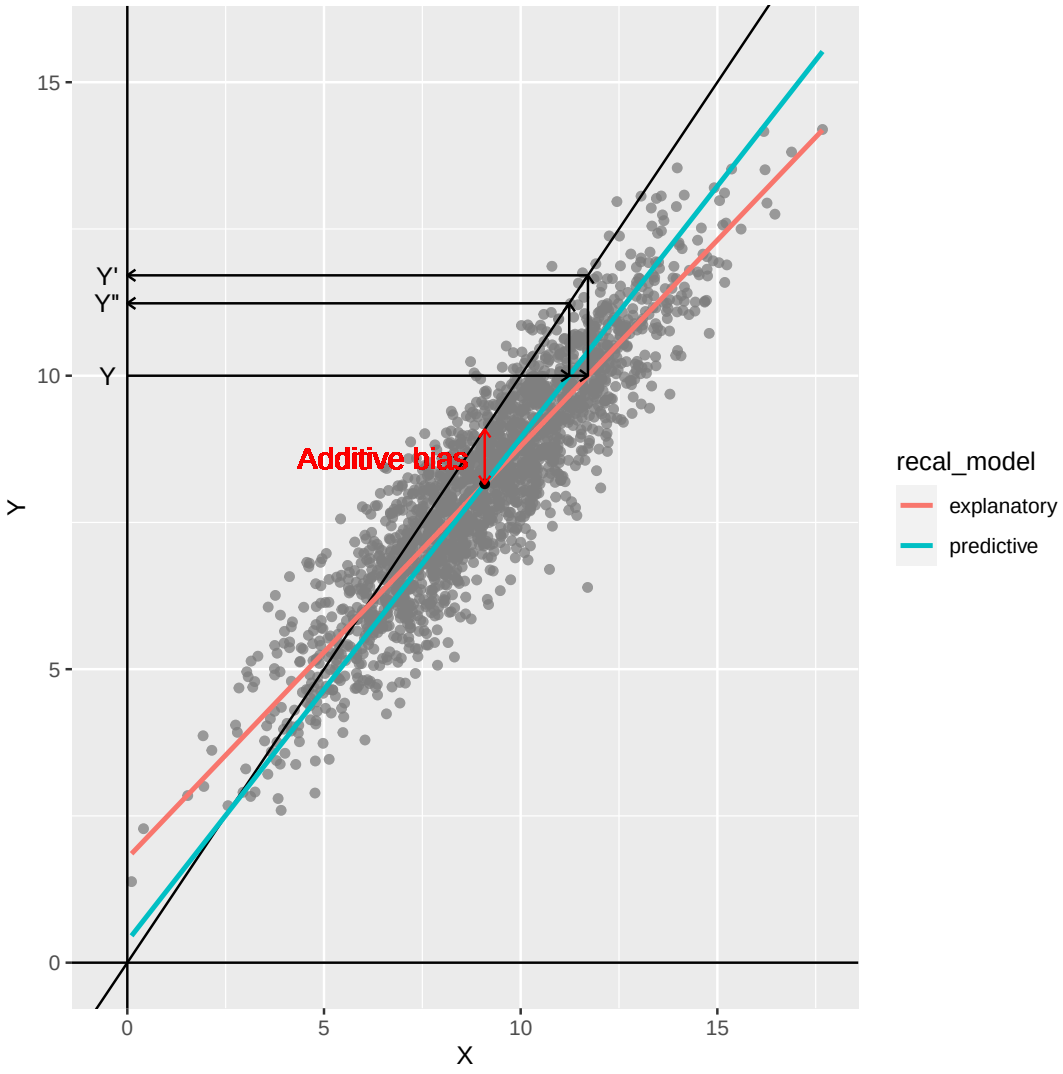


Figure 5: How to re-calibrate from the explanatory ($Y \rightarrow Y'$) and predictive ($Y \rightarrow Y''$) regression lines

Now let's compare the total **R**oot **M**ean **S**quare **E**rrors, or **RMSE**, (and just for comparison also the **MAD**, **M**ean **A**bsolute **D**eviation), since my project uses this KPI:

```
# A tibble: 2 x 3
  metric explanatory predictive
  <chr>        <dbl>      <dbl>
1 RMSE         1.16       1.05
2 MAD          0.929      0.836
```

ML: Aha! It looks like if the $X \sim Y$ model is "more wrong" than the $Y \sim X$ model, then it's also "more useful" in G. Box's sense since we get smaller total errors.

4.2 Minimizing MSTE versus only MSSE: Clarifying the explanatory-predictive differences

PS: Hmm, interesting. I think I know what's happening here, and you were right that this is linked to the MSTE = MSSE + MSRE decomposition. It turns out that the predictive model gives the right answer to a different question:

- What linear re-calibration $Y'' = b_0 + b_1 Y$ minimizes the total sum of squared error $\sum(Y_i'' - X_i)^2$ (or the MSTE, or the RMSE)?

But it is kind of by chance that the right answer to this question is given by the wrong regression.

The explanatory model finds a decomposition of the total errors into systematic errors (corresponding to the regression line) and random errors (corresponding to the residuals) that minimizes the random errors (so maximizes the systematic errors). So, if we constrain ourselves to correct only systematic errors, it optimizes the amount of errors we correct.

In contrast, the predictive model does not yield a clear meaning of the regression line and of the residuals. It just looks wrong.

But let's plot both explanatory and predictive re-calibrations to get a better visual sense of it (figure 6).

ML: I see, the bias for explanatory re-calibration is completely gone, as expected, but there is some residual bias in the predictive re-calibration (smaller that the original one, though, so there has been *some* correction). But at the same time it looks like the cloud of points is slightly denser in the Y-direction for the predictive re-calibration, indicating a reduction in random errors. So, why constraining ourselves to correct only systematic errors, if we can correct both in the predictive way and get a better overall RMSE?

PS: I have mentioned that systematic and random noises are physically two different types of noise and it makes sense to correct them differently, but here we would correct them together and the same way. But even if you don't care about "making sense", there are other problems with that. Before getting to that, though, let's first check the algebra. Considering the explanatory model the truer model, the sum of squared errors for the predictive model is:

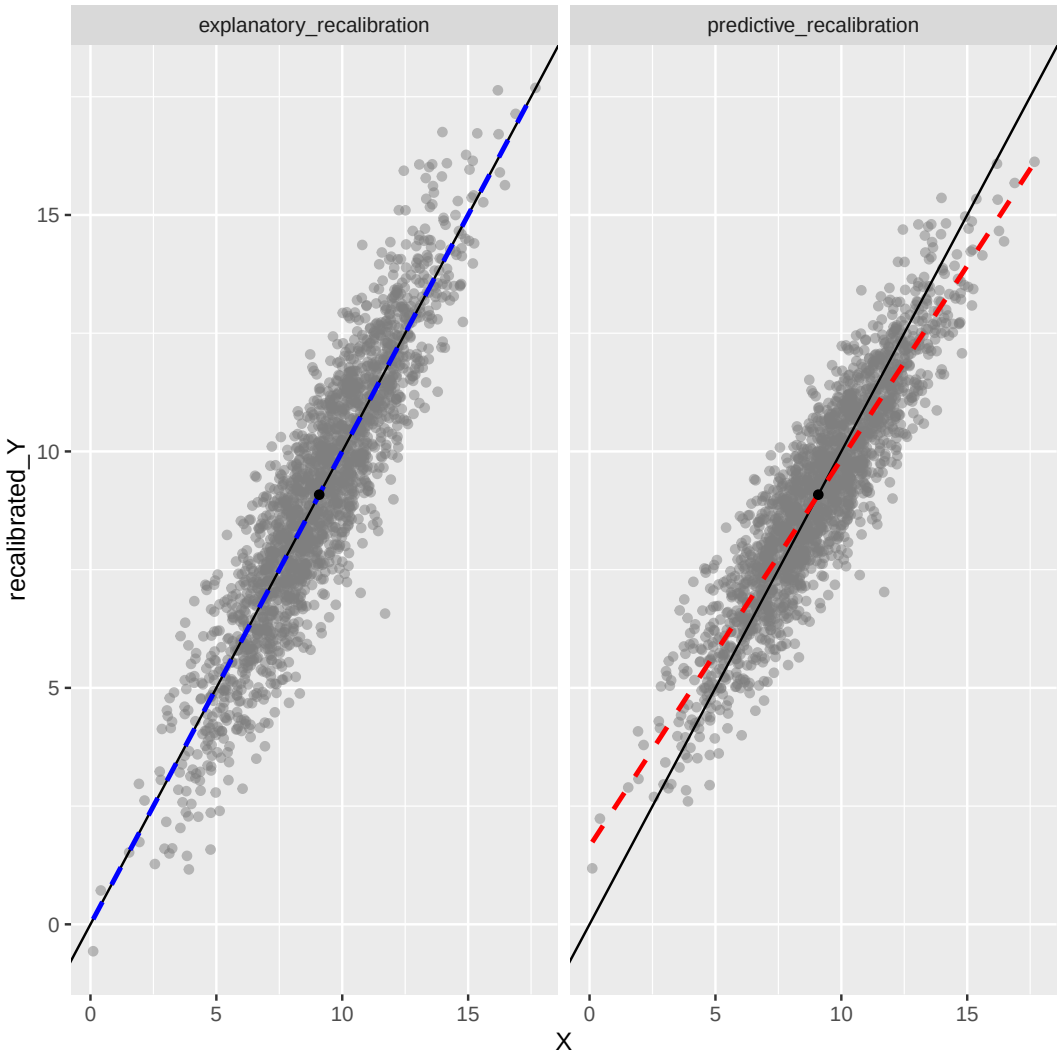


Figure 6: Scatterplots after re-calibration.

$$\begin{aligned}
\sum_i (Y_i'' - X_i)^2 &= \sum_i (b_0 + b_1 Y_i - X_i)^2 \\
&= \sum_i (b_0 + b_1(a_0 + a_1(X_i + \epsilon)) - X_i)^2 \qquad (5) \\
&= \sum_i ((b_0 + b_1 a_0) + (b_1 a_1 - 1)X_i + a_1 b_1 \epsilon)^2 \\
&= \sum_i ((b_0 + b_1 a_0) + (b_1 a_1 - 1)X_i)^2 + a_1^2 b_1^2 \sum_i \epsilon^2 \qquad (6)
\end{aligned}$$

Where in (5) we just replace Y by (2) and in (6) we use the independence of ϵ with X.

The first term in the right side of the last equality corresponds to the bias (it depends only on X) and the last term to the random errors (it depends only on ϵ).

Now, minimizing only the first term, that is setting it to zero, or completely eliminate the bias, yields

$$\begin{aligned}
b_1 &= \frac{1}{a_1} \\
b_0 &= -\frac{a_0}{a_1}
\end{aligned}$$

the inverted coefficients from the "right regression":

$$Y = a_0 + a_1 X \iff X = b_0 + b_1 Y$$

The trouble comes from minimizing the sum of the two terms, a mix of systematic and random errors, which will yield different coefficients b_1 and b_0. So it will correct some bias, not eliminate it (just as you observed), but also artificially shrink the random errors by shrinking the measurements.

4.3 Random errors shrinkage: An artifice

ML: I see that Y will be scaled with b_1, but why "shrink"? Why would $b_1 < 1$?

PS: You're right, b_1 does not have to be smaller than 1, but it is always smaller than $c_1 = 1/a_1$, so the "shrinkage of random errors" is relative to the right coefficient c_1. This was your observation that the cloud of points of the right plot in figure 6 looks denser. In other words the factor $a_1^2 b_1^2$ in the right term of equation (6) is always smaller than 1. Moreover, remember that we defined random noise as the random errors after re-calibration (when re-calibration was suppose to correct bias only). So, assume the bias is already corrected: $Y = X + \epsilon$, that is $a_1 = c_1 = 1$ so $b_1 < 1$. Now the predictive model errors are:

$$\begin{aligned}
Y'' - X &= b_0 + b_1 Y - X \\
&= b_0 + b_1(X + \epsilon) - X \\
&= (b_0 + (b_1 - 1)X) + b_1 \epsilon \qquad (7)
\end{aligned}$$

The first term of the last equality is the residual bias, and the last one is the random error term, which is a shrunk version of the random noise ϵ.

More explicitly, since this is a simple 1-D regression, we know exactly what are the coefficients from the (X, Y)-variance-covariance matrix.

$$V_{XY} = \begin{pmatrix} \sigma^2 & a_1\sigma^2 \\ a_1\sigma^2 & a_1^2(\sigma^2 + \eta^2) \end{pmatrix}$$

So, b_1 is defined as:

$$\begin{aligned} b_1 &= \frac{\operatorname{Cov} XY}{\operatorname{Var} Y} \\ &= \frac{1}{a_1}\frac{\sigma^2}{\sigma^2 + \eta^2} \end{aligned} \tag{8}$$

You see now that $b_1 < c_1 = \frac{1}{a_1}$.

ML: Isn't that called the "bias-variance trade-off"?

PS: Yes, that's exactly that. By not completely correcting for bias, that is, by leaving some residual bias, we can reduce (or shrink) the variance (by $b1^2$).

4.4 Random noise and systematic errors trade-off: Turning the tables

ML: Ok, but why is it bad to shrink the random noise?

PS: Well, let's be facetiously extreme. It's actually very easy to completely eliminate random errors by shrinking them to zero. Just set Y'' to a constant, for example set $Y'' = \overline{X}$ and that way you get rid of both random errors (now there is no variation in Y'') and the additive bias (now $\overline{Y''} = \overline{X}$). Of course, this is useless: Just one measurement for all ground truths.

But besides this extreme, let's emphasize that it is not the random noise that is shrunk, but the random errors in comparison to the random noise. Again, recall that we defined random noise as the random errors after complete bias correction. So, shrinking the random errors by shrinking the measurements is artificial, since it does actually nothing for the random noise. It's like cheating. And by cheating a little bit on random errors and allowing a little bit of bias, you manage overall to reduce the total errors more than if you were just completely eliminating the bias.

ML: But, since my KPI is RMSE, why not cheating a little bit if it helps lowering it?

4.4.1 Dependence on ground truth sampling: First troubles for the predictive model

PS: A simple answer is that total RMSE is not a very good KPI. But before getting to better KPIs, you have to be aware of some of your current KPI pitfalls. For example, if the explanatory model is right, then clearly the $Y \sim X$ regression line (hence the re-calibration) does not depend on the underlying distribution of X. But you see in particular from equation (8) that the $X \sim Y$ regression line does (recall that σ^2 is the variance of X), and ever more so when the random noise, as measured by its variance η^2, is not significantly smaller than σ^2. So, if the sample-set for X changes or gets updated, the distribution might change and for each such change you will have to "re-re-calibrate".

ML: But if the sample is representative of the population, then the distribution of X should remain the same as we get more data. So it should be OK, no?

PS: But that's not necessarily what you have or even what you want. For example it might be more difficult to sample some sub-populations determined for example by socioeconomic status, ethnicity, health conditions, etc. Your initial sample might under-represent such categories. For fairness you might then want to boost their sample size, perhaps to a point where it become actually greater than it's population ratio.

ML: So, if I modify my simulated dataset by adding some data at one extreme, the predictive re-calibration will change but not the explanatory one?

PS: That's right.

ML: Let's visualize that.

Suppose I am not completely happy with my first sample data and I want to sample some extra 200 observations from a sub-population normally distributed, with a mean 1.652 (significantly smaller than the population mean 9), and a standard deviation of 1.225.

```
# A tibble: 200 x 4
       X       EY     Y sampling_batch
   <dbl>    <dbl> <dbl> <chr>
 1 0.742  1.40    3.30  extra
 2 1.60  -0.741   2.40  extra
 3 0.986  0.207   2.64  extra
 4 2.39   0.605   3.90  extra
 5 0.567 -1.87    0.891 extra
 6 2.82  -0.746   3.25  extra
 7 2.94  -0.318   3.63  extra
 8 3.21  -1.11    3.27  extra
 9 1.72  -0.00358 3.00  extra
10 0.490 -0.125   2.06  extra
# ... with 190 more rows
```

We already have the regression lines (intercept and slopes) from the first sampled data. Let's compute the lines from the first plus extra data.

```
# A tibble: 4 x 4
  recalibration_model sampling_data    intercept slope
  <chr>               <chr>                <dbl> <dbl>
1 explanatory         first                1.78  0.702
2 predictive          first                0.365 0.858
3 explanatory         first_plus_extra     1.76  0.704
4 predictive          first_plus_extra     0.968 0.798
```

We see, indeed, that the re-calibration line for the explanatory model did practically not budge, but the line for the predictive model changed significantly. It is even easier to see that on figure 7.

PS: Nice, you also plotted the mean for the first sampled data (the black dot at the intersection of the two solid lines), and the mean for the first plus extra data (the black dot at the intersection of the two dashed re-calibration lines), so we can see the shift.

ML: Well, we actually see only three lines since the solid blue line is on top of the

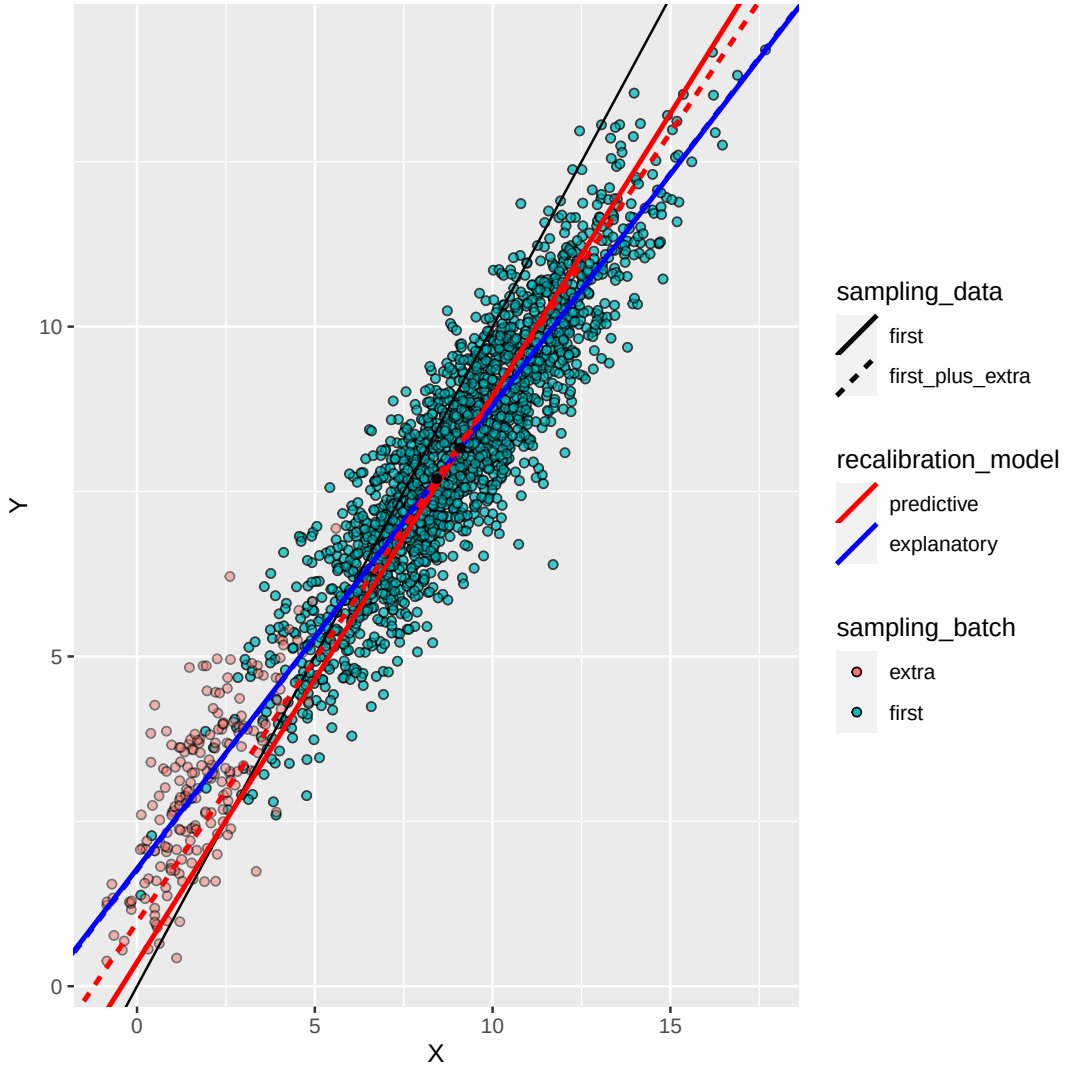


Figure 7: Re-calibration after adding extra samples at one extreme.

dashed blue line, which show that the explanatory re-calibration remains the same. But, you are right that with the predictive model I would have to do amend my initial re-calibration.

PS: Moreover, whether or not you amend it, you see how much greater are the systematic errors for the sub-population than for the overall "mean population". That corresponds to the vertical distance between the blue (which completely corrects the bias) and the red lines (which leaves some "residual bias").

ML: That sounds unfair to this sub-population.

4.5 Explanatory-predictive perspectives through sum-of-square algebra

PS: Absolutely, but before delving in fairness considerations, let's work out the algebra for the MSTE = MSSE + MSRE decomposition in both explanatory and predictive models. We know that regression lines always pass through the mean point $(\overline{X}, \overline{Y})$ (take the expectation on each sides for both $Y = a_0 + a_1 X + \epsilon$ and $X = b_0 + b_1 Y + e$). So, for simplicity, assume that the data has been centered. That means in particular that the additive bias (the "easy" one) has been corrected, since now $\overline{X} = \overline{Y} = 0$. In other words, we assume that $a_0 = b_0 = 0$. Now, the regression $Y'' \sim X$ is given by:

$$\begin{aligned} Y'' &= b_1 Y \\ &= b_1 a_1 (X + \epsilon) \\ &= \frac{\sigma^2}{\sigma^2 + \eta^2}(X + \epsilon) \end{aligned} \tag{9}$$

Where in line (9) we used equation (8). So the residual multiplicative bias corresponds to the slope $\frac{\sigma^2}{\sigma^2+\eta^2}$, and the decomposition MSTE'' = MSSE'' + MSRE'' follows:

$$\begin{aligned} MSTE'' &= \mathrm{Var}(Y'' - X) \\ &= \mathrm{Var}((\frac{\sigma^2}{\sigma^2 + \eta^2} - 1)X + \frac{\sigma^2}{\sigma^2 + \eta^2}\epsilon) \\ &= \left(\frac{\eta^2}{\sigma^2 + \eta^2}\right)^2 \mathrm{Var}(X) + \left(\frac{\sigma^2}{\sigma^2 + \eta^2}\right)^2 \mathrm{Var}(\epsilon)) \\ &= \left[\left(\frac{\eta^2}{\sigma^2 + \eta^2}\right)^2 \sigma^2\right] + \left[\left(\frac{\sigma^2}{\sigma^2 + \eta^2}\right)^2 \eta^2\right] \\ &= MSSE'' + MSRE'' \\ &= \frac{\eta^2 \sigma^2}{(\sigma^2 + \eta^2)^2}(\sigma^2 + \eta^2) \\ &= \frac{\sigma^2}{\sigma^2 + \eta^2}\eta^2 \end{aligned}$$

In comparison, the decomposition for MSTE' is easy:

$$
\begin{aligned}
MSTE' &= \mathrm{Var}(Y' - X) \\
&= \mathrm{Var}(X + \epsilon - X) \\
&= 0 + \eta^2 \\
&= MSSE' + MSRE'
\end{aligned}
$$

We see in particular that $MSTE'' < MSTE$, with the trade-off $MSSE'' > MSSE'$ and $MSRE'' < MSRE'$.

ML: I agree with the algebra. But, in what kind of situation would it be a problem that MSTE' ' < MSTE' even if MSSE' ' > MSSE'?

4.6 Residual bias and unfairness: More troubles for the predictive perspective

PS: Well, it's good timing for going back to fairness consideration. For example in medical or wellness biometric measurements (think for example "percent body fat"), where the "healthy" human subjects are grouped around the mean of the population, and the unhealthy ones are away from the mean. Usually bias is rather small around the mean but grows larger away from the mean, in particular if it is linear (as we observed in the previous plot). Suppose you re-calibrate with minimization of your KPI of overall total errors in mind. You might indeed get a lower total error for the majority of the population (the healthy one), but actually a larger total error at the extremes (for the unhealthy minority population), because there the larger systematic errors will outweigh the shrinkage of random errors.

ML: This does sound unfair for the unhealthy population, specially if the measurements are supposed to determine if the subjects are unhealthy and how much unhealthy they are. But let's check that on our simulated data (without the extra sampling to keep it simpler).

For the sake of validation let's assume a crude threshold and suppose that 80% of the population around the mean is "healthy" and the rest, 10% on each side, is "unhealthy". Now let's split the errors between those two groups, and compute the MSTE = MSSE + MSRE decomposition for both the explanatory $Y \sim X$ and the predictive $X \sim Y$ models.

```
# A tibble: 4 x 5
# Groups:   health [2]
  health    recal_model  MSTE  MSSE  MSRE
  <chr>     <chr>       <dbl> <dbl> <dbl>
1 healthy   explanatory  1.32  0     1.32
2 healthy   predictive   0.98  0.1   0.89
3 unhealthy explanatory  1.45  0     1.45
4 unhealthy predictive   1.7   0.73  0.97
```

PS: We see, indeed, that for the unhealthy sub-population the predictive MSTE (Total Errors) is somewhat larger than the explanatory MSTE, whereas for the healthy sub-population it is smaller. Because predictive re-calibration tends to artificially shrink the overall measurements toward the mean, subjects at the extremes of the distribution will see their measurements artificially less extreme than they should

be. You can see that, for example, on the right plot of figure 6. And for the same reason, health improvements will be less measurably perceptible.

5. Non-linear re-calibration

ML: Yes, this is rather clear when the right model is linear, but the most appropriate regression might not be a linear one.

PS: That's correct. The relation might be a little more complicated. If you have enough data, you can use a more sophisticated regression like Loess[3] or GAM[4] for example. You might also do some appropriate transformations of the variable (exponentiation, log, Box-Cox, logit, ...) in order to "ellipticize" the shape of the scatter plot and straighten the regression line. The principle is the same, though: The curvy regression line still represents the systematic bias. Now you can do interpolation (or careful extrapolation outside the ground truth sample range) to match any measurement value Y with a re-calibrated Y' just like in figure 5. After re-calibration, you can estimate the random noise variance as the variance around the diagonal.

ML: Ok, but in such a case it sometimes happens that the "curvy" regression line is "too curvy", that is, there is more than one pre-image, say (X_1, X_2), for some Y, like in figure 8.

ML: In that case how would you proceed with bias correction when there are two pre-images X_1 and X_2 for a single Y?

PS: Good point. The ends of the regression curve for local regression techniques often act a little wildly. This can usually be mitigated by various technical "hacking" to make the regression curve strictly monotonous (that is, make the regression "invertible"). The problem might be that the number of samples decreases toward the end, to a point where the local regression becomes unreliable. So, you can shave off some samples near the end if their X-density is too low. You can also assume, in the absence of data beyond a certain point, that there are no bias. It's like assuming the data in "bias-innocent" until proven bias-guilty. So, you can add some fake data as a repeated point on the diagonal far enough from the last sample. That will force the regression curve toward the diagonal and avoid it to flare up or down toward the end.

ML: Interesting "hack" ideas. Let's try it here on our simulation by truncating the data to $0.3 <= X < 4$, $Y > 0$, and adding 5 data points on the diagonal at $X = 0$ and 5. Let's plot the result in figure 9.

PS: That seems to work with those parameters. Let's check the re-calibration now in figure 10.

ML: I added the GAM regression curve (in red) on the recalibrated data (truncated, but without fake data), and the linear regression line (in blue). Re-calibration did straighten the data around the diagonal.

[3] or Lowess

[4] Generalized Additive Model

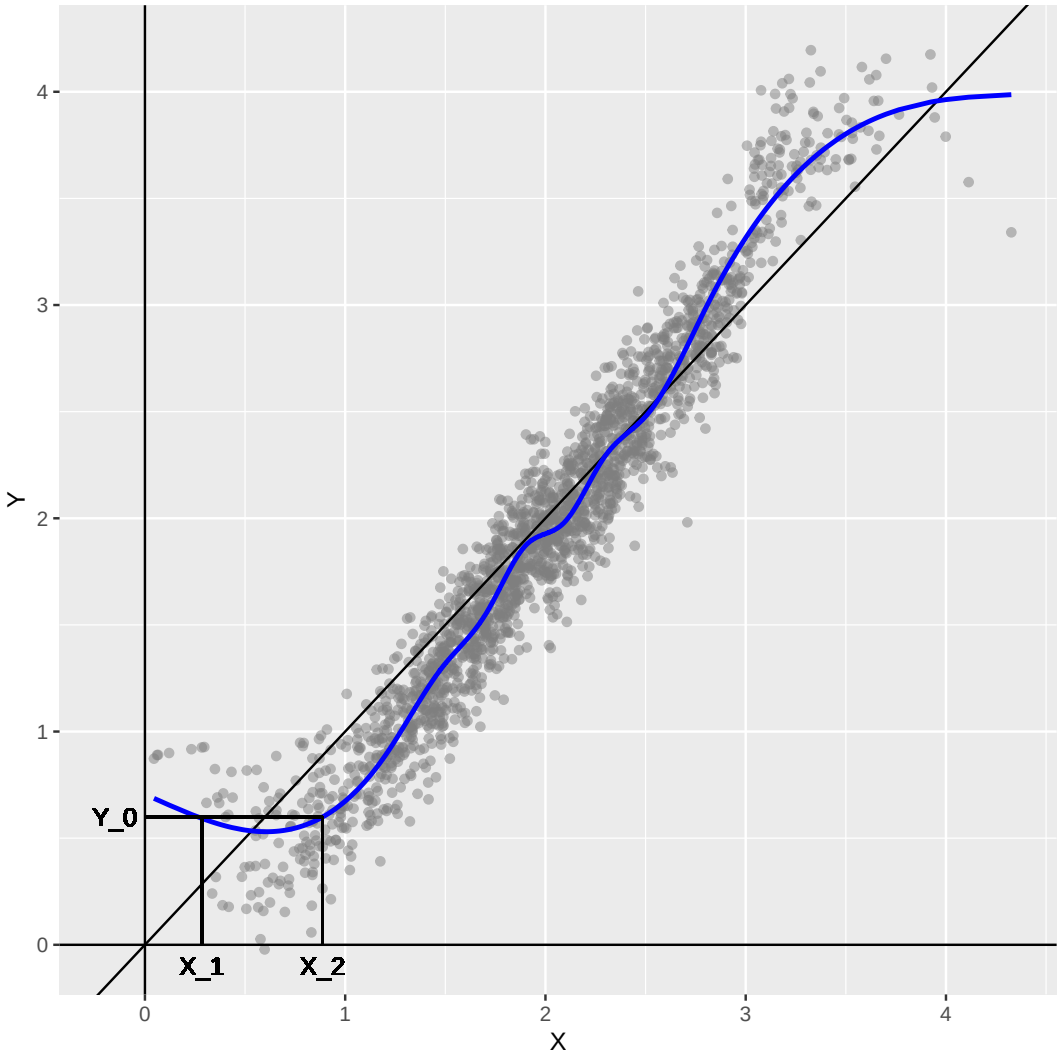


Figure 8: Non-invertible non-linear regression

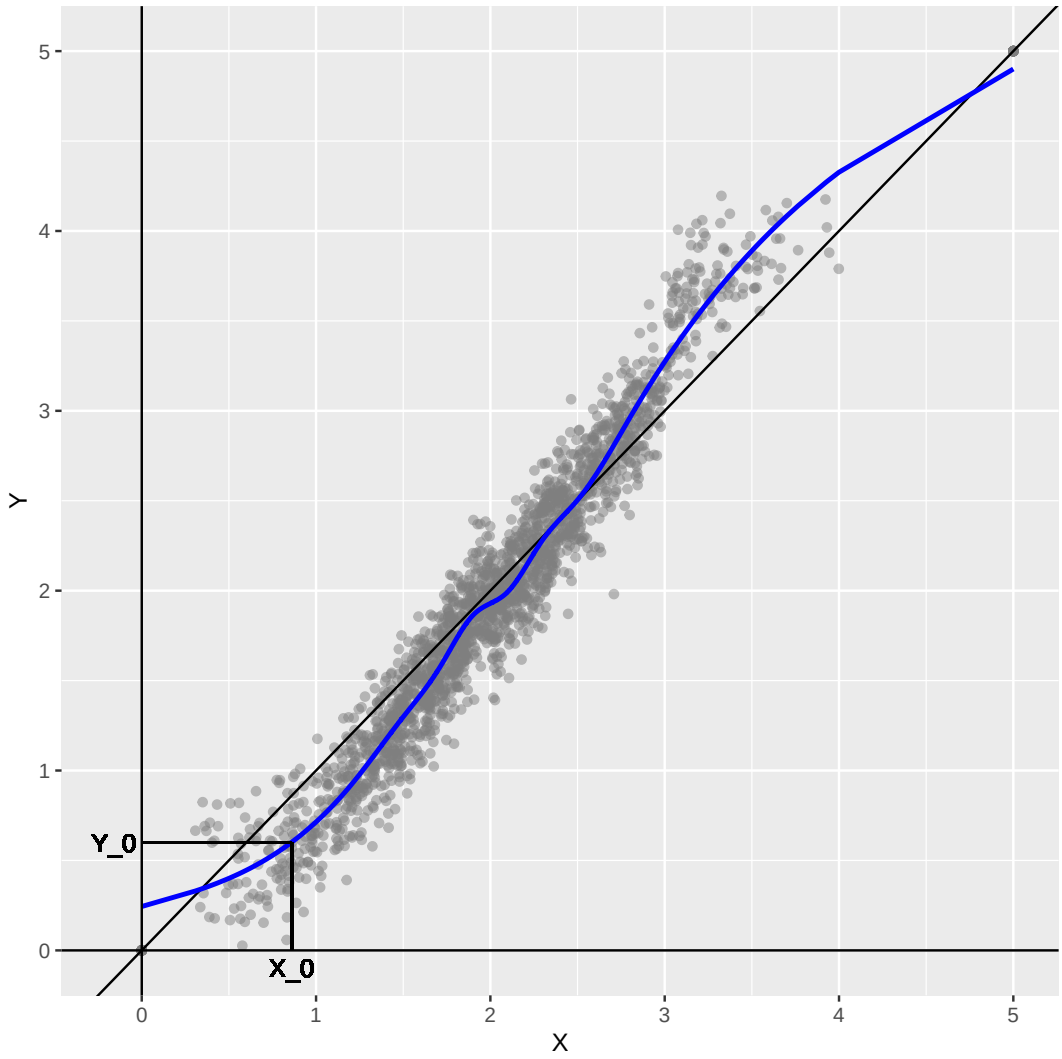


Figure 9: Non-linear regression made invertible by removing some extreme points and adding some extreme points on the diagonal

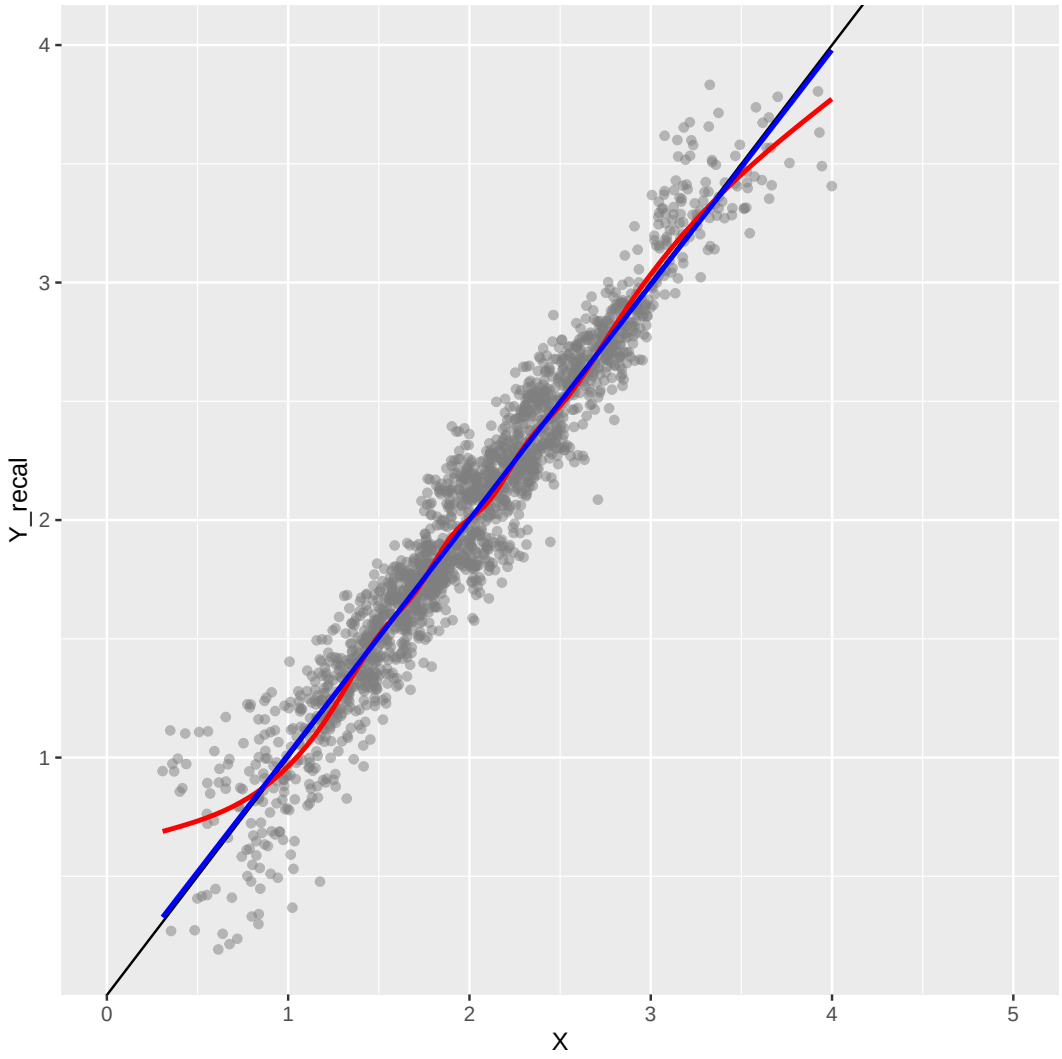


Figure 10: Re-calibration with non-linear regression

6. Toward better KPIs than RMSE

ML: Alright, but we still need KPIs to track and drive our products. So, if total RMSE is not an appropriate KPI, what should we use?

PS: I think it is important to keep the distinction between random and systematic noise, so ideally there should be one KPI for each type of noise. But systematic noise can be easily corrected, so the important one is random noise. So if you have to choose only one, take the one we wrote MSRE' above, that is the MSE after bias correction. That will be more informative for the engineers to fine tune the device in order to correct the random noise.

If you really want only one number that combines random and systematic noise, you could take the sum $MSSE + MSRE'$. It's less mathematically convenient than $MSTE = MSSE + MSRE$, but it is more consistent in the sense that the "wrong" predictive re-calibration now gets a lower KPI than the "right" explanatory re-calibration.

7. Re-calibration versus calibration

ML: We have been talking about **re-calibration** so far, but what about doing the right things first, that is, what about **calibration**?

PS: Calibration is a little more complex. In re-calibration we have only 2 variables, the initial measures Y, and the ground truth X. In calibration, we might have several variables $Y_1, \dots, Y_k$, like for different types of sensors, that combine through some complicated formula to give a measurement Y_0 of a ground truth variable X. In your SpO2 measurement example $k = 2$ with the red and the infra-red sensors.

If the formula is simple, like a linear formula $X + \epsilon = a_0 + a_1 Y_1 + \cdots + a_k Y_k$, then this is the setting for an **inverse regression**, like we just did in the linear case for re-calibration, except that it is more complicated than just regressing $Y \sim X$ and inverting the coefficients, since now we have more than one Y, and they are usually not independent. But that is the subject for another discussion.

ML: So in more complex cases, we could do an ok-calibration followed by a re-calibration. Actually, that makes me think of applying that to artificial neural networks in a more in-meshed way.

7.1 Introduction to an application to neural networks

PS: Well, deep learning has become very popular, so it might indeed be appropriate to talk about how to apply re-calibration to artificial neural networks. How would you do that?

ML: In Neural Networks (NN) the data is usually split between training and testing data, where the former is used, well, to train the NN, and the latter to test or validate the trained NN. This is mainly to avoid over-fitting. But the training data is usually also split into **batches** so that the NN is updated only after each batch.

PS: I have also heard the term **epoch**.

ML: The training data is split into batches, so once the algorithm has gone through all the batches (so all the training data), it has completed one epoch. But that is

usually not enough for good training. So we re-use the training data by splitting it again into batches and repeating the process, thus going through multiple epochs.

PS: Thanks for clarifying. So, how is the NN trained after each batch?

ML: You define a loss function, that computes the errors of your predictions Y with respect to the ground truth X (or **labels**), for a given batch. For example a popular one might be the square loss.

$$L(Y, X, W) = \frac{1}{n} \sum (Y_i - X_i)^2$$

where the sum is over the data from the whole batch, and W are the current **weights** of the NN (or parameters). After each batch, using **gradient descent** on the loss function with **back-propagation** through the layers of the NN, you can fine-tune the weights. There are many places where to get more details on the inner working of NNs, but the important point here is that this algorithm is trying to minimize the loss function. But we have see that this type of loss function is blind to the difference between random and systematic noise, so it has some leftover bias.

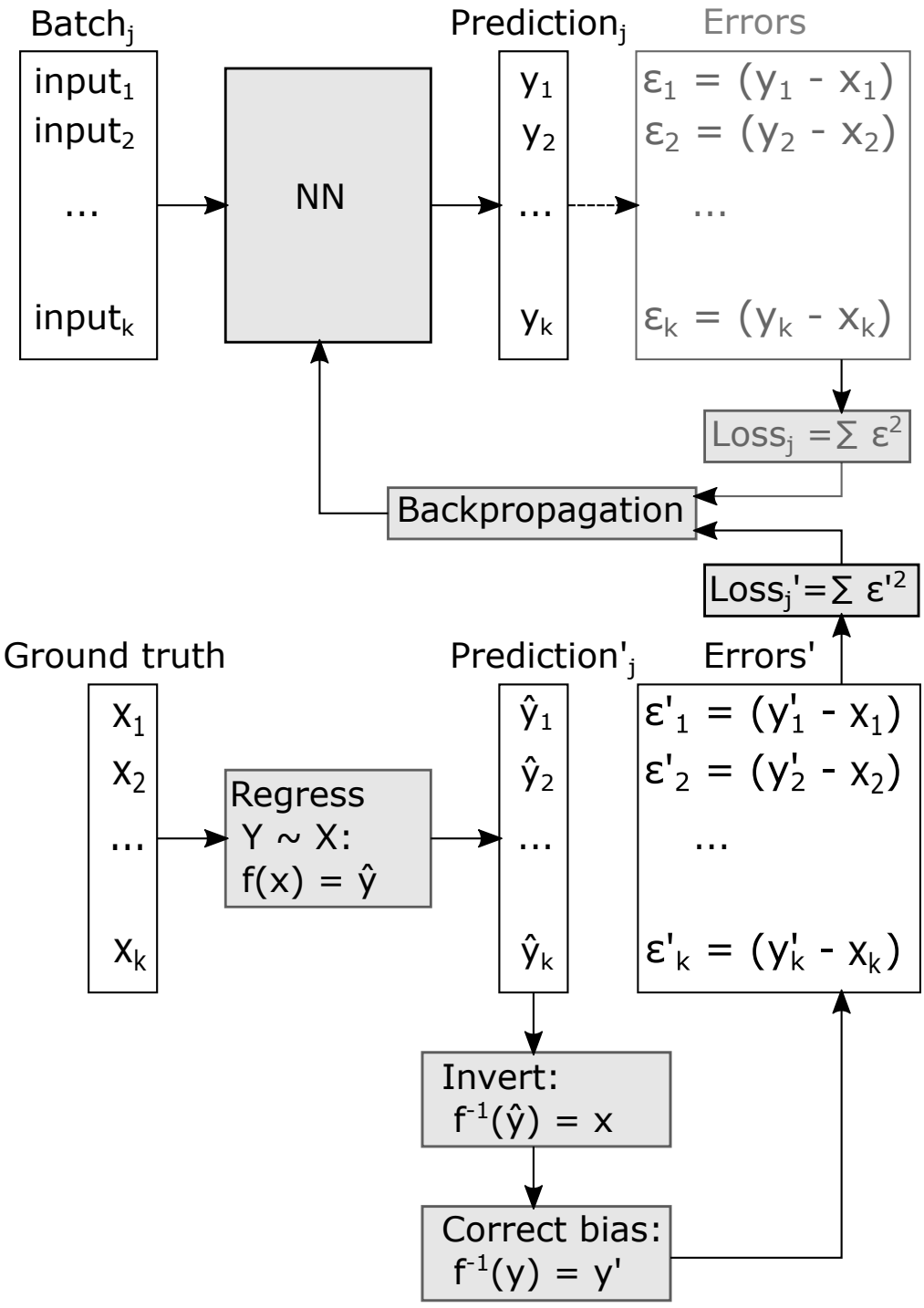
PS: I see. So you would do some bias correction after each batch?

ML: Precisely. After each batch j we obtain a prediction vector $Y_j = (y_1, y_2, \ldots, y_k)$, from the NN, where k is the number of training examples in the batch. Usually, the total errors are computed, $\epsilon_j = Y_j - X_j$, and the loss function summarizes them into one number, like $\frac{1}{n}\|\epsilon\|^2$ for the square loss. Instead, set up the explanatory regression model $Y \sim X$. That might sound a little counter-intuitive from a prediction instead of an explanatory stand-point, since it would seem that the model tries to "predict the predictions". The resulting fitted values are $\hat{Y} = (\hat{y_1}, \hat{y_2}, \ldots, \hat{y_k})$ determined by a function $f(x) = \hat{y}$ (which has to be invertible). This function can be defined either by an analytic formula, like with a linear regression, or by a look-up table (together with interpolation for other values of x), for more involved regressions. Its inverse is defined similarly. Now the bias-corrected values are $y'_i = f^{-1}(y_i)$, the random noise for batch j is the differences $\epsilon'_j = Y'_j - X_j$, and the new loss function is

$$L'(Y, X, W) = \frac{1}{n} \sum (f^{-1}(Y_i) - X_i)^2$$

PS: Doesn't it make the gradient descent algorithm more complicated?

ML: A little, but it's just like adding a last activation function at the end, equal to f^{-1}.



8. Conclusion and next steps

ML: But it is getting late. Let's stop for today and recapitulate the key points of our discussion:

- In general beware to setup your linear regression the right way, that is, with the noisy variable as the Y-variable and the non-noisy (or less noisy) one(s) as the X(s).
- For re-calibration the right and fair way is to have the measurements as the

Y-variable and the ground truth as the X-variable, despite the predictive model temptation of doing it the other way around.

- Then, invert the regression function f to correct for bias, that is, the bias-corrected variable is now $Y' = f^{-1}(Y)$.
- If using a non-linear regression technique, fine-tune your regression function if necessary to make sure your regression function is invertible.
- Report a measure of the bias you just corrected (or "systematic noise") and a measure of the residual errors after bias correction (or "random noise") as 2 separate KPIs.

PS: Yes, and don't forget your novel idea to use re-calibration for calibration in Neural Networks. That begs for further investigation. I also propose that next time we meet we talk about (a) calibration with inverse regression and (b) the re-calibration where we have no ground truth, but only measurements of different quality (usually some precise but costly and some less precise but cheaper). Can we for example combine them in some ways to correct for bias even in the absence of ground truth? It turns out that (a) and (b) are actually linked.

ML: Looking forward to it! Till next time.

References

Brown, Philip J. 1994. *Measurement, Regression, and Calibration.* Vol. Oxford Statistical Science Series, 12. Clarendon Press.

Fuller, Wayne A. 1987. *Measurement Error Models.* Vol. Wiley Series in Probability and Statistics. Wiley.

Shmueli, Galit. 2010. "To Explain or to Predict." *Statistical Science* 25 (3): 289–310.