LSMC Surgery

Abdal Chaudhry, Adel Cherchali, Michael Leitschkis, Julien Vedani

C Milliman

Least Squares Monte Carlo (LSMC) is a widely used proxy modelling technique in the European Insurance industry. It allows its users to model components of their balance sheet via suitable polynomial functions in an automated way. However, users may experience suboptimal goodness-of-fit due to various complexities inherent to the LSMC method while validation is not always straightforward. In this paper, we introduce and apply a new troubleshooting method specifically designed for LSMC work.

LSMC uses a large number of training points ranging anywhere from 10,000 to 100,000. These points typically cover the entire risk factor space evenly. For each training point, LSMC processes rough estimates of dependent variables (e.g., Net Asset Value) relying on a small number—usually between two and 20—of inner simulations. In contrast, traditional curve fitting (TCF) considers a much smaller number of training points, but uses a much more accurate estimate of the dependent variable for each of them.

The advantages of using LSMC over TCF include automated term selection and the ability to scale to a large number of risk drivers—typically up to 50 risk drivers—without losing the ability to capture risk interactions. LSMC has traditionally outperformed TCF-based approaches in the Internal Model space, both in terms of explanatory power and ease of use.

Out-of-sample (OOS) tests performed on fitted proxy models allow users to evaluate goodness-of-fit of their fitted proxy models. OOS validation compares fitted model results against stress valuations generated using the underlying asset liability model (ALM) or the "truth model."

When looking into typical OOS validation results in a high-dimensional risk space, we face a few practical challenges:

- Do the underlying model assumptions hold true, e.g., homoscedasticity of residuals when using ordinary least squares?
- Which deviations represent a good fit? If the overall fit is "poor," how do we find and resolve underlying issues?
- How do we determine whether issues in the underlying asset liability model are distorting the overall fit or if there is a genuine fitting issue?

One of the inherent difficulties with LSMC troubleshooting stems from the method's use of a small number of inner simulations sampled for each outer training data point. In particular, we cannot directly compare training data points, even those that are very close to one another in the risk space, as we estimate the dependent variable for these points using entirely different inner simulations featuring different nominal and real yields, different equity returns etc.

This paper introduces its readers to LSMC Surgery, an approach developed by the authors that is useful in detecting underlying model issues. LSMC Surgery allows its users to understand the behaviour of the dependent variable (e.g., BEL) in one, two or three dimensions of the risk space, thereby allowing users to zoom into the underlying sources of fitting issues.

A common issue seen when using Ordinary Least Squares (OLS) to fit proxy models on LSMC data is heteroscedasticity of residuals, which is a key assumption in regression analysis. This paper shows how LSMC Surgery can help detect heteroscedasticity of residuals and suggests ways to improve regression models.

1. LSMC Surgery approach

Under LSMC Surgery, we operate the (generally stochastic) asset liability model in a deterministic (certainty equivalent) manner—of course, this amounts to ignoring the time value of options and guarantees. This is a conscious omission given that the purpose of this exercise is to detect potential anomalies in the underlying model and/or its configuration for proxy modelling purposes. Furthermore, we restrict ourselves to a low-dimensional subspace of the full risk space—by focusing onto one, two or three risk drivers at a time while "freezing" all other risk drivers.

Under this troubleshooting paradigm, a rather small number of training points (e.g., 50 to 250) can give us a sound basis for our LSMC validation activities, as we eliminate any "sampling noise" while also making neighbouring points in the risk subspace directly comparable with one another in terms of their impact on assets and liabilities.

In the following sections, we demonstrate LSMC Surgery with the help of a few stylised examples.

2. One-dimensional LSMC Surgery

To demonstrate how LSMC Surgery works, we start by looking at a few one-dimensional examples.





2.1. FITTING TO DATA KINKS

Our first example illustrates the issue of fitting a simple polynomial to an underlying data set representing an option-like pattern.

The fitting issue here stems from our limitation in the type of proxy models available to us, i.e., a simple polynomial must be differentiable across the whole of the fitting space, making it impossible to fit to the underlying data.

Simple polynomials therefore cannot closely match data sets featuring kinks or discontinuities.

¹ In each example, we denote the first risk driver as X1.



FIGURE 2: LSMC PROXY MODEL FIT TO DATA WITH DISCONTINUITY (REGRESSION INPUTS IN BLUE, FITTED MODEL IN BLACK)

2.2. FITTING TO DATA DISCONTINUITIES

Our second example illustrates fitting issues stemming from discontinuities in the underlying data.

Here, a few training points in the lower left corner of the plot appear to be out of place when compared against the general trend observed in the entirety of the one-dimensional risk space. Underlying issues in the asset liability models or model configuration errors most likely drive this behaviour. This is contrary to the previous example where a genuine "kink" in the training data was preventing us from achieving a good fit. One-dimensional LSMC Surgery makes it easy for us to detect such issues.



FIGURE 3: LSMC PROXY MODEL FIT MATCHING THE REGRESSION INPUT DATA CLOSELY (REGRESSION INPUTS IN BLUE, FITTED MODEL IN BLACK)

2.3. EXPLORING COUNTERINTUITIVE DATA SHAPES

Our third example illustrates a different issue—we have achieved a good fit as is evident by the fitted (black) curve closely matching the blue fitting data points. However, one must be able to explain the maxima of the dependent variable observed, i.e., the shape of the polynomial or data should be explainable.

It is up to the user to decide whether the fitted model represents genuine business logic and is not the result of a model configuration error.

In summary, LSMC Surgery allows users not just to troubleshoot fitting issues in their proxy models but to also be able to identify issues in the underlying asset liability models.

3. Two-dimensional LSMC Surgery

3.1. DISCOVERING HETEROSCEDASTICITY

In the previous section, we have presented three examples to demonstrate the value of using LSMC Surgery in a single dimension experiment and how troubleshooting methods traditionally used in TCF can help in an LSMC context. However, as the reader will have been suspecting already, some practical challenges will not appear in dimension one but in a higher-dimensional LSMC Surgery.

In the stylised example below, we present a two-dimensional example highlighting a heteroscedasticity issue frequently observed when using LSMC. Heteroscedasticity is the absence of homoscedasticity, which means that the variability of random disturbances in a model is constant across the risk space. Homoscedasticity of residuals is a key assumption in regression analysis.





What we can see in this example is that the size of residuals, plotted on the y-axis, appears to vary significantly depending on the value of a risk driver. These residuals result from an LSMC proxy model fit to a simple example of a training data set featuring 100 points.

This observation is a worrying sign in OLS, which assumes homogeneity of variance of residuals. The above example shows the residuals are heteroscedastic.

Upon closer inspection of the underlying data, we observe that the "residual funnel" goes away when we increase the number of training points. As it turns out, we would need to increase the number of training points to 1,278 in order to overcome the double funnel heteroscedasticity issue.

In addition to the number of points, the appearance of this residual double funnel also depends on the range of risk driver values in the fitting space. If we restrict ourselves to a smaller data cube—by moving from our current data cube size of [-1.5, +1.5] to [-1.2, +1.2]—we only require 168 training points in order to avoid the double funnel.

It is intuitive that we need fewer training points when calibrating LSMC polynomials on a smaller data cube. Yet it is less intuitive that the ratio of the point numbers required in our LSMC Surgery experiment—namely 1,278 / 168 = 7.6—is much larger than the data range ratio of 1.5 / 1.2 = 1.25.

A possible interpretation of this finding is that asset liability model behaviour under extremely large combined stresses can be rather exotic—which can make proxy model calibrations a disproportionately difficult task, especially when dealing with simple polynomial functions.

Yet before interpreting too much into our example, let us address the residual funnel challenge in a mathematically rigorous way.

3.2. SOME MATHEMATICAL NOTATION

Consider the general multiple regression model:

$$Y = X\beta + \epsilon$$

Where:

- Y is a $N \times 1$ vector of observations y_i , i = 1, ..., N
- X is a $N \times p$ matrix of p explanatory variables x_{ik} , i = 1, ..., N; k = 1, ..., p
- ϵ is a $N \times 1$ vector of random errors
- $\beta = (\beta_1, ..., \beta_p)$ is a $p \times 1$ vector of model parameters.

The standard assumptions (\mathcal{H}) on fitting errors are as follows:

$$Var(\epsilon|X) = \sigma^2 I_N$$
$$\mathbb{E}(\epsilon|X) = 0_{N \times 1}$$

Where I_N is an $N \times N$ identity matrix and $0_{N \times 1}$ is the null vector of size N.

The assumption of constant variance of residuals is known as the *homoscedasticity* assumption as we saw in our twodimensional example above. In other words, when the variance of residuals $Var(\epsilon_i) = \sigma_i$ is not constant across all observations, residuals are *heteroscedastic*.

Under (\mathcal{H}) , the OLS estimator $\hat{\beta} = (X^T X)^{-1} X^T Y$ is the Best Linear Unbiased Estimator (BLUE) of β . However, this property does not hold under heteroscedasticity.

3.3. A THEORETICAL APPROACH TO HETEROSCEDASTICITY

After confirming heteroscedasticity of residuals, the variance-covariance matrix takes the more general form:

$$Var(\epsilon|X) = \Omega$$

with, for any one-dimension residuals *i* and *j*, $\Omega_{i,j} = Cov(\epsilon_i, \epsilon_j)$

In this case, the Generalised Least Squares (GLS) estimator $\hat{\beta}^{GLS} = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} Y$

is the minimum variance linear unbiased estimator of β . Indeed this estimator is the OLS estimator of the linearly modified model that keeps our β , $\Omega^{-\frac{1}{2}}Y = \Omega^{-\frac{1}{2}}X\beta + \Omega^{-\frac{1}{2}}\epsilon$ or $Y^{GLS} = X^{GLS}\beta + \epsilon^{GLS}$.

Residuals for this model are no longer heteroscedastic as seen by the variance of residuals: $Var(\epsilon^{GLS}|X^{GLS}) = \Omega^{-\frac{1}{2}}\Omega\Omega^{-\frac{1}{2}} = I_N$

3.4. SIMPLE MODEL EXAMPLES

The remainder of section 4 is dedicated to exploring the sources of heteroscedasticity of residuals, in particular the residual double funnel seen in the examples above. A more detailed analysis of the sources of heteroscedasticity is required to determine whether a solution exists. A number of worked examples follow the mathematical notation in section 3.4.1 and 3.4.2 below.

We consider a simple model of the following form:

$$Y = f(X_1, X_2) + \epsilon = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

with X_1 and X_2 centered and independent (in general, random variables induced by Sobol sequences outcomes, see Glasserman [2013], subsection 5.2.3), ϵ centered with unit variance.

Below we try to explore various reasons of the appearance of residual double funnels, which are mathematically linked to the fact that function $x_1 \mapsto \mathbb{V}(\hat{\epsilon} \mid X_1 = x_1)$ increases as $|x_1|$ increases. This is true for both x_1 and x_2 in our examples, but we consider x_1 without any loss of generality.

Let us now distinguish two model selection outcomes. Firstly, we are going to examine the case where model selection picks up a relevant cross-term successfully. Secondly, we are going to consider the case where our model selection does not detect this cross-term.

It is worth noting here that the double residual funnel can be characterised by the fact that the map $x_1 \mapsto \mathbb{V}(\hat{\epsilon} \mid X_1 = x_1)$ increases *as* $\mid x_1 \mid$ increases.

3.4.1. Case 1: Cross-term detected

We get estimated parameters from the usual OLS estimator

$$\hat{\beta} = \left(\widehat{\beta_1}, \widehat{\beta_2}, \widehat{\beta_3}\right) = (X^T X)^{-1} X^T Y$$

The variance-covariance matrix of the parameters then becomes:

$$\mathbb{V}(\hat{\beta}) = (X^T X)^{-1}$$

This tends to zero as the sample size increases.

Looking at empirical residuals,

$$\begin{aligned} \epsilon \\ &= Y - \hat{Y} \\ &= \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon - \widehat{\beta_1} X_1 - \widehat{\beta_2} X_2 - \widehat{\beta_3} X_1 X_2 \end{aligned}$$

we can calculate their empirical variance subject to some level of the covariates X_1 , assuming covariance between estimators and estimators/covariates being zero for the purpose of illustration:

$$\begin{split} \mathbb{V}(\hat{\epsilon} \mid X_1 = x_1) \\ &= \mathbb{V}(\epsilon \mid X_1 = x_1) + x_1^2 \mathbb{V}(\widehat{\beta_1}) + \mathbb{V}\left(X_2(\beta_2 - \widehat{\beta_2})\right) + x_1^2 \mathbb{V}(X_2\widehat{\beta_3}) \\ &= 1 + x_1^2 \mathbb{V}(\widehat{\beta_1}) + \mathbb{V}(X_2) \mathbb{E}\left(\left(\beta_2 - \widehat{\beta_2}\right)^2\right) + \mathbb{V}(\beta_2 - \widehat{\beta_2}) \mathbb{E}(X_2^2) + x_1^2 \mathbb{V}(X_2) \mathbb{E}\left(\widehat{\beta_3}^2\right) + x_1^2 \mathbb{V}(\widehat{\beta_3}) \mathbb{E}(X_2^2) \\ &= 1 + x_1^2 \mathbb{V}(\widehat{\beta_1}) + 2\mathbb{V}(X_2) \mathbb{V}(\widehat{\beta_2}) + x_1^2 \mathbb{V}(X_2) \left(\mathbb{E}\left(\widehat{\beta_3}^2\right) + \mathbb{V}(\widehat{\beta_3})\right) \\ &= 1 + x_1^2 \left(\mathbb{V}(\widehat{\beta_1}) + \mathbb{V}(X_2) (2\mathbb{V}(\widehat{\beta_3}) + \beta_3^2)\right) + 2\mathbb{V}(X_2) \mathbb{V}(\widehat{\beta_2}) \end{split}$$

Hence, we are going to observe a residual funnel if the sample size N is limited, the estimators' variances are not too low and the cross-term effect is non-negligible, so that $\mathbb{V}(\widehat{\beta_1}) + \mathbb{V}(X_2)\left(\mathbb{E}\left(\widehat{\beta_3}^2\right) + \mathbb{V}(\widehat{\beta_3})\right)$ has an impact.

For small values of x_1 we get $\mathbb{V}(\hat{\epsilon} \mid X_1 = x_1) \sim 1 + 2\mathbb{V}(X_2)\mathbb{V}(\widehat{\beta_2}) = c_N$

For large positive or negative values of x_1 we get $\mathbb{V}(\hat{\epsilon} \mid X_1 = x_1) \sim c_N + x_1^2 \left(\mathbb{V}(\widehat{\beta_1}) + \mathbb{V}(X_2)(2\mathbb{V}(\widehat{\beta_3}) + \beta_3^2)\right)$, which includes the uncertainty in the cross-term estimate $\mathbb{V}(\widehat{\beta_3})$, as well as additional dispersion caused by the second variable $\mathbb{V}(X_2)$. The double funnel size will decrease with the estimation sample size, as $\mathbb{V}(\hat{\beta})$ tends to 0.

3.4.2. Case 2: Cross-term undetected

In our second case, we try to estimate the "truth model" using the following function: $\widehat{Y'} = \widehat{\beta_1}X_1 + \widehat{\beta_2}X_2 + \widehat{\epsilon'}$

where we now get $\epsilon' = \beta_3 X_1 X_2 + \epsilon$ centered with variance

$$\mathbb{V}(\beta_3 X_1 X_2 + \epsilon) = \beta_3^2 \mathbb{V}(X_1 X_2) + 1$$

Then

$$\begin{aligned} \widehat{\epsilon'} \\ &= Y - \widehat{Y'} \\ &= \beta_1 X_1 + \beta_2 X_2 + \epsilon' - \widehat{\beta_1} X_1 - \widehat{\beta_2} X_2 \end{aligned}$$

And

$$\mathbb{V}(\widehat{\epsilon'}|X_1 = x_1) = 1 + x_1^2(\beta_3^2 \mathbb{V}(X_2) + \mathbb{V}(\widehat{\beta_1})) + \mathbb{V}(X_2(\beta_2 - \widehat{\beta_2}))$$

So whatever number N of training points we use, this variance will never decrease below $1 + x_1^2 \beta_3^2 \mathbb{V}(X_2)$, the double funnel will appear, and its size will not depend on N.

A pronounced double funnel appears if both $\mathbb{V}(X_2)$ and the scope of X_1 outcomes are relatively large, such that the term $x_1^2 \beta_3^2 \mathbb{V}(X_2)$ is significant.

3.4.3. Simple numerical illustrations

To illustrate the above, we consider the stylised demo model $Y = X_1 + X_2 + X_1X_2 + \epsilon$, where $X_1, X_2 \sim \mathcal{U}([-2.5, 2.5]), \epsilon \sim \mathcal{N}(0, 1)$ and all random variables are independent.

Firstly, we assume that our model selection algorithm does "find" the cross term (Case 1 above).



In this example, residuals are homoscedastic and increasing the number of training data points does not significantly improve fitting errors as seen when we move from 100 points in Figure 5 above to 1,000 training points in Figure 6.

To see how our model performs outside of the calibration range we can now increase the range of the fitting space from [-2.5, +2.5] to [-25.0, +25.0], keeping the OLS parameters estimated above unchanged. In other words, we are validating our fitted model on a larger risk space. Results of this run are shown in Figure 7 and Figure 8.



The double funnel (heteroscedasticity) appears when we validate our model beyond the initial calibration space, as seen in Figure 7. However, if we increase the number of fitting scenarios, we again obtain a good fit (Figure 8), with homoscedasticity assumption satisfied—significant reduction in residuals even though some heteroscedasticity exists. In this case, OLS residuals become homoscedastic when the numbers of fitting scenarios are increased.

This shows how useful a test of larger covariate scopes can be when looking at residuals. It also shows how dangerous the use of an apparently "good" LSMC curve can be when evaluating risk scenarios that fall outside of the fitting space.

As expected, the double funnel size decreases as the number of fitting scenarios, N, increases. It is therefore essential that LSMC users select the number of training points carefully.

In our next example, we look at the case where cross-terms exist but remain undetected. Figures 9 and 10 are based on the same example as Figures 7 and 8 with the only difference being that the model is set up to purposefully omit a cross-term, This is done to illustrate our point in Case 2, i.e., increasing N will not always result in an improved model.



In this case, the double funnel appears in the initial proxy model and increasing the number of fitting scenarios does not reduce heteroscedasticity. This should be a clear indication of heteroscedasticity due to a missing cross-term or model misspecification. This is quite likely to happen in practice where OLS is used.

A refined example

In our first example (Figures 5-8), we see that residuals are initially heteroscedastic and as we increase the size of the fitting data we observe an improvement in fitting quality up to the point where residuals become homoscedastic. We can use this information to set a criterion for the minimum number of training points, N.

This is not the case in our second example (Figures 9-10) where the fitted model is not a good estimator of the demo model. Here, increasing N cannot reduce the residual funnel, as we have observed.

The demo model used in the examples above is a simple one yet it demonstrates the point under discussion here and our findings can be generalised to a model with higher degrees of individual terms and cross-terms. To demonstrate this, we now consider the following model as our truth model:

$$Y = f(X_1, X_2) + \epsilon = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4 X_1^2 + \beta_5 X_2^2 + \beta_6 X_1^2 X_2 + \beta_7 X_1 X_2^2 + \epsilon$$

Fitting a proxy model to training data produced using this "truth model" results in the shape of residuals shown in Figure 11.



Our OLS proxy-modelling algorithm appears to have picked up all terms of our truth model. The resulting fit is close to perfect!

Our second example demonstrates the case where the fitting algorithm used is unable to detect a particular cross-term. Here, increasing N cannot reduce the resulting heteroscedasticity of residuals. This is in line with previous observations presented in this paper.

Proxy modelling tools commonly used in the insurance industry, e.g., for Solvency II Internal Model purposes, are often applied on ALM models ("truth models") that embed inherent heteroscedasticity (such as higher Net Present Value volatility for extreme scenarios).

There is no standard procedure to identify the root cause of heteroscedasticity in a particular model. In literature, statistical tests like White's, Breush-Pagan or Goldfeld-Quandt (see Breusch and Pagan [1979], Goldfeld and Quandt [1965]) only check the hypothesis that the residual satisfies the homoscedasticity property, but if this is not the case, the test does not tell us what causes this effect. Then several approaches can be developed when trying to render models homoscedastic in practice, like GLS or Quasi-Generalised Least Squares (QGLS, see Shults & Hilbe [2014]).

Conclusions and outlook

LSMC proxy models are not always easy to validate, as a suboptimal fit may be due to a variety of reasons. LSMC Surgery is a useful framework to make LSMC troubleshooting more accurate without requiring any sizeable computing budget.

We have shown how LSMC Surgery facilitates investigations into various fitting issues including the frequently observed heteroscedasticity issue in insurance applications. We have then looked into different heteroscedasticity causes, both theoretically and via stylised numerical examples. It has emerged that an increase of the simulation budget can help remove heteroscedasticity in cases where model selection successfully captures relevant cross-terms, but not in those cases where cross-terms remain undetected. Understanding the cause of heteroscedasticity is therefore important when looking at Insurance-specific applications of LSMC because some "natural causes" do exist, e.g., residual funnels caused by discounting at lower or negative rates in the interest rate risk space.

Because heteroscedasticity cannot always be avoided in LSMC work, more advanced approaches may be worth considering as well. For example, piecewise LSMC only assumes local homoscedasticity and hence relaxes the (global) homoscedasticity assumption underpinning classical LSMC. Piecewise LSMC also improves fitting quality where kinks exist in the training data—something that is currently under investigation and will be the subject of our next LSMC paper.

Literature

- Breusch, T. S. and Pagan, A. R. A simple test for heteroscedasticity and random coefficient variation. Econometrica: Journal of the econometric society, 1979, p. 1287-1294.
- Goldfeld, S. M. and Quandt, R. E. Some tests for homoscedasticity. Journal of the American statistical Association, 1965, vol. 60, no 310, p. 539-547.
- Glasserman, P. (2013). Monte Carlo methods in financial engineering (Vol. 53). Springer Science & Business Media.
- Klein, A. G., Gerhard, C., Büchner, R. D., Diestel, S., & Schermelleh-Engel, K. (2016). The detection of heteroscedasticity in regression models for psychological data. Psychological Test and Assessment Modeling, 58(4), 567.
- Shults, J. and Hilbe J. M. Quasi-least squares regression. CRC Press, 2014.

🕻 Milliman

Milliman is among the world's largest providers of actuarial and related products and services. The firm has consulting practices in life insurance and financial services, property & casualty insurance, healthcare, and employee benefits. Founded in 1947, Milliman is an independent firm with offices in major cities around the globe.

milliman.com

CONTACT

Abdal Chaudhry abdal.chaudhry@milliman.com

Adel Cherchali adel.cherchali@milliman.com

Michael Leitschkis michael.leitschkis@mlliman.com

Julien Vedani julien.vedani@milliman.com

© 2021 Milliman, Inc. All Rights Reserved. The materials in this document represent the opinion of the authors and are not representative of the views of Milliman, Inc. Milliman does not certify the information, nor does it guarantee the accuracy and completeness of such information. Use of such information is voluntary and should not be relied upon unless an independent review of its accuracy and completeness has been performed. Materials may not be reproduced without the express consent of Milliman.