On Valuation of Observations in Linear Regression Models

Bachelor Thesis in Statistics, 15 ECTS

STAH11, Autumn 2019

Mattias Jönsson

Supervisor

Prof. Björn Holmquist

Abstract

In the Machine Learning field, more and more of the data collection is commercialised, even with monetary rewards to people and organisations for providing input data for models. Even if data collection is not associated with direct costs for the researcher, there are many cases where there are indirect, or circumstancial, costs associated with it.

An established concept in game theory is "Shapley Values", which has had a lot of success in the field of statistics and machine learning over the last number of years, for example as a technique for variable importance estimations. Now, researchers have proposed using Shapley Values also to quantify the worth, or value, of an observation in a model (Data Shapley Values). However, little effort has earlier been spent to properly evaluate these in an Ordinary Least Squares setting, especially since there is already a very established way of quantifying an observations influence (Cook's Distance), which should be reasonably well aligned.

Hence, this thesis sets out to explore the use of Data Shapley in Linear Regression models, with the purpose to research if this is a valuable concept for a researcher using OLS models. This thesis will try to approach the topic by answering the following specific questions: * What is a suitable set of parameters for estimating Data Shapley-values for linear regression models? * How well does Data Shapley values and Cooks Distance values agree on the valuation of an observation? * Is it possible to use Data Shapley values to detect outliers also in linear regression models?

Data Shapley is studied in some detail with the use of four different datasets and models, and Data Shapley values that are estimated using three different metrics and four different configurations of the estimation algorithm. Results are compared with Cook's Distance for evaluation.

The main conclusion from this research is that Data Shapley is a serious contender to Cook's Distance in capturing the worth of an observation. It performs better than, or at least as well as, Cook's Distance in capturing the low value observations, but it also performs significantly better than Cook's Distance in capturing good observations as well.

1 Introduction

In April 2019, the authors Amirata Ghorbani and James Zou (2019) released the paper *Data Shapley: Equitable Valuation of Data for Machine Learning*, which has subsequently received some attention in the field. The objective of the authors' study was to find a way to estimate the worth of a specific observation used in a model, and they propose a concept called Data Shapley specifically for this purpose. The paper was accepted to be presented on the International Conference on Machine Learning (ICML) in 2019, has been referenced in other publications several times since its release, and had an entire episode of the very popular podcast *Linear Digressions*.

The paper above is - at least according to the author of this thesis - focused heavily on complex models where the purpose is to perform classification predictions. For example, in all the empirical examples shown in the paper, the measure of quality of the model has been accuracy ($(TP + TN)/(TN + TP + FP + FN)$, where TP=True Positive, TN=True Negative, FP=False Positive and FN=False Negative - a metric only applicable to so called classification problems), and only one case is something other than a Neural Network. Data Shapley should certainly be general enough to also handle simpler regression models, including something as fundamental as Ordinary Least Squares (OLS). However, in the paper there are no mentions of work regarding identification of suitable use or modifications for example for OLS models.

Hence, this thesis sets out to explore the use of Data Shapley in Linear Regression models, with the purpose to research if this is a valuable concept for a researcher using OLS models. This thesis will try to approach the topic by answering the following specific questions:

- What is a suitable set of parameters for estimating Data Shapley-values for linear regression models?
- How well does Data Shapley values and Cooks Distance values agree on the valuation of an observation?
- Is it possible to use Data Shapley values to detect outliers also in linear regression models?

Naturally, there are many mentions of yet undefined concepts in this section, and these will be described in the next chapter.

Worth mentioning, is that this report does not set out to completely exhaust the topic around Data Shapley for linear regression models, and it is likely that many more questions are yet to be explored. The report will for example only look at ordinary least squares models.

Also, the specific models used in the empirical study, does not make any claim of being "the" most suitable model for the specific case. Some effort has been spent in order to get realistic and reasonably well fitting models, but they are still built primarily for the purpose of studying the corresponding Data Shapley values.

All empirical research for this thesis is performed in R (R Core Team 2019).

2 Earlier Research

In this chapter, a brief introduction to some of the foundational concepts underpinning the measurements of influence and valuation of observations will be presented. First and foremost, Cook's Distance will be explained in some detail. Afterwards, the necessary foundation will be laid to understand the Data Shapley values. The understanding of Data Shapley values requires an understanfding about both Monte Carlo-techniques and Shapley values (note the difference between "Shapley values" and "Data Shapley values"), so some time will also be spent on those topics. The author assumes the reader has prior knowledge of multiple linear regression, and foundational related concepts. This includes for example calculating residuals and measurements like Adjusted R^2.

2.1 Measuring Influence of an Observation

In statistics, when one has identified a model and estimated the best fit, an important aspect is to investigate model diagnostics. As the model has been fit, there have most likely been some assumptions made and investigating the model through a variety of methods can for example reveal violations of these assumptions, including non linearlities, heteroskedastic errors and outliers.

However, according to Cook and Weisberg (1982):

> *"A related question that cannot be easily addressed by those methods is that of stability, or the study of the variation in the results of an analysis when the problem formulation [...] is modified. If a case is deleted, for example, results based on the reduced data set can be quite different from those based on the complete data[...]. We call the study of the dependence of conclusions and inferences on various aspects of a problem formulation the study of* ***influence****."*

One of the very established and widely used metrics for diagnosing influential observations, is Cook's Distance, D_i. This can be briefly summarized as the total changes in the regression model when observation i is removed from it. To understand this in more detail, one need to understand a few foundational concepts:

2.1.1 Leverage

Leverage is a metric that tries to capture if an observation x_i 1) is away from the bulk of x's and 2) how attracted the regression line is to this point (Sheather 2009,p. 54). However, Leverage by itself usually does not tell the full story by itself. A model may contain observations of both "good" and "bad leverage". To explain this, let's look at three examples:

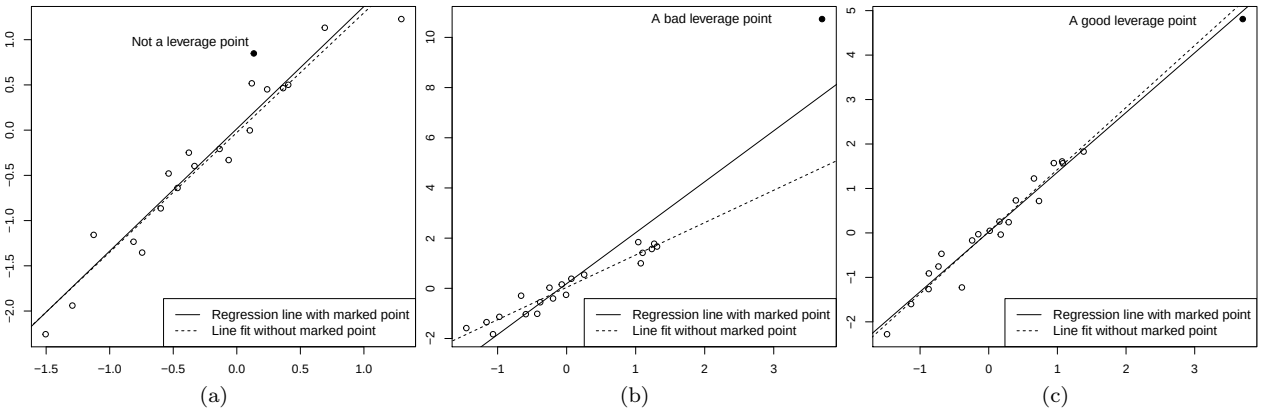


Figure 1: Three examples of Leverage

In Figure 1(a) above, the marked point would not be called a leverage point, since the effect on the regression line is minimal when the observation is omitted, and the observation is not far away from the other *x's*. However, in Figure 1(b) the observation is further away from the other *x's*, and it is also affecting the estimated regression line quite drastically. This would therefore be a "Bad Leverage"-point. In Figure 1(c), the observation is located away from the other *x's*, and the regression line is most likely heavily using this observation. However, since it does not deviate much from the "true" parameter, this would be a "Good Leverage"-point.

Leverage is easiest calculated in the diagonal on the $n \times n$ matrix H, called the Hat matrix. If

$$\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y = HY \tag{1}$$

then

$$H = X(X^T X)^{-1} X^T \tag{2}$$

The diagonal (h_{ii}) now captures the magnitude of the change to the regression line when including observation i.

2.1.2 Standardized Residual

As for example Sheather (Sheather 2009) states, a challenge with using residuals is that these do not have the same variance, due to these being a function of both σ^2 and h_{ii} (Cook and Weisberg 1982,p. 18). This is especially problematic in the case of high leverage points. The advice is therefore to standardize (also called studentized, for example by Cook and Weisberg (1982)) with the following calculation:

$$r_i = \frac{\hat{e}}{s\sqrt{1 - h_{ii}}} \tag{3}$$

where $\hat{e} = y_i - \hat{y}_i$ is the raw residual and $s = \hat{\sigma}$, namely

$$s = \sqrt{\frac{1}{n-p} \sum_{j=1}^{n} \hat{e}_j^2} \tag{4}$$

where p = number of parameters.

2.1.3 Cook's Distance

So, with the definition of both Leverage and Standardized residuals we are ready to define Cook's Distance as defined by Cook (1977). Here, the total influence an observation has is being captured as it is coming from either a high standardized residual, a high leverage value or both. Formally, the Cook's Distance D_i for a particular observation i is defined as

$$D_i = \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}} \tag{5}$$

where p = number of parameters.

2.2 Shapley Values in Game Theory and Statistical Learning

To transition from an established way of measuring influence, we can now introduce a potential alternative - the concept of Data Shapley values. To understand Data Shapley values, the context need to be set from "classical" Shapley values. Again, note that Shapley Values refers to the general concept, whereas Data Shapley and Data Shapley Values (sometimes abbreviated as DSV or just DS) is the specific use of assessing value of observations. Both concepts are described below.

Shapley values is a established concept in game theory and was introduced in 1953 (Shapley 1953). It tries to describe the expected gain of a cooperation, given the individual actors that participate. Also, inversely, what individual value does each actor contribute with to the overall gain of the cooperation.

Formally: for a set N (of $|N|$ actors) a subset of S (i.e. a coalition of $|S|$ actors, also denoted n, which will be described later) has a value function $v(S)$ (the "worth" of the coalition), that describes the total expected sum of payoffs the actors in S can obtain by this specific collaboration. A Shapley value ϕ_j is then, for an individual actor $\{j\}$:

$$\phi_j(v) = \frac{1}{|N|} \sum_{S \subseteq N \setminus \{j\}} \binom{|N|-1}{|S|}^{-1} \left(v\left(S \cup \{j\}\right) - v(S)\right) \tag{6}$$

where $|N|$ and $|S|$ are the sizes (cardinality) of the sets. This formulation is in line with the original formulation by Shapley. However, in this thesis $|S|$ will often be the specific number of observations selected for a model (out of all available observations |N|). Since the number of observations often are notated n in statistics, this will be favored throughout this thesis.

To better understand the Shapley Value concept let us take an example:

A football team wants to set the salary of the players in the team based on their value to the team and a fixed amount is available for distribution to the players. The total set of players include all the regular players and all reserves. That means the set of N has > 11 actors (i.e. $|N| > |S|$), which means there are multiple variations of subsets with $|S| = n = 11$, meaning the players on the field at any given match). One could also quantify the performance of the team (for example by the ratio $\frac{GamesWon}{GamesLost}$ or $\frac{GoalsScored}{GoalsAgainst}$). Some players will then be more valuable to the team than others: With some players active on the field, the team is more likely to perform well according to the definition. Likewise, some other players do likely not contribute as much to the overall performance. Shapley values quantify a particular team composition's marginal contributions to the overall performance and distributes the marginal contribution evenly between the actors, which is then repeated for all different team compositions. That is, how much of the metric $\frac{GoalsScored}{GoalsAgainst}$ can be attributed to each individual player.

Take for example an attacker: In some compositions, there will be too many attackers and to few midfield players or defenders. An attacker that is for example adding more value in a pure forward position than they decrease value in a midfield or defending position would have an overall positive contribution - and a higher Shapley value than a player who is equally good at attacking but worse as a defender or midfielder. Of course, there would have to be many games played with many different subsets of players and team compositions, including suboptimal ones. Therefore, it may be challenging to calculate these values in this specific scenario.

2.2.1 Properties of Shapley values

Properties of Shapley values have been the subject of many studies (Pal and Bharati 2019) and here some of the most central properties are summarized. The properties listed in this section are also the ones that Zou and Ghorbani highlight as key for the Data Shapley concept (2019).

- **Efficiency**: The sum of individual Shapley values for each actor are equal to the total performance of the collaboration:

$$\sum_{i \in N} \phi_i(v) = v(N) \tag{7}$$

- **Symmetry**: If two actors (i, j) have the same Shapley values $\phi_i(v) = \phi_j(v)$, then

$$v(S \cup \{i\}) = v(S \cup \{j\}) \tag{8}$$

- **Linearity**: Not only can actors value be added up within a collaboration (Efficiency), one actors total value across multiple collaboration can also be added, so that:

$$\phi_i(v + w) = \phi_i(v) + \phi_i(w) \tag{9}$$

- **Null player**: The Shapley value of a *null* player is 0 The Shapley value $\varphi_i(v) = 0$ if $v(S \cup \{i\}) = v(S)$ for all coalitions S that do not contain i.

It should be noted that there are several other variations on describing the same fundamental properties. The Nobel Prize winner Aumann (1994) mentions for example that the Linearity (also known as Additivity) and Null Player properties can be replaced with a Monotonicity property. The properties presented here should however be adequate to understand the fundamentals of the Shapley Value concepts and is aligned with the founndation for Data Shapley.

2.2.2 Shapley values for Variable Importance

As a side note, Shapley values have been proposed throughout history to be used to analyze variable importance - the relative importance that each variable adds to an overall model. As Pal and Bharati (2019) note, this method has been proposed with different names for example by Lindeman, Merenda, and Gold (1980), Kruskal (1987), and Lipovetsky and Conklin (2001).

In essence, the proposal from these authors have been the following:

Assume there is a dataset with which one fits a variant of a model like: $y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_p x_p + \beta_0$, where the variant is either the full model or a subset of x_i:s.

One could then see the relative contribution of variable x_i as the average increase in a performance metric (for example R^2 as proposed by Lipovetsky and Conklin (2001) across all the model permutations that include x_i from all the permutations that do not.

Shapley values have also been proposed for model interpretability in the Machine Learning community, by Lundberg and Lee (2017) in a form called SHAP (SHapley Additive exPlanations). This as well is in essence the normal concept, but including some more efficient way to compute or estimate them.

2.3 Monte Carlo-simulations for estimations

This report will include minimal detail on the fundamentals of Monte Carlso methods, and the interested reader would go elsewhere for these details (for example Harrison (2010); or Robert and Casella (2010)). For the reader to have some context, a minimal introduction will however follow here.

Monte Carlo methods could refer to a large variety of approaches where simulations are used as a foundation to numerical analysis. It has been around for hundreds of years, but was more rigorously established during the Manhattan project. It is important to understand that:

> *"There is no single Monte Carlo method – any attempt to define one will inevitably leave out valid examples – but many simulations follow this pattern:*

-model a system as a (series of) probability density functions (PDFs);
-repeatedly sample from the PDFs;
-tally/compute the statistics of interest." (Harrison, 2010)

Harrison (2010) also continues to state *"Monte Carlo simulation is now a much-used scientific tool for problems that are analytically intractable and for which experimentation is too time-consuming, costly, or impractical."*. It will later become obvious that we are facing a challenge where Monte Carlo-methods are a good fit, due to the first piece of this statement - the problem is very challenging to solve analytically.

2.4 Data Shapley for valuation of observations

> "As the legal system moves toward recognizing individual data as property, a natural problem to solve is to equitably assign **value** to this property"

As the quote says, Zou and Ghorbani (2019) identifies a need to assign a quantitative value of a particular observation. Here follows a brief explanation of this concept, first by defining some of the core concepts, and then by going into the principles for the calculation.

Zou and Ghorbani lets D be the set of all observations {1,2,. . . ,n} and S be any subset of D ($S \subseteq D$). Note that D here equals N in more traditional notation (see section 2.2). Other key definitions are

- **Learning Algorithm** $\mathscr{A}$: Different algorithms or model types will have different valuations for the same observation. Therefore, the "learning algorithm" will be the most basic foundation for identifying the valuation. In this study, this is of curse constrained to be linear regression models.
- **Performance Score Value** $V(S, \mathscr{A})$ **or just** $V(S)$: A Performance Score is a measure of model quality, for example R^2, R^2_{Adj}, AIC, etc. Performance score will depend on algorithm and which set of observations is being used to fit the model.
- **Data Value** $\phi_i(D, \mathscr{A}, V)$ **or just** ϕ_i: This is the value of the particular observation $\{i\}$, adhering to the principles mentioned below.

Seen in the light of Shapley values, the Performance Score *V(S)* here corresponds to the function *v(S)* from previously (again, see section 2.2), whereas the Data Value ϕ_i is the Shapley Value ϕ_j.

The Data value should, according to the authors, have a property of equitability, which the authors define as fulfilling the following principles:

1. The valuation ϕ is zero for an observation {i} that does not change the performance
2. The valuation ϕ for an observation {i} is constant for any subset S belonging to the full dataset S-{i}
3. The valuation ϕ is proportional to the performance metric of the model such that $\phi(V_1) + \phi(V_2) = \phi(V_1 + V_2)$. E.g. if observation {i} has $\phi = 10$ and {j} has $\phi = 12$, both measured in SSE (that is, one can expect a decrease of SSE of 10 respective 12 when {i} or {j} is added in), then when both are added in, the SSE decreases with 10+12.

And as one can see here, these principles are based on the mathematical properties of Shapley values.

The authors also summarise their work with saying about Data Shapley method that:

- "It is more powerful than the popular leave-one-out or leverage score in providing insight on what data is more valuable for a given learning task"
- "Low Shapley value data effectively capture outliers and corruptions"
- "High Shapley value data inform what type of new data to acquire to improve the predictor"

There are some more details to cover, specifically regarding the performance metric used, but this is covered in the introductory section to the empirical results.

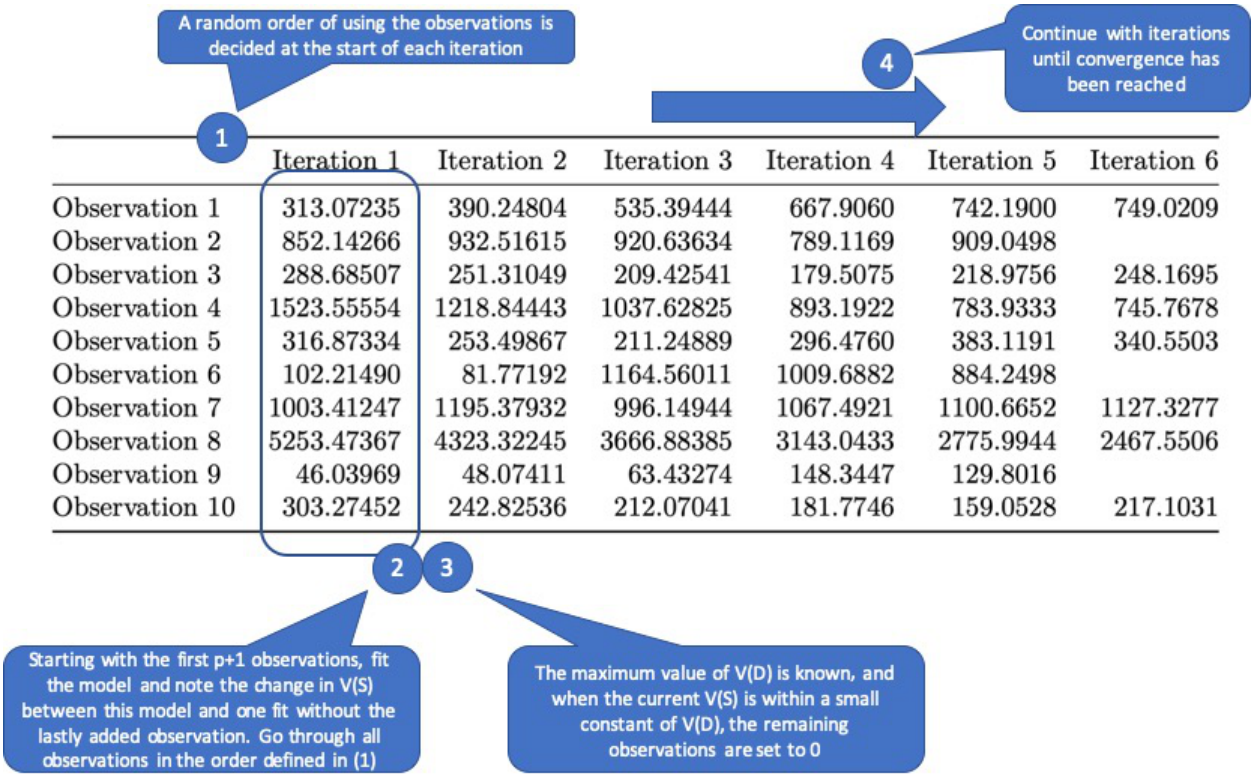
2.4.1 Truncated Monte Carlo

The computation of Data Shapley values will most likely be incredibly heavy, due to the "exploding" combinatorics of datasets with many observations. Therefore Zou and Ghorbani also propose an algorithm (Truncated Monte Carlo) to estimate Data Shapley Values by Monte Carlo approximation methods:

Algorithm 1: Truncated Monte Carlo Shapley

Input : Train data $D = \{1, ..., n\}$, learning algorithm $\mathscr{A}$, performance score V
Output: Shapley value of training points: $\phi_1, ..., \phi_n$
Initialize $\phi_i = 0$ for $i = 1, ..., n$ and $t = 0$
while *Convergence criteria not met* **do**
  $t \leftarrow t + 1$
  π^t : Random permutation of train data points
  $v_0^t \leftarrow V(\emptyset, \mathscr{A})$
  for $j \in \{1, ..., n\}$ **do**
    if $|V(D) - v_{j-1}^t| <$ *Performance Tolerance* **then**
      $v_j^t \leftarrow v_{j-1}^t$
    else
      $v_j^t \leftarrow V(\{\pi^t[1], ..., \pi^t[j]\}, \mathscr{A})$
    end
    $\phi_{\pi^t[j]} \leftarrow \frac{t-1}{t}\phi_{\pi^{t-1}[j]} + \frac{1}{t}(v_j^t - v_{j-1}^t)$
  end
end

One way to understand this is by seeing the outputs of the algorithm as a matrix:



	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5	Iteration 6
Observation 1	313.07235	390.24804	535.39444	667.9060	742.1900	749.0209
Observation 2	852.14266	932.51615	920.63634	789.1169	909.0498	
Observation 3	288.68507	251.31049	209.42541	179.5075	218.9756	248.1695
Observation 4	1523.55554	1218.84443	1037.62825	893.1922	783.9333	745.7678
Observation 5	316.87334	253.49867	211.24889	296.4760	383.1191	340.5503
Observation 6	102.21490	81.77192	1164.56011	1009.6882	884.2498	
Observation 7	1003.41247	1195.37932	996.14944	1067.4921	1100.6652	1127.3277
Observation 8	5253.47367	4323.32245	3666.88385	3143.0433	2775.9944	2467.5506
Observation 9	46.03969	48.07411	63.43274	148.3447	129.8016	
Observation 10	303.27452	242.82536	212.07041	181.7746	159.0528	217.1031

Here, 10 observations (rows) have had 5 completed iterations (columns) generating values and is currently working on the 6th. An iteration starts with randomly reordering the observations. Values are then generated

by adding the observations one by one to a model estimation. For each observation added, the ΔPerformance (i.e. model performance with observation - model performance without observation) is entered into the appropriate place in the matrix. Once the model performance reaches a certain level (maximum identified model performance (with all observations) - model performance with observation $< C$), all remaining cells in the iteration are set to 0 (because the added value would be small).

As one can see, this algorithm is perhaps not a perfect fit with the general principles of Monte Carlo techniques described by Harrison (2010) that were mentioned in section 2.3. Potentially, one could see the Data Shapley values as a value that is being estimated by drawing random samples from a distribution that is both constrained (to retain the propoerties of Shapley values) and very complex (since it would have to be conditioned on which observations are already in the model). However, given that there is a great deal of simulation in the randomization of the order of dealing with the observations, it can most likely be seen as a Monte Carlo method.

3 Method for the empirical study

Throughout this chapter, the Data Shapley concept will be studied from many different perspectives. When evaluating models estimated with reduced set of observations, Adjusted R^2 will be used, and some residual analysis will also be performed.

One important consideration appears when looking at the Data Value-function $\phi_i(D, A, V)$, which is built up from what Ghorbani and Zou (2019) calls a "Black Box Oracle", the Performance Score $V(S, A)$. When A is used to estimate a regression model, it is - as indicated above - no longer possible to use the performance metrics that the authors mainly work with. Hence, A need to be another metric that captures the concept of model performance. In this study, we will investigate the use of the following different metrics: Residual Sum of Squares (SSE), Explained Sum of Squares (SSR) and Adjusted R^2.

This means that the Data Shapley values will have different distributions and different characteristics when describing "high" vs "low" valued observations. If R^2_{Adj} is used, values are likely to be both positive and negative, whereas when SSR or SSE is used, Data Shapley values should mostly be positive. More details on this will follow later in this chapter.

No package for calculating Data Shapley values was found by the author, so the Truncated Monte Carlo algorithm was implemented in R. It is built in a way where it is easy to select different performance metrics for estimating the Data Shapley values for easy experimentation.

Final analysis of Data Shapley values will be performed in a similar fashion to how Ghorbani and Zou (2019) performed their analysis. Given a predefined model for a specific dataset, the following pseudocode will be applied in an iterative process:

Algorithm 2: Valuation Scoring

Set V = Empty Set
Set i = 0

while $i \leq 20$ **do**
    Remove the $n = i$ observations with the lowest Data Shapley values
    Fit model with current set of observations, and get R^2_{Adj}
    Append V with R^2_{Adj}
end

Given that an observation has a specific value in the overall estimation of the correct set of parameters in a model, the quality of the model - in this study identified by R^2_{Adj} - should increase.

3.1 Introduction to Data Shapley for Linear Regression models

Ghorbani and Zou (2019) focus mainly on complex models used for classification tasks, and no or very little emphasis on simple linear regression models. There has been some work performed with a logistic regression model, but the quality of this model was mainly measured in Prediction Accuracy - a specific concept that does not translate well to a linear regression model. Therefore, there is a need to explore the different considerations for using the Data Shapley-method for linear regression models. In this study, the following performance metrics will be used:

Residual Sum of Squares: $\sum_{i=1}^{n} (\hat{y}_i - y_i)^2$, in some cases in this thesis also referred to as *SSE*

Explained Sum of Squared: $\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$, in some cases in this thesis also referred to as *SSReg* or *SSR*

Adjusted R Squared: $1 - \frac{SS_{res}/df_e}{SS_{tot}/df_t}$, where SS_{res} is Residual Sum of Squares from above and $SS_{tot} =$

$\sum_{i=1}^{n}(y_i - \bar{y})^2$, df_e is the error degrees of freedom $n - p - 1$ (number of observations-number of parameters estimated) and df_t is the total degrees of freedom $(n - 1)$

These metrics all qualify as Performace Score metrics, since they can be seen to capture how accurately the model can describe the data - for example a lower Residual Sum of Squares, the better the model fit. This metric is also clearly related to the magnitude of the errors. I.e a model where the residuals are very small will have a better performance tham a model with larger residuals.

For all the three types of performance score, Linearity and Efficiency properties seem to hold since marginal changes can be distributed among the observations. However, the interpretation of the Data Shapley value for an observation may be more or less obvious. For example, one could see that if using Residual Sum of Squares, adding an observation to a subset S will add two things: Added Sum of Square for it's own deviation between $\hat{y}_i$ and y_i, and the change in Residual Sum of Squares for all other observations currently in the subset (as the model parameter potentially changes). This entire change will be associated with the observation, and it must follow that any changes in the end must sum to the total. Notably here, a lower value would be better, but Shapley value properties does not assume anything about the good or bad with high or low values. However, for simplicity and consistency in this research, the sign in this case will be changed so a decrease in Residual Sum of Squares will equate to a higher negative value of the same.

The same as above would hold for both Explained Sum of Squares and Adjusted R^2, but here high values are of course desirable. It is however here that the interpretation of the value becomes more challenging. In the Explained Sum of Squares case, an added observation (x_i, y_i) will change both $\bar{y}$ and $\hat{y}_i$ for every previous observation, plus add its own $\hat{y}_j - \bar{y}$ to the V(S). Adjusted R^2 will then be even more complex, due to changes also to the degrees of freedom and Total Sum of Squares, over and above all the other changing things in Explained Sum of Squares.

Furthermore, Ghorbani and Zou are not specific on how they identified convergence in the algorithm, so this also need to be established up front. The method proposed is a monte carlo method, and therefore the idea is assumed to be using the asymptotic behavior as $n \to \infty$, using the law of large numbers which states that, if $X_1, X_2, X_3, \ldots$ are independent and identically distributed random variables with mean μ and standard deviation σ, then for any constant $\epsilon > 0$ we have

$$\lim_{n\to\infty} \Pr\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| > \epsilon\right) = 0.$$

In this case, convergence can be seen to happen on two levels: As the individual observations get closer and closer to their individual true Data Shapley value, and on the overall dataset level, as more and more Data Shapley values are converging.

This can for example be seen in the following plot, generated for one of the datasets. Each line represents the estimator of the Data Shapley value of a specific observation, and as can be seen in Figure 2, each estimator has a slightly different convergence.

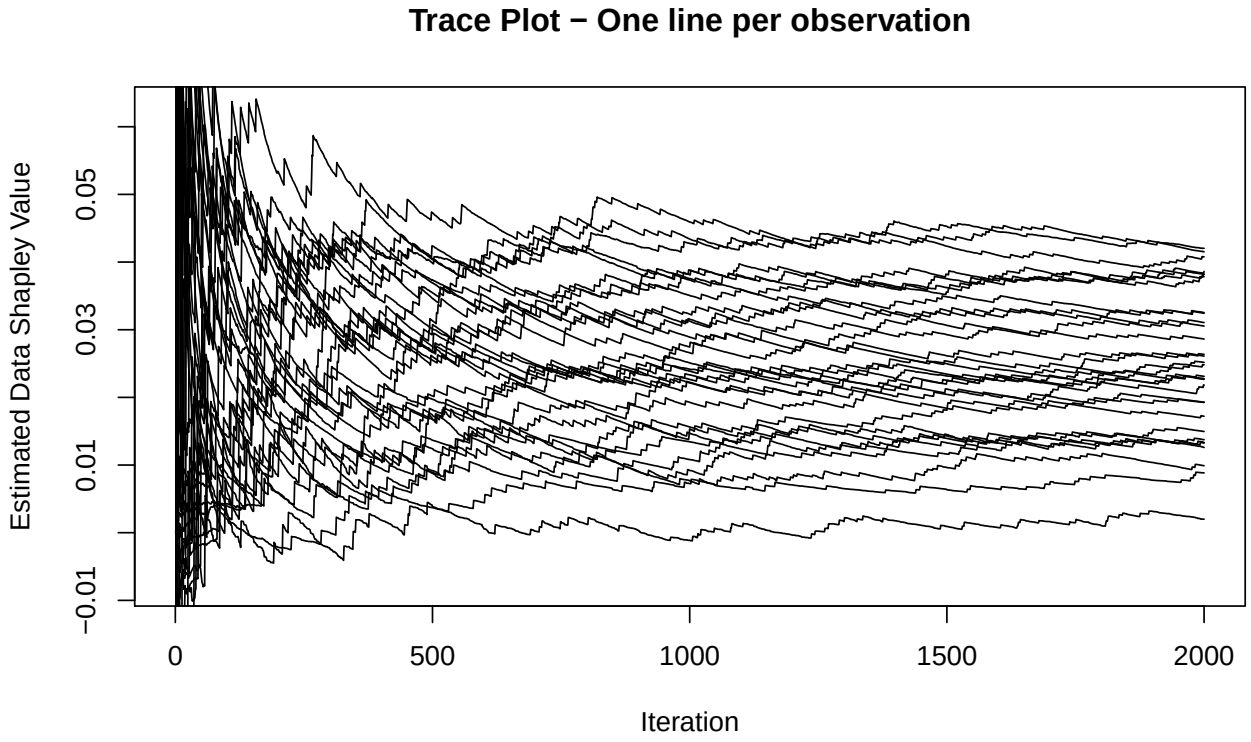


Figure 2: Convergence of Data Shapley-values for individual observations

Convergence is in this study defined as when the standard deviation of $(\bar{x})$: $\frac{\sigma}{\sqrt{n}} \leq c$, where c is a small constant. As more and more of the individual observations converge to their true Data Shapley value, we can look at overall convergence of the algorithm by counting the number of observations that have $\frac{\sigma}{\sqrt{n}} \leq c$.

3.2 Levels of Truncation

Truncation refers to the selected constant C in the Truncated Monte Carlo algorithm. This defines how close the cumulative sum in an iteration can be to the corresponding value calculated for the full set of observation, before setting the remaining observations value to 0. In this study "High Truncation" is used to refer to situations with a large C (i.e. we stop an iteration earlier and set more observations data shapley value to 0). "Low Truncation" refers to the opposite (i.e. small C).

Important to note though, is that the actual magnitude of C will be dependent on the metric used to estimate the Data Shapley value. When using SSE or SSR, C will also be dependent on the magnitudes of the dependent variable, hence the factors below are multiplied with the SSE or SSR for the full model (i.e. with all observations) for the corresponding dataset. For Adjusted R^2 the values presented below are used exactly as presented. The levels of C are as follows:

Truncation.Level	C
Low	0.001
Low-Medium	0.010
High-Medium	0.050
High	0.100

3.3 Datasets

Four different datasets will be used in this thesis. Some of the first observations in these datasets are presented below, together with the model specification being used. Again, this thesis does not make any attempt to model the relationships in the most suitable way, but effort has been spent to find at least something of reasonable quality.

Model formulation is here noted using R syntax. For example, the formula "y~x" means "y is modeled as a function of x" and "y~x+z" means "y is modeled as a function of x and z". Formally, y~x+z means the following model: $y_i = \beta_1 x_i + \beta_2 z_i + \beta_0 + \epsilon_i$. One can also specify "y~." as shorthand for using all the available variables as independent variables.

For more details on the complete models and diagnostic plots, see Chapter 5.

3.3.1 Motor Trend Car Road Tests (mtcars)

A dataset containing information about performance of 32 cars (1973–74 models) (Henderson and Velleman 1981).

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2

Model: hp~vs+cyl+disp+drat

3.3.2 Swiss Fertility

A dataset with fertility and socio-economic indicators for 47 provinces in Switzerland in the late 1800's (Mosteller and Tukey 1977).

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Courtelary	80.2	17.0	15	12	9.96	22.2
Delemont	83.1	45.1	6	9	84.84	22.2
Franches-Mnt	92.5	39.7	5	5	93.40	20.2
Moutier	85.8	36.5	12	7	33.77	20.3
Neuveville	76.9	43.5	17	15	5.16	20.6

Model: Fertility~.

3.3.3 Taiwan Houseprices

A dataset containing real estate valuations in Sindian District, Taipei City, Taiwan (Yeh and Hsu 2018).

TransactionDate	HouseAge	DistanceToMRT	ConvenienceStores	Latitude	Longitude	HousePrice
2012.917	32.0	84.87882	10	24.98298	121.5402	37.9
2012.917	19.5	306.59470	9	24.98034	121.5395	42.2

TransactionDate	HouseAge	DistanceToMRT	ConvenienceStores	Latitude	Longitude	HousePrice
2013.583	13.3	561.98450	5	24.98746	121.5439	47.3
2013.500	13.3	561.98450	5	24.98746	121.5439	54.8
2012.833	5.0	390.56840	5	24.97937	121.5425	43.1

Model: HousePrice~.

3.3.4 Wii MarioKart game prices on Ebay

Auction data from Ebay for the game Mario Kart for the Nintendo Wii video game console. This data was collected in early October, 2009 (openintro.org 2019)

duration	n_bids	start_pr	ship_pr	total_pr	seller_rate	wheels
3	20	0.99	4.00	51.55	1580	1
7	13	0.99	3.99	37.04	365	1
3	16	0.99	3.50	45.50	998	1
3	18	0.99	0.00	44.00	7	1
1	20	0.01	0.00	71.00	820	2

Model: total_pr~.

3.4 Correlations between Data Shapley, Leverage and Cooks Distance

Ghorbani and Zou only briefly mention Cooks distance and Leverage as related work, but do not go into either in detail. Although this could lead to the idea that either would be a good candidate to represent an observations "Value", this is likely not the case. As can be noted in the Earlier Research section above, Leverage may not necessarily explain the full "Value" of an observation. This, since an observation may in theory have either "good" or "bad" Leverage (see section 2.1). Therefore, this study initially looks at the empirical correlations between Data Shapley and Leverage to verify the assumed weaker correlation, and subsequently also Data Shapley and Cooks Distance. Based on the short reasoning here, Cooks distance should be better at capturing the value of an observation than Leverage would.

Data Shapley values are estimated for the datasets described above and is here plotted with both Leverage and Cooks Distance. As can be seen in these plots, the correlations between Cooks Distance and Data Shapley values are far stronger than between Cooks Distance (R approx 0.7) and Leverage (R approx 0.04) values. Similar characteristics are found also for the other three datasets, which seem to indicate a stable correlation with Cooks Distance and a stable non-correlation with Leverage values.

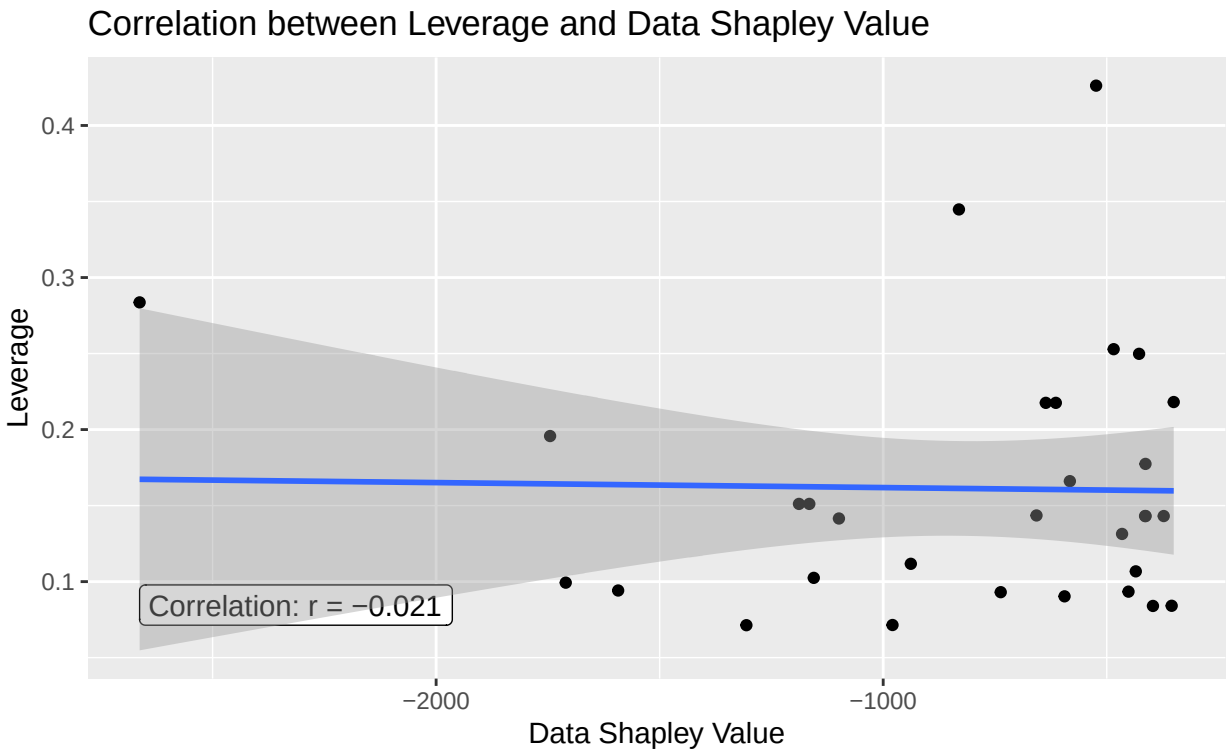


Figure 3: Correlation: Leverage and Data Shapley Values (95% CI)

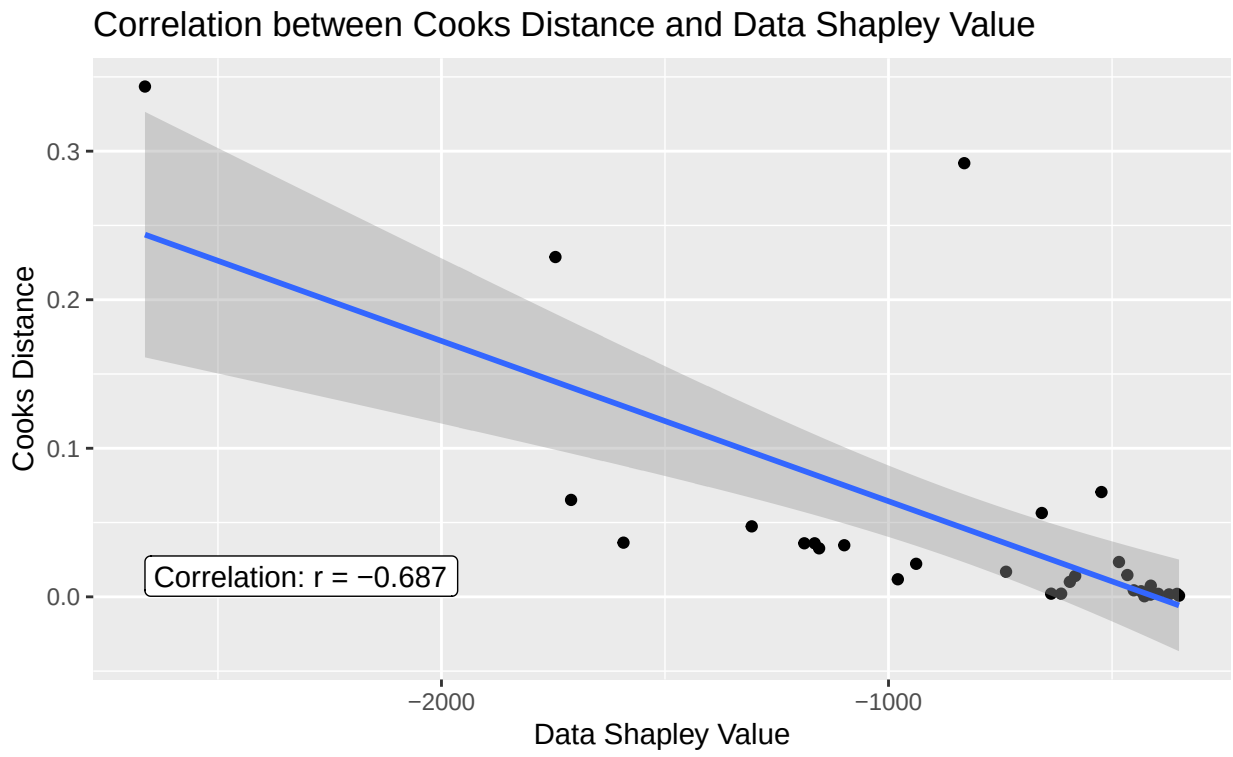


Figure 4: Correlation: Cooks Distance and Data Shapley Values (95% CI)

Given that there is very little correspondence between Data Shapley values and Leverage - as was very much expected - no further investigations will look at Leverage in this study.

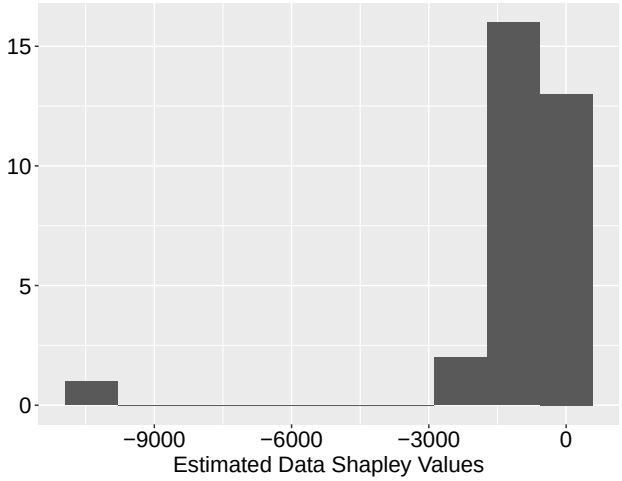
4 Empirical Results

The figures in section 3.3 are generated in the process of empirically studying the behaviour of Data Shapley values and the Truncated Monte Carlo algorithm used to estimate them. Four different datasets are used in the study, described in the Datasets section (3.2) above.

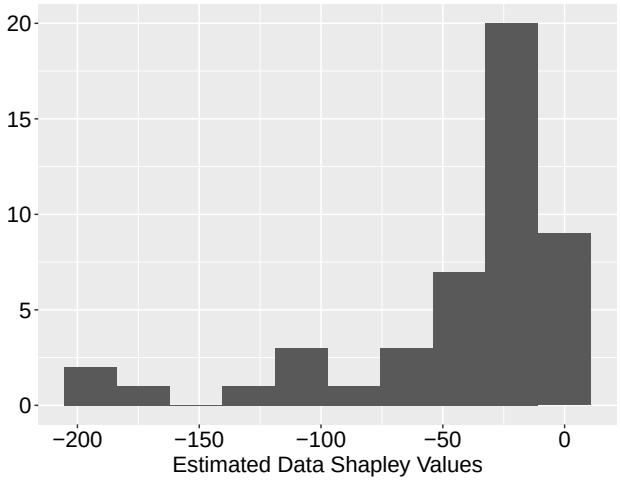
The following aspects are investigated:

- How does different levels of truncation affect the estimation and the estimates
- How does different performance-metrics affect the estimation and the estimates
- How effectively does Data Shapley values capture the value of the observation

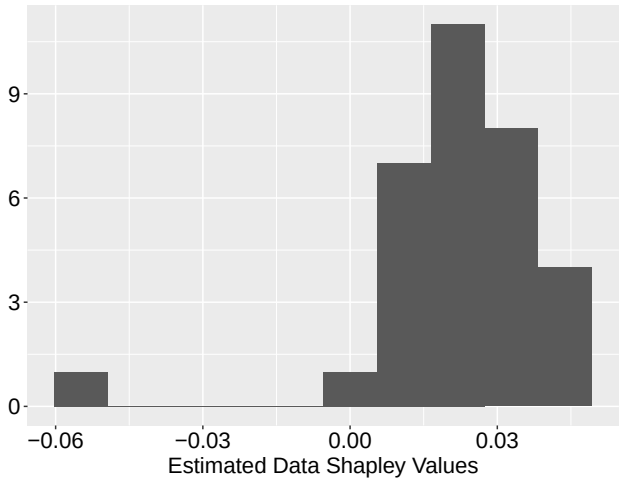
First of all, we can have a look at the most fundamental: what does the distribution of the estimated Data Shapley values look like, when using different performance metrics? To understand this, we can look at histograms for example for two of the four datasets:



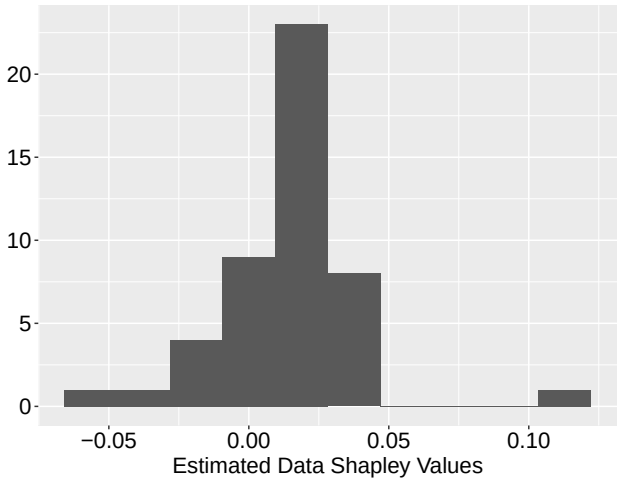
(a) Dataset: mtcars, Metric: Residual Sum of Squares



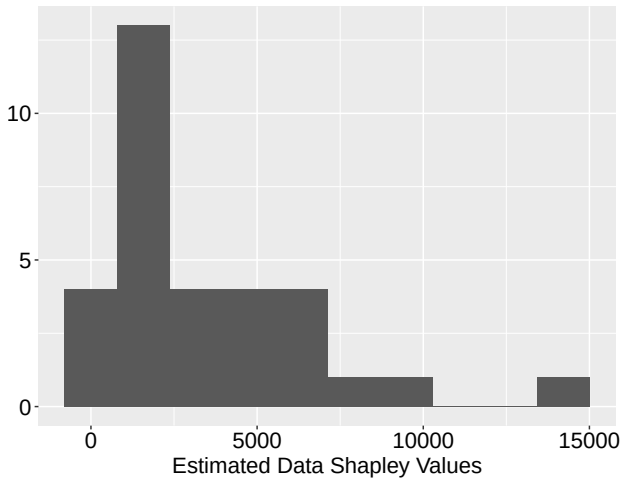
(b) Dataset: Fertility, Metric: Residual Sum of Squares



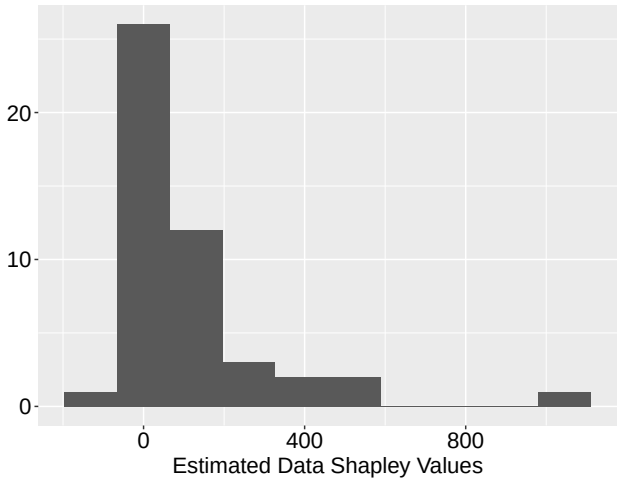
(c) Dataset: mtcars, Metric: Adjusted R Squared



(d) Dataset: Fertility, Metric: Adjusted R Squared



(e) Dataset: mtcars, Metric: Explained Sum of Squares

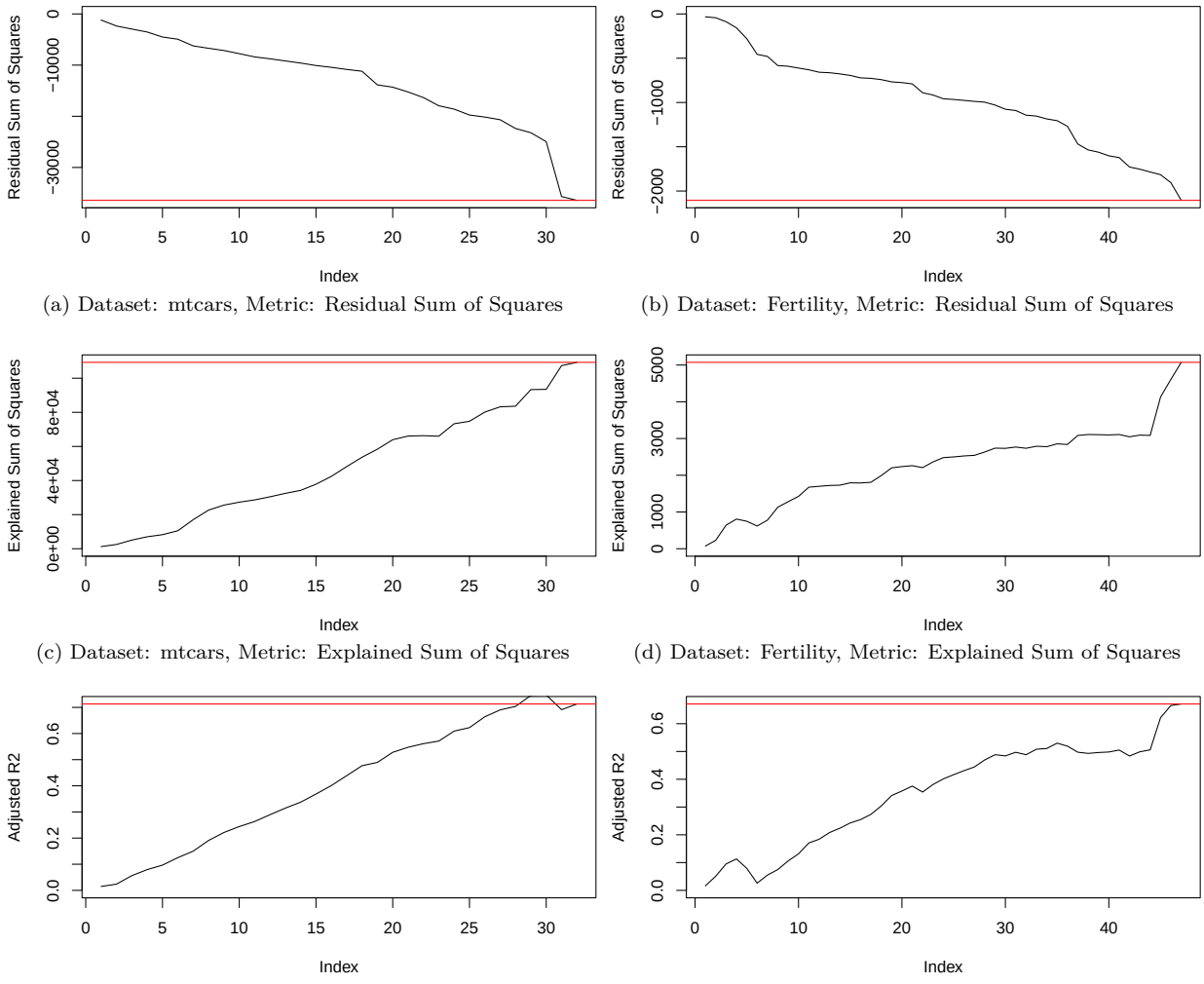


(f) Dataset: Fertility, Metric: Explained Sum of Squares

Figure 5: Histograms for Data Shapley values fit with different Metrics

As can be noted in the histograms in Figures 5(a) to 5(f), Data Shapley values estimated with SSE are strictly positive. Data Shapley values estimated with R^2_{Adj} are often centered just above 0, and with some indications of both symmetry and outliers. Some observations have negative values, which means that they add negatively to the model fit as they are added in.

We can also plot the cumulative sum of estimated (fig Data Shapley values for all configurations, and compare with the corresponding values (SSReg, SSE, Adjusted R^2) for the complete set of observations (the red line shows the performance value for the full set of observations, $V(D)$). As can be seen in Figure 6, the Efficiency property from Shapley values still hold very well.



(a) Dataset: mtcars, Metric: Residual Sum of Squares

(b) Dataset: Fertility, Metric: Residual Sum of Squares

(c) Dataset: mtcars, Metric: Explained Sum of Squares

(d) Dataset: Fertility, Metric: Explained Sum of Squares

(e) Dataset: mtcars, Metric: Adjusted R Squared

(f) Dataset: Fertility, Metric: Adjusted R Squared

Figure 6: Cumulative Sum of Data Shapley values as more and more observations are added

Worth noting, is that when using Adjusted R^2 as metric, the cumulative sum may very well be higher than the final value for the complete set of observations. This is to be expected, based on the fact that negative values are not uncommon in this distribution.

We can also look at the effects of different levels of truncation on the distribution. This can for example be visualised by comparing a denisty plot for the Data Shapley values estimated at a very low level of truncation with the denisty of a medium level of truncation (see Figure 7)

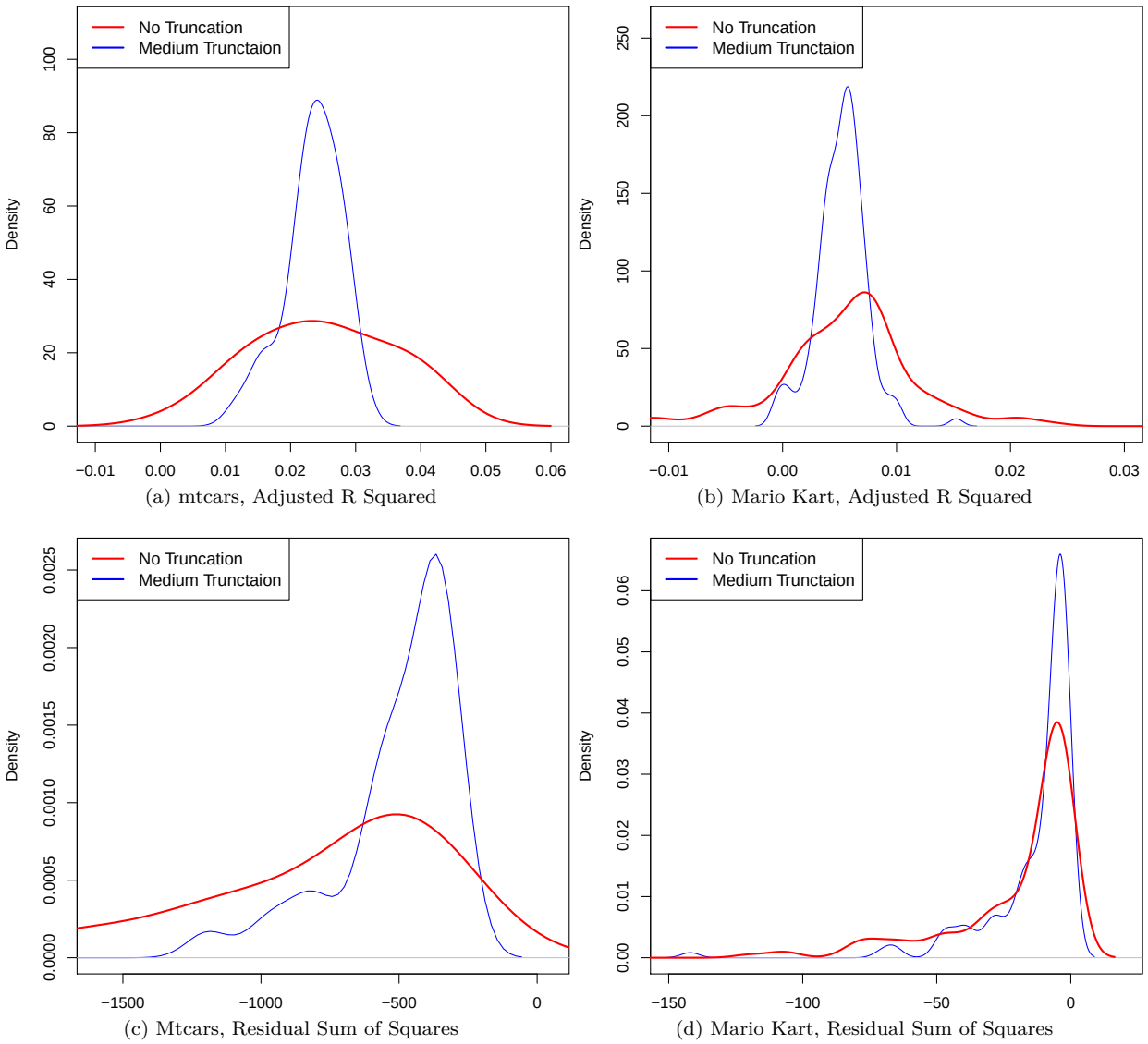
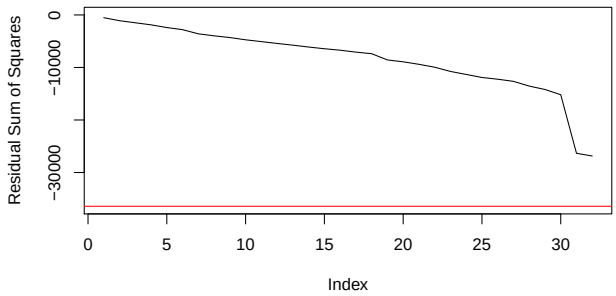


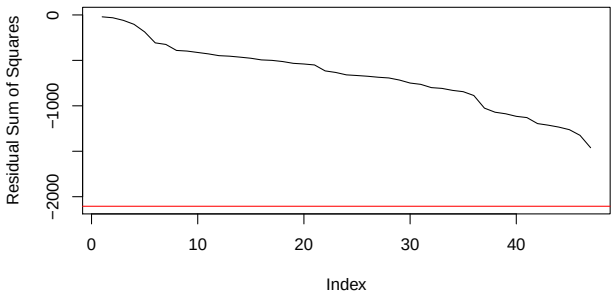
Figure 7: Density plots of Data Shapley values estimated with different levels of truncation

As can be seen in Figure 7, a higher level of truncation actually reduces the variation of the estimated values. This is an interesting finding, which brings us into the topic of truncation and the effect it has on the estimation of Data Shapley values. To study this, we can set the truncation treshold to be a very small value (e.g. 0.0001), and effectively almost disable truncation. We can then estimate the Data Shapley values with different levels of truncation, to see how the estimates vary, by comparing them to the orignial un-truncated estimates.

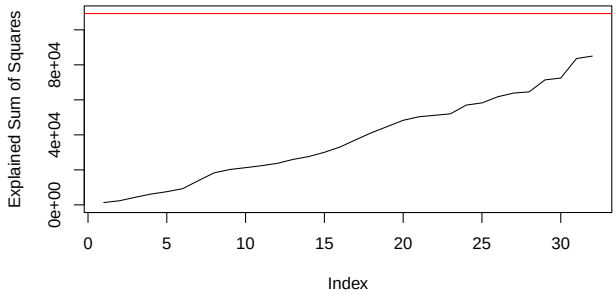
For example, the visualisations above that indicate that the Efficiency property holds, are plotted for the estimates with the lowest level of truncation. However, whether the property still holds for higher levels of truncation, we can also plot the same visuals again for truncated estimates.



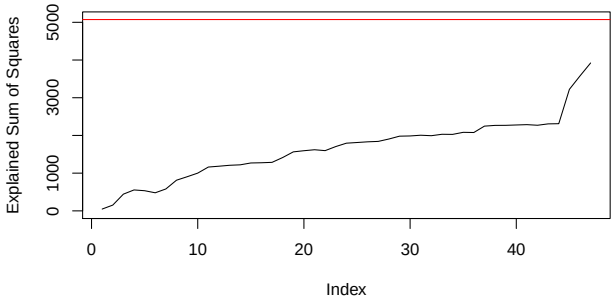
(a) Dataset: mtcars, Metric: Residual Sum of Squares



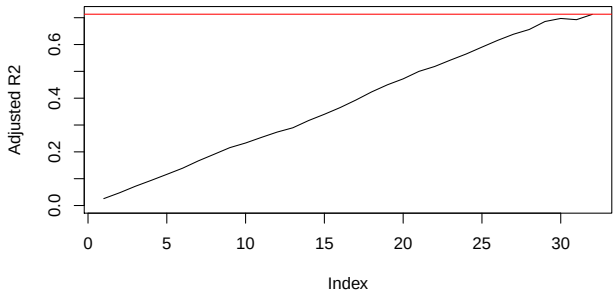
(b) Dataset: Fertility, Metric: Residual Sum of Squares



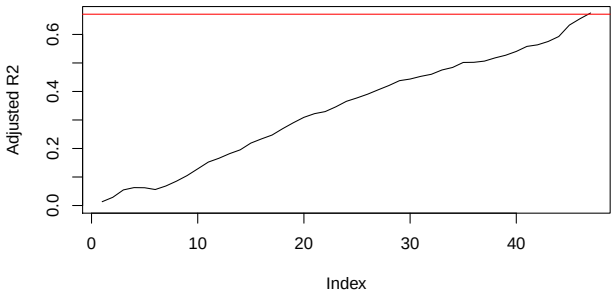
(c) Dataset: mtcars, Metric: Explained Sum of Squares



(d) Dataset: Fertility, Metric: Explained Sum of Squares



(e) Dataset: mtcars, Metric: Adjusted R Squared

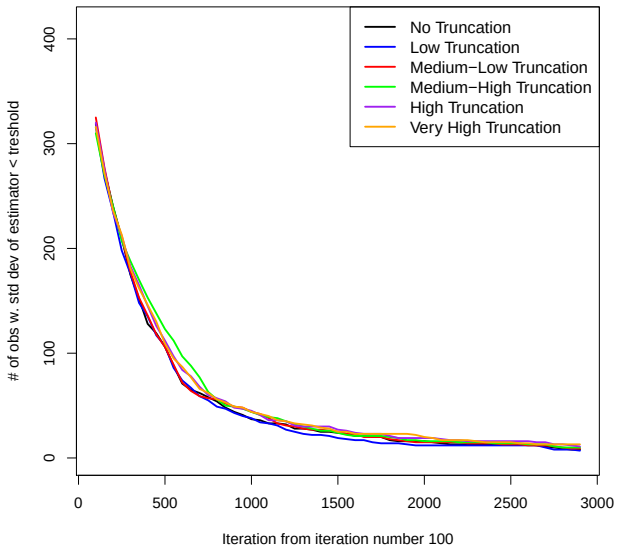


(f) Dataset: Fertility, Metric: Adjusted R Squared

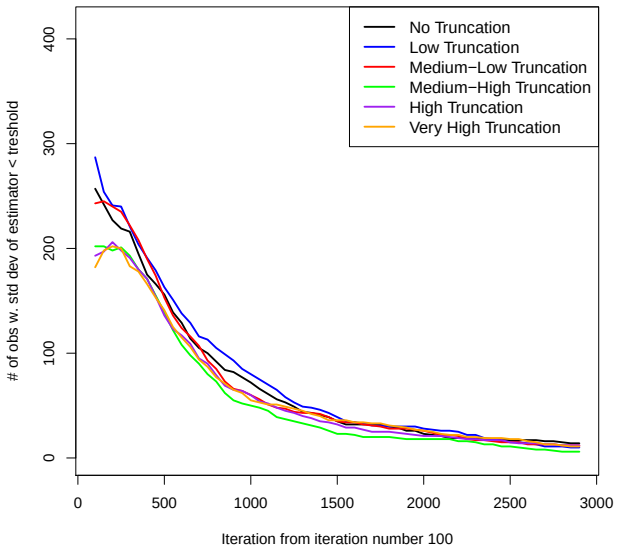
Figure 8: Cumulative Sum of Data Shapley values as more and more observations are added, using estimates from truncated simulation

Here in Figure 8, it turns out that the Efficiency property still seems to hold with Data Shapley values estimated with with adjusted R^2, but seem to break down when truncating Residual Sum of Squares or Explained Sum of Squares. This is certainly a consideration that should be taken into consideration when estimating Data Shapley values. It is also the case that more truncation in the Adjusted R^2 option, seem to reduce the possibility to have a cumulative sum that exceeds that of Adjusted R^2 for all observations. This is logical, since truncation will happen once the estimates are close to the maximum value.

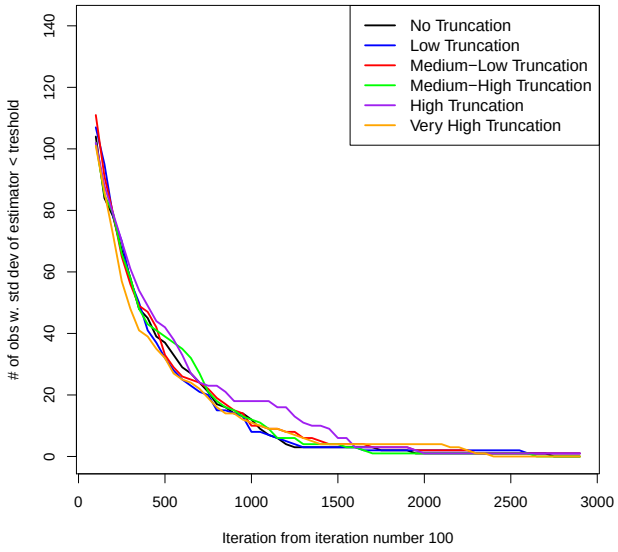
We can also look at the convergence across different levels of truncation. This, by plotting the number of observations that have the $\frac{\sigma}{\sqrt{n}} \leq c$ over the number of iterations. In these plots, the first 100 iterations are omitted (left as "burn-in" due to large variation in the simulations).



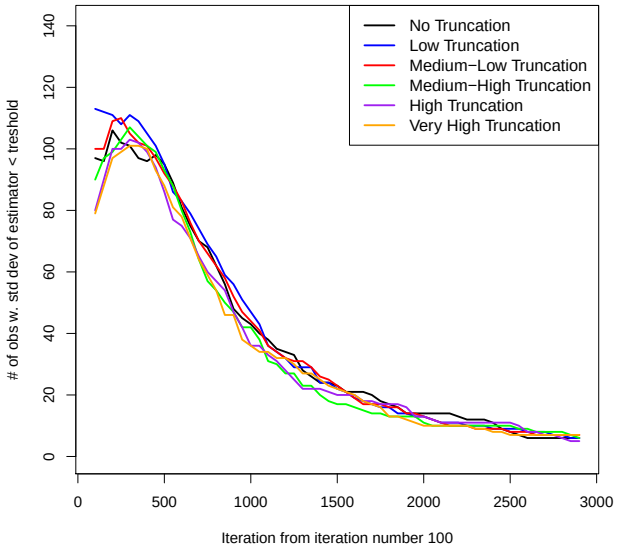
(a) Dataset: Taiwan, Metric: Residual Sum of Squares



(b) Dataset: Taiwan, Metric: Adjusted R Squared



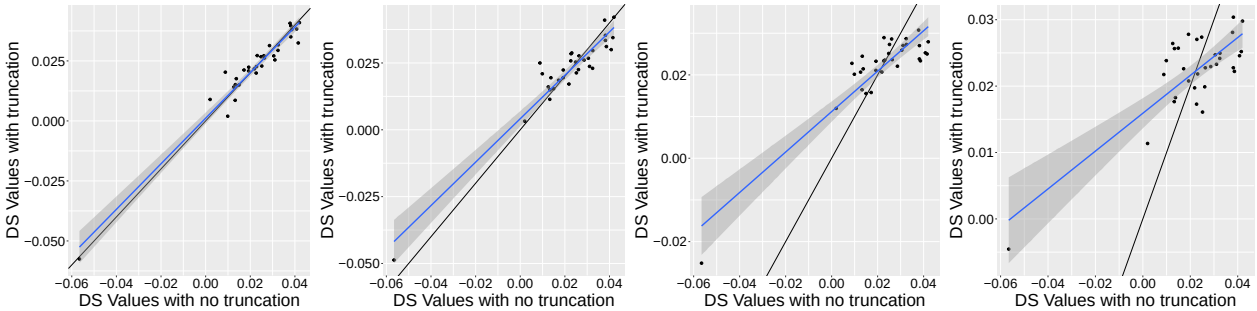
(c) Dataset: Mario, Metric: Residual Sum of Squares



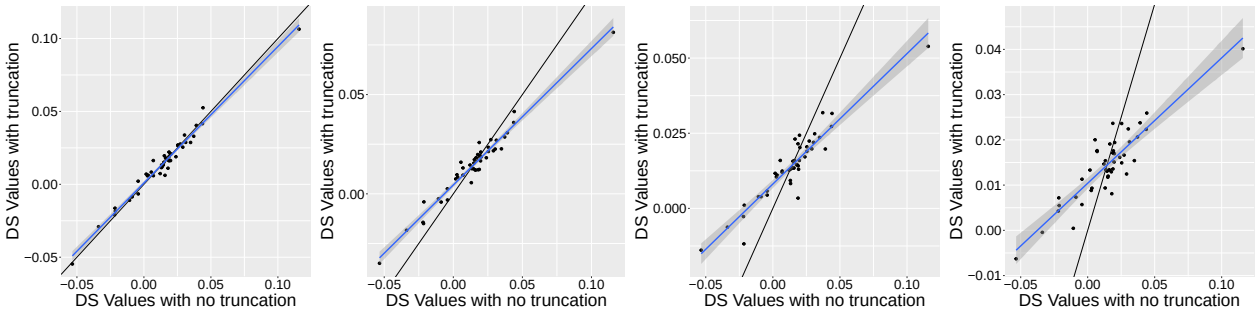
(d) Dataset: Mario, Metric: Adjusted R Squared

Figure 9: Convergence Plots

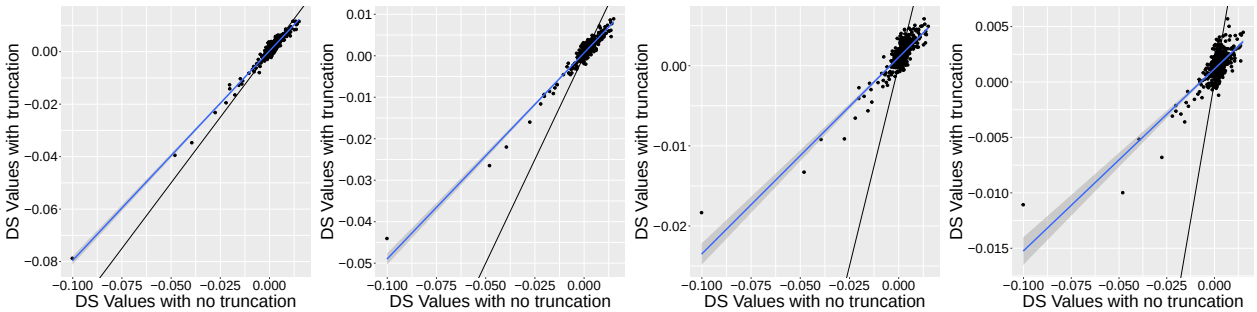
As can be seen in Figure 9, the level of truncation has no significant impact on the convergence. Here, the conclusion would be that truncation is fine and actually preferrable, since the total execution time for the algorithm becomes significantly reduced with more truncation. However, comparing the actual values estimated with very little truncation with values estimated with increasing level of the same, the following pattern emerges (values estimated for the same dataset are on rows, with increased level of truncation the further right one goes):



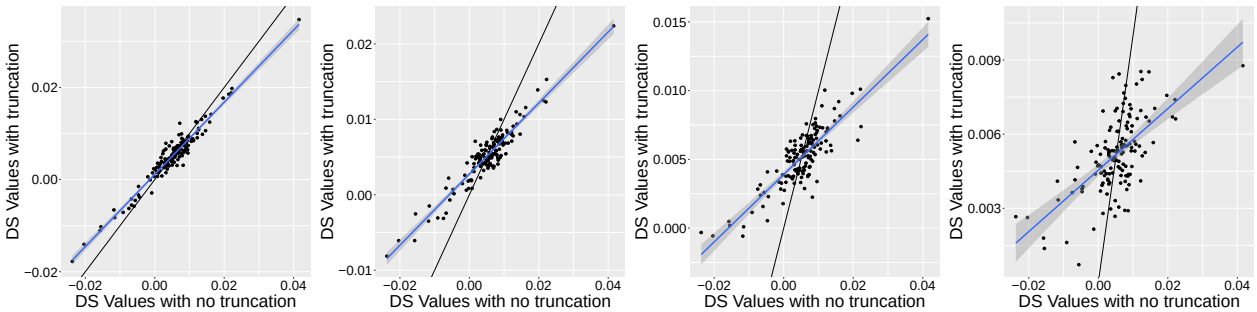
(a) mtcars, Low Trunc. (b) mtcars, Low-Med Trunc. (c) mtcars, Med-High Trunc. (d) mtcars, High Trunc.



(e) Fertility, Low Trunc. (f) Fertility, Low-Med Trunc. (g) Fertility, Med-High Trunc. (h) Fertility, High Trunc.



(i) Taiwan, Low Trunc. (j) Taiwan, Low-Med Trunc. (k) Taiwan, Med-High Trunc. (l) Taiwan, High Trunc.



(m) Mario, Low Trunc. (n) Mario, Low-Med Trunc. (o) Mario, Med-High Trunc. (p) Mario, High Trunc.

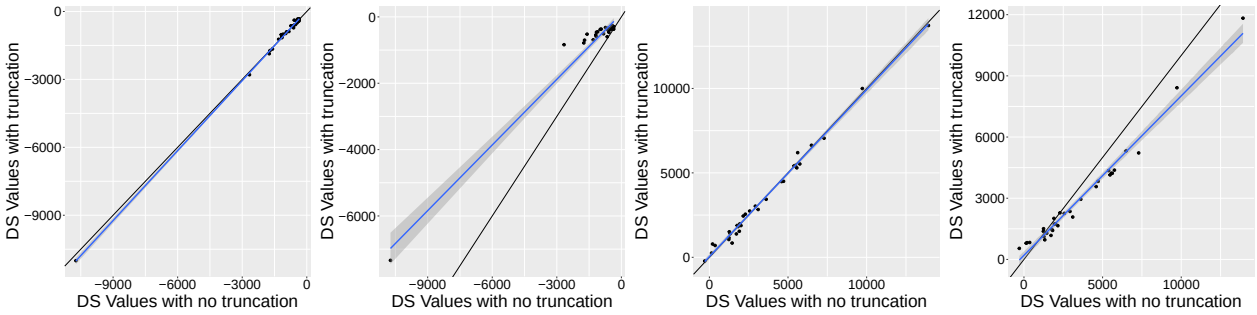
Figure 10: Correlation between untruncated and truncated estimation of Data Shapley values, with increasing level of truncation (Low, Medium/Low, Medium/High and High)

These plots in Figure 10 are presenting values for Data Shapley values estimated with Adjusted R^2. The solid line black represents the x=y line, whereas the blue line represent the least squares estimate of a linear model over the data in each plot.

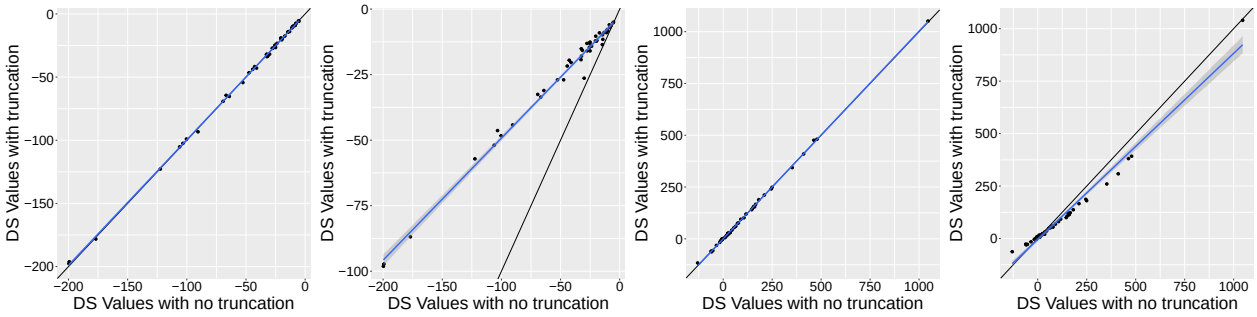
As can be noted in Figure 10, an increase in truncation results in a strong bias in the estimated values. This makes sense because when Data Shapley values are estimated from the generated data, the truncated values will have somwhat of a zero inflated distribution, that will reduce the mean and hence the value estimate. This overdisperion of zero values, could potentially be compensated by a standard b-coefficient in a $y = ax + b$, where b is set based on the level of truncation, but this is left for other researchers to look at.

However, depending on how Data Shapley values are being used in practice, this bias may not be a problem. If for example the value of the observations are only ever to be compared between each other in a particular model specification, then quite a bit more truncation can be allowed, since rank is reasonably well retained. However, if there is a desire to have the Data Shapley value represent something with an actual interpretation, then only very low levels of truncation would be acceptable. For example, a Data Shapley value of 0.05 estimated with low level of truncation with the performance metric set to R^2_{Adj} would mean that this observation has an average impact of raising R^2_{Adj} with 0.05 percentage points everytime it is added in to the model. With truncation that value may instead be for example 0.03, and has then lost much of its original interpretation.

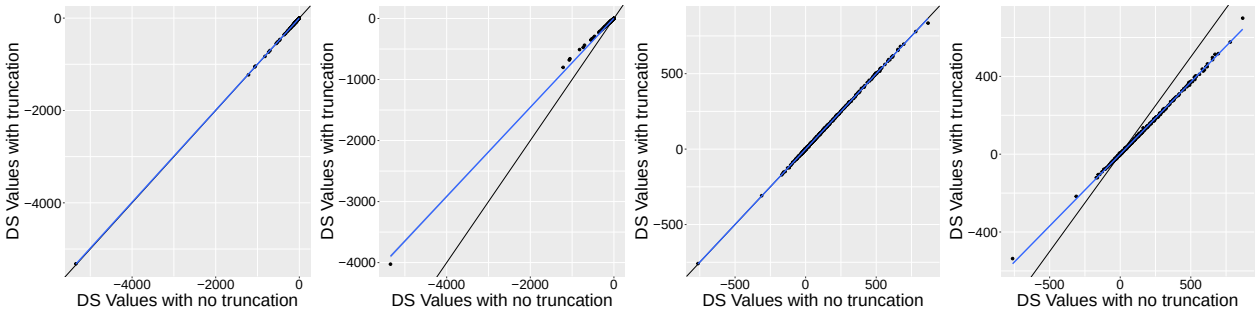
As the plots in Figure 10 above where only for Adjusted R^2, we can also look at the extreme cases (Low and High Truncation for both Residual Sum of Squares and Explained Sum of Squares):



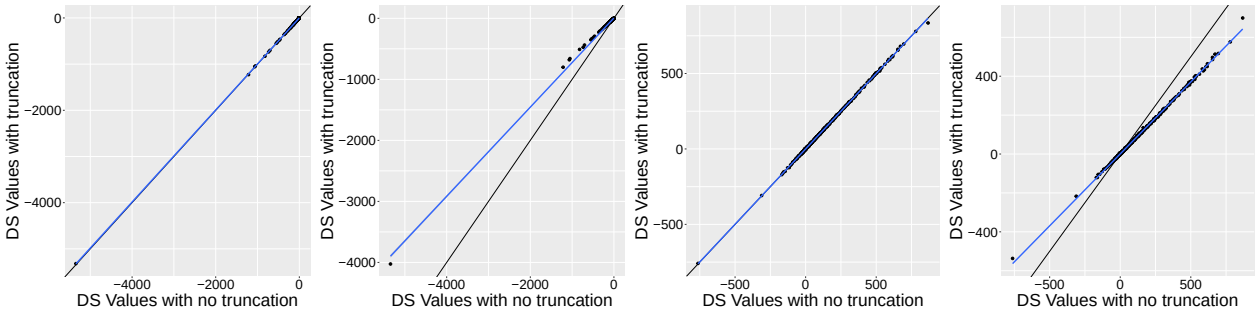
(a) mtcars, SSE, Low Trunc. (b) mtcars, SSE, High Trunc. (c) mtcars, SSReg, Low Trunc. (d) mtcars, SSReg, High Trunc.



(e) Fertility, SSE, Low Trunc. (f) Fertility, SSE, High Trunc. (g) Fertility, SSReg, Low Trunc. (h) Fertility, SSReg, High Trunc.



(i) Taiwan, SSE, Low Trunc. (j) Taiwan, SSE, High Trunc. (k) Taiwan, SSReg, Low Trunc. (l) Taiwan, SSReg, High Trunc.

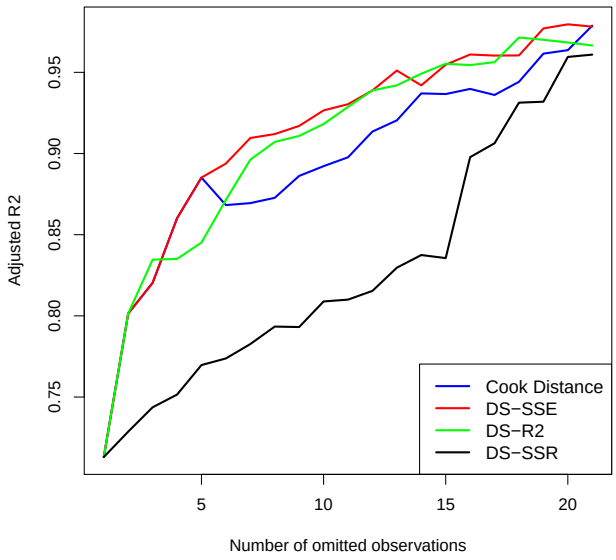


(m) Mario, SSE, Low Trunc. (n) Mario, SSE, High Trunc. (o) Mario, SSReg, Low Trunc. (p) Mario, SSReg, High Trunc.

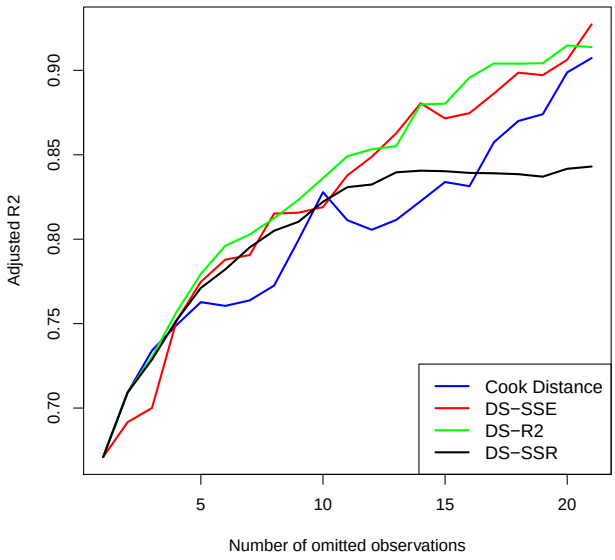
Figure 11: Correlation between untruncated and truncated estimation of Data Shapley values.

And as can be seen in Figure 11, the introduction of bias is visible here as well, although Explained Sum of Squares seem to be a bit more robust than the other two options.

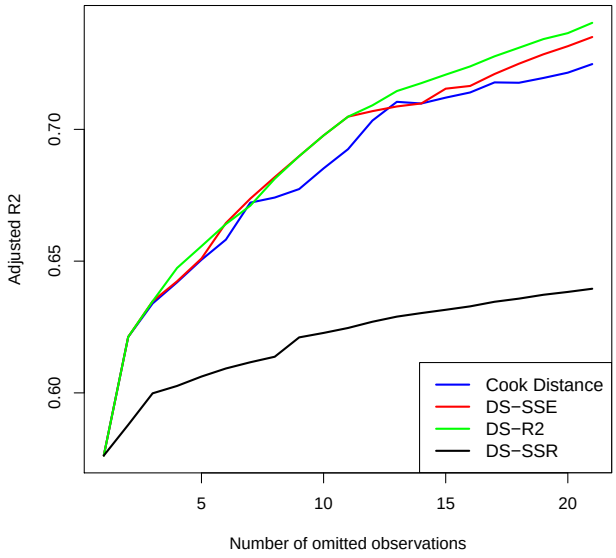
Finally, we can also investigate the value empirically, by refitting the models while gradually removing (i.e. one by one) the 20 observations with the lowest estimated Data Shapley values.



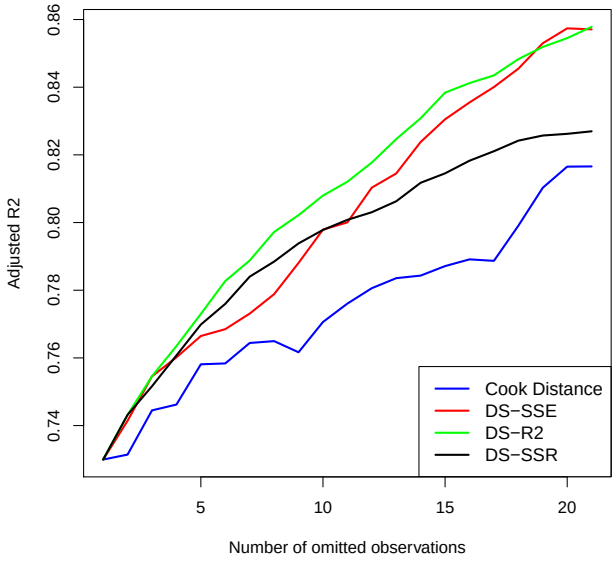
(a) Performance Increase, Mtcars dataset



(b) Performance Increase, Fertility dataset



(c) Performance Increase, Taiwan dataset

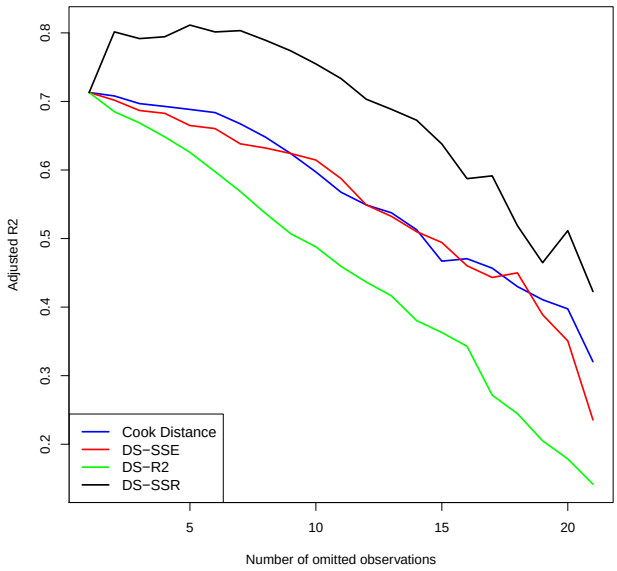


(d) Performance Increase, Mario dataset

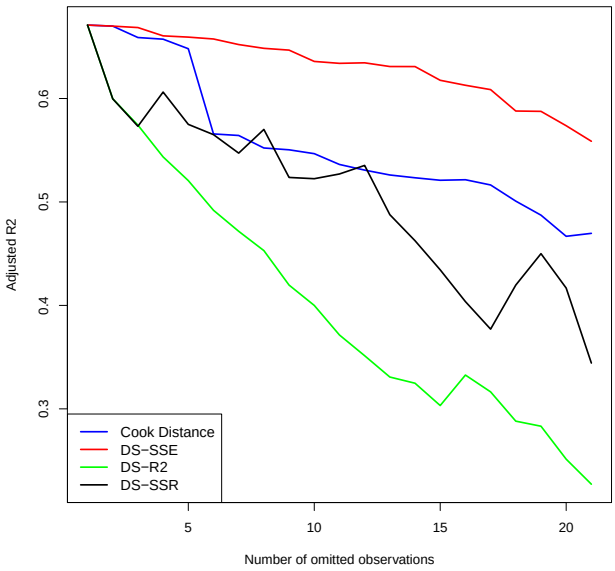
Figure 12: Increase in Adjusted R2 when removing low value observations

In Figure 12, Data Shapley values estimated using Explained Sum of Squares are mostly not performing as well as they could, whereas when Adjusted R Squared or Residual Sum of Squares are used, the Data Shapley value is as good as, or even better than Cooks Distance in capturing the value of the observations.

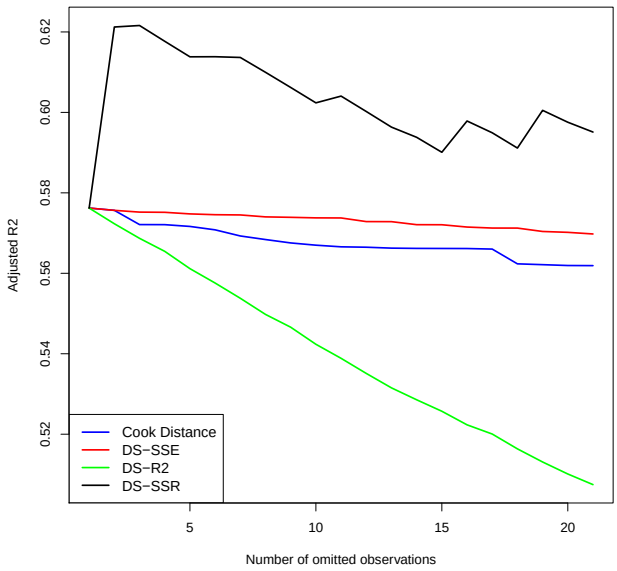
We can of course also look at this from the other direction: How does the performance decrease as high valued observations are removed?



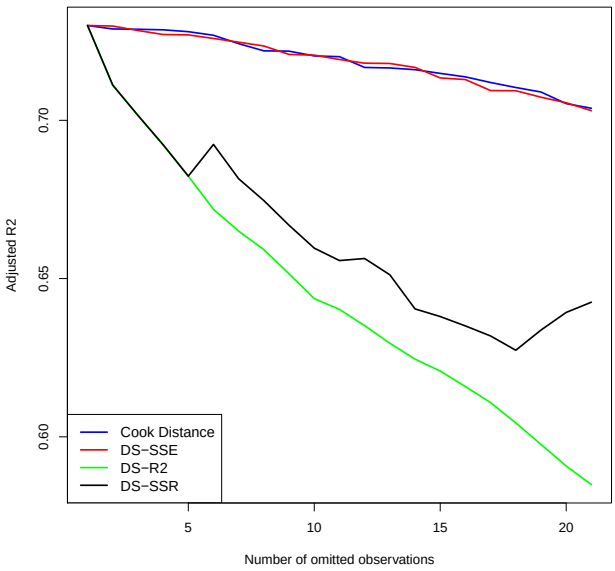
(a) Performance Decrease, Mtcars dataset



(b) Performance Decrease, Fertility dataset



(c) Performance Decrease, Taiwan dataset



(d) Performance Decrease, Mario dataset

Figure 13: Decrease in Adjusted R2 when removing high value observations

Here in Figure 13, when focusing on the high-value observations, it can be noted that Data Shapley Values using Adjusted R^2 is consistently and significantly outperforming Cook's Distance. Using Explained Sum of Squares seem to result in very erratic estimates, whereas using Residual Sum of Squares seem to reasonably well align with Cook's Distance (apart from in the Fertility dataset where one high value observation seems to have been missed).

4.1 Outlier Detection with Data Shapley

Ghorbani and Zou's proposal for Data Shapley includes using it for outlier detection, which the paper also empirically successfully shows is possible. If this is a true property of the Data Shapley values, this may very

well still hold in a regression model. Therefore, we can estimate the Data Shapley values for a dataset that is known to contain outliers. The Mario dataset that has been used previously has two observations that so far has been omitted. These two observations are known to be Ebay-entries for entire Nintendo Wii consoles sold together with the Mario Kart game and not only the game itself. Therefore, these would be outliers due to incorrect measurements, and should probably be omitted.

If again, Data Shapley values are estimated with little to no truncation and using all three metrics, the performance plot looks as follows:

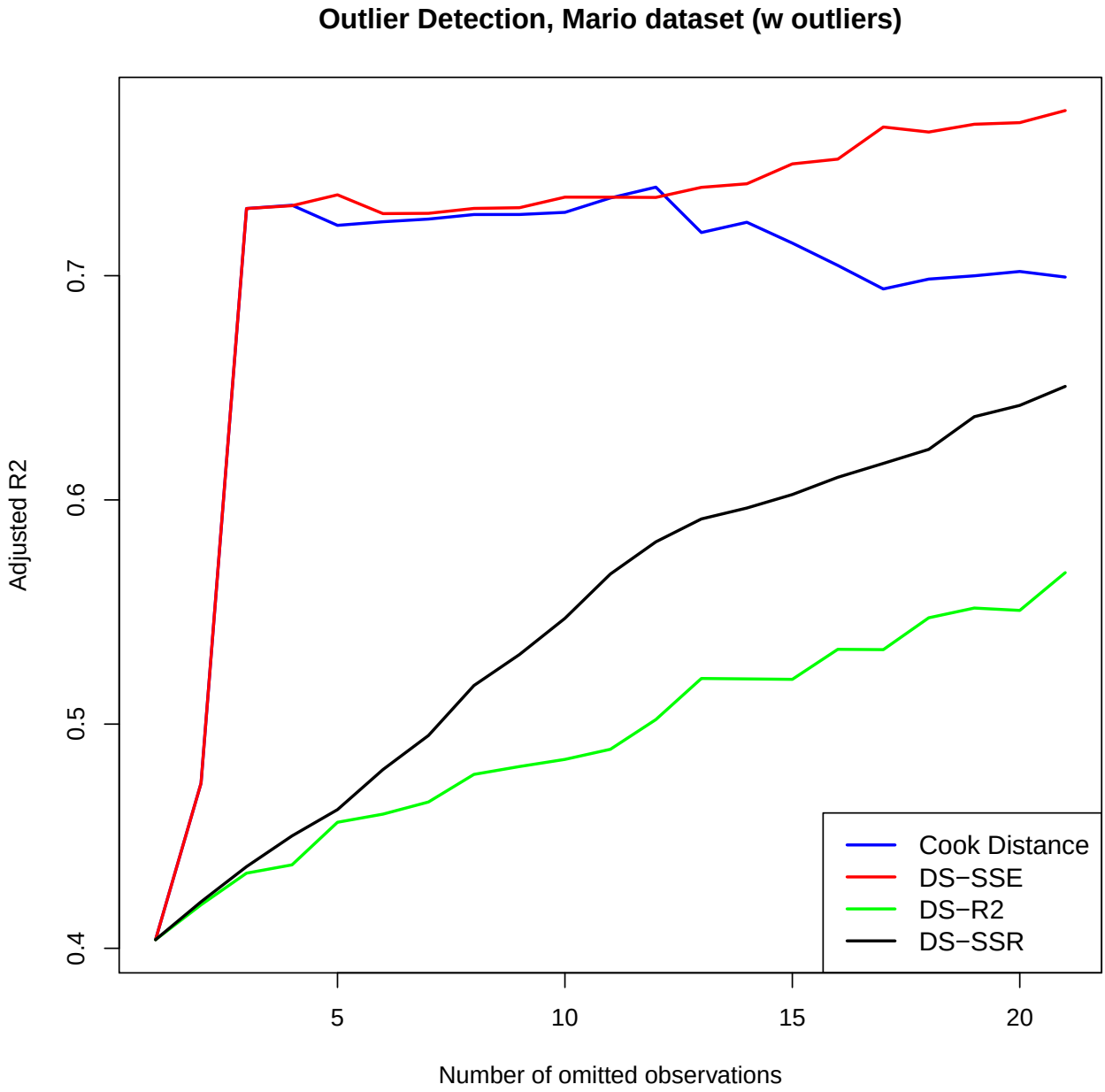


Figure 14: Detecting outliers by gradually removing low valued observations, Mario dataset

Interestingly, neither the R^2_{Adj} nor the Explained Sum of Squares variants seem to capture the information that the two observations are outliers, but the SSE variant certainly matches Cooks Distance in capturing this property.

5 Full Model Diagnostics

This chapter shows the residual plots and presents the model summaries for the four different models and datasets (in their final form with all observations). As can be noted, these are in cases neither parsimonious nor perfect models, and certainly include outliers that have not been dealt with. One example to point at is the deviation from normality of residuals in the two last models, as can be seen in the corresponding QQ-plots. However, this issue is probably not the most critical one, and it does provide a broader set of models for the purposes of this study. All models also include a combination of significant and non-signoficant relationships (at $p < 0.05$).

5.1 Mtcars

Model: hp~vs+cyl+disp+drat

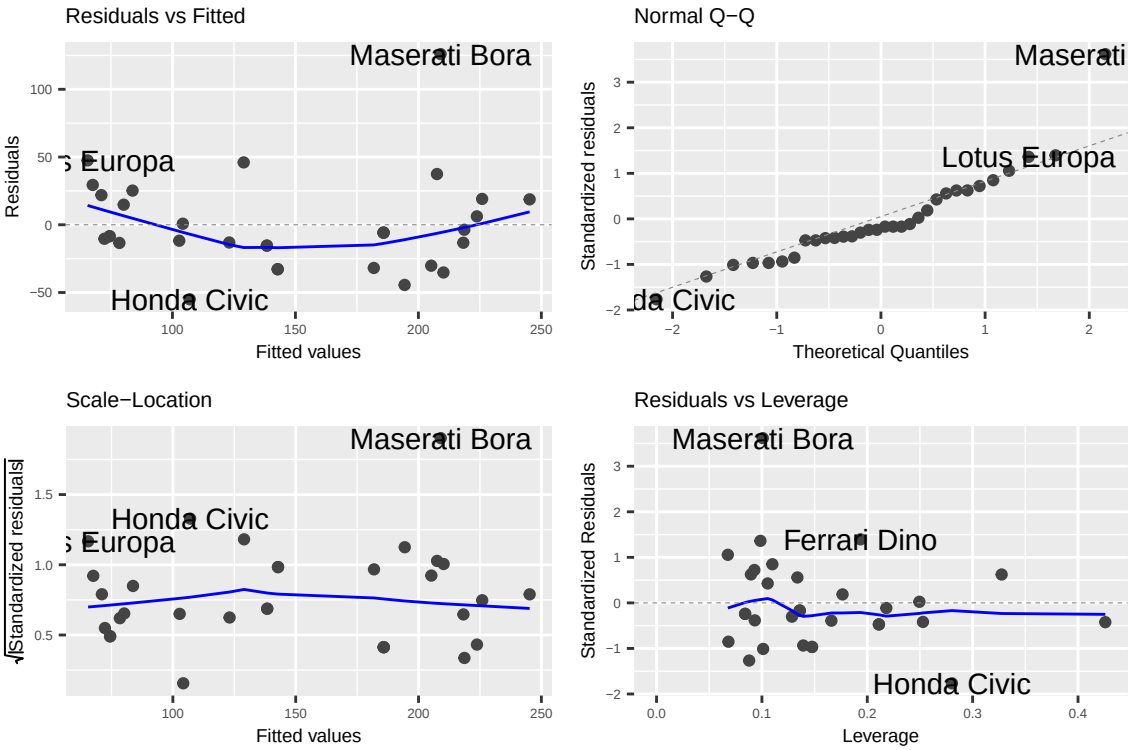


Figure 15: Model Diagnostic Plots, Mtcars

Table 6: Model Estimates using all observations

Term	Coefficient	SE	T-statistic	P-value
(Intercept)	-199.42	112.63	-1.77	0.088
vs	-6.65	23.48	-0.28	0.78
cyl	26.52	10.92	2.43	0.022
disp	0.19	0.13	1.51	0.14
drat	39.02	18.69	2.09	0.046

5.2 Swiss Fertility

Model: Fertility~.

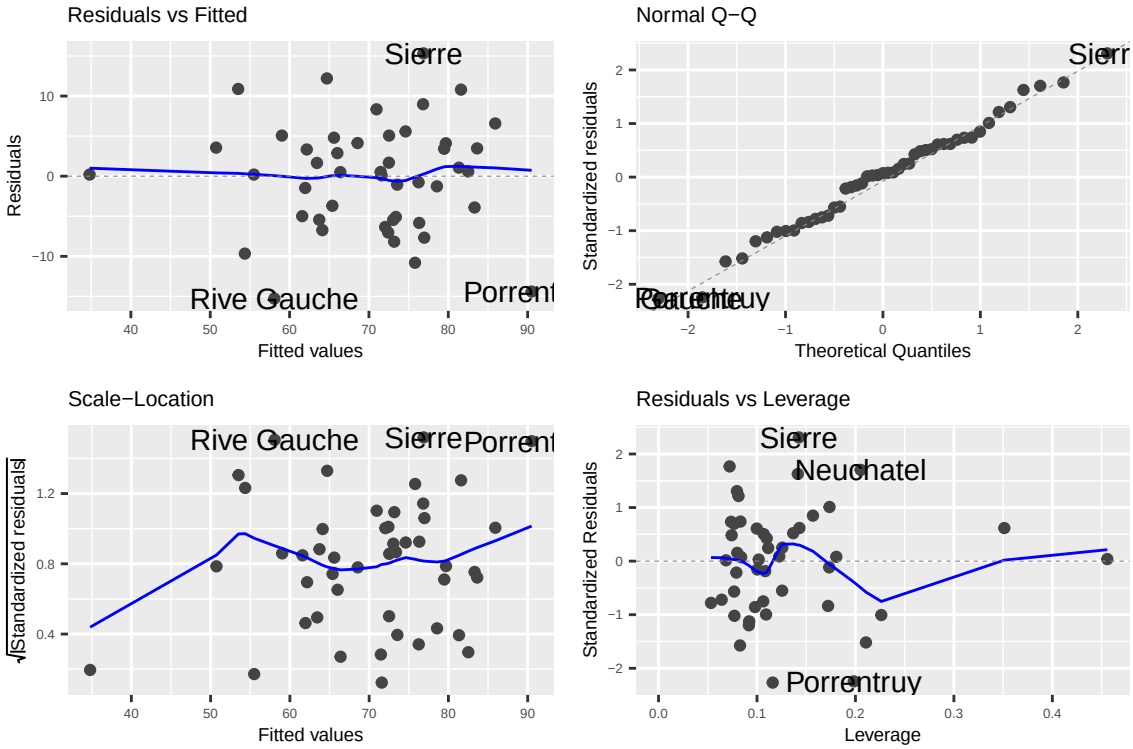


Figure 16: Model Diagnostic Plots, Fertility

Table 7: Model Estimates using all observations

Term	Coefficient	SE	T-statistic	P-value
(Intercept)	66.92	10.71	6.25	< 0.001
Agriculture	-0.17	0.07	-2.45	0.019
Examination	-0.26	0.25	-1.02	0.32
Education	-0.87	0.18	-4.76	< 0.001
Catholic	0.1	0.04	2.95	0.005
Infant.Mortality	1.08	0.38	2.82	0.007

5.3 Taiwan Houseprices

Model: HousePrice~.

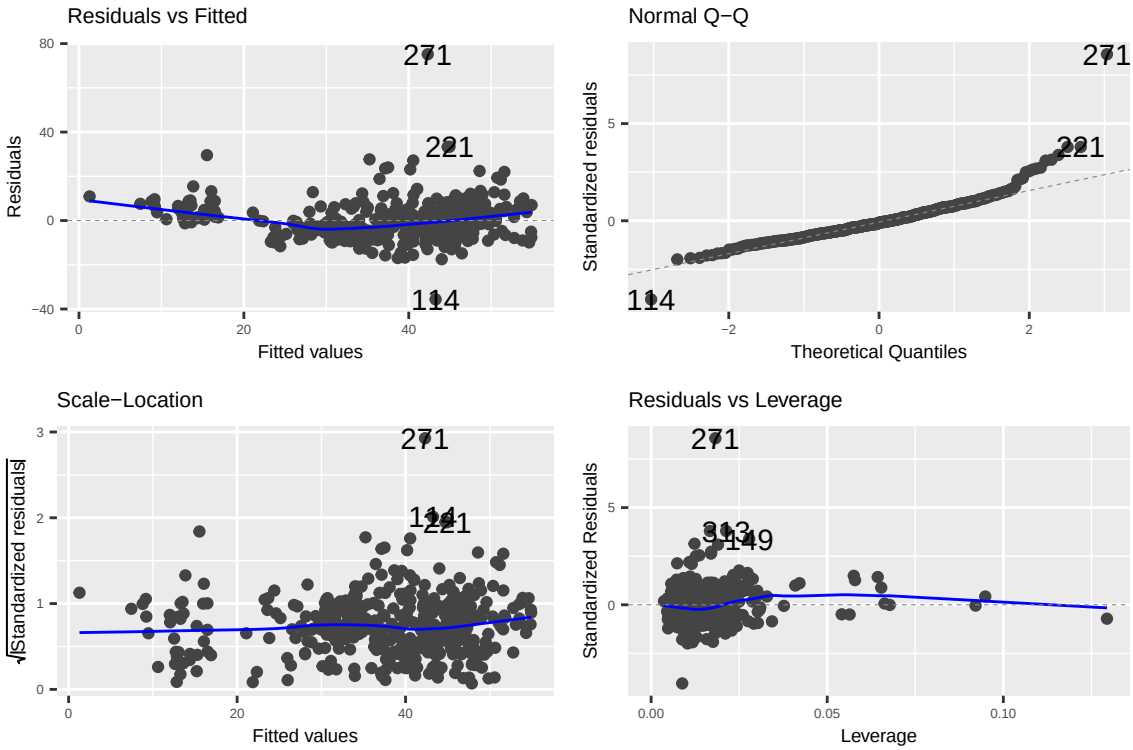


Figure 17: Model Diagnostic Plots, Taiwan

Table 8: Model Estimates using all observations

Term	Coefficient	SE	T-statistic	P-value
(Intercept)	-14437.1	6775.67	-2.13	0.034
TransactionDate	5.15	1.56	3.31	0.001
HouseAge	-0.27	0.04	-7	< 0.001
DistanceToMRT	0	0	-6.25	< 0.001
ConvenienceStores	1.13	0.19	6.02	< 0.001
Latitude	225.47	44.57	5.06	< 0.001
Longitude	-12.42	48.58	-0.26	0.8

5.4 Mario Kart Prices

Model: total_pr ~ .

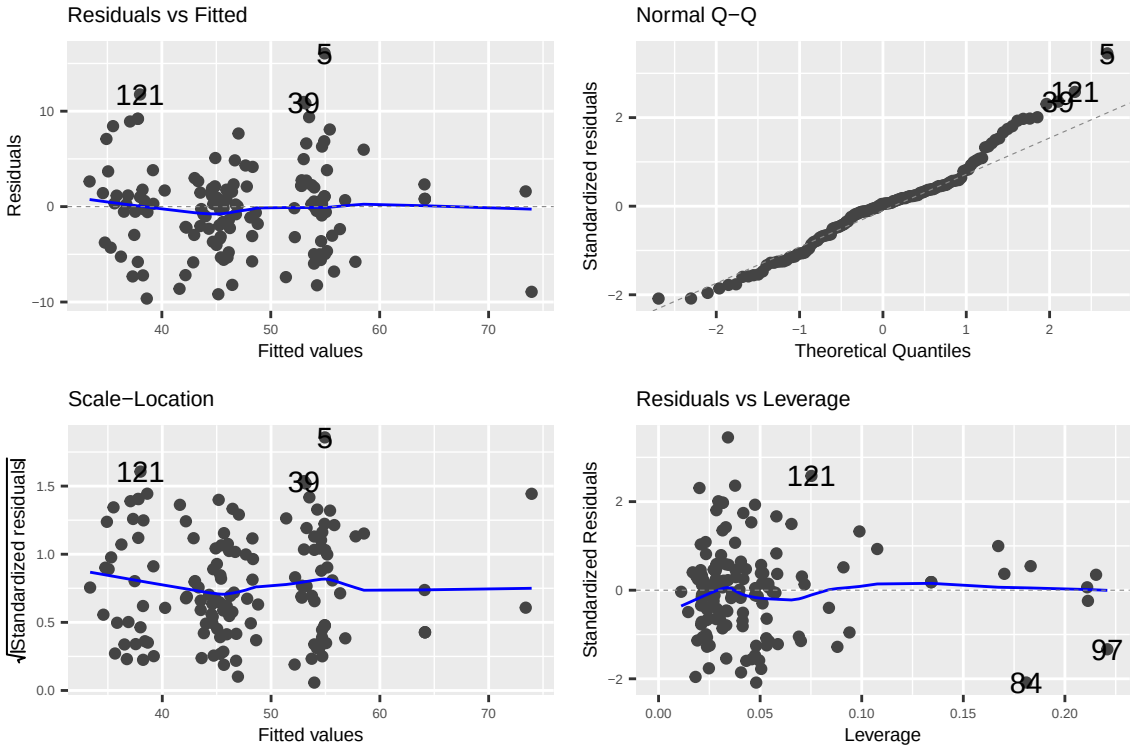


Figure 18: Model Diagnostic Plots, Mario

Table 9: Model Estimates using all observations

Term	Coefficient	SE	T-statistic	P-value
(Intercept)	34.83	1.8	19.36	< 0.001
duration	-0.53	0.18	-2.94	0.004
n_bids	0.24	0.09	2.6	0.011
start_pr	0.17	0.04	4.48	< 0.001
ship_pr	0.14	0.16	0.91	0.37
seller_rate	0	0	2.67	0.008
wheels	7.94	0.54	14.65	< 0.001

6 Conclusions and Summary

Looking at the empirical results, it seems a Data Shapley value estimated with Adjusted R^2, generally outperforms Cook's Distance, both in identifying low value observations and in identifying the high value ones, in the case of small to moderate outliers.

If the objective is to detect large outliers or extreme values, one should probably instead use Residual Sum of Squares i the TMC algorithm, but more experimentation is most likely required for better support for that decision. However, seeing that the Data Shapley value in this case is pretty much on par with Cook's Distance, it is perhaps difficult to argue for a method as resource intensive and complex as this.

It is important to understand that the conclusions from this study are somewhat dependent on the interpretation and "objective" of Cook's Distance. Cook himself motivates the Cook's Distance measure as *"an easily interpretable measure that [...] will naturally isolate 'critical' values"* (Cook 1977). In Cook and Weisberg (1982), the motivation is a little richer:

> *"The ability to find influential cases can benefit the analyst in at least two ways. First, the study of influence yields information concerning reliability of conclusions and their dependence on the assumed model. [...] Second, we shall see that cases in the p-dimensional observation space that are far removed from other cases will tend to have, on the average, a relatively large influence on the analysis. This, in turn, may indicate areas in the observation space with inadequate coverage for reliable estimation and prediction."*

Empirically, Data Shapley values and Cook's Distance seem to capture the same inherent property of an observation, although it is referred to Value or Worth in one case and Influence in the other. One could of course argue that these two are not the same thing, and that it therefore does not make much sense comparing them. Even so, consider this study to have been to establish a "Shapley Influence" (instead of exploring "Data Shapley"), and the results are still both valid and relevant.

Also, as a short reflection on the "Value" or "Worth" of an observation in a set of data: It is a very challenging problem. Both Data Shapley values and Cook's Distance assumes that an "agreement" with the general pattern in the data is something good, and of course quite often this will be the case. There is however no guarantee for this. If, due to poor experiment design, poor measurements, challenging phenomena being measured or something completely different, some observations are of "higher true value", these may not be captured correctly. Let us for example consider a scenario that Ghorbani and Zou (2019) specifically reference: A setup where multiple parties are involved in providing subsets of data for a common estimation of a model. If in this case, one party were to do measurements with higher quality and smaller bias but at the cost of being able to provide fewer than other involved parties, these could potentially be classed as "non-conforming" and get a low Data Shapley value and a high Cook's Distance.

To summarize:

We are no closer to bringing automated insights into the connection between "Influence" and some "True Value" of an observation, but the Data Shapley method, gives a more accurate measurement of the influence of an observation.

References

Aumann, R. J. (1994), "Data shapley," in *Game-theoretic methods in general equilibrium analysis*, Dordrecht: Kluwer Academic Publishers.

Cook, R. D. (1977), "Detection of Influential Observation in Linear Regression," *Technometrics*, 19, 15–18.

Cook, R. D., and Weisberg, S. (1982), *Residuals and influence in regression*, London: Chapman & Hall.

Ghorbani, A., and Zou, J. (2019), "Data Shapley: Equitable Valuation of Data for Machine Learning," *https://arxiv.org/pdf/1904.02868.pdf*.

Harrison, R. L. (2010), "Introduction to Monte Carlo Simulation," *AIP Conference Proceedings*, 1204, 17–21. https://doi.org/10.1063/1.3295638.

Henderson, H. V., and Velleman, P. F. (1981), "Building multiple regression models interactively," *Biometrics*, 37, 391–411.

Kruskal, W. (1987), "Relative Importance by Averaging over Orderings," *The American Statistician*, 41, 6–10.

Lindeman, Richard Harold, Merenda, Peter Francis, and Gold, R. Z. (1980), *Introduction to bivariate and multivariate analysis*, Glenview, Ill.: Scott, Foresman.

Lipovetsky, S., and Conklin, M. (2001), "Analysis of regression in game theory approach," *Applied Stochastic Models in Business and Industry*, 17, 319–330.

Lundberg, S. M., and Lee, S.-I. (2017), "A unified approach to interpreting model predictions," in *Advances in neural information processing systems 30*, eds. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Curran Associates, Inc., pp. 4765–4774.

Mosteller, F., and Tukey, J. (1977), *Data analysis and regression: A second course in statistics*, Addison-Wesley series in behavioral science, Addison-Wesley Publishing Company.

openintro.org (2019), "MarioKart ebay auctions dataset," *Available at: https://www.openintro.org/stat/data.php ?data=mariokart*, Accessed 2019–11–01.

Pal, M., and Bharati, P. (2019), "Relative contribution of regressors," in *Applications of regression techniques*, Singapore: Springer Singapore, pp. 155–169.

R Core Team (2019), *R: A language and environment for statistical computing*, Vienna, Austria: R Foundation for Statistical Computing.

Robert, C., and Casella, G. (2010), *Introducing Monte Carlo Methods with R*, New York: Springer New York.

Shapley, L. S. (1953), "A value for n-person games," *Contributions to the Theory of Games*, 2, 307–317.

Sheather, S. (2009), *A Modern Approach to Regression with R*, Springer texts in statistics, New York: Springer New York.

Yeh, I.-C., and Hsu, T.-K. (2018), "Building real estate valuation models with comparative approach through case-based reasoning," *Applied Soft Computing*, 65.