ST 542
Final Project

Aaron Brake
Robert West

June 13, 2021

Contents

1 Introduction

The client for this project is Dr. Robert Hayes. He is an Associate Professor of Nuclear Engineering at NCSU. His research mostly involves the subjects of Retrospective Dosimetry and Nuclear Assay. Professor Hayes pitched this project looking for a theoretically rigorous justification for a statistical method he has used in prior publications.

1.1 Research Background

1.1.1 Nuclear Forensics and Bricks

In the field of accident analysis and nuclear forensics, it is of vital importance to determine the extent to which an area has been exposed to potentially harmful radiation. Traditional techniques are very accurate, however, they are time consuming and expensive. Two proposed methods have been developed to address both of those issues: an experimentalist method, and a novel statistical analysis method. The first method involves using standard, untreated bricks as the capture medium instead of standard expensive material. Usually, the capture method must be specially treated, which again adds to the cost and time of the analysis. Additionally, analysis of brick material thus far has involved complicated chemical treatment to isolate the quartz located in the brick. Experimentalists have developed a method to measure the energy resolution of the radiation source coming from untreated brick material, using thermoluminescence and/or optically stimulated luminescence [1].

In an attempt to estimate, model, and predict the dose of exposure, a Monte-Carlo N-Particle (MCNP) model was built on photos taken of bricks. The MCNP is the industry standard method and code for modeling and analyzing the transport of neutrons and gamma rays using the Monte Carlo method [2]. Because we are unable to employ this model, we do not provide a more robust explanation. We simply introduce this model to present a comprehensive review of the previous work conducted.

The core of the research being conducted is to determine if dose deposition profiles can be sufficiently determined with untreated brick material. Along with using traditional lab equipment to measure these profiles, modeling is done. An analysis method was developed by Professor Hayes to determine accuracy and stability of these models. As stated above, the primary goal of this project is to justify this novel analysis method.

1.1.2 Novel Statistical Method Under Consideration

Assume a setting with data $\{Y,x\}_{i=1}^{n}$, and some true model of the data $Y = f(x)$. The fit model will be denoted as $\widehat{f(x)}$. Recall, residuals for any supervised model are defined as $e_i = Y_i - \hat{Y}_i$, where $\hat{Y}$ are the resulting estimated values produced from a fit model (ie. $\hat{Y} = \widehat{f(x)}$). The Residual Sum of Squares (RSS) is defined in equation (1)

$$RSS \equiv \sum_{i=1}^{n} e_i^2 \tag{1}$$

The plot of RSS as a function of e_i is quadratic by definition, and will open upwards. If this parabola were inverted to open downwards, it is claimed this graph looks like a normal distribution. This inverted parabola is now fit to the normal distribution using a non-linear least squares algorithm.

An applied example can be seen in [1]. This example uses the MCNP model presented above. The RSS vs Perturbation plot can be seen in Figure 1. Note these RSS values do not follow a parabola exactly, which we attempt to recreate. Further, notice the minimum of the RSS is not at the estimated values, but instead at a perturbation of +2. The data are then inverted and fit to a normal non-linear least squares regression; Figure 2 shows this fit. After this normal model is fit, it is used to make claims about the original data and MCNP model. The propose of this normal fit is to determine the optimal energy estimate (i.e. estimate of the parameter) but also an uncertainty estimate of this parameter [3].

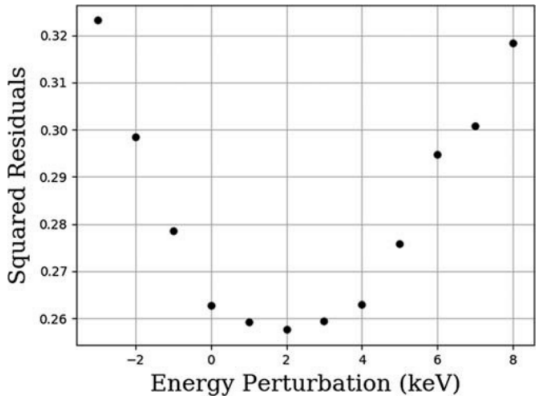

Figure 1: Plot of the SSR vs the Energy Perturbation, which is the change in the parameter estimates. Zero is the original estimated values [1].

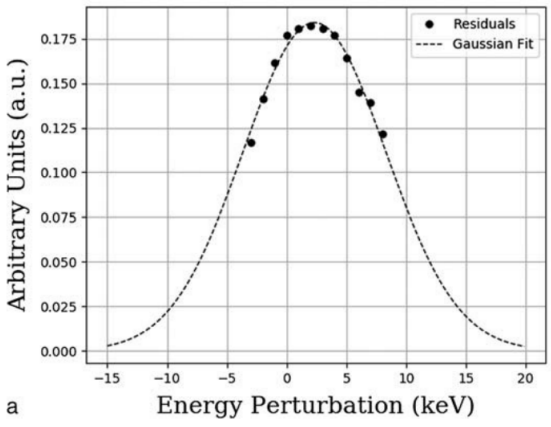

Figure 2: Plot of the fit normal model to the SSR curve in Figure 1 [1].

2 Project Goals

There are a few open-ended questions in the developed method that need to be investigated. We will do this by breaking this proposed method into smaller parts, and testing each individually and separately. This way, we can critique each part of the method, and hopefully provide some suggestions or alterations to the method. We do not believe it possible to actually provide much statistical theory to the foundations of this method, but we can at least test this method's limitations. Since it is not possible for us to use any actual data for this analysis, all data used will be simulated.

3 Methods

We note this project seems to have a special structure we can utilize to help the analysis. This method has two distinct, disjoint parts: an investigation to examine the pattern in RSS when parameter estimates are subject to perturbations, and fitting quadratic data with probability distributions. Thus, we will investigate these two parts independently, where the results from one part will not influence the other part.

3.1 RSS Perturb Method

Data was simulated form the normal distribution, with 3 predictors. The data generating process is as follows: $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, where $X_i \sim N(0,1)$, and $\sim N(0,1)$. Note: the true beta parameters are not of concern since we are not doing any testing of the estimated values. First, we answer the question about RSS being quadratic. Second, we will attempt to replicate similar results to Figure 1. To investigate these two claims, we will investigate multiple different perturbation types, and two different model types.

The two model types are Ordinary Least Squares and Bayesian estimation. The Bayesian model used the following framework, with uninformative priors:

$$\begin{aligned} Y|\boldsymbol{X},\boldsymbol{\beta} &\sim Normal(\boldsymbol{X}\boldsymbol{\beta},\sigma^2) \\ \beta_i &\sim Normal(0,100) \\ \sigma^2 &\sim InvGamma(0.1,0.1) \end{aligned} \tag{2}$$

The OLS estimators were calculated using the lm() command in **R** [4]. The Bayesian estimators were calculated using the *rjags* package in **R** [4]. The Bayesian Model is included since it involves Monte Carlo sampling, which is present in the MCNP model. After obtaining estimates of beta, we will use different methods for perturbing this estimate, and plot the

resulting RSS. The perturbations will be done simultaneously to each parameter. Table 1 gives the different perturbation types.

Perturbation Type
Equally Split
Unequally Split
Random Uniform

Table 1: Perturbation Types

Both the equally and unequally split perturbation are known constants. The Uniform perturbations are random.

3.2 Distributional Fits Methods

Quadratic data was generated using the following steps:

1. $X_i \in (-1,1)$, where the X_i's are evenly spaced, and $i = 1,2,\ldots,9$
2. $\boldsymbol{Y} = -\boldsymbol{X}^2$
3. $\boldsymbol{Y} = \boldsymbol{Y} - min(\boldsymbol{Y})$
4. $\boldsymbol{Y} = 5 * \boldsymbol{Y} / sum(\boldsymbol{Y})$

The second step assures the data is a quadratic opening downward. The third and fourth steps are done to assure the data follows the rules of the probability distributions; namely $Y > 0$. We also set bounds $Y \in (0,1)$ to achieve better convergence. After this, $\boldsymbol{Y}$ was multiplied my 5, which was the largest value that would still keep the $\boldsymbol{Y}$ strictly less than 1, but larger than the small values provided by the standardization process. Having $\boldsymbol{Y}$ values less than one helped with the fits, but without the multiplication, the fits did not converge.

If the $\boldsymbol{Y}$ data were not standardized using the method proposed, a constant c could have been added to each distribution. This was avoided for two reasons. First, adding extraneous, non-essential variables to estimate is poor practice. By allowing the distributions to arbitrarily shift upwards or downwards, the variance of the estimated distribution could be over/under estimated at random by this new constant. Second, the constant would not longer make the distributions proper by statistical definitions, which could lead to issues with estimation of the final variance and mean. So, standardization made the fits more accurate to the underlying distribution shape.

Notice no variation was added to this data. Using the nls() function in **R** [4], which fits non-linear least squares models, different probability density functions were fit to the data. Table 2 shows all the distributional fits calculated. For distributions that have strictly positive support, X was transformed to positive values by $X_{pos} = X + 1$. For distributions that have support $x \in [0,1]$ (ie. Beta distribution), X was transformed using $X_{0,1} = X_{pos}/2$. Two

measures of fit will be used to determine the "best" fits: $Cor(Y, \hat{Y})$, and RSS for each fit. The distributions in Table 2 used the regular parameterization in **R**.

Distribution	Support
Normal	Real
Cauchy	Real
Logistic	Real
Gamma	Positive
Exponential	Positive
χ^2	Positive
Log-Normal	Positive
Weibull	Positive
Beta	$[0,1]$

Table 2: Distribution fits used, and their supports

4 Results

4.1 Investigation of RSS

4.1.1 Residuals Theory

Assume data takes the form $\{X_i, Y_i\}$, and a true parametric relationship $Y = f(X)$, with parameter vector $\boldsymbol{\beta}$. The estimated relationship is denoted $\hat{Y} = \widehat{f(X)}$, and the estimated parameter vector is $\hat{\boldsymbol{\beta}}$. Let $p \in R^n$. Then, perturbations are defined as $p + \hat{\boldsymbol{\beta}}$, which results in a new estimate of Y, $\hat{Y}_{ptb} = \widehat{f(X)}_{ptb}$. When $p^T = (0, \ldots, 0)$, this would return the usual estimates. RSS can be defined as:

$$RSS_{ptb} = (Y - \widehat{f(X)}_{ptb})^T (Y - \widehat{f(X)}_{ptb}) \tag{3}$$

If $\hat{Y}$ is fixed, then $\widehat{f(X)}_{ptb}$ becomes a function of p alone, and then RSS_{ptb} is a funciton of p only.

If p^T known, then $Var(RSS_{ptb}) = 0$ (assuming $\widehat{f(X)}$ is fixed, so it has no variance). So, regardless of the form of p^T (ie equal, or unequal), the plot of RSS_{ptb} vs $\sum_{i=1}^n p_i$ will be exactly quadratic.

If p^T is a random variable, with mean μ, and variance $\Sigma \neq 0$. Then, the $Var(RSS_{ptb}) \neq 0$. So, the resulting curve RSS_{ptb} vs $\sum_{i=1}^n p_i$ will have some variation, and not be a perfect quadratic.

We have plotted the RSS Perturbation curves for the three types of perturbations investigated. Figure 3 shows the plot of RSS where the perturbations were split evenly. Figure 4 shows the plots of RSS where the perturbations were unevenly split. Figure 5 shows the plot of RSS where the perturbations were randomly generated from a Uniform distribution. All

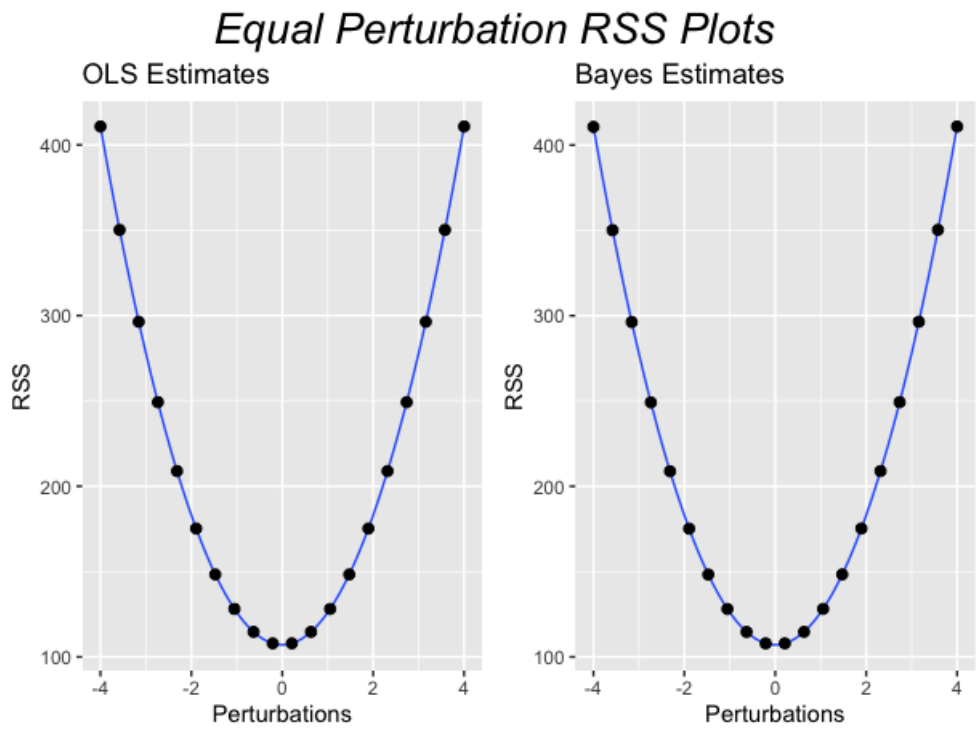

Figure 3: Plot of the Change in RSS for total perturbation Values. Perturbations were split evenly. Left: OLS $\boldsymbol{\beta}$ Estimates. Right: Bayesian $\boldsymbol{\beta}$ Estimates

three plots have an estimated quadratic fit overlaid. Each plot was calculated for both the Bayes and OLS estimators, and the plots look similar since the estimated parameter values were the same out to two decimal places. The perturbations were kept the same for both the OLS and Bayesian plots.

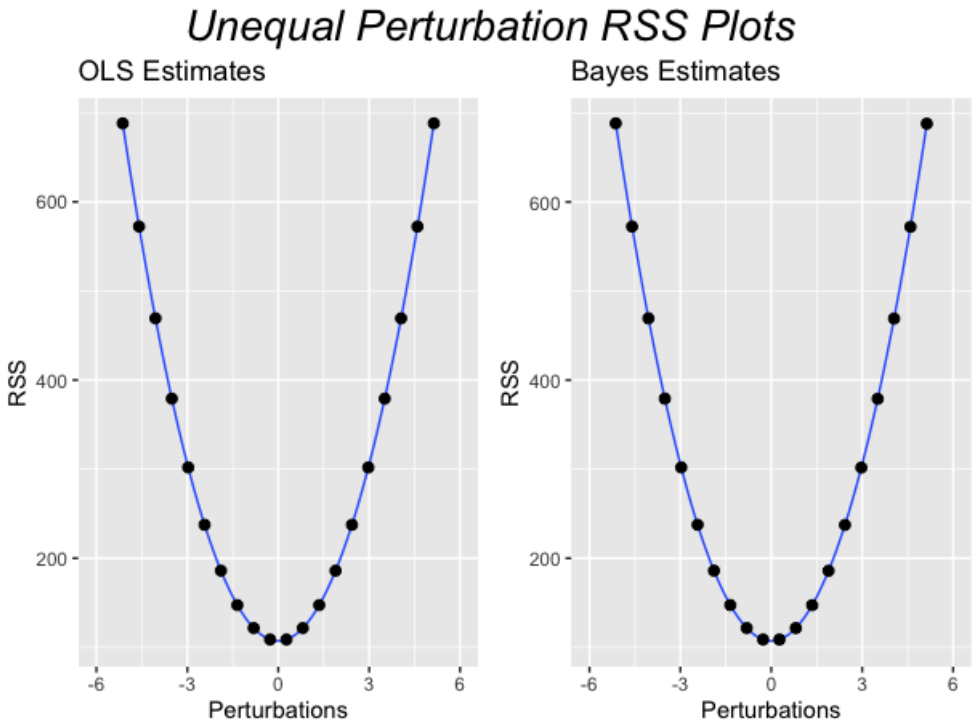

Figure 4: Plot of the Change in RSS for total perturbation Values. Perturbations were unevenly split. Left: OLS $\boldsymbol{\beta}$ Estimates Right: Bayesian $\boldsymbol{\beta}$ Estimates

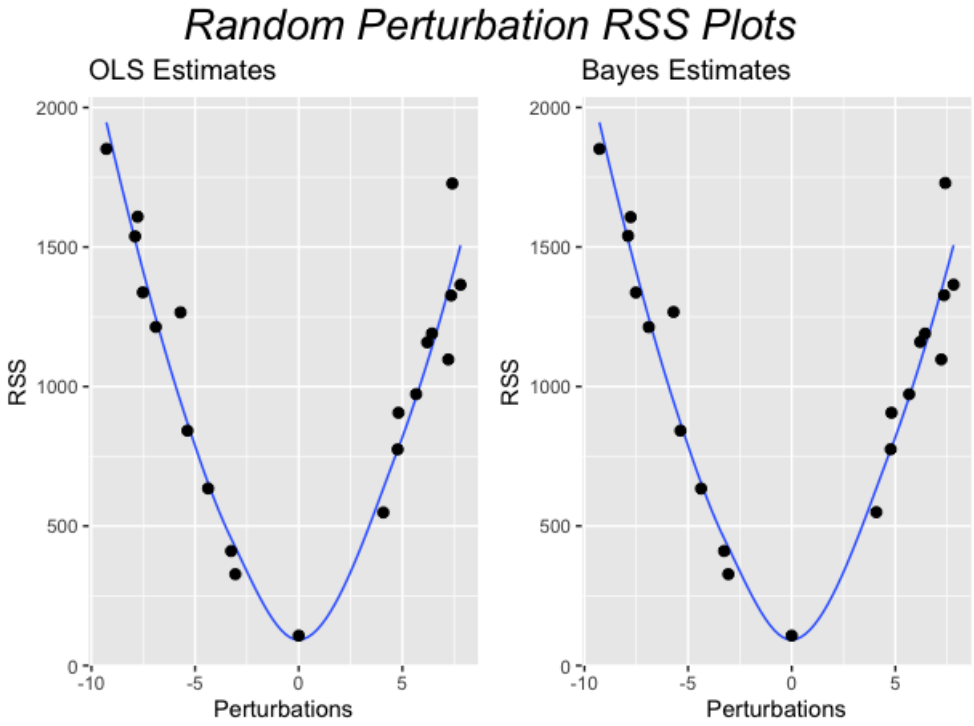

Figure 5: Plot of the Change in RSS for total perturbation Values. Perturbations were generated from Random Uniform. Left: OLS $\boldsymbol{\beta}$ Estimates. Right: Bayesian $\boldsymbol{\beta}$ Estimates

4.2 Distributional Fits

The data, along with the estimated curve overlaid corresponding to the X, are plotted below. Figure 6 shows the $\{X_i, Y_i\}$ data, with the Normal, Cauchy, and Logistic fits, which all take a real valued support. Figure 7 shows the $\{X_{pos_i}, Y_i\}$ fits: Gamma, Exponential, χ^2, Log-Normal, and Weibull, which all take positive support. Figure 8 shows the Beta fit, which has support between 0 and 1.

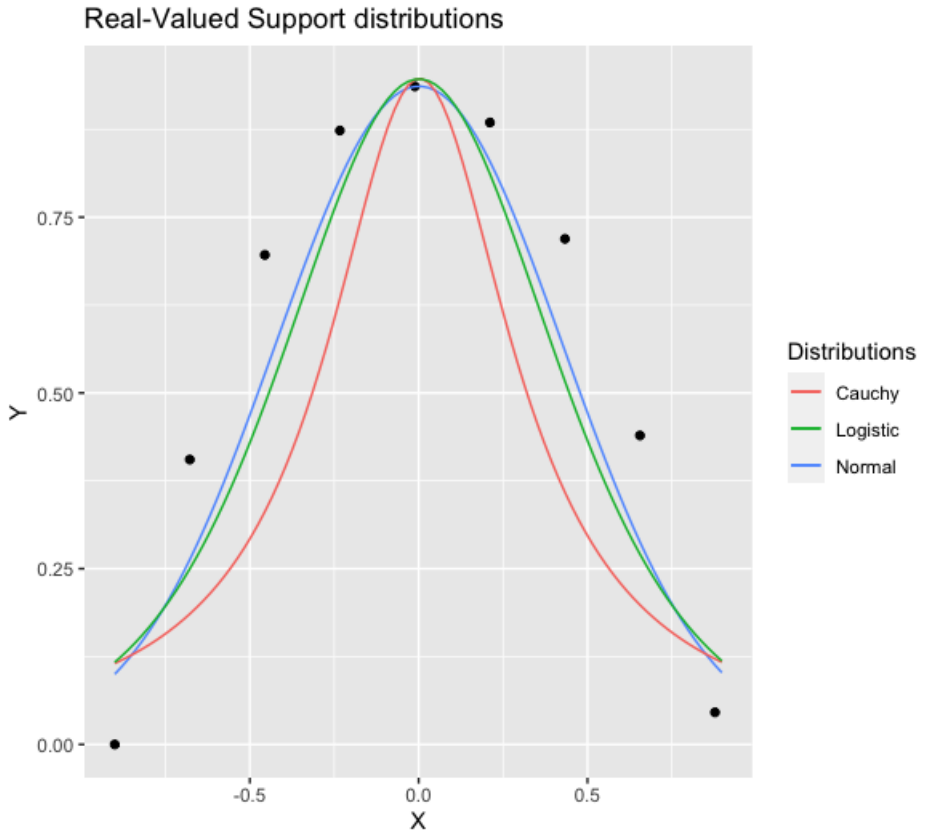

Figure 6: Plot of the data vs the estimated fits (real valued distributions)

Table 3 presents the mean and variance of each estimated fit, along with the two measures of fit described in the methods section. Since there were three different "X" values used, some standardizing of these estimates needed to be done in Table 3.

- X_{pos} distributions had the mean scaled $\hat{\mu} - 1$ to be on the same scale as X
- $X_{0,1}$ distributions had the mean scaled $\hat{\mu} - 0.5$, and $\sigma^2 * 4$ to be on the same scale as X
- Median and MAD are presented, since the mean and variance are undefined for Cauchy Distribution

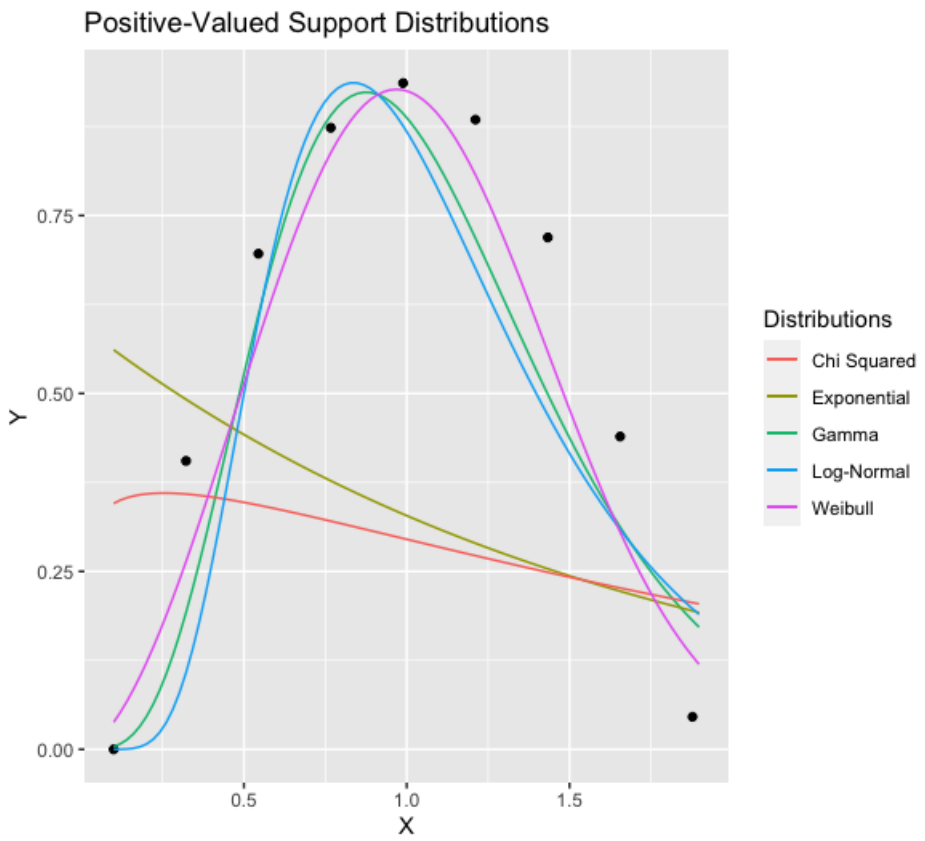

Figure 7: Plot of the data vs the estimated fits (Positive Support distributions)

Distribution	Mean	Variance	Cor	RSS
Normal$^+$	0.001	0.182	0.96	0.120
Cauchy	0.003	0.336	0.861	0.486
Logistic	0.001	0.229	0.941	0.173
Gamma$^+$	0.081	0.222	0.943	0.166
Exponential	0.679	2.818	-0.19	1.646
χ^2	1.254	4.508	0.101	1.629
Log-Normal	0.146	0.307	0.912	0.249
Weibull$^+$	0.025	0.170	0.971	0.094
Beta$^+$	-0.001	0.239	0.997	0.013

Table 3: Summary Statistics of each Distributional Fit. $^+$ indicates top four fits

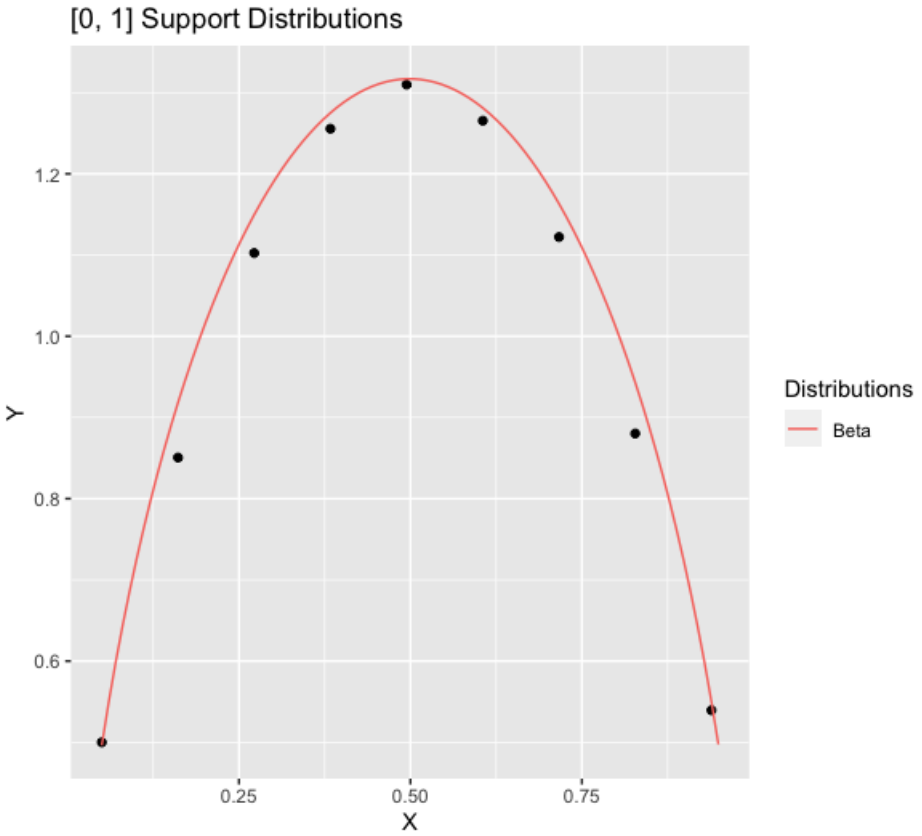

Figure 8: Plot of the data vs the estimated Beta fit

	Mean	Standard Deviation
Estimated Mean	0.027	0.00146
Estimated Variance	0.203	0.00106

Table 4: Average and Standard Deviation of the best 4 distributional fits estimated Mean and Variance

5 Conclusion

5.1 RSS Perturbation

The theory developed showed known perturbations would be perfectly quadratic, and random perturbations would not. Figures 3, 4, which are known, follow a quadratic perfectly. The random perturbations in Figure 5 are far from fitting the perfect quadratic. The plots created in Dr. Hayes research (Figure 1) are not perfectly quadratic, thus we would expect the perturbations are random. However, we know this is not the case, since [1] specifies the perturbation are equally split, known constants. So, there are two possible possibilities of this abnormality: either the $\widehat{f(X)}$ is not fixed, given perturbations, or a mistake was made in the analysis. Since we were unable to get data or access to the MNCP modeling system, we cannot make a claim as to which happened.

5.2 Distributional Fits

The four best distributional fits, as measured by both RSS and Cor, were the Beta, Weibull, Normal, and Gamma. The estimated means and variances for these fits were averaged, and the sample variation of these four fits was calculated. Table 4 shows these results. We argue, since our data had no variation between X and Y, the fits retain similar qualities regardless of data. As long as the Y data is generated in the same fashion as described, the fits should retain the "rankings" they currently have.

5.3 Recommendations

5.3.1 RSS Curves

We cannot make any clear recommendations about the RSS curves from the method, beyond the theory we developed. RSS curves which are not perfectly quadratic should be highly investigated, to check for errors or mistakes.

5.3.2 Distributional Fits

One clear problem with this method is the arbitrary nature of the Normal fit. Many of the distributional fits were similar in terms of both estimates of the mean, variance, and quality of fit. Even though the original X is centered at 0, this data had no sense of a true mean and

variance. However, the fits must have mean values close to 0 to be accurate. The distribution fit to perfectly quadratic data will not result in massive changes to the estimates of mean and variance, which, recall, is the overall goal of this estimation. The amount of points investigated can greatly affect the estimates of mean and variance, since a wider curve is made steeper when scaling to $[0, 1]$. However, as more perturbations are added, there is the squared term multiplied by the variance to re-scale. Thus, this problem is negated. As the range of X increases, the Variance will increase quadratically. Since the amount of points is extremely arbitrary, we suggest using the Beta distribution. The Perturbations could be scaled to be between $[0, 1]$, and then the Beta could be fit. Then the mean and variance could be re-scaled to the original values. Since the Beta distribution is extremely flexible, it has the "best" fit. Further, of the four best fits, it has the largest Variance estimate, and we argue overestimating the Variance for this method is better than underestimating.

References

[1] Ryan O'Mara and Robert Hayes. Dose deposition profiles in untreated brick material. *Health Physics*, 114(4):414–420, 2018.

[2] R. E. Faw J. K. Shultis. *An MNCP Primer.* Dept. of Mechanical and Nuclear Engineering Kansas State University, December 2011.

[3] RYAN PATRICK O'MARA. *Retrospective Dosimetry for Nuclear Nonproliferation and Emergency Response.* PhD thesis, NCSU, 2020.

[4] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.