Bulletproof Math

Marco Bijvank*
Vrije Universiteit, Amsterdam

Maria Caterina Bramati
Free University of Brussels

Leila Mohammadi, Fabio Rigat, Peter van de Ven
EURANDOM, Eindhoven

Roel Braekers, Tetyana Kadankova
Universiteit Hasselt

Sergei Anisov
Universiteit Utrecht

Aad Schaap
Teijin Twaron

Abstract

Teijin Twaron is a company that develops and produces fibres for bullet proof vests. In order to compare two fibres (or two bullet proof vests made of two different fibres for that matter) they want to determine the velocity V_p at which p percent of the bullets pass through the vest, in particular when $p = 50\%$. The objective of this research is to find good estimates for V_p. The available data have been analysed to examine which aspects influence the probability of perforation and have to be taken into consideration to determine V_p. Next, a general framework has been developed in which the notation is introduced. Several approaches are proposed to find good estimates for V_p. All methods are numerically illustrated. We recommend Teijin Twaron to use loss functions. But a logistic model or an istonic regression approach with linear interpolation performs also well. The paper ends with a new procedure how the data should be gathered to determine V_p.

KEYWORDS: quantile estimation, classification tree, generalised linear models, isotonic regression, smoothing splines, bootstrap method

1 Introduction

Having a good bullet proof vest to protect your body is important if you are in a gun fire or battle field. A few years ago a bullet proof vest contained metal and was

*corresponding author

therefore rather heavy to wear. The current vests are made of multiple layers of fibre, which make the vests much lighter. Teijin Twaron B.V. is a producer of these fibres. When they develop a new kind of fibre they make bullet proof vests out of it and test the quality of the vest. The easiest way to compare the quality of two vests is to determine for both vests the velocity at which fifty percent of the bullets pass through. This velocity will be denoted by V_{50}. A higher V_{50} means a better quality vest. The determination of the V_{50} is rather easy, since the event for a bullet to perforate the vest is equal to non-perforating the vest at this velocity. When this is simulated the number of times either event happens is the same on the long run.

The procedure Teijin Twarom uses to gather the data in determining V_{50} is to shoot at a particular vest 6 times with the same speed of the bullet and repeat this for 7 vests made from the same fibre but for different velocities. The first velocity they shoot with, is based on the estimate where they think V_{50} is at. When more than 3 of the 6 shots perforate the vest, the next velocity they shoot with is decreased with 50 m/s. Otherwise, it will be increased with 50 m/s. This procedure is repeated 7 times. So, it is some kind of evolution towards V_{50}.

Based on these 42 data records Teijin Twaron estimates V_{50}. First they group the data records in intervals based upon the velocity, with an interval length of 5 m/s. They average out the observations of perforating the vest within each interval. Next, a cumulative density function of the normal distribution is fitted into the new data points. How to fit such a function is explained in more detail in Section 4. Based upon the inverse of this function they determine V_{50}. Teijin Twaron wants to use the same function to determine the velocity at which an arbitrary percentage p of the bullets perforates the vest, in particular for p equal to 1%. This velocity is denoted by V_{01} and for a general p by V_p.

In toxicology we find studies that are similar to this research. A general introduction can be found in Agresti [1], Agresti [2] and Emmens [8]. In these toxicology studies, the interest lies in determining models to describe the relationship between the probability of reacting to a certain toxic chemical as a function of the given dose of this chemical. More specifically, for different dose levels, the researchers observe whether the dose results in a toxic reaction.

In this paper we will improve the procedure to determine V_p once the data is provided. But we will also design a new test procedure to gather the data that is used to determine V_p. In Section 2 we start with an analysis of the data Teijin Twaron currently uses. A general framework is presented in Section 3, in which we also introduce notation and a general set-up to compare different techniques. In Section 4 until Section 6 we present different techniques to derive a function that maps a velocity on the probability of perforating a particular vest. In Section 4 we will use Generalised Linear Models (GLMs), while the techniques discussed in Section 5 do not impose a predefined functional form. The last technique to estimate V_p is a bootstrap method. This approach determines V_p based upon a characteristic at this velocity instead of finding the inverse of a function. In Section 7 we propose a new procedure to gather the data. In Section 8 we compare the different solution techniques and give a conclusion which method we recommend.

2 Data analysis

The data set provided by Teijin Twaron contains the 42 data records as explained in Section 1 for 9 different vest types. Each data record consists of whether the bullet perforated the vest (this is also called the perforation status), the velocity of the bullet which was shot at the vest, the shot number (1 to 6), the vest number (1 to 7) and the vest type. For one particular vest type the data set contains 126 data records.

The objective of the statistical analysis of the data is to provide an overview of the relationships between the perforation probability (also called the response variable) and its four explanatory variables or covariates, i.e. the bullet velocity, the shot number, the vest number and the vest type. The model employed in this analysis explains the observed variability of the data without making any assumption on the physical or chemical mechanisms which might have played a role in generating the samples. This means that all data records for the different vest types under investigation are used all together in this analysis.

In the analysis the entire sample is modeled according to a classification tree (Breiman *et al.* [4]). This is a semiparametric statistical model in which the data is partitioned among several subsamples with significantly different perforation probabilities. The data subgroups are defined by a binary tree where the splits are functions of the covariates. For instance, two groups can be obtained by considering the samples with a bullet velocity smaller than 400 m/s and those with a velocity larger than or equal to 400 m/s. Within the latter group, two clusters of data points can be formed by dividing the samples associated to a particular vest type versus those corresponding to all other vest types, and so on. We will refer to the groups of data generated by a given tree structure as its leaves. In this analysis we do not assume any specific distribution on the space of tree structures, whereas within each leaf we model the perforation status as a Bernoulli random variable with a leaf-specific perforation probability.

In order to estimate the tree structure we perform a stochastic search using the probability of the tree given the data as the score function. This is a simulation-based computationally intensive method which evaluates the uncertainty on the specification of the tree structure conditionally on the sample (Breiman *et al.* [4], Chipman *et al.* [5], Chipman *et al.* [6], Denison *et al.* [7], Holmes *et al.* [13]). Given the best tree structure found by the stochastic search, we estimate the leaf-specific perforation probabilities in a Bayesian fashion. In particular, for each leaf we assume a uniform prior perforation probability. By combining this prior with the Bernoulli likelihood we obtain a Beta perforation probability given the samples falling in the leaf. The Beta distribution can be summarised analytically, providing both a point estimate of the leaf-specific perforation probabilities and their confidence intervals.

Figure 1 shows the results of the estimation of the tree structure when all data is analysed. The tree has a total of eight leaves, which cluster the samples as a function of the bullet velocity and shot number. Notice that this tree structure does not depend on the vest type. This surprising result is emphasised in Table 1. For each of the four available covariates, the table shows its estimated probability of inclusion in the tree structure. The covariates with the highest predictive power

are the bullet velocity and the shot number but the vest type does not appear to discriminate groups of samples with significantly different perforation probabilities.

Figure 1: The tree structure which fits the data the best.

	covariate			
	velocity	shot number	vest type	vest number
Estimated inclusion probability	1	0.92	0.08	0.02

Table 1: The estimated inclusion probabilities for the four covariates to incorporate them into the classification tree.

Finally, the estimated perforation probabilities for the eight leaves of the tree are presented in Table 2. It can be noted that the estimated perforation probability appears to be increasing in the bullet velocity. Moreover, at any given bullet speed, the estimated perforation probability of the first shot is lower than of the second shot, which in turn is higher than any of the other shot numbers.

Leaf number	Estimated perforation prob.	95% Conf. interval
1	0.95	$[0.87; 0.98]$
2	0.21	$[0.12; 0.35]$
3	0.77	$[0.65; 0.87]$
4	0.60	$[0.46; 0.72]$
5	0.48	$[0.34; 0.63]$
6	0.28	$[0.17; 0.42]$
7	0.25	$[0.17; 0.35]$
8	0.56	$[0.44; 0.69]$

Table 2: The estimated perforation probabilities including the 95% confidence intervals for the eight leaves of the tree.

Although Figure 1 and Table 1 make clear that only the velocity and shot number are of relevance to determine the probability of perforation, we consider the velocity and vest type as explanatory covariates in this research. The main reason to take the vest type into consideration is because Teijin Twaron wants to compare different vest types. The shot number is not used as covariate, since the shooting pattern should result in independent shots. Appearantly this is not the case. Therefore, we advise Teijin Twaron to look at this. But in the remainder of this paper, we will use all data records of one particular vest type to determine V_{50} based upon the speed of the bullet. A motivation not to take the shot number into consideration is to have a bigger sample set. Otherwise there are only 7 data records available.

3 General Framework

A procedure has to be developed to determine V_p; the velocity at which p percent of the bullets perforates the vest. Since we do not take the shot number or the vest number into account (see Section 2), the data records of one vest type are presented by (X_i, Y_i)-pairs. The velocity of the i-th shot (expressed in m/s) is denoted by X_i and the event of a perforation by Y_i, where

$$Y_i = \begin{cases} 1, & \text{if shot } i \text{ perforated the vest,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The number of data records is denoted by N, so $i = 1, \ldots, N$. Whenever the data set is rearranged into intervals with an interval length of 5 m/s, the (X_i, Y_i)-pairs are transformed into (X'_j, Y'_j)-pairs where $X'_{j+1} = X'_j + 5$. The new response variabe Y'_j becomes the average probability of perforating the vest where the velocity of the bullet is in interval j:

$$Y'_j = \left[\sum_{i=1}^{N} 1_{X_i \in \left[X'_j - 2\frac{1}{2}, X'_j + 2\frac{1}{2}\right)} \right]^{-1} \sum_{i: X_i \in \left[X'_j - 2\frac{1}{2}, X'_j + 2\frac{1}{2}\right)} Y_i,$$

where $1_{\text{condition}}$ is the indicator function:

$$1_{\text{condition}} = \begin{cases} 1, & \text{if } \textit{condition} \text{ is satisfied,} \\ 0, & \text{otherwise.} \end{cases}$$

Figure 2 shows the (X_i, Y_i)-pairs as well as the (X'_j, Y'_j)-pairs for the data set where 126 data records are available for a particular vest type.

For every vest type we are interested in finding a function $f(v) : \mathbb{R}_+ \to [0, 1]$ that maps a velocity v onto the probability of perforating the vest when the bullet has speed v. By taking the inverse $(f^{-1}(p) : [0, 1] \to \mathbb{R}_+)$ we find V_p. In Section 4 and Section 5 different approaches are proposed to find an appropriate $f(v)$. A bootstrap method is described in Section 6 to determine V_p directly.

In order to compare the different techniques we have to define a measure of fitness that relates the differences between the function $f(v)$ and the data (X_i, Y_i) for $i = 1, \ldots, N$. A classical measure of discrepancy is the mean squared error (MSE) as defined in Equation (2).

$$MSE = \frac{1}{N} \sum_{i=1}^{N} \left[f(x_i) - y_i \right]^2, \tag{2}$$

where (x_i, y_i) are the observed realisations of the stochastic variables X_i and Y_i. Small deviations are not penalised as much as large deviations in this definition for the fitness measure.

4 Generalized Linear Model

Generalized linear models (GLMs) are generalizations of the linear model (see McCullagh, *et al.* [14]). In its simplest form, a linear model specifies the linear relationship between a dependent (or response) variable, and a set of predictor variables (or

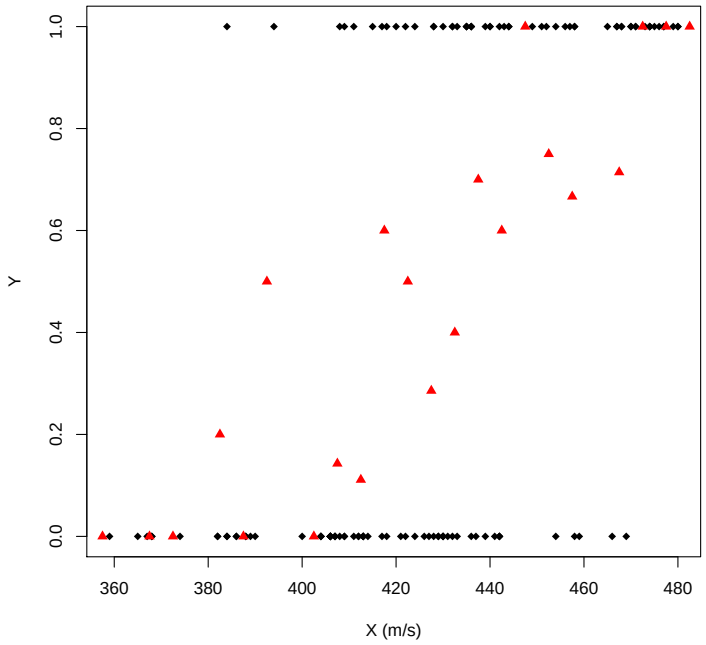

Figure 2: The rough data (i.e. the (X_i, Y_i)-pairs) and the rearranged data (i.e. the (X'_j, Y'_j)-pairs) are represented by block dots and red triangles respectively, for one particular vest type.

covariates). In this research it is inadequate to describe the observed data (perforation status Y_i) with a linear relationship between the variables (bullet speed X_i). The main reason for this is that the effect of the velocity on the perforation status is not linear in nature.

4.1 Link function

In generalized linear models a so-called link function, denoted by g, specifies the connection between the response variable Y_i and the covariate X_i. For each experiment, the response Y_i can take only one of two possible values, denoted for convenience by 0 and 1 (see also Equation (1)). Therefore, we may write

$$P(Y_i = 0) = 1 - \pi_i \qquad P(Y_i = 1) = \pi_i \tag{3}$$

for the probabilities of non-perforation and perforation respectively. Linear models play an important role in both applied and theoretical work. We suppose therefore that the dependence of Y on X occurs through the linear predictor η_i given by

$$\eta_i = \beta_0 + \beta_1 X_i, \quad i = 1, \ldots, N$$

for unknown coefficients β_0 and β_1. For binary random variables the link function g should map the interval $[0,1]$ onto the whole real line $(-\infty,\infty)$. So,

$$g(\pi_i) = \eta_i = \beta_0 + \beta_1 X_i, \quad i = 1, \ldots, N.$$

A wide choice of link functions is available. Three link functions commonly used in practice are

1. the logit or logistic function

$$g(\pi) = \log\big(\pi/(1-\pi)\big),$$

2. the probit or inverse Normal function

$$g(\pi) = \Phi^{-1}(\pi),$$

3. the complementary log-log function

$$g(\pi) = \log\big(-\log(1-\pi)\big).$$

The first two functions are symmetrical in the sense that

$$g(\pi) = -g(1-\pi).$$

All three functions are continuous and increasing on $(0,1)$. This last characteristic is exactly what is required for this research.

To give an example, we look at the logit function

$$\begin{aligned} g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \beta_0 + \beta_1 x_i \\ \pi_i &= \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}. \end{aligned}$$

This expression equals $f(v)$ based on Equation (3). By inverting this expression we can determine V_p:

$$f(V_p) = p = \frac{\exp(\beta_0 + \beta_1 V_p)}{1 + \exp(\beta_0 + \beta_1 V_p)} \quad \Leftrightarrow \quad V_p = \frac{\log\left(\frac{p}{1-p}\right) - \beta_0}{\beta_1}.$$

The same can be done for the other link functions. The results are summarised in Table 3.

When we combine the probit model with the rearranged data (see Section **??**) we get the current approach to determine V_p (see Section 1).

link function	π_i	V_p
logit	$\frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}$	$\frac{\log\left(\frac{p}{1-p}\right)-\beta_0}{\beta_1}$
probit	$\Phi(\beta_0+\beta_1 x_i)$	$\frac{\Phi^{-1}(p)-\beta_0}{\beta_1}$
complementary log-log	$1-\exp\big(-\exp(\beta_0+\beta_1 x_i)\big)$	$\frac{\log\big(-\log(1-p)\big)-\beta_0}{\beta_1}$

Table 3: The probability of perforation π_i as a function of the observed velocity x_i for the different link functions. The inverse gives an expression to determine V_p.

4.2 Alternative Predictor

A disadvantage of all three link functions is that the inverses have support on the entire real axis. This means that a velocity of 0 m/s results in a strictly positive probability of perforating the vest. This phenomena is absolutely not true in the experimental setting. A possible solution is to define an alternative predictor η_i as

$$\eta_i = \beta_0 + \beta_1 \log(X_i).$$

For example, the alternative logit function results in

$$\pi = \frac{\exp\big(\beta_0 + \beta_1 \log(x_i)\big)}{1 + \exp\big(\beta_0 + \beta_1 \log(x_i)\big)},$$

and

$$V_p = \exp\left(\frac{\log\left(\frac{p}{1-p}\right) - \beta_0}{\beta_1}\right) = \left(\frac{p}{1-p}\right)^{1/\beta_1} \exp(-\beta_0/\beta_1).$$

Having selected a particular model, it is required to estimate the parameters β_0 and β_1. The parameter estimates are the values that minimize the goodness of fit between the observed data (y_i) and the fitted values generated by the model (π_i). We are concerned with estimates obtained by maximizing the likelihood of the parameters for the data observed. This principle is explained in the next subsection.

4.3 Maximum Likelihood

The likelihood of the data is the probability of observing the data for certain parameter values (Ross [16]) and is expressed by Equation (4).

$$L(\beta_0, \beta_1; y_1, \ldots, y_N) = \prod_{i=1}^{N} p_{\pi_i}(y_i|\beta_0, \beta_1), \tag{4}$$

where $p_{\pi_i}(y_i|\beta_0, \beta_1)$ is the probability of observing y_i when the probability of perforation equals π_i if β_0 and β_1 are the parameter values. Because of the definition in Equation (3), we should get

$$p_{\pi_i}(y_i|\beta_0, \beta_1) = \begin{cases} \pi_i, & \text{if } y_i = 1, \\ 1 - \pi_i, & \text{if } y_i = 0 \end{cases} \tag{5}$$

and, therefore,

$$p_{\pi_i}(y_i|\beta_0, \beta_1) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}. \tag{6}$$

This expression can be substituted into Equation (4) to define the likelihood function. A same kind of expression can be derived when the variable Y_j' is used in stead of Y_i. Only then, Expression (3) is not valid anymore.

The objective is to find the values of the two estimators, which maximize the likelihood funtion. Often it is easier to maximize the log-likelihood function because it is easier to manipulate mathematically. Therefore, we derive this by taking the natural logarithm of the likelihood function. When we use Equation (6) in Equation (4) and take the natural logarithm, we get

$$l(\beta_0, \beta_1; y_1, \ldots, y_N) = \log L(\beta_0, \beta_1; y_1, \ldots, y_N) = \sum_{i=1}^{N} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i) \right].$$

For the logit function, the log-likelihood function equals

$$l(\beta_0, \beta_1; y_1, \ldots, y_N) = \sum_{i=1}^{N} \left[y_i(\beta_0 + \beta_1 x_i) - \log\left(1 + \exp(\beta_0 + \beta_1 x_i)\right) \right].$$

The log-likelihood function is differentiable in this study. Therefore, the values of β_0 and β_1 that maximize the log-likelihood can be found by solving the first order conditions for the two parameters.

4.4 Numerical Results

The classical GLMs (Section 4.1) and alternative GLMs (Section 4.2) are implemented. Figure 3 shows the results when the dataset represented in Figure 2 is used for the different link functions. The dots in this figure are the original observed (X_i, Y_i)-pairs (or the rearranged (X_i', Y_i')-pairs). The classical GLMs are represented by a plain line and the alternative GLMs by a dashed line. We notice that there is not much difference between the two models. However, in the tails of the curves the alternative model always has a lower probability of perforation at the same velocity in comparison to the classical models. This is to be expected since the alternative model only allows strictly positive velocities. Therefore, it should have a tighter tail at low velocities and a thicker tail at high velocities.

The maximum likelihood estimators for β_0 and β_1 (e.g., $\hat{\beta}_0$ and $\hat{\beta}_1$ respectively) of each model are presented in Table 4, as well as the mean squared error (MSE).

We notice that the value of $\hat{\beta}_1$ is strictly positive in both models. This implies that the curves are strictly increasing. We expected such a result because we know that for an increasing speed of the bullet the probability of perforation will also increase. Based upon these results, we can also conclude that a logit model performs the best (lowest MSE). However, these are the results for only one vest type. Therefore, we compare all techniques and all data sets (of the different vest types) in Section 8.

Based upon the estimated parameter values we determined V_p for the different models with the expressions formulated in Table 3. Table 5 gives the estimated

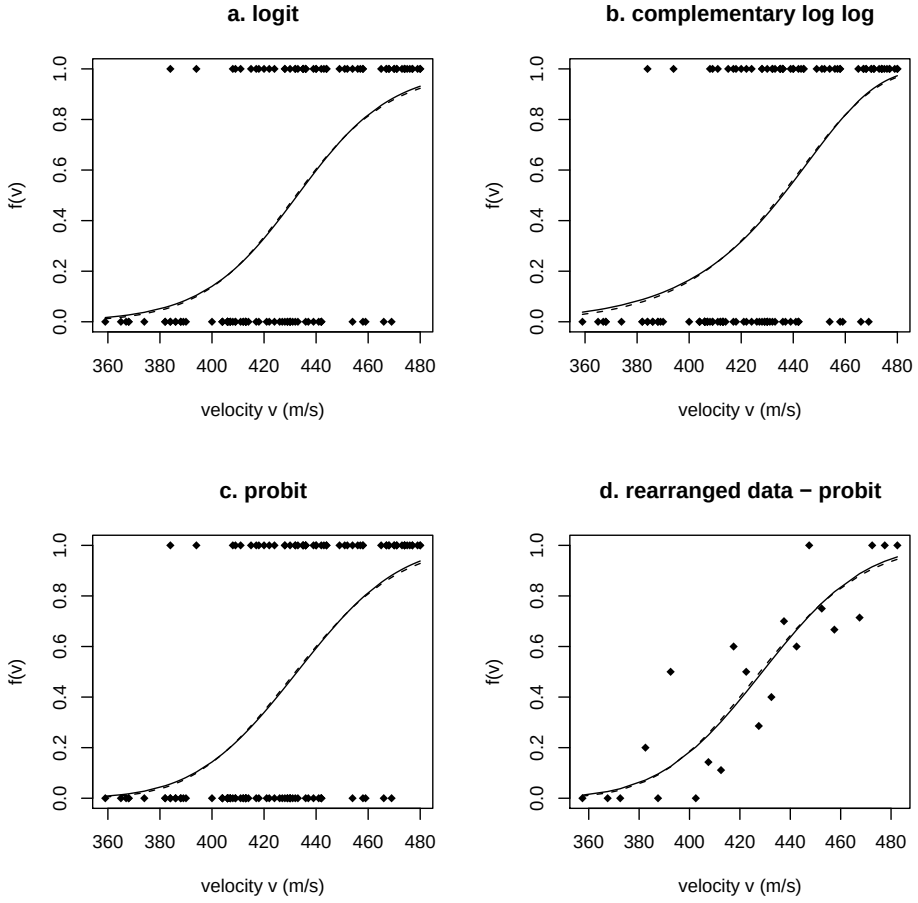

Figure 3: The GLM estimates for the perforation probability $f(v)$ as a function of velocity v for different link functions. The straight line is the classical link function and the dashed line is the alternative link function. The dots are the input data records.

model	$\hat{\beta}_0$	$\hat{\beta}_1$	MSE
logit	-23.93	0.0553	0.1650172
alt. logit	-144.2	23.760	0.1650141
probit	-14.09	0.0326	0.1650477
alt. probit	-84.61	13.942	0.1650921
log-log	-16.70	0.0375	0.1659732
alt. log-log	-99.71	16.350	0.1656465
current model	-13.52	0.0315	0.1667188

Table 4: The likelihood estimators for the different models including the measure of fitness of the model.

velocities at which 1% (V_{01}) and 50% (V_{50}) of the bullets perforate the vest. This table also presents a 95% confidence interval for V_p. This interval is generated with the bootstrap method (or resampling): randomly selecting N observations from the data with replacement and obtaining estimates for V_p for the resulting bootstrap sample. We repeated this procedure many times, calculating estimates for each bootstrap replication. This gives a distribution for the estimate of V_p.

model		estimation	95% Confidence Interval
logit	V_{50}	432.81	$[425.04; 441.04]$
	V_{01}	349.71	$[312.22; 376.21]$
alt. logit	V_{50}	432.28	$[424.77; 439.77]$
	V_{01}	356.27	$[327.58; 380.58]$
probit	V_{50}	432.74	$[424.57; 440.83]$
	V_{01}	361.28	$[332.69; 384.28]$
alt. probit	V_{50}	432.12	$[424.15; 440.18]$
	V_{01}	365.71	$[341.26; 386.59]$
c log-log	V_{50}	435.95	$[428.00; 443.59]$
	V_{01}	322.94	$[281.11; 356.47]$
alt. c log-log	V_{50}	435.29	$[427.42; 442.87]$
	V_{01}	335.99	$[302.86; 363.45]$
current	V_{50}	428.92	$[419.90; 437.09]$
	V_{01}	355.12	$[325.27; 382.83]$

Table 5: The estimated velocities including their 95% confidence intervals.

For V_{01} we notice that the estimates from the classical models are smaller compared to those from the alternative models. This is not a surprise since we mentioned already that the curves for $f(v)$ show lower values in the tails for the alternative models in comparison to the classical models (see Figure 3). The confidence intervals in the alternative models are smaller than in the classical models. This is because of the same reasoning.

5 Non-Parametric Models

In the previous section we fitted a curve according to generalized linear models. Such a technique assumes something about the form of the curve. When we do not want to make such an assumption, we have to fit a curve entirely based upon the data. The only restriction we have is that the curve should be monotonic increasing (i.e. non-decreasing). Most of the time, the data does not have this property (see Figure 2). Therefore, smoothing has to take place. This can be done in two different ways: either smooth the data first and then find the curve or find a curve on the rough data with the use of smoothing. An example of the first approach is isotonic regression and for the second approach smoothing splines can be used. The third approach we mention in this section is the use of loss functions, which are based upon empirical distributions. All three applications are discussed in this section and

we end with numerical results on the three methods.

5.1 Smoothing Splines

Splines are piecewise polynomial functions that fit together (Eubank [9]). In particular, for cubic splines, the first and second derivatives are also continuous in every point. Smoothing splines are curves that get reasonably close to the data in a graceful manner such that it gives the appearance of a single curve.

Smoothing splines arise as the solution to the following simple-regression problem: Find the function $\hat{f}(x)$ with two continuous derivatives that minimizes the penalised sum of squares,

$$SS^*(h) = \sum_{i=1}^{n} [y_i - f(x_i)]^2 + h \int_{x_{\min}}^{x_{\max}} \left[f''(x)\right]^2 dx, \tag{7}$$

where h is a smoothing parameter (Fox [11]). The first term in Equation (7) is the residual sum of squares. The second term is a roughness penalty, which is large when the integrated second derivative of the regression function $f''(x)$ is large. The endpoints of the integral enclose the data. At one extreme, when the smoothing constant is set to $h = 0$ (and if all the x-values are distinct), $\hat{f}(x)$ simply interpolates the data. So, small values of h correspond to more emphasis on goodness-of-fit. Conversely, when h is large it places a premium on smoothness. Typically $h \in (0, 1]$. Since we are interested in a monotonically increasing function, we set h to the smallest smoothing parameter such that this restriction is satisfied.

5.2 Isotonic Regression

Isotonic regression is a non-parametric method that is used when a dependent response variable is monotonically related to an independent predictor variable (Barlow *et al.* [3] and Robertson *et al.* [15]). We are indeed looking for an isotonic (i.e., non-decreasing) function where the probability of perforation $f(v)$ depends on the velocity v of the bullet. A commonly used algorithm for computing the isotonic regression is the pair-adjacent violators algorithm (PAVA), which calculates the least squares isotonic regression of the data set (Barlow *et al.* [3] and Robertson *et al.* [15]).

The basic idea of PAVA is the following: sort the (x_i, y_i)-data pairs such that $x_1 \leq x_2 \leq \ldots \leq x_N$. If $y_1 \leq y_2 \leq \ldots \leq y_N$, then all points are increasing and the algorithm stops. Otherwise, select the first data pair i for which $y_i > y_{i+1}$. In that case replace (x_i, y_i) and (x_{i+1}, y_{i+1}) by their weigthed average (x_i^*, y_i^*), where

$$\begin{aligned} x_i^* &= \frac{w_i x_i + w_{i+1} x_{i+1}}{w_i + w_{i+1}}, \\ y_i^* &= \frac{w_i y_i + w_{i+1} y_{i+1}}{w_i + w_{i+1}}, \\ w_i^* &= w_i + w_{i+1}. \end{aligned}$$

This procedure is repeated until the algorithm terminates. The algorithm starts with weights equal to one ($w_i = 1$ for $i = 1, 2, \ldots, N$). The algorithm is applied upon the data of Teijin Twaron and represented in Figure 4.

Now the new data set is such that it is non-decreasing. We can easily find an interpolation scheme to connect the data points and find $f(v)$. We make use of two interpolation schemes in Section 5.4: stepwise interpolation and piecewise linear interpolation.

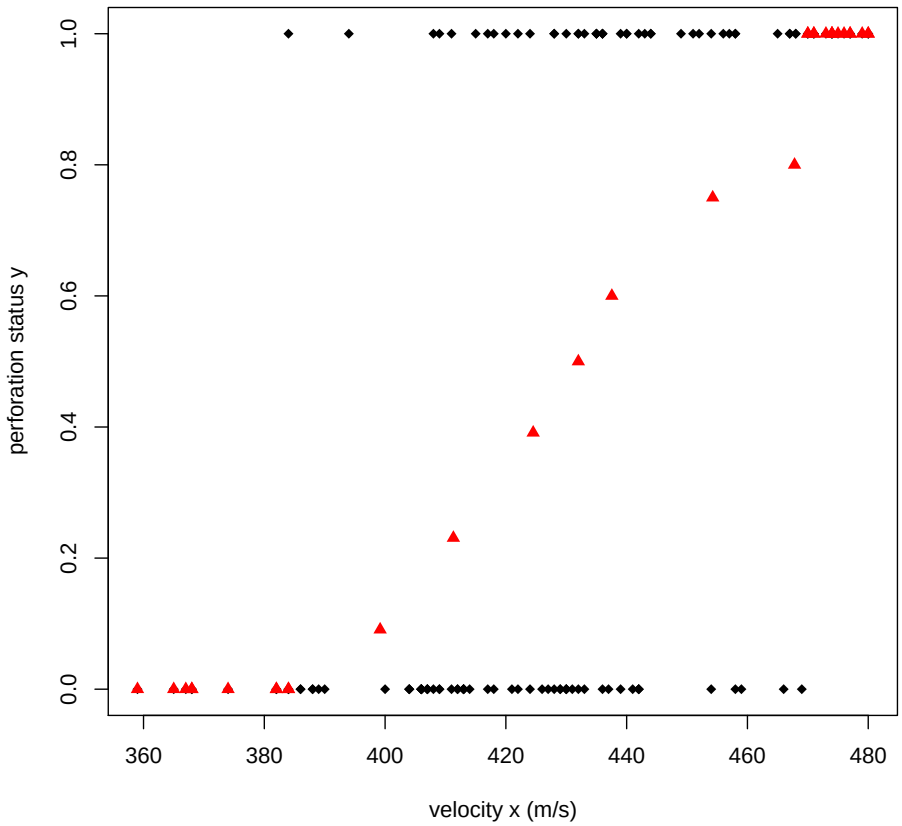

Figure 4: We transformed the data from rough data (dots) to monotone increased data (triangle).

5.3 Loss Function

In this section we describe a method that determines the probability of perforation (or the function $f(v)$) entirely based on empirical distributions. Besides this function $f(v)$, this approach also requires a probability density function of the velocity v, denoted by $g(v)$. Based on the data we can consider the empirical density function of the velocity (denoted by G) and the empirical distribution of $f(v)$ (denoted by F). So,

$$F(x_i) = \begin{cases} 1, & \text{if } y_i = 1, \\ 0, & \text{otherwise.} \end{cases}$$

We want to minimize the result to obtain an estimator for $f^{-1}(p)$. This function is called a loss function. Select positive α and β and define the loss function as

$$L(a) = \alpha \int_0^a f(v)g(v)\partial v + \beta \int_a^\infty (1 - f(v))g(v)\partial v.$$

To minimize L, we take the derivative to a and set it equal to 0:

$$\frac{\partial}{\partial a} L(a) = \big(\alpha f(a) - \beta\big(1 - f(a)\big)\big)\, g(a) = 0, \tag{8}$$

Equation (8) is solved by a^* with

$$f(a^*) = \frac{\beta}{\alpha + \beta}.$$

Note that

$$\frac{\partial^2 L(a)}{\partial a^2} = (\alpha + \beta)g(a)\frac{\partial}{\partial a} f(a) + \big(\alpha f(a) - \beta\big(1 - f(a)\big)\big)\, \frac{\partial}{\partial a} g(a),$$

such that

$$\left.\frac{\partial^2 L(a)}{\partial a^2}\right|_{a=a^*} = \big(\alpha + \beta\big) \left.\frac{\partial}{\partial a} f(a)\right|_{a=a^*} g(a^*) \geq 0$$

because $f(v)$ is increasing in v and $\alpha f(a^*) - \beta\big(1 - f(a^*)\big) = 0$, using Equation (8). It means that a^* is the minimizer of L. We may set $p = \beta/(\beta + \alpha)$. For simplicity, we take $\beta = 1$ and $\alpha = 1/p - 1$. To estimate the inverse of $f(v)$ (i.e., $f^{-1}(p)$), it is now enough to minimize the empirical counterpart of L, namely

$$\begin{aligned} L(\alpha) &= \left(\frac{1}{p} - 1\right) \int_0^a f(x)g(x)\partial x + \beta \int_a^\infty \big(1 - f(x)\big)g(x)\partial x \\ &= \left(\frac{1}{p} - 1\right) \sum_{i=1}^n 1_{x_i < a, y_i = 1} + 1_{x_i \geq a, y_i = 0}. \end{aligned}$$

We select $m \in N$ and for $j = 1, \ldots, m$, we set $p = 1/j$ and minimize L to obtain $\hat{f}^{-1}(1/j)$.

5.4 Numerical Results

All three non-parametric approaches are implemented and the resulting functions $f(v)$ for each approach are presented in Figure 5. The inverses of these functions yield the estimator for V_p. With the use of resampling we constructed a 95% confidence interval (see also Section 4.4). The results are represented in Table 6. This table also represents the MSE for each technique.

6 Bootstrap Method

In the previous two sections we were interested in finding a function $f(v)$ in order to determine V_p. In this section we rewrite $f(v)$ as a conditional probability

$$f(v) = P(Y = 1 | X = v),$$

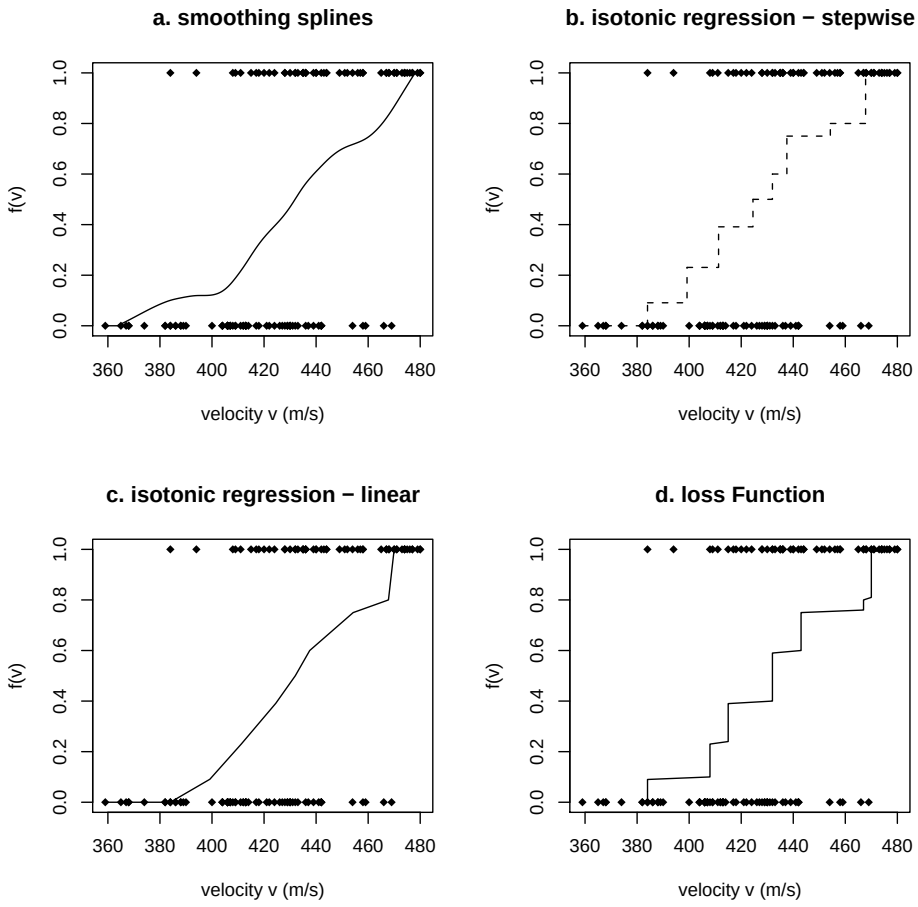

Figure 5: The estimates for the perforation probability $f(v)$ as a function of velocity v for different non-parametric approaches. The dots are the input data records.

model		estimation	95% Confidence Interval	MSE
smoothing spline	V_{50}	431.92	[;]	0.1618964
	V_{01}	365.64	[;]	
isotonic regression	V_{50}	432	$[410.8182; 440.6667]$	0.17325
- stepwise	V_{01}	384	$[374; 409]$	
isotonic regression	V_{50}	432	$[418.2733; 442.7394]$	0.1643158
- linear	V_{01}	385.67	$[376.2300; 409.1175]$	
loss function	V_{50}	432	$[432; 443]$	0.157155
	V_{01}	384	$[384; 384]$	

Table 6: The estimated velocities including their 95% confidence intervals.

the probability of perforation under the condition that the velocity X equals v. Now,

$$P(Y = 1|X = V_p) = p \qquad P(Y = 0|X = V_p) = 1 - p. \tag{9}$$

Using Bayes rule, we can rewrite Equation (9) as

$$P(Y = 1|X = V_p) = \frac{P(X = V_p|Y = 1)P(Y = 1)}{P(X = V_p)}, \tag{10}$$

and

$$P(Y = 0|X = V_p) = \frac{P(X = V_p|Y = 0)P(Y = 0)}{P(X = V_p)}. \tag{11}$$

Dividing Equation (10) by Equation (11) and use Equation (9), we get the following expression

$$\frac{p}{1-p} = \frac{P(X = V_p|Y = 1)P(Y = 1)}{P(X = V_p|Y = 0)\,(1 - P(Y = 1))}. \tag{12}$$

Now we have to compute each of the components of Equation (12). Let us first look at $P(Y = 1)$, i.e. the proportion of data records of which the bullet perforates the vest. This can be estimated directly from the data. The two other probabilities can also be derived directly from the data for every observed velocity x_i, $i = 1, \ldots, N$, where

1. $P(X = x_i|Y = 1)$ is the frequency at which we observe velocity x_i when the vest is perforated, and
2. $P(X = x_i|Y = 0)$ is the frequency at which we observe velocity x_i when the vest is not perforated.

Since the property of Equation (12) holds for V_p we calculate the ratio

$$\frac{P(X = x_i|Y = 1)P(Y = 1)}{P(X = x_i|Y = 0)\,(1 - P(Y = 1))},$$

for each observed velocity x_i and the one that is closest to $p/(1-p)$ is the estimate for V_p.

The main problems with this approach are the few data points in each conditional distribution of the velocity and the fact that we can find different velocities that are closest to the property of V_p. To overcome the first problem we propose to use the bootstrap method to get a distribution for V_p, which allows us to estimate V_p with the average and to construct a 95% confidence interval. The second problem (of multiple velocities satisfying Equation (12)) is solved for V_{50} by taking the median and for V_{01} the minimum value of those velocities is selected.

6.1 Verification

In order to verify whether the algorithm performs well, we can generate samples from known distributions (like normal or weibull) that can be used as input. For known distributions, we know what the outcome of the algorithm should be. With the use of a small Monte-Carlo simulation experiment we can test the performance. Table 7 shows the deviation of the result from algorithm with the true outcome for different distributions. The parameters for the distributions are such that the mean and variance are equal as the data set of Teijin Twaron. Based on these results, we can conclude that the algorithm works well for most distributions.

distribution	percentage deviation	
	V_{50}	V_{01}
chi-square	5.35%	6.30%
gamma	5.51%	6.30%
logistic	5.21%	9.17%
log-normal	72.1%	3616.41%
normal	5.34%	6.89%
student	0.33%	1.07%
uniform	5.78%	1.90%

Table 7: To verify the bootstrap method, we performed the method with known distributions and therefore the actual outcome is known as well.

6.2 Numerical Results

When we apply the proposed procedure to the dataset, the resulting estimates for both V_{50} and V_{01} are presented in Table 8 including the 95% confidence intervals. Based on these results we conclude that the estimates for V_{50} have a large 95% confidence interval and for V_{01} a rather small interval. This is because the data is collected in a way to determine V_{50}. As a result, not much different velocities are detected satisfying the property as defined in Equation (12) for V_{01}.

sample	size	V_{50}		V_{01}	
		estimation	95% conf. int.	estimation	95% conf. int.
0	126	426.81	[407; 458]	361.90	[359; 368]
1	42	422.03	[412; 429]	398.82	[398; 402]
2	42	458.42	[438; 471]	414.68	[413; 418]
3	42	423.45	[418; 432]	397.94	[397; 404]
4	42	466.59	[448; 483]	434.45	[433; 439]
5	42	459.54	[445; 468]	438.20	[438; 440]
6	42	479.00	[459; 501]	458.60	[458; 460]
7	42	491.87	[471; 499.5]	454.95	[454; 458]
8	42	392.68	[373; 402]	346.68	[346; 351]
9	36	383.88	[361; 406]	350.92	[350; 353]

Table 8: Estimates on different samples, $B = 200$.

7 Experimental Set-Up

The current design of the experiment set-up was originally developed to determine V_{50} (see Section 1). With the same data Teijin Twaron wants to make statements about V_p for arbitrary p. Also in other fields where quantile estimation plays an important role we see a shift towards generalization. In this section we give some recommendations on the design of future experiments.

The median ($p = 50\%$) is the most commonly used measure of characteristic of the response curve. In some situations this estimation is of intrinsic interest, but more often it is because this quantile is the easiest to estimate (Wu [18]). Recently, several designs have been proposed for estimating quantiles where $10\% \leq p \leq 90\%$ (Wu [18], Stylianou and Flournoy [17]). The designs that are typically suggested are so-called adaptive or sequential designs where the velocity for a run is based on the response (perforation or no perforation) in the previous run(s). So in the current design used by Teijin Twaron is also adaptive (see Section 1). Except in the extreme tails of the quantile response function, the optimal design for estimating a particular quantile is a one-point design at the (unknown) target quantile (Ford *et al.* [10]). Hence, a good adaptive strategy should result in taking relatively much observations around the velocity V_p of interest. An adaptive strategy, that has been shown to work fine for values of p between 10% and 50% is discussed in section 7.1. We end this section with addressing the problems that arise when this probability of interest is small.

7.1 Adaptive Desigs

Stylianou and Flournoy [17] proposed an adaptive design called the up-and-down Biased Coin Design (BCD). Giovagnoli and Pintacuda [12] showed that the BCD is optimal within a large class of generalised up-and-down biased coin designs in the sense that the distribution of the velocities considered in the experiment is most peaked around V_p.

Before the BCD procedure starts performing the experiments, a collection of velocities of interest $\Omega = \{v_1 < v_2 < \ldots < v_K\}$ is set. The target velocity V_p should be in the range of Ω. In the first experiment a bullet is shot at velocity $v \in \Omega$. The velocity v may be fixed (e.g. the velocity that is thought to be closest to the target value V_p) or random. If the bullet perforated the vest, the next velocity to shoot with is one slower from Ω. However if the bullet did not perforate the vest, the procedure randomizes: Since we only consider cases where $p \leq 50\%$, the velocity becomes higher according to Ω with probability $p/(1-p)$ and with probability $(1-2p)/(1-p)$ the same velocity is used in the next shot. Appropriate adjustments need to be made at the lowest and highest velocities in Ω.

7.2 Small perforation probabilities

Not much is known about the design of experiments when the percentage p is smaller than 10%. A major problem is that the response is binary, which means that the amount of information that we gather each run is very small. Most of the bullets fired at velocities around V_p will be stopped for small values of p. However, some perforations for velocities around V_p are needed in order to estimate the probability of perforation at these velocities and eventually to help locating velocity V_p. Let us denote N_p^r as the number of shotes fired at the vest with velocity V_p until the r-th perforation occurs. Under the assumption that bullets are fired independently, this random variable has a negative binomial distribution with parameters p and r. The

probability distribution function is given by Equation (13).

$$P(N_p^r = n) = \begin{pmatrix} n-1 \\ r-1 \end{pmatrix} p^r (1-p)^{n-r}, \quad n \geq r. \tag{13}$$

The expection and the variance of N_p^r equals

$$E\left[N_p^r\right] = \frac{r}{p}, \quad Var\left[N_p^r\right] = \frac{r(1-p)}{p}.$$

Table 9 gives some details of the distribution of N_p^r for some small perforation probabilities p and different values of r. When the number of experiments is limited to 150 (or even 500) it will be difficult to locate the velocities V_p for small values of p using only the binary response. Using the depth of the perforation (e.g. the number of perforated layers) as a response variable may be a better way to gather more information from a single shot and to reduce the total number of shots required to determine V_p for small values of p.

p	r	$\mathrm{E}\left[N_p^r\right]$	$\sqrt{\mathrm{Var}\left(N_p^r\right)}$	95% Confidence Interval
5.0%	1	20	4.36	$[1; 59]$
	3	60	7.55	$[3; 124]$
1.0%	1	100	9.95	$[1; 299]$
	3	300	17.23	$[3; 628]$
0.1%	1	1000	31.61	$[1; 2995]$
	3	3000	54.75	$[3; 6294]$

Table 9: Some statistics of the negative-binomial distribution of N_p^r for different values of perforation probabilities p and number of perforations r.

8 Conclusions

In this paper we investigated the factors that influence the probability that a bullet perforates a bullet proof vest. Section 2 made clear that the velocity of the bullet and how many times you shoot at the vest are most important. This is something new for Teijin Twaron, since they assumped independence between the different shots with the shooting pattern they use. Therefore, we recommend to look into this phenomenon and investigate their shooting pattern. The data analysis also showed that the vest type is of less importance. This is a bit strange, since Teijin Twaron wants to compare different vest types by their perforation probabilities. It is, however, very well possible that the influence of the other aspects suppress the influence of the vest type.

8.1 Comparing Techniques to Estimate V_p

In the remainder of the paper we investigated the relationship between the velocity v and the probability of perforation for every vest type. In particular we have

developed several procedures to determine the velocity at which p percent of the bullets go through the vest (denoted by V_p). In the different methods a function $f(v)$ is established which describes this relationship.

When we would like to compare the different approaches that are proposed in this paper, we use the mean squared error as measure of fitness (see Section **??**). This measure can be computed for all data sets corresponding with different vest types. In total there are ten data sets. Table 10 shows the performance of the current model on all samples.

sample	size	MSE current method
0	126	0.166719
1	42	0.127233
2	42	0.162312
3	42	0.068530
4	42	0.145247
5	42	0.131702
6	42	0.120194
7	42	0.113723
8	42	0.121351
9	36	0.163257

Table 10: The MSE of the current model for the different vest types (or data sets).

The same can be done for the parametric approaches (GLMs) and the non-parametric approaches, presented in Table 11 and Table 12 respectively. In order to retrieve one number for the performance of a method, we looked at the deviation of each MSE with the lowest MSE of each sample and averaged this over all samples. The results are shown in Table 13.

sample	logit	alt. logit	probit	alt. probit	c log-log	alt. c log-log
0	0.165017	0.165014	0.165048	0.165092	0.165973	0.165647
1	0.126806	0.127032	0.127494	0.127774	0.125250	0.125344
2	0.161179	0.161406	0.161132	0.161374	0.159999	0.160042
3	0.068192	0.068333	0.069782	0.069937	0.068547	0.068493
4	0.141198	0.140954	0.141223	0.140913	0.144903	0.144382
5	0.131761	0.131928	0.132066	0.132292	0.131089	0.131081
6	0.121305	0.121472	0.120707	0.120875	0.119120	0.119242
7	0.111732	0.110691	0.114492	0.113428	0.123853	0.122826
8	0.117431	0.117986	0.118233	0.118817	0.112520	0.112991
9	0.155910	0.153498	0.164483	0.160903	0.179272	0.175300

Table 11: The MSE of the classical and alternative GLMs for the different vest types (or data sets).

Based on these results we see the smoothing spline technique to have the lowest average percentage deviation from the lowest MSE. Smoothing splines however tend

sample	smoothing spline	isotonic regression (stepwise)	isotonic regression (linear)	loss function
0	0.161896	0.173250	0.164316	0.157155
1	0.120854	0.154894	0.136643	0.127269
2	0.151425	0.212950	0.160480	0.167148
3	0.069995	0.099286	0.070823	0.064706
4	0.128712	0.137205	0.153609	0.155883
5	0.123416	0.190476	0.137557	0.148669
6	0.117319	0.160788	0.118012	0.135566
7	0.089776	0.090624	0.097533	0.090461
8	0.111329	0.125800	0.104811	0.111609
9	0.145758	0.160601	0.161300	0.140037

Table 12: The mean squared error as deviation measure from the real data for the different non-parametric approaches.

technique	average deviation (%)
current	12.78%
logit	8.94%
alt. logit	8.77%
probit	10.22%
alt. probit	9.97%
comp. log-log	11.45%
alt. comp. log-log	40.77%
smoothing spline	2.15%
isotonic regression (stepwise)	26.61%
isotonic regression (linear)	8.83%
loss function	8.01%

Table 13: The average deviation as percentage of the lowest MSE for each vest type

to perform better around the data points. Therefore, we recommend Teijin Twaron to use loss functions. Bu also the logistics model (logit model) and the isotonic regression approach with linear interpolation perform well. Especially when the confidence interval is of interest, we recommend the later two techniques.

The final technique we developed is a bootstrap method in which a particular characteristic at V_p is determined based upon conditional probabilities. A disadvantage of this procedure is that it will only work nicely for particular values of p ($p = 1\%$ and $p = 50\%$ work fine). This procedure will probably give the same results for $p = 1\%$ and $p = 10\%$. This is not likely to happen in reality.

8.2 Experimental Design

Besides the technique to determine V_p, Teijin Twaron also has to change their set-up of the experiments (the shootings at the vest). Their design is currently developed

to determine V_{50}. When they want to determine V_{01} (or any other V_p for a particular value of p) they need different data records. The data records to determine V_p should concentrate on the influence of the velocity on the perforation probability around V_p. Therefore, we propose a Biased Coin Design, that has been proven to work well in practice for values of p between 10% and 50%. If, however, Teijin Twaron wants to estimate V_{01}, they will not find a good predictor with this design where they only look at perforation of the vest of no perforation. They need to use other information as well, like the number of perforated layers. Otherwise, the number of experiment to perform becomes more than thousand.

References

[1] A. Agresti. *An introduction to categorical data analysis*. Wiley, New York, 1996.

[2] A. Agresti. *Categorical Data Analysis*. Wiley, New York, 2nd edition edition, 2002.

[3] R.E. Barlow, D.J. Bartholomew, J.M. Bremmer, and H.D. Brunk. *Statistical inference under order restrictions: the theory and application of isotonic regression*. Wiley, London, 1972.

[4] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and Regression Trees*. Chapman and Hall, New York, 1984.

[5] H.A. Chipman, E.I. George, and R.E. McCulloch. Bayesian cart model search. *Journal of the American Statistical Association*, 93(443):935–947, September 1998.

[6] H.A. Chipman, E.I. George, and R.E. McCulloch. Bayesian treed generalized linear models. *Bayesian Statistics*, 7, 2003.

[7] D.G.T. Denison, B.K. Mallick, and A.F.M. Smith. A bayesian cart algorithm. *Biometrika*, 85, 1998.

[8] C.W. Emmens. The dose/response relation for certain principles of the pituitary gland and of the serum andurine of pregnancy. *Journal of Endocrinology*, 2:194–225, 1940.

[9] R.L. Eubank. *Spline smoothing and nonparametric regression*. Dekker, New York, 1988.

[10] I. Ford, B. Torsney, and C.F.J. Wu. The use of a canonical form in the construction of locally optimal designs for nonlinear problems. *Journal of the Royal Statistical Society, Ser. B*, 54:569–583, 1992.

[11] J. Fox. *Non-parametric simple regression: smoothing scatterplots*. Sage, Thousand Oaks, 2000.

[12] A. Giovagnoli and N. Pintacuda. Properties of frequency distributions induced by general 'up-and-down' methods for estimating quantiles. *Journal of Statistical Planning Inference*, 74(1):51–63, October 1998.

[13] C.C. Holmes, D.G.T. Denison, and B.K. Mallick. Bayesian partitioning for classification and regression. *Technical Report, Imperial College London*, 1999.

[14] P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Chapman & Hall, London, 2nd edition edition, 1989.

[15] T. Robertson, F.T. Wright, and R.L. Dykstra. *Order restricted statistical inference*. John Wiley & Sons, New York, 1988.

[16] S.M. Ross. *Introduction to Probability Models*. Academic Press, San Diego, eighth edition edition, 2003.

[17] M. Stylianou and N. Flournoy. Dose finding using the biased coin up-and-down design and isotonic regression. *Biometrics*, 58(1):171–177, 2002.

[18] C.F.J. Wu. Efficient sequential designs with binary data. *Journal of American Statistical Association*, 80:974–984, 1985.