Discussion Paper

Reconstruction of a regression line from specifically censored data

Leon Willenborg

29 November 2025

This paper deals with a particular kind of regression problem, where the data have been censored in a specific way. Through some misunderstanding the original data cloud was partly obliterated: all data points above the regression line have been discarded. And also the regression line itself was lost. The question now is: can the regression line be reconstructed from the available, censored data and by using certain information about the original data? We assume that the original data were 'neat', that is without outliers and other irregularities, and where the lost regression line was obtained using OLS regression, fitted them well. In the paper various methods are presented to reconstruct the regression line. Some methods estimate slope and intercept of the regression line simultaneously. Other methods estimate these parameters separately. Some methods are direct, others are iterative. Some methods use statistical tests, others use imputed data. Some methods are non-parametric, others are parametric.

Keywords: linear regression, estimation, nonparametric methods, parametric methods, statistical testing, order statistics, spacings, linear programming, quadratic programming, computational geometry, missing data, imputation, iterative procedures, EM algorithm.

Contents

1 Introduction

In this paper we deal with a particular kind of regression problem. The aim is to reconstruct a regression line from a specifically censored data set. Through some misunderstanding the original data cloud was partly obliterated: all the data points above the regression line. And also the regression line itself was lost. The question now is: can the regression line be reconstructed from the censored data? Reconstructed in a statistical sense, that is, estimated from the available data and using certain information about the original data set. In general the answer to this question would be negative, as all sorts of data clouds could have been the original data, with various problematic aspects such as the presence of outliers, data with measurement errors, a sample that was severely skewed (due to e.g. selective response or due to a sample that severely over- or underrepresented certain subpopulations), etc. But for the sake of simplicity we assume this is not the case: we assume that the original data were 'neat' and the lost regression line fitted them well, in the sense of estimation by using OLS.[1)] In the paper various estimation techniques are presented with the aim of finding a satisfactory reconstruction of the original regression line.

This problem was dealt with in a note written at the Methodology Department at CBS in 1973 ([10]).[2)] This note in turn came to the notice of the present author and he also was captivated by the statistical problem addressed.[3)] The author was primarily interested in the problem, not so much in the solution provided in the note. Studying the problem over some period of time he came up with several methods to reconstruct the regression line. Most of these methods were nonparametric. Only one is parametric and it is the capstone of the present paper.

The paper is organized as follows. Techniques applied to the censored data are discussed in the main text. These techniques, when more generally applicable, are discussed for the complete data, in various appendices following the main text. For the complete data it is easy once the slope of the regression line has been estimated to derive the intercept, using a simple formula. However, in case of the censored data this is not possible. In this case estimates for the slope and the intercept have to be constructed separately.

However slope and intercept can be estimated simultaneously from the censored data by different methods. We consider two such approaches. They consider (candidate) regression lines 'floating on top' of the cloud of censored data. A regression line is then estimated which has minimal total distance to the points in the data set (or a specific subset thereof). Sometimes they can be estimated in a single 'sweep', but using a quite different approach. Therefore we have divided the main text into a part which is entirely devoted to estimating the slope of the regression line, and parts which, by a different approach, estimate the regression line 'on top' of

1) Ordinary Least Squares.

2) The problem originated from an experience of the author of that note. On the blackboard in his office, in the lower right-hand side corner, he had drawn a picture of a cloud of points together with a regression line. Because he wanted to keep this picure for some time he had written next to it: do not erase! This was a message for the cleaner who then apparantly also had the task of cleaning blackboards. But this person had misunderstood the message and thought that it only pertained to the information in the bottom right corner below the line. So everything above this line, as well as the line itself, was erased. The researcher quickly realised that an interesting statistical problem was created as a result of the misunderstanding of the message. He then set out to study it , come up with a solution and write this down in a note.

3) For which it is not easy to find a serious application. But it is interesting in its own right, as a statistical problem. Maybe some day a serious application will be found.

the censored data separately using techniques from operations research (OR). A (minor) problem encountered in this approach is that, without special precautions, an estimated regression line will intersect with at least one of the censored data points. If one wants to avoid that, certain precautions can be taken, two of which are discussed in the paper.

To give a more detailed picture of this paper, we describe it section by section. We start in Section 2 with a description of the data which is used, both the complete set as well as the censored part of this set. The description of the complete set is to exclude wild specimens (with outliers, dependencies between data points, missing data, etc.) which could prohibit the reconstruction of the regression line to a high degree of accuracy. The censoring of the data has occurred in a very special way, which, in a sense, enhances the reconstruction of the original regression line (in particular the slope), and also the missing part of the data cloud, in a statistical way. In Section 3 we discuss the basic statistical technique for this paper (namely OLS regression) applied to the censored data. In fact it is only used to estimate the slope of the regression line. This is actually only one way to estimate this parameter. In Section 4 we discuss several alternative slope estimators that are all distribution free. These estimators were first proposed by Wald and Theil, in different contexts. Some additional slope estimating methods are proposed in the same spirit as these Wald and Theil estimators, as well as inspired by Principal Components Analysis (PCA). We continue this theme of ‘fabricating’ new slope estimators by making variations on a theme. In Section 5 we propose some slope estimators that are obtained by using variations of a formula obtained in Section 3, involving inner products of vectors with part of the data. Therefore we call these estimators heuristic slope estimators. As it turns out, however, some of these heuristic estimators are actually part of an estimator inspired by PCA . So these estimators are not merely figments of the imagination.

To complement the estimators for the slope of the regression line, we consider estimators for the intercept in the next section, i.e. Section 6. We now exploit the fact that the data below the regression line are intact and have not been censored. In particular we are now interested in spacings, i.e. distances between subsequent data points. The idea is to look at bundles of lines with the same slope above the observed data, and choose the one with the ‘right distance’ to it. This ‘right distance’ is obtained from certain spacings of the censored data, i.e. the observed data, which are situated below the regression line. Several possibilities are considered. Now we consider different kinds of estimators from the ones considered to far. We look at lines above the data (implying certain constraints for the parameters of these lines) and consider the vertical distances of these lines in the observed x values (as in case of ordinary least squares (OLS) regression).

Then, in Section 7, we consider the reconstruction of the regression line by means of constrained optimization techniques, namely linear programming (LP) and quadratic programming (QP). The object functions used are linear (in the LP case) and quadratic (in the QP case) and involve all censored data. The constraints in both cases are the same and are linear. Their role is to make sure that the regression line to be reconstructed is situated above the censored data. The regression lines are allowed to cut the top of the boundary of the avoidance area $\mathcal{A}$, called the roof. So in both approaches the optimal regression line always cuts at least one point of the roof (and hence of the censored data set). This may be somewhat undesirable as the regression line to be reconstructed should be allowed to avoid all censored data points, albeit narrowly. But in most cases this would require only marginal adjustments. And the remedies are easy to implement. Therefore we do not bother too much about this issue.

Next, in Section 8, a parametric approach is sketched to the problem of reconstructing a regression line from the censored data, making use of maximum likelihood estimation and imputation of missing observations. In this section an iterative algorithm is sketched that alternatively estimates model parameters on the one hand and imputes missing data, using the most recent parameter estimates. The algorithm presented is in the same spirit as the EM-algorithm (see [3]) but is different in the way it is applied.

Section 9 contains a discussion of the main results obtained in the present paper. And also it offers some suggestions as to what could be (further) investigated by someone interested in the problem considered in the present paper. An obvious goal would be to apply the methods suggested here to some data sets and compare the results.

The main text is completed with a list of references, pointing to relevant background information. The paper is supplemented by eight appendices, on a variety of subjects. Some of them introduce certain estimators in a 'normal' context, that is, when applied to a complete data set and not the (specifically) censored subset we consider. They will be applied to the censored data set, if possible, and with cautions and restrictions: the slope estimators can be used for this, but not the intercept estimators (or the intercept estimator derived from the slope estimator by using a simple formula as in case of OLS). Other appendices are about the roof (its definition and computation) , about certain statistical tests, about spacings, about barrier functions. Together they take up quite some space. Deliberately we kept all material not directly related to making estimates for the lost regression line based on the censored data out of the main text and relegated it to the appendices.

2 Remarks on the data, complete and censored

The aim of the present section is to describe the data that we are dealing with. We have the original data on the one hand and a subcloud of observed data on the other hand. Part of the original data is censored, namely all points above the regression line that was fitted using OLS applied to the original data. For simplicity we assume that the original data set is 'well-behaved' in the sense that it does not contain outliers, partially missing data, data points with gross observation errors, etc. For if this would be the case our goal could not be possible as relevant information would be missing and cannot be estimated from the observed data.

It should be noted that the special censoring applied to the data is actually helpful in estimating the slope of the original regression line. This is so because the regression line in the original data set can be viewed as a separator between two parts of the data cloud, which in a sense can be viewed as a kind of mirror images of each other. This implies that observing one of them gives a lot of information about the missing part of the data cloud and also about the missing regression line that we want to estimate.

The points of the observed part of the data cloud nearest to the regression line (and situated at its roof[4)]) are close to being situated on a line, in particular if the original data cloud contained

4) See Appendix F, where the roof of the censored data set is discussed and also how to compute it.

many points, and hence also the observed part, approximately half this number.

As a result of the particular type of censoring applied to the original data, one should expect to provide good estimates of the slope of the original regression line, provided there is enough data in the censored data set. For the complete data set using OLS this would imply that once the slope of the regression line is known, the intercept follows immediately. A simple formula yields an estimate of the intercept as a function of the estimate of the slope.[5)] However for the censored data the intercept has to be estimated separately from the slope. In this case one knows that the regression line is situated above the censored data.

Therefore we separate the estimation of the slope of the regression line from that of its intercept for the censored data. In fact, estimating both parameters on the basis of censored data can generally be expected to work well, because of the special censuring of the data, based on the regression line estimated from the original, complete data.

3 OLS slope estimator for the censored data

In the present section we discuss the application of OLS to the censored data set. OLS applied to the original data is discussed in Appendix A.

We can apply OLS to the censored data, but only to estimate the slope of the regression line. The estimate for the intercept based on these data that this method yields is useless, because the regression line is not supposed to separate the censored data, but should lie entirely above those data. However, due to the special type of censoring the censored data should be useful to estimate the slope of the original regression line.

As in the case of complete data, for the censored data it is useful to shift its center of gravity to the origin $\mathcal{O}$ of the coordinate axes , as it simplifies computations and the slope is not affected by this transformation.

We now consider the censored data. We indicate the observed (and shifted) data by a superscript 'w' . So we have x^w and y^w instead of x and y in Appendix A, satisfying $\langle x^w, \iota \rangle = 0$ and $\langle y^w, \iota \rangle = 0$ because the center of gravity of the censored cloud was shifted.[6)] Hence $\bar{x}^w = 0$ and $\bar{y}^w = 0$. As we assume that the censored cloud has n points, x, y and ι have length n, that is, they consist of n components.[7)]

When censored data are used the question is how well the vectors x^w and y^w, which are the equivalents of the vectors x and y in case of the complete data, are good 'proxies' of these vectors.

5) It would be possible to use this formula to estimate the intercept using an estmate of the slope not based on OLS.

6) Where $\iota = (1, \ldots, 1)$ is the all one vector of length N.

7) In order not to overburden the formulas below we have not used superscripts w for the components of x^w and y^w. However, below we write ξ^w to distinguish it explicitly from the variable ξ defined in (A.21) which applies to the original data cloud. For the same reason we write $\bar{x}^w$ and $\bar{y}^w$ to distinguish from $\bar{x}$ and $\bar{y}$, defined in (A.10) and (A.11), also applying to the original cloud.

We focus in the present section on the slope parameter b of the regression line. So the question is how well the estimate

$$\hat{b}^w = \frac{\langle x^w, y^w \rangle}{\langle x^w, x^w \rangle} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} (x_i)^2} = \sum_{i \in NZ_n} \xi_i^w \frac{y_i}{x_i}, \tag{1}$$

with $NZ_n = \{i \in V_n \mid x_i \neq 0\}$, where $V_n = \{1 \ldots, n\}$, approximates $\hat{b}$ in (A.16). The ξ_i^w are defined in analogy to the ξ_i in (A.21) as

$$\xi_i^w = \frac{x_i^2}{\sum_{j=1}^{n} x_j^2}. \tag{2}$$

Looking at (1) one would expect the estimate $\hat{b}^w$ to be too small, as the observed ($=$ non-censored) y-values tend to be smaller than the y values of the data in the original, complete set, for comparable x values. But the only way to assess the quality of this estimate is to apply the estimation method to (real) data and see what estimates are actually produced.

We shall delve deeper into the problem of finding acceptable (or perhaps even satisfactory) estimates for the slope b, as well as for the intercept a, in the sequel.

4 Alternative slope estimators for the censored data

The methods discussed in the present section are more tailor made and less conventional than the ones discussed in the Section 3. Instead of estimating the intercept and the direction simultaneously as in the OLS approach, here only the slope estimator is estimated from the censored data.

To estimate the slope for the censored data two methods proposed by Wald are considered in Sections 4.1 and 4.2. These methods are introduced for the complete data set in Appendix B. In Section 4.3 a similar method due to Theil is presented. In Section 4.4 an iterative method to estimate the slope for the censored data using statistical tests is presented. In Section 4.5 the PCA[8)] method is applied to the censored data to find a slope estimator. The PCA method is explained in Appendix D where it is applied to the original, uncensored data. A few other geometric methods are proposed in Section 4.6, but not elaborated. They are intended as suggestions to be explored elsewhere.

8) PCA = Principal Component Analysis.

The idea is to apply these methods to (real) data and to compare the results, rather than to make judgements about their merits (or demerits) on the basis of symbolic expressions, i.c. the formulas, only.

We present the estimators first for the complete, original data set, as it is the ideal case. Then the estimators for the censored data set are stated. They are basically applied the same way as in case of the completely data, but this time to the censored data. As in Section 3 we use the superscript w to indicate that we are dealing with the observed (and shifted) data set.

4.1 Wald's estimator for two groups

Adaptation of the method described in Appendix B.1.1 to the censored data set is straightforward. Now the groups G_1^w and G_2^w satisfy $G_1^w \leq G_2^w$ and $G_1^w \cup G_2^w = V_n$, where $V_n = \{1, \dots, n\}$ is the set labelling the observed points. Correspondingly, the centers of gravity of G_1^w and G_2^w are

$$z_i^w = \frac{1}{|G_i^w|} \sum_{j \in G_i^w} p_j, \tag{3}$$

for $i = 1, 2$. We take the vector (z_1^w, z_2^w) as an estimator of the slope of the original regression line, defined for the complete data set.

Not only is the adaptation of the method to the censored data set straightforward, it is also to be expected that this results in a good estimate of the slope, due to the nature of the censoring: the roof should give a good idea of the slope of the original regression line.

If there are enough data in the censored data set one may find a variant of this method useful, by restricting the points to be considered to the roof of the observed data. As was already remarked elsewhere the roof can be expected to give a good idea of the slope of the original regression line.

4.2 Wald's estimator for three groups

Adapting the method explained in Appendix B.1.2 to the uncensored part of the data cloud is also straightforward, as was the case in Section (B.1.1). We now have three groups H_1^w, H_2^w and H_3^w of about equal size, such that $H_1^w \leq H_2^w \leq H_3^w$. Let ζ_1^w and ζ_3^w be the centers of gravity of H_1^w, H_3^w and use (ζ_1^w, ζ_3^w) as the estimator of the slope of the original regression line for the complete data cloud.

4.3 Theil's slope estimator

Adapting the method in Section B.2 to the censored data cloud is trivial and requires only new notation but no new ideas. Instead of S_{ij} as defined in (B.2), we now have

$$S_{ij}^{w} = \frac{y_j - y_i}{x_j - x_i}, \tag{4}$$

for $i < j$ with $i, j \in V_n = \{1, \dots, n\}$. So the slopes are defined for the points in the censored data cloud only.

If there are enough data points, one could consider a variant of this method by picking only the data of the roof of the censored data set. One can expect that the roof of a sizeable data set (of observed data), being close to the original regression line, should produce a good estimate of the slope of the original regression line. The only problem is how close a candidate regression line is supposed to be to the roof. If there is not too much room, one can expect a rather good estimate.[9)]

4.4 Estimating the slope iteratively, using statistical testing

In Appendix G some distribution free tests are presented (taken from [4]) concerning the fit of a regression line with slope b_0 to the data. We will use two of these to compute an estimate for the slope of a regression line iteratively. It starts with some inital estimate and proceeds to find a better one through stepwise improvement. For details about the tests, in particular about Kendall's statistic, see Appendix G and possibly also [4], which was the source of the material in the appendix.

The tests we are interested in are the following ones in Appendix G: two tests concerning the slope of a regression line to test whether a regression line with slope b_0 fits that data or not:

$$H_0 : b = b_0, \tag{5}$$
$$H_1 : b > b_0, \tag{6}$$

with acceptance criteria:

$$\text{accept } H_0 \text{ if } C < k(\alpha, n), \tag{7}$$
$$\text{accept } H_1 \text{ if } C \geq k(\alpha, n), \tag{8}$$

and

9) In the original data the regression line can be viewed as a separator between the (future) observed part of the data (below the original regression line) and the (future) unobserved part of the data (above the original regression line). In case of a sizeable amount of data one can expect that this separator has only a very small 'channel' into which it can fit. But with only the observed data we do not have such a channel.

$$H_0 : b = b_0, \tag{9}$$
$$H_2 : b < b_0, \tag{10}$$

with acceptance criteria:

$$\text{accept } H_0 \text{ if } C > -k(\alpha, n), \tag{11}$$
$$\text{accept } H_2 \text{ if } C \leq -k(\alpha, n). \tag{12}$$

We can use these test criteria to improve the initial estimate of the slope b by b_0. We first pick a suitable acceptance parameter α, compute $k(\alpha, n)$ and quantities D_i as defined in (G.1), the test statistic C as defined in (G.2) and apply both tests.

The following possibilities may occur:

- If H_0 is accepted by both tests we are done: b_0 is the slope we are looking for, depending on the acceptance parameter α that was chosen.
- If H_1 is accepted then the slope should be bigger than b_0. We then take a $b_1 > b_0$ and apply the procedure again, starting with recomputing all the abovementioned parameters for b_1 instead of b_0.
- If H_2 is accepted then the slope should be smaller than b_0. We now take a $b_1 < b_0$ and apply the procedure again, starting with recomputing all the abovementioned parameters for b_1 instead of b_0.

Ideally this procedure should stop at some point when for some $b = b_f$ the hypothesis H_0 is accepted. Ideally, because this (also) requires the right steps to be taken to the next choice for the parameter b. If these steps are too small, it may take a long time to reach the final state (H_0 accepted). However, if these steps are too big then there may be no convergence but rather some kind of oscillating, alternating between slopes that are too small or too big.

How much bigger or smaller b_1 should be taken than b_0 is hard to tell in abstracto. In practice one should try some values. The challenge is to find a step from b_0 to b_1 that is just right, neither too big nor too small. But the testing approach is intended to indicate in which direction to move to improve the slope.

4.5 PCA inspired approach

In the present section we apply the machinery of Principal Components Analysis (PCA) to the censored data in order to find an estimate of the slope of the original regression line. The idea is to use the direction of the main axis of the covariance matrix of the censored data set. In Appendix D the computations needed to obtain the eigenvalues and corresponding eigenspaces for the original data set are presented. In the present section we therefore can focus on the censored data set and adapt the results in Appendix D to these data.

The censored data in matrix form is the following $n \times 2$ matrix:

$$D^w = \begin{pmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{pmatrix}, \tag{13}$$

with the n observed data points.

From (13) we compute the covariance matrix for the censored data:

$$\mathcal{C}^w = (D^w)' D^w = \begin{pmatrix} f^w & g^w \\ g^w & h^w \end{pmatrix}, \tag{14}$$

where

$$\begin{aligned} f^w &= \langle x^w, x^w \rangle = ||x^w||^2, \\ g^w &= \langle x^w, y^w \rangle, \\ h^w &= \langle y^w, y^w \rangle = ||y^w||^2. \end{aligned} \tag{15}$$

$\mathcal{C}^w$ is of similar structure as $\mathcal{C}$, the coveriance matrix for the full data set, defined in (D.2) in Appendix D, in the sense that it also is a nonnegative symmetric matrix of order 2, with real nonnegative eigenvalues and orthogonal eigenspaces corresponding to different eigenvalues. In this case we have two eigenvalues and hence two 1D eigenspaces, represented by single vectors. Similar to the case of $\mathcal{C}$ we have that the eigenvalues of $\mathcal{C}^w$ are the roots of the characteristic polynomial with a similar structure as (D.3) but with different coefficients:

$$\det\left(\mathcal{C}^w - \lambda\, I_2\right) = \lambda^2 - (f^w + h^w)\,\lambda + f^w h^w - (g^w)^2 = 0, \tag{16}$$

where I_2 is the identity matrix of order 2. The roots of (16), that is, the solutions of $\det\left(\mathcal{C}^w - \lambda\, I_2\right) = 0$, which are also the eigenvalues of $\mathcal{C}_n$, can be expressed as

$$\lambda_\pm^w = \tfrac{1}{2}\,(f^w + h^w) \pm \tfrac{1}{2}\sqrt{(f^w - h^w)^2 + 4(g^w)^2}, \tag{17}$$

analogous to (D.7) in Appendix D. So the largest eigenvalue of $\mathcal{C}^w$ is

$$\lambda_+^w = \tfrac{1}{2}\,(f^w + h^w) + \tfrac{1}{2}\sqrt{(f^w - h^w)^2 + 4(g^w)^2}, \tag{18}$$

with corresponding eigenvector

$$v_{\lambda_+^w} = \left[\begin{pmatrix} 1 \\ (\lambda_+^w - f^w)/g^w \end{pmatrix}\right] = \left[\begin{pmatrix} g^w \\ \lambda_+^w - f^w \end{pmatrix}\right], \tag{19}$$

provided $g^w \neq 0$. $[v]$ represents the linear space generated by the vector $v \in \mathbb{R}^2$ as is defined in Appendix D.

So the tangent of the line defined in (19) equals $(\lambda_+^w - f^w)/g^w$, which can be viewed as an estimate of the slope of the regression line for the censored data, in anology to the result for the complete data set, that is (D.10). We find

$$\begin{aligned}\frac{\lambda_+^w - f^w}{g^w} &= \frac{1}{2}\left(\frac{h^w}{g^w} - \frac{f^w}{g^w} + \sqrt{\left(\frac{h^w}{g^w} - \frac{f^w}{g^w}\right)^2 + 4}\right) \\ &= \frac{1}{2}\left(\frac{\langle y^w, y^w\rangle}{\langle x^w, y^w\rangle} - \frac{\langle x^w, x^w\rangle}{\langle x^w, y^w\rangle} + \sqrt{\left(\frac{\langle y^w, y^w\rangle}{\langle x^w, y^w\rangle} - \frac{\langle x^w, x^w\rangle}{\langle x^w, y^w\rangle}\right)^2 + 4}\right).\end{aligned} \tag{20}$$

Adapting a result from Appendix D (namely (D.11)) to the case of the partially observed data cloud as considered here, we have:

$$\frac{\lambda_+^w - f^w}{g^w} \approx \begin{cases} h^w/g^w & \text{if} \quad f^w \ll h^w, \\ 1 & \text{if} \quad f^w \approx h^w, \\ 0 & \text{if} \quad f^w \gg h^w. \end{cases} \tag{21}$$

which can also be expressed in more easily understandable notation as

$$\hat{\hat{b}}^w \approx \begin{cases} \langle y^w, y^w\rangle/\langle x^w, y^w\rangle & \text{if} \quad ||x^w|| \ll ||y^w||, \\ 1 & \text{if} \quad ||x^w|| \approx ||y^w||, \\ 0 & \text{if} \quad ||x^w|| \gg ||y^w||. \end{cases} \tag{22}$$

It is interesting to compare the eigenvectors obtained from the censored cloud with those of the original cloud theoretically (using results from section H) and numerically (on the basis of simulation studies), to get an idea of the error involved.

To finish the present discussion we want to point out that another variant to estimate the slope may be of interest, namely one where the data used consist of the roof data only, which, obviously, form a subset of the censored data. The points on the roof are close to the original regression line. In case of an elongated cloud with quite some points one can expect that these points yield a very good PCA inspired estimator of the slope.[10)] Even in case the original data cloud was close to a ball like shape, the censuring by the original regression line would in fact help in restoring it using this method of estimation, using the roof data only. Of course there are fewer of them generally, but that also reduces ‘noise’ when estimating the slope using the censored data, caused by the non-roof data in the censored data set.

So instead of (13) we have our matrix with roof data

[10)] One could even consider an OLS estimator to be fitted to these data, and use only the slope estimate.

$$D^r = \begin{pmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_{n_1} & y_{n_r} \end{pmatrix} \tag{23}$$

as the basis for estimates.

As above, we can estimate the roof equivalents of (20) and (21). We shall not explicitly spell out the results, as in symbolic form they are clear. What is interesting is to compare the numerical results of the estimates in simulations, for the original data, the censured data and the roof data.

4.6 Slope estimators based on other geometric ideas

In this section we present a few ideas that could be further explored to see if they produce useful estimates of the missing regression line. These ideas are geometric in nature and require computational geometry to apply them. Some of them are just suggestions, others are developed a bit more.

4.6.1 Diameter of convex hull and implied direction

The idea is to look at the convex hull of the censored data and determine its diameter. In particular the extreme points in the convex hull (and hence of the censored data set itself) determine a direction that can be used as an estimate for the slope of the original regression line.[11] We do not want elaborate this type of problems here as they are studies in the area of computational geometry. Results about diameters of polygons and finite sets of points and about convex hulls of such sets, in the plane, can be found in e.g. [8], Chapters 3 and 4. We only want to point out that slope estimators can be found using these kind of results from computational geometry.

4.6.2 Smallest rectangle containing the censored data

In Section 4.5 we considered a method inspired by PCA. Here we consider another method inspired by this technique. To motivate it we quote from the Wikipedia lemma on PCA:[12] ”PCA can be thought of as fitting a p-dimensional ellipsoid to the data, where each axis of the ellipsoid represents a principal component.” In our case $p = 2$ The axes (‘the principal axes’) are orthogonal. The idea we propose is the following: instead of fitting an ellipsoid to the data cloud the idea is to find the smallest rectangle containing it.[13]

What ‘minimum rectangle’ is supposed to mean has to be specified, of course. Various options are possible, such as ‘minimum area’ or ‘minimum circumference’. What operationalization to use depends on the results obtained (are they in tune with our expectations?) and how easy it is to compute solutions for these methods. It is also of interest to find out how different these solutions are and which solution yields the intuitively best fitting box.

11) That one such pair of points exist is the ideal case. In case several such pairs exist there are several directions, which can be averaged to yield a slope estimate.

12) `https://en.wikipedia.org/wiki/Principal_component_analysis`.

13) A refinement would be to look for such a rectangle that contains most of the data. This is to avoid that outliers mess things up.

The purpose of this exercise is similar as that of PCA (see Section 4.5), namely to find an estimate of the direction of the data cloud, which will be used as an estimate for the regression line. This corresponds to the direction of the longest side of the fitted rectangle.[14)]

We shall stop here discussing this approach, as it would become quite technical at this point. Elaboration should be done elsewhere.

4.6.3 Using the roof points to estimate the slope with OLS

The linear regression line can be viewed as a separator of the original data. So one would expect the data nearest to the regression line to provide a lot of information about the slope of this line, especially if there is a tight fit. The points we mean here are the roof points. In Appendix F the roof is discussed, as well as a method how to compute it for a given set of censored data. The roof points are part of this set. They are roughly the points at the top of the censored data.

A simple, but possibly good way to estimate the slope of the original regression line is the following. Take the roof points only and apply OLS to estimate a regression line. The expectation is that the slope of this regression line should be rather similar to the slope of the original regression line. This should certainly be the case if the original data set was large.

As before ((1)) we first shift the roof data $(x_1^r, y_1^r), \ldots, (x_\mathfrak{r}^r, y_\mathfrak{r}^r)$ in $\mathfrak{R}$ (see Section 7) so that their center of gravity coincides with the origin of the coordinate system $\mathcal{O}$. Hence we may assume that the averages $\bar{x}^r = \bar{y}^r = 0$ for the roof points. Then the intercept of the regression line $\hat{a}^r = 0$ and the slope

$$\hat{b}^r = \frac{\langle x^r, y^r \rangle}{\langle x^r, x^r \rangle}, \tag{24}$$

where $x^r = (x_1^r, \ldots, x_\mathfrak{r}^r)$ and $y^r = (y_1^r, \ldots, y_\mathfrak{r}^r)$ so that $\langle \bar{x}^r, \iota \rangle = 0$ and $\langle \bar{y}^r, \iota \rangle = 0$. It is interesting to find out how well $\hat{b}^r$ performs in practice as an estimator of the slope of the original regression line, compared to its rivals presented in this paper, in particular the one presented in Section 4.6.4.

4.6.4 Using only the roof points to fit a regression line on top

Here we present a variant of the method suggested in Section 4.6.3 But it will be elaborated in Section 7.3 where it better fits. In the case we consider in this paper we only have the data situated below the (unknown) regression line. So if we would fit a line above these data, one would expect to have a very good idea of the slope of the original regression line. In fact one then has to consider only the data at the top of the observed data cloud, which we call the roof points (see Appendix F). The approach suggested here is elaborated in Section 7.3. This method, as the one in Section 7.1 (and in Section 7.2), can in fact be used not only to estimate the slope of the regression line but also its intercept. So it is more versatile than the method described in Section 4.6.3.

[14)] If the rectangle would turn out to be a square there is no question of an elongated shape of the cloud and hence there is no pronounced direction of this cloud.

5 Heuristic slope estimators for the censored data

In Appendix C some estimators were derived by using heuristic ideas to find variations of the slope estimator in case of applying OLS to the entire data set. This estimator is $\hat{b}_1$ in Appendix C. There we derive five more estimators of b, denoted as $\hat{b}_2, \ldots, \hat{b}_6$, one of which is an iterative estimator. In the present section we want to apply them to the observed data set. This requires modifications in notation, but no new ideas. We now obtain the estimators $\hat{b}_2^w, \ldots, \hat{b}_6^w$, which are the analogues of the aforementioned estimators, this time to be applied to the observed data set.

The alternative estimators of the parameter b of the slope of the regression line as given in (A.16) are obtained by 'playing with the algebraic form of the estimator'. We shall assume that the center of gravity of the censored data cloud has been shifted to the origin $\mathcal{O}$ of the axis system used. This leaves b as the only parameter of the regression line to bother about, as the intercept $a = 0$.

5.1 Parade of slope estimators

In Appendix C a set of heuristic estimators for the slope b is derived from the estimator (A.16). These estimators are derived for the original data set. In the present section we present their counterparts which are applied to the censored data to yield estimators that can actually be used to find estimates for the slope. The adaptations are straightforward. We parade the estimators here, without further introduction (for which is referred to Appendix C). Throughout this section we assume $\langle x^w, x^w \rangle > 0$ and $\langle y^w, y^w \rangle > 0$.

We define the first slope estimator as:

$$\hat{b}_1^w = \frac{\langle x^w, y^w \rangle}{\langle x^w, x^w \rangle}. \tag{25}$$

If $\langle x^w, y^w \rangle \neq 0$, we define the second slope estimator as:

$$\hat{b}_2^w = \frac{\langle y^w, y^w \rangle}{\langle x^w, y^w \rangle}. \tag{26}$$

The third slope estimator is defined as:

$$\hat{b}_3^w = \begin{cases} \sqrt{\frac{\langle y^w, y^w \rangle}{\langle x^w, x^w \rangle}} & \text{if } \langle x^w, y^w \rangle > 0, \\ -\sqrt{\frac{\langle y^w, y^w \rangle}{\langle x^w, x^w \rangle}} & \text{if } \langle x^w, y^w \rangle < 0. \end{cases}$$

If $\langle x^w, y^w \rangle \neq 0$, we define the fourth slope estimator as:

$$\hat{b}_4^w = \frac{1}{2}\left(\frac{\langle x^w, y^w\rangle}{\langle x^w, x^w\rangle} + \frac{\langle y^w, y^w\rangle}{\langle x^w, y^w\rangle}\right). \tag{27}$$

If $\tilde{b}_1^w \geq 0$ and $\tilde{b}_2^w \geq 0$, we define the fifth slope estimator as:

$$\hat{b}_5^w = M(\hat{b}_1^w, \hat{b}_2^w), \tag{28}$$

where M is the AGM function (see Appendix C).

The sixth, and final, heuristic slope estimator we present is:

$$\hat{b}_6^w = \frac{\langle g^w, y^w\rangle}{\langle g^w, x^w\rangle}, \tag{29}$$

where $g^w \neq \kappa\, \iota_n$ ($\kappa > 0$ constant) is a weight vector of length n, with ι_n the all one vector of length n .

No claim is made about the quality of any of these estimators. How they perform in practice should be found out by applying them to numerical data. It should be noted, however, that they are not entirely ‘out of the blue’. Some appear in (20), which is in the context of the PCA inspired approach.

6 Intercept estimators for the censored data

Once the direction of the original regression line $\hat{b}^w$ has been estimated from the censured data, then its intercept a has to be estimated from these same data. We know that the original regression line is situated above these data, so we cannot use the OLS approach for this. But we can proceed differently as we will show in the present section. We first consider a horizontal regression line in Section 6.1 and then the case with a sloped regression line in Section 6.2.

6.1 Horizontal regression line

Assuming that the regression line is horizontal, we only have to deal with its intercept a. We first derive a simple estimate of this parameter. This may be somewhat naive as it allows the regression point to cut a point in the censored data. We can easily mend this shortcoming by adding a (usually small) value based on spacings derived from the order statistics of the y values of the observed points.

So let us assume a constant regression line:

$$y = y(x) = a, \tag{30}$$

for all x in the domain, where a is the unknown intercept that has to be estimated. We only know about a that it should be larger than any y value in the censored data set, because the regression line is above all data points in the censored data set. Hence

$$\max_{i \in V_n} y_i \leq a, \tag{31}$$

with $V_n = \{1, \dots, n\}$. So a straightforward estimate of a is:

$$\hat{a}^0 = \max_{i \in V_n} y_i = y_{(n)}, \tag{32}$$

where $V_n = \{1, \dots, n\}$ and where:

$$y_{(1)} \leq \cdots \leq y_{(n)}, \tag{33}$$

where $y_{(1)}$ is the smallest and $y_{(n)}$ the largest of the values $y_1, \dots, y_n$, with other order statistics $y_{(m)}$, $1 < m < n$, in between.

Can $\hat{a}^0$ be improved (a little bit)? To answer this question we can look at the ordered y values in the censored data set. To simplify the discussion of this problem[15)] we assume that there are no ties in the y-values. So instead of (33) we actually have that the stronger set of inequalities holds:

$$y_{(1)} < \cdots < y_{(n)} \tag{34}$$

Actually this is too strong. We could settle for less, only requiring that $y_{(n-1)} < y_{(n)}$.[16)] But there is no reason to bother about the minimal conditions required. Even the bolder one (34) can be simply made to hold by perturbing some y values slightly, if necessary.

In particular at the largest y value in the censored data set, $y_{(n)}$ and the next largest y value, $y_{(n-1)}$ say. The smallest y value of the point in the missing part of the complete data set could be estimated by $y_{(n)} + y_{(n)} - y_{(n-1)} = 2y_{(n)} - y_{(n-1)}$ It seems reasonable to assume that the original regression line was situated halfway between the largest y value in the censored data and the smallest y value in the missing data. So[17)] an improved estimate of the intercept a is

15) Without essentially simplifying things. The theory can cope with situation with ties in the y values, but describing it requires a bit more text. In case of ties, we have to be careful in distinguishing various situations in order to avoid using intervals consisting of a single point.

16) And possibly also $y_{(n-2)} < y_{(n-1)}$. And no further requirements for the remaining order statistics.

17) We assume that the spacing in the y of the observed and the missing part of the original data have the same distribution. There is no reason to assume otherwise.

$$\hat{a}^1 = y_{(n)} + \frac{y_{(n)} - y_{(n-1)}}{2} = \tfrac{3}{2}y_{(n)} - \tfrac{1}{2}y_{(n-1)}. \tag{35}$$

Of course, other improved estimates would be possible by taking more spacings into account, such as $y_{(n-1)} - y_{(n-2)}$, $y_{(n-2)} - y_{(n-3)}$, etc. to estimate the gap between the censored data cloud and the missing data cloud. An example of an alternative adjusted intercept estimator is the following one:

$$\hat{a}^2 = y_{(n)} + \tfrac{1}{2}\frac{y_{(n)} - y_{(n-2)}}{2} = \tfrac{5}{4}y_{(n)} - \tfrac{1}{4}y_{(n-2)}. \tag{36}$$

But as the matter of correcting the slope (a bit) by using some spacings information concerns a minor issue, we leave its discussion at this point.

6.2 Sloped regression line

We now consider the more general case of a sloped regression line, assuming that the slope has been estimated (in some way) from the censured data. This given estimate is denoted by $\hat{b}^w$. The aim is now to estimate the intercept a for these data.

We consider the equation for the regression line

$$y = a^w + \hat{b}^w x, \tag{37}$$

where $\hat{b}^w$ is given and a^w is to be estimated. If we evaluate this at each data point (x_i, y_i) of the censored data and rearrange we get

$$a_i^w = \hat{b}^w x_i - y_i, \tag{38}$$

where the a_i^w are the kind of data that are considered in Section 6.1. So an estimate of a_i^w is, with an eye to Section (6.1):

$$a_{(n)}^w = \max_{i=1,\dots,n} a_i^w, \tag{39}$$

where the a_i^w are given by (38). Then the estimated regression line is

$$y = a_{(n)}^w + \hat{b}^w x, \tag{40}$$

with the estimated intercept $a_{(n)}^{w}$ as in (39).[18)]

Now, as in Section (6.1) we can add a (small) value to make sure that the reconstructed regression line is slightly above the highest point in the censored data set (when considering the bundle of lines in the plane with slope $\hat{b}^{w}$. In this case we have to look at the order statistics $a_{(1)}^{w} \leq \cdots \leq a_{(n)}^{w}$ for the values a_{i}^{w} as given is (38). For the regression line (40) we know that all data from the censored data are situated below it. The line with the smallest intercept with this property (and the one defined in (40)) is likely to contain only one point of this data set. There may be more than one point on it but that would be purely coincidental.

As in case of a horizontal regression line we can add a small value to make sure that the estimated (reconstructed) regression line is strictly above each point in the censored data set. We then have to look at the set of a_{i}^{w}, $i = 1, \ldots, n$ and in particular their spacings. For convenience (again, as before), we assume that holds $a_{(1)}^{w} < \cdots < a_{(n)}^{w}$. Looking at (35) we could use

$$a_{s}^{w} = \frac{3}{2} a_{(n)}^{w} - \frac{1}{2} a_{(n-1)}^{w} \tag{41}$$

as an estimate of the intercept, obtaining as the reconstructed regression line

$$y = a_{s}^{w} + \hat{b}^{w} x. \tag{42}$$

As remarked in Section (6.1) alternative correctors for the intercept, using more neighbouring rank statistics, could be defined. An example is shown in Section 6.1 (see (36)), which can be easily modified to apply to the case of a sloped regression line in the present section.

Remark Ideally the a_{i}^{w} as defined in (38) should be independent of i. One can test whether this is true (or not) by ordering the a_{i}^{w} as $a_{(1)}^{w} \leq \ldots \leq a_{(n)}^{w}$ and investigate whether the indices i and the ranks (i) are randomly linked (or not). This is actually a test of the randomness of a permutation. (See [5] for an interesting discussion on randomness of sequences and permutations in the context of algorithms to generate random numbers). If there is a (nonrandom) link between indices and ranks then the reconstructed regression line is not a good fit to the data. ☐

7 LP and QP models

In the present section we consider two formulations for finding a regression line fitting the censored data. These formulations differ in terms of the object functions used: a linear function of the residues or a quadratic one. The constraints in both cases are the same and derive from

[18)] Another way to arrive at the same result follows from the observation that the regression line to be derived passes through a point in the uncensored data. This regression line has the property that all data in the censored data should lie below it. So to find it we can determine, for each point in the censored data set, the line with slope $\hat{b}$ that passes through it, and pick from all these lines the one with the property mentioned. The intercept for this line is precisely the intercept given in (39).

the requirement that the regression line should be above all censored data. The formulations lead to a linear programming (LP) problem in the first case and a quadratic programming (QP) problem in the second to find a solution for the intercept a and the slope b of the regression line. In Section 7.1 we consider the LP problem and in Section 7.2 the QP problem. In both approaches all censored data are used to contribute to the obsject function, whereas only a subset of these data (consisting of the roof points) are used to define the linear constraints. These constraints guarantee that the regression line that is reconstructed is situated above the censored data, as should be the case.

7.1 LP regression for the censored data

In the present section we consider the problem of estimating the parameters slope and intercept of the reconstructed regression line. We start using a linear loss function: sum of the residues instead of the sum of squares of the residues. We also use linear constraints , derived from a certain subset of points, namely the roof of the censored data set (see Appendix F). We know that the original regression line is above all the points in the censored data set, so in particular above the roof points. The sum of the residues $\varepsilon_i = a + bx_i - y_i$, for all points i in the set of censored points, is used as the object function. For the moment we consider the sum of the residues rather than the sum of the squares ε_i^2, for simplicity. In a later step this object function is considered.

Instead of taking the squares of the residues, we take the residues themselves, as they are always nonnegative. This approach leads to an object function of the type $\sum_i \varepsilon_i$. This is an object function that is linear, not quadratic. It seems interesting enough to consider it in its own right, leading to a linear programming (LP) formulation. It is, of course, possible to consider a quadratic object function, i.e. $\sum_i \varepsilon_i^2$, instead of $\sum_i \varepsilon_i$.[19]

The linear object function $\sum_i \varepsilon_i$ translates into

$$\begin{aligned}\sum_{i=1}^{n} \left(a^w + b^w x_i^w - y_i^w\right) &= na^w + \left(\sum_{i=1}^{n} x_i^w\right) b^w - \sum_{i=1}^{n} y_i^w \\ &= n\left(a^w + \bar{x}^w b^w - \bar{y}^w\right), \end{aligned} \tag{43}$$

where we have superscripted the parameters a and b by 'w' to indicate that they are based on the censored data. Also the data have been superscripted similarly.

The object function (43) can be simplified by removing the irrelevant parameters n and $\bar{y}^w$ because they are constant. The simplified, but equivalent, object function that results is defined by:

$$(a^w, b^w) \mapsto a^w + b^w \bar{x}^w, \tag{44}$$

[19] In a remark at the end of the present section we consider briefly a quadratic loss function.

where $\bar{x}^w$ is a constant for a given set of censored data.

Somewhat surprisingly this target function only depends on the parameter $\bar{x}^w$. It should be noted that (44) can even be further simplified by first shifting the censored data cloud such that $\bar{x}^w = 0$, so that the resulting object function is defined by

$$(a^w, b^w) \mapsto a^w, \tag{45}$$

and hence depends only on the intercept of a candidate regression line.

We must consider some linear constraints due to the fact that all points of the censored data are situated below the original regression line. However, it is sufficient to require this only for the roof points and not for all the n points in the censored data set. We obtain the following set of $\mathfrak{r} \leq n$ constraints:

$$\begin{aligned} a^r + b^r x_1^r &\geq y_1^r, \\ &\vdots \\ a^r + b^r x_{\mathfrak{r}}^r &\geq y_{\mathfrak{r}}^r, \end{aligned} \tag{46}$$

where $\mathfrak{r}$ is the number of points $(x_1^r, y_1^r), \ldots, (x_{\mathfrak{r}}^r, y_{\mathfrak{r}}^r)$ in the roof $\mathfrak{R}$, which are part of the censored data.

So the LP problem we are dealing with in this case can be stated as

$$\begin{aligned} \min \ & a^w + b^w \bar{x}^w \\ & \text{such that} \\ a^w + b^w x_1^r &\geq y_1^r, \\ &\vdots \\ a^w + b^w x_{\mathfrak{r}}^r &\geq y_{\mathfrak{r}}^r. \end{aligned} \tag{47}$$

the object function in (47) simplifies to a^w if all points in the censored data set are shifted in the x direction, such that the average x values are 0. The corresponding shift[20]

$$\mathring{x}_i^w = x_i^w - \bar{x}^w, \tag{48}$$

for $i = 1, \ldots, \mathfrak{r}$, where

[20] We consider the shift introduced here explicitly, with dedicated notation. In Section 7.3 we consider a similar, but different, shift. The different notations should prevent equating the two shifts and hence the results of the optimization problems to be solved later.

$$\bar{x}^w = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{49}$$

Then indeed

$$\bar{\mathring{x}}^w = \frac{1}{n}\sum_{i=1}^{n} \mathring{x}_i^w = 0. \tag{50}$$

With this shift of the censored data (47) transforms into:

$$\begin{aligned} &\min\ a^w \\ &\quad\text{such that} \\ &a^w + b^w \mathring{x}_1^r \geq y_1^r, \\ &\quad\vdots \\ &a^w + b^w \mathring{x}_{\mathfrak{r}}^r \geq y_{\mathfrak{r}}^r. \end{aligned} \tag{51}$$

Now (51) defines a linear programming (LP) problem. Such problems are typically solved by applying the simplex algorithm or some interior point method, using a dedicated sotware package. But (51) is fairly simple and we can present its solution explicitly. From the constraints of (51) it follows, for b^w given, that

$$\begin{aligned} &a^w \geq y_1^r - b^w \mathring{x}_1^r, \\ &\quad\vdots \\ &a^w \geq y_{\mathfrak{r}}^r - b^w \mathring{x}_{\mathfrak{r}}^r, \end{aligned}$$

so that

$$a^w \geq \max_{i=1,\dots,\mathfrak{r}} \{y_i^r - b^w \mathring{x}_i^r\}. \tag{52}$$

We can take the smallest possible value satisfying (52) which means taking a^w equal to the righthand side of (52), that is:

$$\mathring{a}^w(b^w) = \max_{i=1,\dots,\mathfrak{r}} \{y_i^r - b^w \mathring{x}_i^r\}, \tag{53}$$

where we have stressed that $\mathring{a}^w$ is a function of b^w. Hence the solution $\hat{\mathring{a}}^w$ and $\hat{\mathring{b}}^w$ to (51) is:

$$\hat{\mathring{b}}^{w} = \arg\min_{b^w} \max_{i=1,\ldots,\mathfrak{r}} \{y_i^r - b^w \mathring{x}_i^r\}, \tag{54}$$

$$\hat{\mathring{a}}^{w} = \min_{b^w} \max_{i=1,\ldots,\mathfrak{r}} \{y_i^r - b^w \mathring{x}_i^r\} = \max_{i=1,\ldots,\mathfrak{r}} \{y_i^r - \hat{\mathring{b}}^{w} \mathring{x}_i^r\}. \tag{55}$$

(79) and (80) present similar estimates, but this time entirely based on roof data, whereas (54) and (55) use roof data for the constraints and data from th entire censored data set for the object function.

In Figure 7.1 the following situation is depicted: a potential regression line outside the avoidance area $\mathcal{A}$ and residues for the points in the cloud indicated.[21] The avoidance area $\mathcal{A}$ is bounded at the top by the roof $\mathfrak{R}$ of the censured data set (see Appendix F), with domain defined by the smallest x value $x_{(1)}$ and the largest x value $x_{(n)}$ of the censured data. The formulation (47) or (51) yields as a solution a regression line that cuts the roof in at least one point. This can be adjusted by adding a 'bit of margin' to the computed intercept in the same way as in Section 6.2 by considering lines with the same slope as have been calculated from (47) or (51).

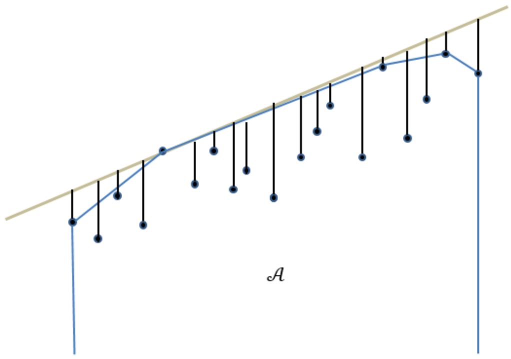

Figure 7.1 An example of a censored data set in the plane with its roof indicated (blue line) and a line touching it, which may be a candidate for the regression line. All data points in the data cloud are involved in the estimation, as indicated by the line segments between each of these points and the regression line. These line segments symbolize the residues. The longer the segments are the harder the pull from the point to which it is attached on the regression line. The roof will provide the only pushback to the pulls from the points in the censored data set. Once an equilibrium has been reached the regression line has been found.

7.2 QP regression for the censored data

We now want to consider a quadratic loss function like

$$\sum_{i=1}^{n} \varepsilon_i^2, \tag{56}$$

[21] A is defined as a closed set. But actually its interior $\mathcal{A}^\circ$ is the area to be avoided. A regression line is allowed to contain points of the roof $\mathfrak{R}$, which belong to the boundary of $\mathcal{A}$.

i.e. the sum of squares of the residues, instead of the linear function $\sum_i \varepsilon_i$ leading initially to the object function given in (43) and ultimately to the one in (45) depending only on the intercept of the candidate regression line, assuming that $\bar{x}^w = 0$, which may require shifting the uncensored data in the x direction.

So we now consider the object function

$$\sum_{i=1}^{n} \left(a + bx_i^w - y_i^w\right)^2, \tag{57}$$

with the additional assumptions that $\bar{x}^w = 0$ and $\bar{y}^w = 0$, which may require first shifting the data in both the x and the y direction. Elaborating (57) taking these conditions into account and removing constant parts leads to the following object function

$$b^2 \sum_{i=1}^{n} \left(x_i^w\right)^2 - 2b \sum_{i=1}^{n} x_i^w y_i^w = b^2 \langle x^w, x^w \rangle - 2b \langle x^w, y^w \rangle, \tag{58}$$

where we have used the inner product $\langle \cdot, \cdot \rangle$ introduced before. Note that (58) depends on the slope b but not on the intercept a. We can simplify (58) even further by first dividing the expression by $\langle x^w, x^w \rangle$ and then completing the square, to find the following equivalent object function:

$$\left(b - \frac{\langle x^w, y^w \rangle}{\langle x^w, x^w \rangle}\right)^2 - \left(\frac{\langle x^w, y^w \rangle}{\langle x^w, x^w \rangle}\right)^2, \tag{59}$$

from which the last term, being constant, can be deleted, so that we get the following object function

$$\left(b - \frac{\langle x^w, y^w \rangle}{\langle x^w, x^w \rangle}\right)^2. \tag{60}$$

Note that the constant term in (60) is equal to the OLS estimate of the slope (1), which is no surprise, of course.

If we put

$$b^\dagger = b - \frac{\langle x^w, y^w \rangle}{\langle x^w, x^w \rangle}, \tag{61}$$

we find the following equivalent formulation of the original QP problem, where the same linear inequalities hold as in the LP case:

$$
\begin{gathered}
\min \ \left(b^\dagger\right)^2 \\
\text{such that} \\
a + b^\dagger x_1^r \geq y_1^{r,\dagger}, \\
\vdots \\
a + b^\dagger x_\mathfrak{r}^r \geq y_\mathfrak{r}^{r,\dagger},
\end{gathered}
\tag{62}
$$

where

$$
y_i^{r,\dagger} = y_i^r - \frac{\langle x^w, y^w \rangle}{\langle x^w, x^w \rangle} x_i^r, \tag{63}
$$

for $i = 1, \dots, \mathfrak{r}$. We can replace $\left(b^\dagger\right)^2$ in the object function of (62) by $|b^\dagger|$ to obtain the equivalent optimization problem:

$$
\begin{gathered}
\min \ |b^\dagger| \\
\text{such that} \\
a + b^\dagger x_1^r \geq y_1^{r,\dagger}, \\
\vdots \\
a + b^\dagger x_\mathfrak{r}^r \geq y_\mathfrak{r}^{r,\dagger}.
\end{gathered}
\tag{64}
$$

We can rewrite the constraints in (64) when taking into account the signs of the x values.[22)]Suppose that

$$
x_1^r \leq x_2^r \leq \cdots \leq x_\nu^r \leq x_{\nu+1}^r \leq \cdots \leq x_\mathfrak{r}^r \tag{65}
$$

where the first ν x values are negative and the remaining ones positive. We then can replace (64) by the following equivalent problem:

$$
\begin{gathered}
\min \ |b^\dagger| \\
\text{such that} \\
\max_{j \in \{\nu+1,\dots,\mathfrak{r}\}} \left\{ \frac{y_j^{r,\dagger} - a}{|x_j^r|} \right\} \leq b^\dagger \leq \min_{i \in \{1,\dots,\nu\}} \left\{ \frac{a - y_i^{r,\dagger}}{|x_i^r|} \right\},
\end{gathered}
\tag{66}
$$

22) In case one of these values is 0 the corresponding point(s) could be left out the estimation procedure sketched here. Or the value could be slightly perturbed so that it is not equal to 0 anymore.

We can derive a constraint for a from the condition in (66), by noting that the interval for $b^{\dagger}$ should not be empty. This implies that for every i in $\{1, \dots, \nu\}$ and every j in $\{\nu + 1, \dots, \mathfrak{r}\}$ should hold:

$$\frac{y_j^{r,\dagger} - a(i,j)}{|x_j^r|} \leq \frac{a(i,j) - y_i^{r,\dagger}}{|x_i^r|}, \tag{67}$$

where we have written $a(i,j)$ to stress that a depends on i and j. From (67) we derive the following inequality for $a(i,j)$:

$$a(i,j) \geq \frac{|x_i^r|\, y_j^{r,\dagger} + |x_j^r|\, y_i^{r,\dagger}}{|x_i^r| + |x_j^r|}. \tag{68}$$

As a should be larger than each of the $a(i,j)$, we have:

$$a \geq \max_{\substack{i \in \{1,\dots,\nu\},\\ j \in \{\nu+1,\dots,\mathfrak{r}\}}} \frac{|x_i^r|\, y_j^{r,\dagger} + |x_j^r|\, y_i^{r,\dagger}}{|x_i^r| + |x_j^r|}. \tag{69}$$

Let

$$a^* = \max_{\substack{i \in \{1,\dots,\nu\},\\ j \in \{\nu+1,\dots,\mathfrak{r}\}}} \frac{|x_i^r|\, y_j^{r,\dagger} + |x_j^r|\, y_i^{r,\dagger}}{|x_i^r| + |x_j^r|}. \tag{70}$$

In (66), a should be as small as possible, that is, should be equal to a^*. Then the solution $b^{\dagger,*}$ to (66) is

$$b^{\dagger,*} = \begin{cases} \alpha & \text{if} \quad 0 \leq \alpha \leq \beta, \\ 0 & \text{if} \quad \alpha \leq 0 \leq \beta, \\ \beta & \text{if} \quad \alpha \leq \beta \leq 0. \end{cases} \tag{71}$$

where

$$\alpha = \max_{j \in \{\nu+1,\dots,\mathfrak{r}\}} \left\{ \frac{y_j^{r,\dagger} - a^*}{|x_j^r|} \right\}, \tag{72}$$

$$\beta = \min_{i \in \{1,\dots,\nu\}} \left\{ \frac{a^* - y_i^{r,\dagger}}{|x_i^r|} \right\}. \tag{73}$$

So the regression line with intercept a^* and slope $b^{\dagger,*}$ is the solution the the QP problem considered in this section.

7.3 Using roof data only

The approach that we elaborate here was suggested in Section 4.6.4, where it was introduced as a variant of the method described in Section 4.6.3. The method elaborated here can also be seen as a variant of the method dealt with in Section 7.1, with the difference that the roof data are the only data used in the object function.

As a variant of (47) we have the following optimization problem:

$$\begin{aligned} &\min\ a^r + b^r \bar{x}^r \\ &\qquad\qquad \text{such that} \\ &a^r + b^r x_1^r \geq y_1^r, \\ &\qquad\quad \vdots \\ &a^r + b^r x_{\mathfrak{r}}^r \geq y_{\mathfrak{r}}^r, \end{aligned} \tag{74}$$

where a^r and b^r are the parameters to be estimated (intercept and slope) in the current situation. As before the roof $\mathfrak{R}$ consists of the points $(x_1^r, y_1^r), \ldots, (x_{\mathfrak{r}}^r, y_{\mathfrak{r}}^r)$, with $\mathfrak{r} = |\mathfrak{R}|$, the number of roof points.

$$\bar{x}^r = \frac{1}{\mathfrak{r}} \sum_{i \in \mathfrak{R}} x_i. \tag{75}$$

The problem simplifies if we shift the points so that the average of the x values of the roof points is 0, corresponding to the shift

$$\tilde{x}_i^r = x_i^r - \bar{x}^r, \tag{76}$$

for $i = 1, \ldots, \mathfrak{r}$. Then indeed

$$\bar{\tilde{x}}^r = \frac{1}{\mathfrak{r}} \sum_{i \in \mathfrak{R}} \tilde{x}_i^r = 0. \tag{77}$$

Now (74) simplifies to the LP problem

$$\begin{aligned} &\min\ a^r \\ &\qquad \text{such that} \\ &a^r + b^r \tilde{x}_1^r \geq y_1^r, \\ &\qquad\quad \vdots \\ &a^r + b^r \tilde{x}_{\mathfrak{r}}^r \geq y_{\mathfrak{r}}^r, \end{aligned} \tag{78}$$

As solutions $\hat{\hat{b}}^r$ and $\hat{\hat{a}}^r$ to (78) we find as before (i.c. (54) and (55))

$$\hat{\hat{b}}^r = \text{arg}\ \text{min}_{b^r} \max_{i=1,\dots,\mathfrak{r}} \{y_i^r - b^r \tilde{x}_i^r\}, \tag{79}$$

$$\hat{\hat{a}}^r = \min_{b^r} \max_{i=1,\dots,\mathfrak{r}} \{y_i^r - b^r \tilde{x}_i^r\} = \max_{i=1,\dots,\mathfrak{r}} \{y_i^r - \hat{\hat{b}}^r \tilde{x}_i^r\}. \tag{80}$$

These estimates are comparable to those in (54) and (55) as already remarked earlier. But it is likely that in practice they are not very different. These difference depend on the average of the x values of the points in the entire censored data set, i.e $\bar{x}^w$ (see (49)) on the one hand and on the average of the x values of the roof points, i.e. $\bar{x}^r$ (see (75)). In practice these averages might be rather close, in which case the regression lines estimated would be very similar.

8 Parametric approach

The approaches so far in this paper have been nonparameric. In the present section we want to present a parametric approach by assuming a normal distribution for the residues in the original data. Apart from this we have to deal with an estimation problem that involves missing data. We propose an approach like the EM-algorithm where alternative steps are taken to estimate a tentative regression line, use this to improve the pseudo-complete data set, and use this to estimate the next version of a reconstructed regression line. Repeat this procedure ‘until convergence’. A similar approach is suggested in Appendix H for a nonparametric setting,[23)] whereas in the present case we have the normal distribution that is supposed to describe the residues adequately.

8.1 Setting the scene

We assume that the original data consisting of N points are described well by a linear model

$$y = a + bx + \varepsilon, \tag{81}$$

where, as before, a and b are the parameters of the (linear) model and where ε is a ‘noise term’ with a normal distribution with expectation 0 and an (unknown) variance σ^2. That is, the random variable ε has a probability density of the following type

$$f_\sigma(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x/\sigma)^2}, \tag{82}$$

for $x \in \mathbb{R}$.

23) Or rather a general setting, not further specified.

We assume that the original data consists of data points (x_i, y_i) for which holds that

$$y_i = a + bx_i + \varepsilon_i, \tag{83}$$

for $i = 1, \dots, N$, where the residues ε_i are assumed to be independently drawn from the density in (82).

Now, our 'disaster' occurs, where all the data points above the regression line, as well as the regression line itself are 'wiped out' and hence the values of a and b are assumed to be lost as well. The remaining data (= the censored data) consists of points below the regression line, that is, points (x_i, y_i) with $\varepsilon_i < 0$. After the 'disaster' we do still know that the original data were well described by a linear regression line (with unknown parameters) and where the residues were described by (82), a density with mean 0 and (unknown) variance σ^2.

We can consider these censored data in a new light, namely as points (x_i^w, y_i^w) such that

$$y_i^w = a + bx_i^w + \eta_i, \tag{84}$$

for $i = 1, \dots, n$, where η_i is viewed as a random variable drawn from a following truncated normal distribution with density:

$$f_\sigma^t = \begin{cases} \frac{1}{\sigma}\sqrt{\frac{2}{\pi}}e^{-\frac{1}{2}(x/\sigma)^2} & \text{if } x < 0, \\ 0 & \text{if } x \geq 0. \end{cases} \tag{85}$$

In (85) σ is a shape parameter of the density function, not the square root of its variance, as in (82).

(85) is a probability density with its own expectation and variance. To compute these parameters (as functions of σ) we first consider its first and second moment, M_1^t and M_2^t respectively. Its first moment (and also its mean) is

$$M_1^t = \frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\int_{-\infty}^{0} xe^{-\frac{1}{2}(x/\sigma)^2}dx = -\sigma\sqrt{\frac{2}{\pi}}, \tag{86}$$

which is strictly negative if $\sigma > 0$, which is typically the case. Its second moment is

$$M_2^t = \frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\int_{-\infty}^{0} x^2e^{-\frac{1}{2}(x/\sigma)^2}dx = \sigma^2. \tag{87}$$

From (86) and (87) the variance V_t of (85) follows:

$$V_t = M_2^t - (M_1^t)^2 = \left(1 - \frac{2}{\pi}\right)\sigma^2 \tag{88}$$

8.2 Iterative reconstruction of the regression line

About half of the original data are missing, and these data cannot be viewed as randomly chosen. They have been censored in a very specific way. In order to reconstruct the original regression line we use the idea to impute missing values in order to create a pseudo complete file. The idea is to proceed in an iterative way: we produce a first estimate of the regression line, compute estimates of residues, use these to impute a set of (virtual) missing data (situated above the regression line), used the pseudo-completed data to make a second reconstruction of the regression line (using ML), and repeat this procedure until the reconstructed regression line does not essentially change (‘convergence has been reached). The basic ideas used here reflect those of the EM algorithm, although the execution is somewhat different. In particular, there are no missing parameter values but missing data. For these missing data there is no information about any of the coordinates, such as the x coordinates.

We now can proceed to estimate a first reconstruction of the regression line from the censored data using all the information on the data we have. We only use the censored data themselve to make the first estimate of the parameters. These are the intercept a, the slope b and the shape parameter σ for the censored distribution. We employ maximum likelihood (ML) estimation to compute these parameters.[24)][25)] We use the fact that the residues are assumed to be independent drawings from the density (85). Then the likelihood function (with all irrelevant constants discarded) is:

$$\begin{aligned}\prod_{i=1}^{n} \frac{1}{\sigma} \exp\left(-\frac{\eta_i^2}{2\sigma^2}\right) &= \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} \eta_i^2\right) \\ &= \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} \left(y_i^w - a - bx_i^w\right)^2\right).\end{aligned} \tag{89}$$

Taking (minus) the logarithm of the final expression in (89) yields the function $\mathcal{L}$ of the parameters a, b and σ that we want to optimize:

$$\mathcal{L}(a,b,\sigma) = n \log\sigma + \frac{1}{2\sigma^2}\sum_{i=1}^{n} \left(y_i^w - a - bx_i^w\right)^2 . \tag{90}$$

24) It should be noted that both the original normal distribution f_σ (see (82)) and its truncated version f_σ^t ((85)) give exactly the same ML estimates.

25) It should be noted that ML estimation is ideal for the problem we are considering in the present paper. The censoring process does not prohibit the use of the observed variables to estimate σ, the only parameter in the model used.

A necessary condition for $\mathcal{L}$ to have a minimum in a point is that the partial derivatives of $\mathcal{L}$ in that point should be 0.[26] This yields:

$$\frac{\partial \mathcal{L}}{\partial a} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i^w - a - bx_i^w\right) = 0, \tag{91}$$

$$\frac{\partial \mathcal{L}}{\partial b} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i^w y_i^w - ax_i^w - b\left(x_i^w\right)^2\right) = 0, \tag{92}$$

$$\frac{\partial \mathcal{L}}{\partial \sigma} = \frac{n}{\sigma} - \frac{1}{\sigma^3}\sum_{i=1}^{n}\left(y_i^w - a - bx_i^w\right)^2 = 0. \tag{93}$$

From (91) and (92) it follows that, using notation introduced in Section 3:

$$\langle\iota, y^w\rangle = \langle\iota,\iota\rangle a + \langle\iota, x^w\rangle b, \tag{94}$$

$$\langle x^w, y^w\rangle = \langle\iota, x^w\rangle a + \langle x^w, x^w\rangle b, \tag{95}$$

which can also be written as, after dividing the left-hand and right-hand sides of both equations by n:

$$\bar{y}^w = \hat{a} + \bar{x}^w\hat{b}, \tag{96}$$

$$\overline{xy}^w = \bar{x}^w\hat{a} + \overline{xx}^w\hat{b}, \tag{97}$$

where we have used the notation:

$$\overline{xy}^w = \langle x^w, y^w\rangle/n, \tag{98}$$

$$\overline{xx}^w = \langle x^w, x^w\rangle/n \tag{99}$$

In matrix form we can write (96) and (97) as:

$$\begin{pmatrix} \bar{y}^w \\ \overline{xy}^w \end{pmatrix} = \begin{pmatrix} 1 & \bar{x}^w \\ \bar{x}^w & \overline{xx}^w \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix}. \tag{100}$$

from which follows

[26] Secondary partial derivatives at the solutions of the first order equations should also be checked to be positive semidefinite so as to make sure that we are indeed dealing with minima. We leave this—somewhat tedious—work to the interested reader to check.

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} 1 & \bar{x}^w \\ \bar{x}^w & \overline{xx}^w \end{pmatrix}^{-1} \begin{pmatrix} \bar{y}^w \\ \overline{xy}^w \end{pmatrix} \tag{101}$$

$$= \frac{1}{\overline{xx}^w - \bar{x}^w \bar{x}^w} \begin{pmatrix} \overline{xx}^w & -\bar{x}^w \\ -\bar{x}^w & 1 \end{pmatrix} \begin{pmatrix} \bar{y}^w \\ \overline{xy}^w \end{pmatrix}. \tag{102}$$

Elaborating (101) we find as estimates of a and b:[27)]

$$\hat{a}_0 = \frac{\overline{xx}^w \bar{y}^w - \bar{x}^w \overline{xy}^w}{\overline{xx}^w - \bar{x}^w \bar{x}^w}, \tag{103}$$

$$\hat{b}_0 = \frac{\overline{xy}^w - \bar{x}^w \bar{y}^w}{\overline{xx}^w - \bar{x}^w \bar{x}^w}. \tag{104}$$

Furthermore, (93) implies the following estimate of σ:

$$\hat{\sigma}_0 = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(y_i^w - \hat{a}_0 - \hat{b}_0 x_i^w\right)^2}, \tag{105}$$

where $\hat{a}_0$ and $\hat{b}_0$ are as in (103) and (104), respectively. These estimates are actually the same as the OLS estimates applied to the censored data (see Section 3), which were criticized there: the estimate of the slope b can be considered acceptable but certainly not that for the intercept a.

As to the method we are concentrating on now, this criticism does not apply as it concerns only a first estimate of the regression line. It is a start of an iterative process that should ultimately lead to a reconstructed regression line of better quality and that is beyond this criticism. That the estimates in (103) and (104) are just initial values is indicated by the subscripts ‘0’. For subsequent estimates the values of these subscripts value will be increased.

So with (103) and (104) we have the initial estimates of the parameters of the reconstructed regression line. With (105) we have an estimate of the variance of the normal distribution describing the residues of the original data. Equipped with this we can set out to impute a missing cloud above the regression line that should be like the cloud below it, consisting of observed data. We do this following in essence the ideas presented in Appendix H. However, because we are now using a parametric method we do some things slightly different.

We start our description of the approach with our first approximation of the reconstructed regression line:

$$y = \hat{a}_0 + \hat{b}_0 x, \tag{106}$$

27) Presented in a form that highlights the symmetry of the components that ‘build’ the estimates. The hats are used to stress that we are dealing with estimates. The subscript ‘0’ is explained later.

with $\hat{a}_0$ and $\hat{b}_0$ given by (103) and (104). We also have our initial estimate of the normal density of the residues:

$$f_{\sigma_0}(x) = \frac{1}{\sigma_0\sqrt{2\pi}} e^{-\frac{1}{2}(x/\sigma_0)^2}, \tag{107}$$

with σ_0 as in (105).

We now generate a cloud of data $D_0^{\uparrow}$ above the regression line (106) that looks very much like the observed (i.e. censored) data, mirrored in this regression line, as follows. First we generate n points in the interval $\left(x_{(1)}^w, x_{(n)}^w\right)$ by simple random sampling with replacement (srswr).[28] We now consider the following virtually complete data D_0 set that will be used to obtain the next estimate of the reconstructed regression line.

$$D_0 = D_0^{\uparrow} \cup D_0^{\downarrow}, \tag{108}$$

where $D_0^{\uparrow}$ is defined above, consisting of imputed points above the regression line (106) and $D_0^{\downarrow}$ is the part of the censored data consisting of points that are below the regression line (106):

$$D_0^{\downarrow} = \{\left(x_j^w, y_j^w\right) \in D \mid y_j^w \leq \hat{a}_0 + \hat{b}_0 x_j^w\}. \tag{109}$$

The idea is now to use D_0 to obtain the next estimate for the reconstructed regression line

$$y = \hat{a}_1 + \hat{b}_1 x, \tag{110}$$

as well as an improved estimate of the standard deviation:

$$\hat{\sigma}_1 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i^w - \hat{a}_1 - \hat{b}_1 x_i^w\right)^2}, \tag{111}$$

and corresponding normal distribution for the residues of the original data, and in particular of the observed data in the censored data set:

[28] There is no reason to assume that the number of points above the regression line is exactly equal to n, the number of points below it. Instead we could also use the idea that for each point with first coordinate x_i the corresponding second coordinate y_i is with probability $1/2$ above the regression line or below it with the same probability. With this in mind we can generate a Bernoulli sequence, starting from a srswr from $\left(x_{(1)}^w, x_{(n)}^w\right)$ of size $2n$. For each of these numbers it will be tossed whether the number is kept or discarded (each with probability $1/2$). This yields a net sample of random size and with approximately n numbers. In this way additional variability can be brought into the estimation. If we happened to know how many points N the original data set had then we know the number of missing data points exactly, namely $N - n$.

$$f_{\sigma_1}(x) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{1}{2}(x/\sigma_1)^2}. \tag{112}$$

Remark It should be noted that if $D_0^{\downarrow}$ does not have all points from D (in other words, $D_0^{\downarrow}$ is a proper subset of D) then it is likely that the next estimate of the reconstructed regression line is pushed up a bit if there are more points above the regression line than below it. So this would be a move of the estimate of the reconstructed regression line in the right direction. □

Equipped with the new estimates of the parameters $\hat{a}_1$ and $\hat{b}_1$ for the improved estimate for the reconstructed regression line and the improved density f_{σ_1} we can repeat the procedure of the first step to produce an imputed set of points, this time $D_1^{\uparrow}$, and a new selection of observed points from the original data set, namely $D_1^{\downarrow}$. Together these sets form $D_1 = D_1^{\uparrow} \cup D_1^{\downarrow}$. Then a new cyle is started. These alternating process steps can be repeated until convergence is (hopefully) achieved for the sequence of estimates for the regression line to be reconstructed.[29)]

Remark Convergence of the iterative process proposed here has not been investigated, let alone be proved. It is felt that the current paper is also not the appropriate place to do so. It seems to be more appropriate to find out first if, and if so, how well the method works in some simulations. If it does in the cases investigated then there is hope that convergence applies more generally, and this fact can actually be proved (under certain conditions). It is probably a good idea to look at how this is done in case of the EM-algorithm (see [3]). □

9 Discussion

In this paper we have discussed several methods to reconstruct a regression line from incomplete data that are censored in a very specific way. A line is specified by two parameters, an intercept a and a slope b. Most of the methods presented to estimate a or b were nonparametric, only one was parametric. Some methods estimate the parameters simultaneously, whereas other do it separately, first estimating b and then a. Some methods used standard statistical theory (OLS or ML), others techniques from OR such as LP or QP. Some techniques yielded the estimates directly, some others required an iteration. Some techniques required missing data methods to virtually complete the censored data, others did not. So there was quite a variety of approaches. How they perform in practice remains an interesting open question, waiting for an answer.

A surprising result (to the present author) is that three apparently different methods (discussed in Sections D, E.1 and E.2) yield exactly the same estimator for the slope b of the reconstructed regression line. It is argued (in Appendix E, Section E.3) that this is partly a coincidence and partly a necessity.[30)] Maybe on second thought the 'mystery' disappears and it becomes obvious that the three methods yield the same estimators.

29) To be more precise, 'convergence achieved' should be interpreted as: when there are no significant changes anymore for the estimates of the intercept and slope of the series of regression lines computed.

30) Admittedly, the argument supporting coincidence is somewhat contrived.

The estimators based on linear programming (LP) and quadratic programming (QP) are quite different from OLS (or weighted least squares) or maximum likelihood. The regression line is hovering above the data instead of cutting through them. Maybe these (or similar) estimators can be applied in other situations encountered in practice.

Unfortunately there was no time to perform some simulation studies so that numerical results would have been available for the methods presented. This would give an insight in how the methods performed in practice, to which extent the results differed or agreed. If someone is interested in filling this lacuna it would be a useful contribution. Of course, research into alternative estimation methods would be welcome too.

The convergence of the iterative method presented in Section 8 has not been proved to be true, given suitable conditions. It is presented there as an intuitively appealing variant of the EM-algorithm. So this lack of proof is obviously a glaring omission. So it would be nice if someone would fill this hiatus.

A special case of the EM-algorithm in Example 15.7 in [6] applies to the problem considered in the present paper, more particularly to the parametric model considered in Section 8.[31)]

The notation used in [6] and in the present paper differ. Table 9.1 serves as a translation table for the notations used.

[6]	Section 8
x_i	$(1, x_i)'$
y_{1i}	y_i
β_1	$(a, b)'$
y_{2i}	$-\varepsilon_i/\sigma$
β_2	$(0, 0)'$
σ_1^2	σ^2

Table 9.1 Linking notation used in [6] and the present paper.

We elaborate this link a bit, as it is not immediately clear. Now the variables in Table 9.1 satisfy equation (15.11) in [6], with $\rho = -1$. Equation (15.12) in [6] describes the censoring mechanism used in the present paper. Equation (15.13) in [6] with $\rho = -1$ transforms into an expression equivalent to (15.12) in [6]:

$$P\left[M_{i1} = 1 \mid y_i, x_i\right] = \begin{cases} 1 & \text{if } y_i \geq a + bx_i, \\ 0 & \text{if } y_i < a + bx_i. \end{cases} \tag{113}$$

The EM-algorithm on p.323 in [6] can now be applied to this problem, where it can be used that $\rho = -1$ and hence ρ does not have to be estimated.

The regression problem studied in this paper was inspired by [10]. In the introduction of the present paper there is a brief background to [10], in particular concerning its origin. This application, although curious, is not of great practical importance. The reason that this problem is studied in [10] as well as in the present paper was not for its practical relevance, but for the

31) This reference was pointed out to the author by Sander Scholtus, who also provided the information provided here to link both methods.

statistical challenges it offered. It is like a problem in recreational mathematics, which is driven by the challenge it offers to solve it.[32)] In this vein the current problem was viewed by the present author and, no doubt, also by the author of [10], Gerard van de Kant. Nevertheless it would be nice if a relevant statistical application could be found, so that the solutions obtained here (and in [10]) become useful. Looking at the econometrics literature it can be remarked that Tobit models (see [1], chapter 10), in particular those of type 2 (see [1], Section 10.7) are relevant. We leave it to the interested reader to further explore this connection.

The problem was considered for data involving two variables, x and y of which one (y) is assumed to be dependent on the other. However, the problem could easily be generalized to more variables, that is to higher dimensional data. It is unclear that this yields generalizations of problems that we have considered that are quite straightforward to formulate and solve. Surely, if the number of variables is too high with respect to the number of data one runs into problems concerning the sparsity of data. Possibly there is, for some, a problem with the motivation with such a generalization, in case no real application can be found. But it is perhaps as with mountains: their mere existence is reason for some of us to want to climb them.

Acknowledgments

The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

I thank Sander Scholtus for reviewing this paper. He provided me with a number of suggestions to improve it which I gratefully made use of.

References

[1] T. Amemiya (1985). *Advanced Econometrics*, Blackwell.

[2] J. Borwein & P. Borwein (1987). *Pi and the AGM*, Wiley.

[3] A. Dempster, N. Laird & D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B, 39 (1)*, pp. 1–-38.

[4] M. Hollander & D. Wolfe (1973). *Nonparametric Statistical Methods*, Wiley.

[5] D. Knuth (1981). *The Art of Computer Programming, Vol. 2, Seminumerical Algorithms (2nd ed.)*, Addison-Wesley.

[6] R. Little & D. Rubin (2002) *Statistical Analysis with Missing Data (2nd ed.)*, Wiley.

[7] K. Mardia, J. Kent & J. Bibby (1982) *Multivariate Analysis*, Academic Press.

[8] F. Preparata & M. Shamos (1985). *Computational Geometry - An Introduction*, Springer.

[9] H. Theil (1971). *Principles of Econometrics*, Wiley.

[32)] 'Recreatonal statistics' is perhaps an apt name that presents itself for this (as of yet nonexisting) area.

[10] G. van de Kant (1973). Regression analysis with non-response for observations with positive residuals (in Dutch), Nota 6207-73-SA, CBS (The Hague).

[11] A. Wald (1940). The fitting of straight lines if both variabes are subject to error, *Annals of Mathematical Statistics*, 11, pp. 284–300.

Appendix
A OLS for the complete data

Given is a set of points $\{(x_1, y_1), \dots, (x_N, y_N)\}$ in the plane $\mathbb{R}^2$. This is supposed to be the complete data set, which in this paper is only partially observed, because some points have been censored. Nevertheless the complete data set is the natural starting point for our discussion of ordinary least squares (OLS) theory as a basic technique to consider. The theory in case of the censored data is the subject of the rest of the paper. In Section 3 we consider OLS applied to the censored data set, in particular to estimate the slope of the (also censored) regression line.

Suppose that we want to fit a regression line through these points. So we assume that holds

$$y_i = a + bx_i + \varepsilon_i, \tag{A.1}$$

for certain parameters $a, b \in \mathbb{R}$. We can estimate the parameters by minimizing the sum of the squared residues, that is

$$\begin{aligned}\langle \varepsilon, \varepsilon \rangle &= \langle y - a\iota - bx, y - a\iota - bx \rangle \\ &= \langle \iota, \iota \rangle a^2 + \langle x, x \rangle b^2 + 2\langle x, \iota \rangle ab - 2\langle y, \iota \rangle a - 2\langle x, y \rangle b + \langle y, y \rangle,\end{aligned} \tag{A.2}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product, that is,

$$\langle x, y \rangle = \sum_{i=1}^{N} x_i y_i. \tag{A.3}$$

with $\varepsilon = \langle \varepsilon_1, \dots, \varepsilon_N \rangle'$, $x = (x_1, \dots, x_N)'$, $y = (y_1, \dots, y_N)'$ and $\iota = (1, \dots, 1)'$ is the all 1 vector (of length N, which is dependent on the context), with all its N components equal to 1.[33) 34)] Note that $\langle \iota, \iota \rangle = N$.

The advantage of using inner products in (A.2) is simplified notation (no summation symbols) and a greater appreciation of symmetry properties of the expressions built with them. This is convenient for the derivation that we are about to deal with. But when the results of the derivation have been obtained they can be readily introduced.

33) The apostrophe (') indicates transposition, interchanging rows and columns in vectors and matrices.
34) The length N of the vectors x, y, ι should be specified to be precise, for instance by using a subscript, but we do not do this to avoid cluttered formulas. Instead we expect that the length is clear from the context when the formulas are used.

Minimizing the expression (A.2) yields the solutions to the ordinary least squares (OLS) problem, where the residues have equal weight 1. The minimum can be found by computing the partial derivatives of $\langle\varepsilon,\varepsilon\rangle$ and setting them to 0. We find the solutions $\hat{a}$ and $\hat{b}$ to be:

$$\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \begin{pmatrix} \langle\iota,\iota\rangle & \langle\iota,x\rangle \\ \langle\iota,x\rangle & \langle x,x\rangle \end{pmatrix}^{-1} \begin{pmatrix} \langle\iota,y\rangle \\ \langle x,y\rangle \end{pmatrix} \tag{A.4}$$

$$= \frac{1}{\langle\iota,\iota\rangle\langle x,x\rangle - \langle\iota,x\rangle^2} \begin{pmatrix} \langle x,x\rangle & -\langle\iota,x\rangle \\ -\langle\iota,x\rangle & \langle\iota,\iota\rangle \end{pmatrix} \begin{pmatrix} \langle\iota,y\rangle \\ \langle x,y\rangle \end{pmatrix}, \tag{A.5}$$

so that

$$\hat{a} = \frac{\langle x,x\rangle\langle\iota,y\rangle - \langle\iota,x\rangle\langle x,y\rangle}{\langle\iota,\iota\rangle\langle x,x\rangle - \langle\iota,x\rangle^2}, \tag{A.6}$$

$$\hat{b} = \frac{\langle\iota,\iota\rangle\langle x,y\rangle - \langle\iota,x\rangle\langle\iota,y\rangle}{\langle\iota,\iota\rangle\langle x,x\rangle - \langle\iota,x\rangle^2}, \tag{A.7}$$

which, after some rearranging, can be represented in more familiar form as

$$\hat{a} = \frac{\sum_{i=1}^N x_i^2 \sum_{i=1}^N y_i - \sum_{i=1}^N x_i \sum_{i=1}^N x_i y_i}{N\sum_{i=1}^N x_i^2 - \left(\sum_{i=1}^N x_i\right)^2} = \frac{\bar{y}\sum_{i=1}^N x_i^2 - \bar{x}\sum_{i=1}^N x_i y_i}{\sum_{i=1}^N x_i^2 - N\bar{x}^2}, \tag{A.8}$$

$$\hat{b} = \frac{N\sum_{i=1}^N x_i y_i - \sum_{i=1}^N x_i \sum_{i=1}^N y_i}{N\sum_{i=1}^N x_i^2 - \left(\sum_{i=1}^N x_i\right)^2} = \frac{\sum_{i=1}^N x_i y_i - N\bar{x}\bar{y}}{\sum_{i=1}^N x_i^2 - N\bar{x}^2}, \tag{A.9}$$

where

$$\bar{x} = \frac{1}{N}\sum_{i=1}^N x_i, \tag{A.10}$$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^N y_i. \tag{A.11}$$

Direct calculation shows that

$$\bar{y} = \hat{a} + \hat{b}\bar{x}, \tag{A.12}$$

from which trivially follows an estimator of the intercept a, given that of the slope b:

$$\hat{a} = \bar{y} - \hat{b}\bar{x}. \tag{A.13}$$

(A.13) is an insightful equation as it indicates that the important parameter to estimate in a linear regression problem is the slope parameter b, from which the intercept a can be derived in the manner indicated by this formula.[35)]

The formulas simplify if we assume that $\langle x, \iota \rangle = 0$, which is equivalent to the requirement $\bar{x} = 0$. This can always be achieved by shifting the data cloud over a distance $\bar{x}$ in the x-direction, that is by applying the linear transformation

$$x_i \rightarrow x_i - \bar{x}. \tag{A.14}$$

We then obtain instead of (A.6) and (A.7):

$$\hat{a} = \frac{\langle \iota, y \rangle}{\langle \iota, \iota \rangle} = \frac{1}{N} \sum_{i=1}^{N} y_i = \bar{y}, \tag{A.15}$$

$$\hat{b} = \frac{\langle x, y \rangle}{\langle x, x \rangle} = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}, \tag{A.16}$$

Additional simplication is obtained by assuming that also

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i = 0. \tag{A.17}$$

Both assumptions actually mean that the center of gravity of the data cloud is assumed to coincide with the origin $\mathcal{O}$ of the coordinate axes used. This situation can always be obtained by applying the following shift in the y-direction as well:

$$y_i \rightarrow y_i - \bar{y}. \tag{A.18}$$

Then

$$\hat{a} = \bar{y} = 0. \tag{A.19}$$

So (A.19) indicates that the (estimated) regression line intersects the origin $\mathcal{O}$ of the coordinate axes used. This is convenient, as this leaves now only one parameter, namely the direction (coefficient), i.e. the coefficient b of the regression line. This is unaffected by shifting the data cloud.

35) Or, as will be clear from the following, this is not necessary if it is equal to 0. However, it is the case if the center of gravity of the data cloud coincides with the origin $\mathcal{O}$ of the axis system used.

We can rewrite the expression in $\hat{b}$ in (A.16) as follows:

$$\hat{b} = \frac{\langle x, y \rangle}{\langle x, x \rangle} = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2} = \sum_{i \in NZ_N} \xi_i \frac{y_i}{x_i}, \tag{A.20}$$

where

$$\xi_i = \frac{x_i^2}{\sum_{j=1}^{n} x_j^2}, \tag{A.21}$$

for $i = 1, \ldots, N$, and

$$NZ_N = \{i \in V_N \mid x_i \neq 0\}, \tag{A.22}$$

where $V_N = \{1 \ldots, N\}$.

The summation in the last expression in (A.20) involves only those points i for which $x_i \neq 0$. This is not a real restriction because for such points i the weight $\xi_i = 0$.

The final expression for the estimate $\hat{b}$ in (A.20) is meaningful, because it is a weighted average of slopes. So it can be directly interpreted as an estimator of the slope of the regression line. The fact that each slope is weighted proportional to the square of the x-coordinate of a point according to (A.21) stands out. So a point further away from the origin, in terms of its x-value, is weighted (quadratically) more heavily than a point that is closer to the origin in the sense indicated. That the weighting is proportional to the square of the x_i is also remarkable. So, the sign of x_i is unimportant for the weighting, as it intuitively should be.

Remark In fact (A.21) can also be the source of inspiration for alternative estimates of the direction of the regression line. We could for instance look at estimates that are more robust in the sense that the weights associated with points further away from the origin are more tempered. So instead of (A.21) one could use weights of the following type:

$$\xi_i^{\alpha} = \frac{|x_i|^{\alpha}}{\sum_{j=1}^{n} |x_j|^{\alpha}}, \tag{A.23}$$

for $0 \leq \alpha < 2$. But other choices to achieve this effect are also possible. To decrease the influence of points far to the left or right (that is, with small or large x values) would not make sense if these values can actually be freely chosen by the statistician. The worry then should rather be about the corresponding y values that are out of kilter, that is, outlying. But in the present paper we shall not consider such situations, as we are actually dealing with a subset of the complete data set, which is censored in a specific way. This makes it difficult to deal with outliers in the (original, complete) data set. □

A.1 Recapitulation

Looking back at the above, we can make the following observations. It is clear that shifting the data cloud so that its center of gravity coincides with the origin $\mathcal{O}$ of the coordinates axis used has only advantages: it simplifies formulas and makes the interpretation of the simplified formulas straightforward. In fact, it immediately yields the intercept of the regression line ($\hat{a} = 0$). Secondly, estimating the slope of the regression line is a matter of (weighted) averaging the tangentia of the angles at the origin $\mathcal{O}$ of the triangles defined by a data point, its foot points (orthogonal projection on the x-axis) and the origin $\mathcal{O}$. The weight used for a data point is proportional to the square of its x coordinate.

Therefore such a shift of the data set is assumed, unless stated otherwise. It should be noted, however, that the center of gravity depends on the data cloud being used. It is typically different for the entire (original) cloud, the censored cloud, or a censored cloud with imputed missing points (in which case the result of the imputation is of influence as well). But as the shifting does not alter the slope of the regression line, it is a very useful simplifying transformation.

B Wald and Theil estimators for the slope

In this appendix we discuss some distribution free methods to estimate the slope of a regression line. We assume the methods to be applied to the original, uncensored, data set. Many more variations of the methods discussed are possible. But we will not go into this, in order not to increase the size of this paper even more.

B.1 Wald's methods of fitting straight lines

We start presenting Wald's methods for the complete data set and then switch to the case of the censored data set. Wald's method is described in [9], pp. 610 ff., which we use as a reference for this method.

We consider two variants of Wald's method in Sections B.1.1 and B.1.2. The idea underlying both methods is that the various subsets defined are used to estimate points of gravity, which in turn are used to form a vector. All the vectors obtained in this way are used to estimate the direction of the regression line, by averaging the directions of the various vectors.

B.1.1 Wald's method for two groups

In this case we divide the group of points $p_i = (x_i, y_i)$ into two groups G_1 and G_2 of (approximately) equal size, such that $G_1 \leq G_2$, meaning that for every $u = (u_1, u_2) \in G_1$ and $v = (v_1, v_2) \in G_2$ we have $u_1 \leq v_1$. So $G_1 \cup G_2 = V_N$, where $V_N = \{1, \dots, N\}$. First the centers of gravity of G_1 and G_2 are computed:

$$z_i = \frac{1}{|G_i|} \sum_{j \in G_i} p_j, \tag{B.1}$$

for $i = 1, 2$. As an estimate of the slope of the regression line we can take the vector (z_1, z_2).

Remark Instead of using the centers of gravity of each of the two groups, one can consider the vectors obtained by randomly picking a point p_l from the left group G_1 and, independently, also randomly a point p_r from the right group G_2 . The vector (p_l, p_r) can be used as an estimate of the direction. This can be used in a bootstrapping application to find out the sensitivity of various estimates. The points are randomly selected with replacement to make things easy. Of course, one can also use averages of such vectors as direction estimates with smaller variances. □

B.1.2 Wald's method for three groups

This methods partitions the set of points $p_i = (x_i, y_i)$ in the complete data cloud into three groups, H_1, H_2 and H_3 of about equal size, such that $H_1 \leq H_2 \leq H_3$, with $\leq$ as defined in Section B.1.1. This time we only use the points in the two extreme groups, i.e. H_1 and H_3, which are furthest apart. We compute the centers of gravity of these sets (as in (B.1)), say ζ_1 and ζ_3 , and use them to estimate the slope of the original regression line, by taking (ζ_1, ζ_3) as its estimator.

As in Section B.1.1 was suggested for the two groups case, instead of using the centers of gravity of each of the two extreme groups H_1 and H_3, one can consider the vectors obtained by picking a point $\mathfrak{p}_l$ from the left-most group H_1 and a point $\mathfrak{p}_r$ from the right-most group H_3 and consider the vector $(\mathfrak{p}_l, \mathfrak{p}_r)$ as an estimate of the direction vector. This can be used in a bootstrapping application to find out the sensitivity of various estimates. The points are randomly selected with replacement to make things easy. Of course, one can also use averages of such vectors as direction estimates with smaller variances.

B.2 Theil's slope estimator

The methods suggested here generalise those in the remarks in Sections B.1.1 and B.1.2 by using more direction vectors than Wald's methods employ.

In fact, the method discussed here considers certain pairs of points from the observed part of the complete cloud of points to estimate the slope of the regression line. To simplify the discussion we assume that no ties occur in the sequences $x_1, \dots, x_N$ and $y_1, \dots, y_N$.[36] For $i < j$ with $i, j \in V_N = \{1, \dots, N\}$ we can compute fractions S_{ij} defined as

$$S_{ij} = \frac{y_j - y_i}{x_j - x_i}, \tag{B.2}$$

which can be viewed as estimators of the slope b. If we take the median of the S_{ij} in (B.2) we obtain a slope estimator named after Theil (see [4], Section 9.2):

$$\hat{b}_T = \text{median}\,\{S_{ij} \mid i < j \text{ and } i, j \in V_N\}. \tag{B.3}$$

The median of an ordered sequence s_i of $2m + 1$ numbers is the one at the middle, i.e. $s_{(n+1)}$. In case the sequence consists of $2m$ numbers, the median is defined as $\left(s_{(n)} + s_{(n+1)}\right)/2$. The $s_{(j)}$ are rank statistics, with $s_{(1)} \leq \dots \leq s_{(N)}$.

It should be noted that variants of Theil's slope estimator can easily be defined, using two ideas. One is to use weights that favour pairs of points (x_i, y_i) and (x_j, y_j) that imply longer vectors $(x_i - x_j, y_i - y_j)$. Such weights could be taken proportional to $\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$.[37] The other idea is to not consider all possible combinations of pairs of points, but only pick a

36) This is only an assumption for convenience: it avoids taking care of exceptions. But nothing substantial is lost by making this assumption.

37) Or, more simply, proportional to $|x_i - x_j|$.

specific selection of combinations of points. For example, one could select points that are sufficiently far apart, which is the idea underlying Wald's estimators. Or one could select only roof points (see Section F.1). Or roof points that are sufficiently far apart, that is, with a distance above some suitably chosen threshold value. We shall not pursuit these ideas in the present paper, but leave them to be explored in future research.

C Slope estimators based on heuristics

We want to point out some (possible) alternative estimators of the parameter b of the slope of the regression line as given in (A.16). They follow from 'playing with the form of the estimator' rather than from applying some alternative optimization criteria. We assume that the center of gravity of the data cloud has been shifted to the origin $\mathcal{O}$ of the axis system used. This leaves b as the only parameter of the regression line to estimate, as the intercept a is known to be 0.

Throughout this section we assume that $\langle x, x\rangle > 0$ and $\langle y, y\rangle > 0$. These are not serious restrictions (in the sense that if they do not hold we are dealing with a nonproblem. Additional restrictions may be required for certain estimators.

So (A.16) gives the estimator of b following from applying the least squares criterion. For convenience we repeat this estimator here so that it can be immediately compared to the alternative estimators that are presented. So we have

$$\hat{b}_1 = \frac{\langle x, y\rangle}{\langle x, x\rangle}. \tag{C.1}$$

We can look at (C.1) heuristically, where (virtually) we can cancel an x in the numerator and denominator and are left with a fraction with y as denominator and x as numerator, which looks like a tangent, which is exactly what the direction coefficient is.[38] We can produce similar expressions when applying this heuristic. An example of this is presented in the next estimator:

$$\hat{b}_2 = \frac{\langle y, y\rangle}{\langle x, y\rangle}. \tag{C.2}$$

For it to be well-defined it should hold that $\langle x, y\rangle \neq 0$, which we may assume to be the case in practice. A vertical regression line is not what we expect to use in practical applications.

A third estimator for b can be produced by taking the geometric mean of the estimators in (C.1) and (C.2):

$$\hat{b}_3 = \begin{cases} \sqrt{\frac{\langle y,y\rangle}{\langle x,x\rangle}} & \text{if } \langle x, y\rangle > 0, \\ -\sqrt{\frac{\langle y,y\rangle}{\langle x,x\rangle}} & \text{if } \langle x, y\rangle < 0. \end{cases}$$

If we take the arithmetic mean of estimators $\hat{b}_1$ and $\hat{b}_2$ we obtain our next estimator for the slope b:

[38] Another way to look at it is by assuming that $y = bx$, discarding the vector ϵ of residues. Subsituting this for all estimators below (some requiring $b > 0$) will yield b as a result.

$$\hat{b}_4 = \frac{1}{2}\left(\frac{\langle x,y\rangle}{\langle x,x\rangle} + \frac{\langle y,y\rangle}{\langle x,y\rangle}\right), \tag{C.3}$$

which requires $\langle x,y\rangle \neq 0$.[39)]

It is possible to use other averaging operators (beside geometric and arithmetic mean), such as harmonic mean to produce more alternative estimators of b, and take average combinations of such estimators to obtain yet other one. But we will not elaborate on this here.

We can use the estimators $\hat{b}_1$ and $\hat{b}_2$ to produce an estimator based on AGM, the arithmetic geometric mean. This is a convergent iterative procedure starting with the estimates $\tilde{b}_1$ and $\tilde{b}_2$, yielded by the estimators $\hat{b}_1$ and $\hat{b}_2$ respectively, for given values of $\langle x,y\rangle$, $\langle x,x\rangle$ and $\langle y,y\rangle$. We assume $\tilde{b}_1 \geq 0$ and $\tilde{b}_2 \geq 0$. If $\tilde{b}_1 = \tilde{b}_2$ we are done. The AGM estimator yields the same result. If $\tilde{b}_1 \neq \tilde{b}_2$, order them, say $0 \leq \tau < \chi$, with $\tau = \min\{\tilde{b}_1, \tilde{b}_2\}$ and $\chi = \max\{\tilde{b}_1, \tilde{b}_2\}$. It is easy to verify that the geometric mean of two (nonnegative) numbers is always smaller than the arithmetic mean of these numbers.[40)] That is

$$\sqrt{\tau\chi} \leq \frac{1}{2}(\tau + \chi). \tag{C.4}$$

So when we have computed $\sqrt{\tau\chi}$ and $\frac{1}{2}(\tau+\chi)$ for the next step in the iteration we replace τ by $\sqrt{\tau\chi}$ and χ by $\frac{1}{2}(\tau+\chi)$. We then repeat this step until the results are sufficiently close (smaller than some prior chosen value $\delta > 0$. As stated, the procedure converges and the limit value is inside the segment $[\tau, \chi]$. So this is our fifth estimator of b, which we shall denote as $\hat{b}_5$. So if $\tilde{b}_1^w \geq 0$ and $\tilde{b}_2^w \geq 0$, we define the fifth slope estimator as:

$$\hat{b}_5 = M(\hat{b}_1, \hat{b}_2), \tag{C.5}$$

where $M(x,y)$ is the AGM of $x, y \geq 0$, as defined in [2].[41)] We refer readers who are interested in the AGM (and its link to π, in particular its numerical approximation) to this fascinating book.

After the introducion of this somewhat elaborate estimator we now switch to the arguably simplest one in our assortment:

$$\hat{b}_6 = \frac{\langle g,y\rangle}{\langle g,x\rangle}, \tag{C.6}$$

39) $\langle x,x\rangle > 0$ is also required, of course, but we take this for granted, as otherwise there is no problem.

40) This follows , for instance,from the inequality $(\sqrt{\tau} - \sqrt{\chi})^2 \geq 0$ for $\tau, \chi \geq 0$.

41) It is possible to define this AGM estimator for any (real) values of input slopes b_1 and b_2, but we shall limit ourselves to the case where both slopes are nonnegative. The extension hinges in particular on that for the geometric mean $\mu_g(x,y) = \sqrt{xy}$ (as the extension for the arithmetic mean is straightforward). We can define this extension of μ_g as follows. If $x, y \in \mathbb{C}$ and $x = |x|e^{i\varphi_x}$ and $y = |y|e^{i\varphi_y}$ one can define $\mu_g(x,y) = \sqrt{|x||y|}e^{i(\varphi_x+\varphi_y)/2}$. This implies, for instance, that $\mu_g(-x,-y) = -\mu_g(x,y) = -\sqrt{xy}$, for $x, y \geq 0$. Also for $x, y \geq 0$ we have $\mu_g(-x,y) = \mu_g(x,-y) = \sqrt{xy}\, i$ (with $i^2 = -1$) is purely imaginary, that is, is in $i\mathbb{R} \subset \mathbb{C}$. But this case would obviously be useless for the AGM slope estimator.

where $g \neq \kappa\iota$, where κ is a constant, (because $\langle \iota, x \rangle = 0$) is a weight vector of length N. g could for instance be used to produce more robust estimates of b by dampening the influence of points with extreme or outlying y values. For instance the components of g could be defined as:

$$
\begin{aligned}
g_i &= s \quad \text{if } |y_i/x_i - M| > \theta, \\
g_i &= 1 \quad \text{if } |y_i/x_i - M| \leq \theta,
\end{aligned} \tag{C.7}
$$

where $s < 1$[42] and

$$
M = \text{median}\,\{y_j/x_j \mid j = 1, \ldots, N, x_j \neq 0\}, \tag{C.8}
$$

for some prior chosen threshold value $\theta > 0$. Compare this to (A.21) and (A.23) where the weights associated with the data points depend on the $|x_i|$, the is, the x distance of data point i to the origin $\mathcal{O}$ of the axis system.

This ends our little digression on possible alternative and heuristic estimators of b. No claim is made about the quality of any of these estimators. There is only an invitation to consider them in practice, experiment with them and see how (well or badly) they do. Actually, the estimators that really matter in the present paper would be those applied to the observed data set. This case is discussed in Section 5.

[42] Or even $s = 0$ in extreme cases.

D PCA and spectral analysis

Principal Component Analysis (PCA)[43] is a multivariate analysis technique which, informally stated, seeks to fit a p-dimensional ellipsoid to the data, where each of its axes corresponds to a principal component. More formally expressed, the goal of PCA is to find a standardized linear combination (SLC)[44] which has maximum variance. As a secondary goal it wants to find an SLC with the second largest variance. Likewise there may be an interest in one or more additional lesser eigenvalues. But the idea is to keep the number of main eigenvalues (and their eigenspaces) small, in the hope that they will give sufficient insight into the main features of the data.

To illustrate PCA see Figure D.1.[45] It shows an elongated cloud of data points, and two vectors. The larger one of these vectors indicates the direction of the largest variance. Or, put differently, indicates the eigenspace of the largest eigenvalue. This is the direction we are interested in. The other vector is orthogonal to this one. It is associated with an eigenspace that corresponds to a smaller eigenvalue.

For a full discussion of PCA refer to a book on multivariate analysis, such as [7], where Chapter 8 is entirely devoted to this method. In this appendix we illustrate PCA using the full data set. In Section 4.5 we apply PCA to the observed data and discuss potential issues that may arise.

Our interest is somewhat different: we want to estimate the direction of the eigenspace (in our case generated by a single vector) corresponding to the largest eigenvalue of a matrix summarising the data, to be introduced later (see (D.2)).

Of particular interest is how the solution to these problems is found, namely by diagonalising the covariance matrix, which is symmetric and hence only has real eigenvalues. The eigenvector corresponding to the largest eigenvalue is the object of our interest, because it can serve as an estimate of the direction of the regression line fitting the data, from which its slope is derived.

In our case, the data only has two variables, namely x and y. If we consider the original data in fact, it consists of N points (x_i, y_i), which can be arranged in matrix form as follows:

$$D = \begin{pmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_N & y_N \end{pmatrix} = \begin{pmatrix} x & y \end{pmatrix}, \tag{D.1}$$

43) There are other names for this technique as it is applied in various disciplines: the principal axis theorem in mechanics, the discrete Karhunen–Loève transform (KLT) in signal processing, the Hotelling transform in multivariate quality control, proper orthogonal decomposition (POD) in mechanical engineering, singular value decomposition (SVD) of X, eigenvalue decomposition (EVD) of $X'X$ in linear algebra, where X is a data matrix like D in (D.1), etc. See `https://en.wikipedia.org/wiki/Principal_component_analysis` for further discussion on this issue.

44) Let $v_1, \dots, v_k$ denote the variables. In the weighted sum $\sum_{i=1}^{k} w_k v_k$ the w_i are standardized (nonnegative) weights, for which $\sum_{i=1}^{k} w_k^2 = 1$.

45) The picture was taken from the Wikipedia lemma on PCA : `https://en.wikipedia.org/wiki/Principal_component_analysis`. It is of no particular importance that the data is generated using a Gaussian distribution to illustrate the idea behind PCA.

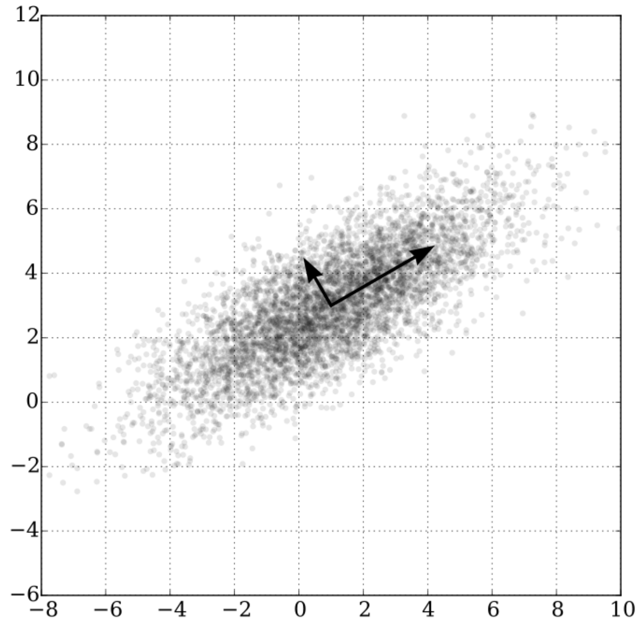

Figure D.1 Principal components drawn in a 2D data cloud. They indicate the principal axes of the ellipsoid associated with the covariance matrix of this data cloud.

where $x = (x_1, \ldots, x_N)'$ and $y = (y_1, \ldots, y_N)'$. In fact, we assume that for these data the center of gravity coincides with the origin O of the coordinate system. We therefore have $\langle x, \iota \rangle = 0$ and $\langle y, \iota \rangle = 0$, where $\iota = (1, \ldots, 1)$, the all one vector of length N. Hence $\bar{x} = 0$ and $\bar{y} = 0$.

The covariance matrix of D is

$$\mathcal{C} = D'D = \begin{pmatrix} \langle x, x \rangle & \langle x, y \rangle \\ \langle x, y \rangle & \langle y, y \rangle \end{pmatrix} = \begin{pmatrix} f & g \\ g & h \end{pmatrix}. \tag{D.2}$$

The final matrix of (D.2) introduces symbols f, g and h as shorthand notation for $\langle x, x \rangle$, $\langle x, y \rangle$ and $\langle y, y \rangle$, respectively. It is more convenient to use these letters when manipulating formulas instead of the expressions they stand for. But they are only auxiliary means and in the end results they will be replaced by the quantities (i.e. inner products) that they represent.

Note that (D.2) only contains information about the length of x, the length of y and the inner product of these vectors.

The eigenvalues of $\mathcal{C}$ are the roots of the characteristic equation:

$$\det(\mathcal{C} - \lambda I_2) = \lambda^2 - (f + h)\lambda + fh - g^2 = 0, \tag{D.3}$$

where I_2 denotes the identity matrix of order 2.

We have a closer look at the coefficients in (D.3) in order to understand them better. First the coefficient of the linear term:

$$f + h = \langle x, x\rangle + \langle y, y\rangle = ||x||^2 + ||y||^2. \tag{D.4}$$

Now (D.4) is the total length of vectors x and y. If these vectors should be orthogonal (that is, $\langle x, y\rangle = 0$) it could be interpreted as the length of the hypotenuse of the orthogonal triangle with sides of length $||x||$ and $||y||$. But in general this will not be the case, and we will have that $\langle x, y\rangle \neq 0$. In fact, this aspect is what the constant term is about. The constant part of (D.3) can be expressed as follows:

$$\begin{aligned} fh - g^2 &= fh\left(1 - \frac{g^2}{fh}\right) = fh\left(1 - \left(\frac{g}{\sqrt{f}\sqrt{h}}\right)^2\right) \\ &= ||x||^2\,||y||^2\left(1 - \langle\frac{x}{||x||}, \frac{y}{||y||}\rangle^2\right). \\ &= ||x||^2\,||y||^2\,\left(1 - \cos^2\varphi\right), \end{aligned} \tag{D.5}$$

where φ can be interpreted as the angle between the unit vectors $x/||x||$ and $y/||y||$.

Substituting the expressions for the coefficients in (D.4) and (D.5) into (D.3) and after some reshuffling, yields the following expression for the characteristic equation:

$$\left(\lambda - ||x||^2\right)\left(\lambda - ||y||^2\right) - ||x||^2||y||^2\cos^2\varphi = 0. \tag{D.6}$$

The roots $\lambda_\pm$ of (D.2), which are also the eigenvalues of C, can be expressed as follows:

$$\lambda_\pm = \tfrac{1}{2}(f + h) \pm \tfrac{1}{2}\sqrt{(f - h)^2 + 4g^2}. \tag{D.7}$$

The largest eigenvalue

$$\lambda_+ = \tfrac{1}{2}(f + h) + \tfrac{1}{2}\sqrt{(f - h)^2 + 4g^2}. \tag{D.8}$$

is of special interest to us, in particular the associated eigenspace, which is one-dimensional and spanned by the vector $v_{\lambda+}$:

$$[v_{\lambda_+}] = [\begin{pmatrix} 1 \\ (\lambda_+ - f)/g \end{pmatrix}] = [\begin{pmatrix} g \\ \lambda_+ - f \end{pmatrix}], \tag{D.9}$$

assuming $g \neq 0$. $[v]$ denotes the linear space generated by vector $v \in \mathbb{R}^2$.

is the linear space generated by the vector v. Of particular interest is the expression $(\lambda_+ - f)/g$, which can be viewed as the slope parameter (direction coefficient) of the line defined by v_{λ_+}. We find the following slope estimator using PCA:

$$
\begin{aligned}
\hat{\hat{b}} = \frac{\lambda_+ - f}{g} &= \frac{1}{2}\left(\frac{h}{g} - \frac{f}{g} + \sqrt{\left(\frac{h}{g} - \frac{f}{g}\right)^2 + 4}\right) \\
&= \frac{1}{2}\left(\frac{\langle y, y\rangle}{\langle x, y\rangle} - \frac{\langle x, x\rangle}{\langle x, y\rangle} + \sqrt{\left(\frac{\langle y, y\rangle}{\langle x, y\rangle} - \frac{\langle x, x\rangle}{\langle x, y\rangle}\right)^2 + 4}\right) \\
&= \frac{||y||^2 - ||x||^2}{2\langle x, y\rangle} + \sqrt{\left(\frac{||y||^2 - ||x||^2}{2\langle x, y\rangle}\right)^2 + 1}, \qquad \text{(D.10)}
\end{aligned}
$$

where (D.2) was used to obtain the expressions containing the vectors x and y. Exactly the same estimator is obtained in Appendix E (see (E.10)) which is derived in a different way, using complex numbers.

From (D.10) it follows that

$$
\hat{\hat{b}} = \frac{\lambda_+ - f}{g} \approx \begin{cases} h/g & \text{if} \quad f \ll h, \\ 1 & \text{if} \quad f \approx h, \\ 0 & \text{if} \quad f \gg h. \end{cases} \qquad \text{(D.11)}
$$

We can also express (D.11) in a more directly understandable form as

$$
\hat{\hat{b}} \approx \begin{cases} \langle y, y\rangle / \langle x, y\rangle & \text{if} \quad ||x|| \ll ||y||, \\ 1 & \text{if} \quad ||x|| \approx ||y||, \\ 0 & \text{if} \quad ||x|| \gg ||y||. \end{cases} \qquad \text{(D.12)}
$$

These approximations are in agreement with intuition. For (D.10) this cannot be claimed as it is a rather complicated expression.

Note that (D.10) contains expressions also introduced in Appendix C where some slope estimators of the regression line were derived from using (what were called) ‘heuristic methods’ derived from the estimator (C.1).

E Other ways to estimate slopes

in this appendix we consider two regression techniques for the complete data set, which [46] yield the same slope estimator as PCA in Section D. In both cases we assume complete data where the center of gravity coincides with the origin of the xy-coordinate system used, so $\langle \iota, x\rangle = 0$ and $\langle \iota, y\rangle = 0$ are supposed to hold.[47]

The approach in Section E.1 uses the view of the data (x_j, y_j) for $j = 1, \ldots, N$ as complex numbers $x_j + y_j i$), with $i^2 = -1$ to find a slope estimator. The approach in Section E.2 uses orthogonal regression to fit the regression line to the data. From a geometrical perspective this approach is more attractive one (than, say in the OLS case). But statistically considered this is not necessarily the case, as the x values may be without error.

E.1 Viewing the data as complex numbers

The method presented here is presented in the Wikipedia lemma on ‘Demming regression’ [48], as a special case. However, from the discussion there the choice of this term is not clear at all (to us). We therefore do not use it here. Instead we reserve ‘Orthogonal regression’ for the method discsussed in Section E.2.

We assume a complete data cloud with N points in the plane. But this time we assume they are complex numbers. So a point $z_j = x_j + iy_j$, for $j = 1, \ldots, N$, where $i^2 = -1$. We assume that the center of gravity of the data set coincides with the origin $\mathcal{O}$ of the axis system used, so:

$$\bar{z} = \sum_{j=1}^{N} z_j = 0. \tag{E.1}$$

Now the sum of (complex) squares with respect to the center of gravity equals

$$S^2 = \sum_{j=1}^{N} \left(z_j - \bar{z}\right)^2 = \sum_{j=1}^{N} z_j^2. \tag{E.2}$$

Suppose $S^2 \neq 0$. Then we can take S as an estimate of a regression line through the origin (or center of gravity). The intercept of this line is 0, of course. The slope of this line can be easily computed. If $S^2 = |S|^2 e^{2i\varphi}$ then $S = |S|e^{i\varphi}$, where $|S|$ is the norm of S and φ the angle, which is also the slope of the regression line we are looking for. It can be expressed as the arctangent of an expression with x and y values of the points in the data set, as is done below.

46) Somewhat surprisingly. But closer inspection undoubtedly cast light on the mystery.
47) As noted before, if the original data do not have this property they can be shifted so that they will.
48) See `https://en.wikipedia.org/wiki/Deming_regression#Orthogonal_regression`.

We can elaborate these ideas. If we represent the point (x_j, y_j) as the complex number $x_j + iy_j$, for $j = 1, \ldots, n$, where $i^2 = -1$ we have

$$
\begin{aligned}
S^2 &= \sum_{j=1}^{N} \left(x_j + iy_j\right)^2 = \sum_{j=1}^{N} x_j^2 - \sum_{j=1}^{N} y_j^2 + 2i \sum_{j=1}^{N} x_j y_j \\
&= \langle x, x\rangle - \langle y, y\rangle + 2i\langle x, y\rangle
\end{aligned}
\tag{E.3}
$$

Now

$$
|S|^2 = |S^2| = (\langle x, x\rangle - \langle y, y\rangle)^2 + 4(\langle x, y\rangle)^2, \tag{E.4}
$$

so

$$
|S| = \sqrt{(\langle x, x\rangle - \langle y, y\rangle)^2 + 4\langle x, y\rangle^2}. \tag{E.5}
$$

Furthermore, assuming $\langle x, x\rangle \neq \langle y, y\rangle$, we have

$$
\tan(2\varphi) = \frac{2\langle x, y\rangle}{\langle x, x\rangle - \langle y, y\rangle}. \tag{E.6}
$$

From (E.6) we can compute $\tan\varphi$, which is the slope of the regression line we want to reconstruct, in two ways. The first is by employing the function arctan. We then get

$$
\tan(\varphi) = \tan\left(\frac{1}{2}\arctan\left(\frac{2\langle x, y\rangle}{\langle x, x\rangle - \langle y, y\rangle}\right)\right), \tag{E.7}
$$

which is a somewhat formal expression for the solution. Another, more useful one, is obtained by employing the well-known trigonometric formula

$$
\tan(2\varphi) = \frac{2\tan\varphi}{1 - \tan^2\varphi}. \tag{E.8}
$$

from which follows

$$
\begin{aligned}
\tan\varphi &= \frac{-1 + \sqrt{1 + \tan^2(2\varphi)}}{\tan(2\varphi)} \\
&= \frac{-1}{\tan(2\varphi)} + \sqrt{\frac{1}{\tan^2(2\varphi)} + 1},
\end{aligned}
\tag{E.9}
$$

where the other root (with minus the square root in the numerator) was excluded as it is irrelevant. Applying (E.9) to (E.6) yields

$$\begin{aligned}\tan\varphi &= \frac{\langle y,y\rangle - \langle x,x\rangle}{2\langle x,y\rangle} + \sqrt{\left(\frac{\langle y,y\rangle - \langle x,x\rangle}{2\langle x,y\rangle}\right)^2 + 1} \\ &= \frac{||y||^2 - ||x||^2}{2\langle x,y\rangle} + \sqrt{\left(\frac{||y||^2 - ||x||^2}{2\langle x,y\rangle}\right)^2 + 1},\end{aligned} \tag{E.10}$$

which is the same as (D.10), the PCA slope estimator.

E.2 Orthogonal regression

The idea is to consider a centered data cloud. All regression lines cut this center. We wish to measure the distance of a data point to a regression line by the shortest distance between that point and the line, not in the way OLS does it, by looking at the difference in y values for a given data point. This actually makes it depend on the coordinate system, by considering orthogonal regression (using shortest distances between a set of points and a line) we use a coordinate free approach. We can compute these distances easily by first rotating the data cloud in such a way that the rotated regression line co-incides with the x axis and then compute the residues, which are the y values of the rotated data points.

Let

$$\bar{b} = \begin{pmatrix} \cos\vartheta \\ \sin\vartheta \end{pmatrix} \tag{E.11}$$

be a unit vector representing the slope of a regression line. Its orthocomplement is

$$\bar{b}^{\perp} = \begin{pmatrix} -\sin\vartheta \\ \cos\vartheta \end{pmatrix} \tag{E.12}$$

The linear transformation from the unit vector

$$\bar{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{E.13}$$

to $\bar{b}$ and

$$\bar{e}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{E.14}$$

to $\bar{b}^{\perp}$ is

$$T = \begin{pmatrix} \cos\vartheta & -\sin\vartheta \\ \sin\vartheta & \cos\vartheta \end{pmatrix} \tag{E.15}$$

We need the reverse transformation of T in (E.15):

$$T^{-1} = T' = \begin{pmatrix} \cos\vartheta & \sin\vartheta \\ -\sin\vartheta & \cos\vartheta \end{pmatrix}. \tag{E.16}$$

If we apply (E.16) to the regression line and the data, we get a horizontal regression line through the origin and transformed data set, as follows

$$\begin{pmatrix} x_i' \\ y_i' \end{pmatrix} = \begin{pmatrix} \cos\vartheta & \sin\vartheta \\ -\sin\vartheta & \cos\vartheta \end{pmatrix} \begin{pmatrix} x_i \\ y_i \end{pmatrix}, \tag{E.17}$$

for $i = 1, \ldots, N$. From (E.17) we take the y_i' as residues. The quadratic sum of these residues is the loss function

$$\sum_{i=1}^{N} (y_i')^2. \tag{E.18}$$

Minimizing (E.18) for ϑ yields the optimum slope of the regression line for this type of regression, namely orthogonal regression. So we first elaborate (E.18) which yields

$$\sum_{i=1}^{N} (y_i')^2 = \langle x, x\rangle \sin^2\vartheta + \langle y, y\rangle \cos^2\vartheta - 2\langle x, y\rangle \sin\vartheta\cos\vartheta. \tag{E.19}$$

Differentiating (E.19) for ϑ and equating the result to 0 yields the equation

$$2\left(\langle x, x\rangle - \langle y, y\rangle\right) \sin\vartheta\cos\vartheta + 2\langle x, y\rangle \left(\sin^2\vartheta - \cos^2\vartheta\right) = 0. \tag{E.20}$$

Dividing (E.20) by $2\cos^2\vartheta$ (which we may assume to be strictly positive) yields

$$\left(\langle x, x\rangle - \langle y, y\rangle\right) \tan\vartheta + \langle x, y\rangle \left(\tan^2\vartheta - 1\right) = 0. \tag{E.21}$$

Rewriting (E.21) produces the following quadratic equation in $\tan\vartheta$:

$$\langle x,y\rangle \tan^2\vartheta + (\langle x,x\rangle - \langle y,y\rangle)\tan\vartheta - \langle x,y\rangle = 0. \tag{E.22}$$

Solving (E.22) yields as the largest root, and largest eigenvalue (corresponding to the most important eigenspace):

$$\tan\vartheta = \frac{\langle y,y\rangle - \langle x,x\rangle + \sqrt{(\langle y,y\rangle - \langle x,x\rangle)^2 + 4\langle x,y\rangle^2}}{2\langle x,y\rangle} \tag{E.23}$$

$$= \frac{\langle y,y\rangle - \langle x,x\rangle}{2\langle x,y\rangle} + \sqrt{\left(\frac{\langle y,y\rangle - \langle x,x\rangle}{2\langle x,y\rangle}\right)^2 + 1}$$

$$= \frac{||y||^2 - ||x||^2}{2\langle x,y\rangle} + \sqrt{\left(\frac{||y||^2 - ||x||^2}{2\langle x,y\rangle}\right)^2 + 1}, \tag{E.24}$$

which is precisely the result obtained in (E.10) as well as in (D.10). Why this is the case may not be immediately clear. In Section E.3 we consider this remarkable fact.

E.3 Coincidence or necessity?

Here we consider the fact that the slope estimators obtained in Sections D, E.1 and E.2 are the same. Is this coincidence or necessity?

It should be noted that the estimator derived in Section E.2 is the result of solving an optimization problem. The target function is (E.18), which is a free choice. A loss function could have been defined differently and one still would be dealing with a (kind of) orthogonal regression. But the resulting slope estimator would have been different with a different choice of loss function. So the fact that (E.23) is the same as the other two slope estimators is a coincidence, can be argued.

Now the estimator in Section D is the result of the fact that it depends on the covariance matrix (D.2). The fact that matrix multiplication can be described in terms of standard inner products implies that the entries of the covariance matrix consist of the right ingredients, namely $\langle x,x\rangle$, $\langle x,y\rangle$ and $\langle y,y\rangle$ in the right way. It would be possible to define matrix multiplication using a more general kind of inproduct, such as

$$\langle x,y\rangle_w = \sum_{j=1}^{N} w_j x_j y_j, \tag{E.25}$$

where $w = (w_1,\ldots,w_N)$ is a vector of fixed weights $w_j \geq 0$ of length N and $x = (x_1,\ldots,x_N)$ and $y = (y_1,\ldots,y_N)$.[49] Using the inner product (E.25) we can define an alternative matrix product of $m \times N$ and $N \times k$ matrices, X and Y, respectively:

[49] The 'w' used here to denote weights should not be confused with the superscript 'w' denoting observed data after censoring the orignal data (inspired by the Dutch word 'waargenomen', 'observed' in English.

$$X = \begin{pmatrix} x'_{1\cdot} \\ \vdots \\ x'_{m\cdot} \end{pmatrix}, \tag{E.26}$$

where $x'_{1\cdot}, \dots, x'_{m\cdot}$ are row vectors of length N, and

$$Y = (y_{\cdot 1}, \cdots, y_{\cdot m}), \tag{E.27}$$

with $y_{\cdot 1}, \dots, y_{\cdot m}$ column vectors of length n.

We now define the matrix product $\circ_w$ of X and Y as follows:

$$X \circ_w Y = \begin{pmatrix} \langle x_{1\cdot}, y_{\cdot 1}\rangle_w & \cdots & \langle x_{1\cdot}, y_{\cdot k}\rangle_w \\ \vdots & & \vdots \\ \langle x_{m\cdot}, y_{\cdot 1}\rangle_w & \cdots & \langle x_{m\cdot}, y_{\cdot k}\rangle_w \end{pmatrix}. \tag{E.28}$$

If $D' \circ_w D$ with $w \neq \iota$ would have been used instead of $D'D$ in (D.2) then the resulting slope estimator would probably have been different from both estimators derived in Sections E.1 and E.2.

Note that the eigenvalues of the covariance matrix C in (D.2) are computed from the characteristic equation (D.3), which happens to be similar to the normal equation (E.21) resulting from an optimization condition. If we consider the expression S^2 in (E.3) we see the same ingredients appear as its real and imaginary parts, as a result of the definition of algebraic operations such as multiplication and addition for complex numbers. This has nothing to do with the determination of eigenvalues or the computation of a best fitting (regression) line. So one could consider this also to be a lucky coincidence.

So the overall conclusion is that the three slope estimators could have been different if different definitions of matrix multiplication and a different object function had been chosen in the orthogonal regression case. So these different procedures result in the same slope estimator is a result of coincidence. No matter how unlikely the alternatives mentioned are, they cannot formally be excluded.

F The roof

Roughly speaking the roof consists of the points at the top of the censored data cloud. They are the closest to the original regression line. And they are all below this line. This information is useful for deriving constraints for optimization problems that can be formulated to reconstruct such a regression line.

Also the roof points are closest of the censored data to the original regression line. For this reason one would expect that they carry sufficient information to help reconstruct this line to good approximation.

The roof of the censored data is defined in Section F.1. In Section F.2 it is indicated how the roof can be computed by presenting links to the computational geometry literature, where a similar problem (computation of convex hulls) has been studied for a long time and useful algorithms have been found.

F.1 The roof defined

Let $\mathcal{P}^w$ be the set of observed data points.[50] Consider $\mathcal{P}^w$ as part of an area $\mathcal{A}$ bounded by the line $x = x_1^w$ and the line $x = x_n^w$, where it is assumed that $x_1^w \leq \cdots \leq x_n^w$. The third part of the boundary of $\mathcal{A}$ is the top part of the convex combination of the points in $\mathcal{P}^w \subseteq \mathcal{A}$, connecting the points (x_1^w, y_1^w) on the left and (x_n^w, y_n^w) on the right. This is the roof of $\mathcal{P}^w$.

Below the roof is the avoidance area $\mathcal{A}$ that potential regression lines cannot intersect, except for the roof itself.[51]

Figure F.1 is an example of an observed data set and its roof, as well as the avoidance area $\mathcal{A}$ for regression lines (defined in Section 7.1). $\mathcal{A}$ is a convex set.

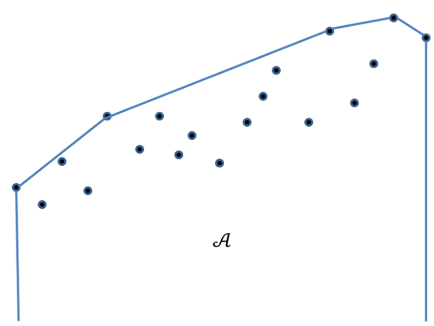

Figure F.1 Roof of an observed data set of points at the top of avoidance area $\mathcal{A}$.

In Section F.2 we shall briefly consider the problem of computing the roof in practice. We shall present an intuitive (and informal) description how to do this. For details one is referred to the area where such problems are typically studied, namely computational geometry.

50) The roof can be defined for any data set of points in the plane. But as we apply this concept specifically to a set of observed data points we have made the restriction with an eye to this application.

51) Excluding the roof would create a situation where regression lines can be arbitrarily close to the avoidance area, implying that there would not be a closest one. So the interior of $\mathcal{A}$, often denoted $\mathcal{A}^\circ$, is the set that actually should be avoided, and not the roof.

F.2 Computing the roof

Computing the roof is obviously of importance when carrying out some of the computations proposed in this paper. Fortunately this does not involve a new problem. In computational geometry similar problems have been studies for a long time and algorithms have been developed to solve them. One such problem is the computation of the convex hull of a (finite) set of points in the plane. One of these algorithms is Jarvis's march (see [8], Section 3.3.3). Intuitively this algorithm can be described as finding the convex hull by wrapping a string tightly around the set of points in question, which can be imagined as a set of pegs hammered into a wooden board at the locations where the points are situated. In our case, the roof is only part of the convex hull of the censored data. It is the only part of interest for our problem of reconstructing the regression line.

The roof consists of extreme points in the censored data set. To find them is easiest by considering pairs of points in the data set, each pair defining a line. If both points are extreme points (that is roof points in our case) all other points should be below this line.

We now consider an approach that is very suitable to be used in practice. It constructs parts of the roof step by step, moving from left to right through the censored data (from small x coordinates to bigger ones). Once a part is completed, the next part is computed in a manner very similar to a previous part, in particular the first part. As the method finds (at least) one new roof point, the number of points left to process strictly decreases at each step, so that the method will finish after a finite number of steps, which is less than the number of unprocessed points remaining.

To start the process we note that the point $\pi_1 = (\xi_1, \eta_1)$ in the censored data set with the smallest x value is a roof point. To find the next roof point $\pi_2 = (\xi_2, \eta_2)$ we consider all points $p \neq \pi_1$ in the censored data set and consider the line defined by π_1 and p. π_2 is the point for which the line through π_1 and π_2 is above all the points in the censored data with $x \geq x_1$. It actually has the largest slope:

$$\tau^1 = \max_{j=2,\dots,n} \{\tau_{1j}\}. \tag{F.1}$$

where

$$\tau_{1j} = \frac{y_j - \eta_1}{x_j - \xi_1}, \tag{F.2}$$

where $p_j = (x_j, y_j)$, for $j = 2, \dots, n$ such that $x_j \geq \xi_1$.

Now the line segment $[\pi_1, \pi_2]$ dominates all points in the censored data with x values in $[\xi_1, \xi_2]$ which therefore is part of the roof. We now repeat the procedure with π_2 in the role of π_1 and look for the next roof point π_3. And then continue this process until we reach the final roof point, which is the one with the largest x value (and if there are multiple ones, the one with the largest y value among these).

G Distribution free tests for slope estimators

In [4], Chapter 6, various procedures and tests are discussed for regression problems involving slopes. In Section 9.1 of this book a distribution free test is discussed for the slope of the regression line with a specified slope, say b_0 against various alternatives that the slope is larger, i.e. $b > b_0$. The test involves quantities

$$D_i = y_i - b_0 x_i, \tag{G.1}$$

for $i = 1, \ldots, n$. The test statistic used is

$$C = \sum_{i<j} c(D_i - D_j), \tag{G.2}$$

where

$$c(a) = \begin{cases} 1 & \text{if } a > 0, \\ 0 & \text{if } a = 0, \\ -1 & \text{if } a < 0. \end{cases} \tag{G.3}$$

The statistic C is Kendall's K statistic computed between the values x and $y - b_0 x$. See [4], Section 8.1 for a discussion of Kendall's distribution-free test for independence of random variables.

Consider the following null hypothesis and a one-sided alternative hypothesis concerning the slope of the regression line, where b_0 denotes a specified slope:

$$H_0 : b = b_0, \tag{G.4}$$
$$H_1 : b > b_0. \tag{G.5}$$

At the α level of significance we have the following acceptance criteria for these hypotheses:

$$\text{accept } H_0 \text{ if } C < k(\alpha, n), \tag{G.6}$$
$$\text{accept } H_1 \text{ if } C \geq k(\alpha, n). \tag{G.7}$$

The understanding is that accepting H_1 is synonymous to rejecting H_0, as there is no other alternative available.

The parameters $k(\alpha, n)$, which satisfy the equation $P_0[C \geq k(\alpha, n)] = \alpha$, can be found in Table A21 in [4].[52)]

Another test with a one-sided alternative is the following one:

$$H_0 : b = b_0, \tag{G.8}$$
$$H_2 : b < b_0. \tag{G.9}$$

At the α level of significance we have the following acceptance criteria for these hypotheses:

$$\text{accept } H_0 \text{ if } C > -k(\alpha, n), \tag{G.10}$$
$$\text{accept } H_2 \text{ if } C \leq -k(\alpha, n). \tag{G.11}$$

Again, it is understood that accepting H_2 is the same as rejecting H_0, due to lack of an alternative.

A two-sided test set-up is the following one

$$H_0 : b = b_0, \tag{G.12}$$
$$H_3 : b \neq b_0. \tag{G.13}$$

At the α level of significance we have the following acceptance criteria for these hypotheses:

$$\text{accept } H_0 \text{ if } -k(\alpha_1, n) < C < k(\alpha_2, n), \tag{G.14}$$
$$\text{accept } H_3 \text{ if } C \leq -k(\alpha_1, n) \text{ or } C \geq k(\alpha_2, n), \tag{G.15}$$

where $\alpha = \alpha_1 + \alpha_2$ and $\alpha_1, \alpha_2 > 0$.

To gain insight into the large sample behaviour of C, consider the statistic

$$C^* = \frac{C - E_0(C)}{\mathrm{var}_0(C)^{1/2}} = \frac{C}{(n(n-1)(2n+5)/18)^{1/2}}. \tag{G.16}$$

It is known that the statistic C^* converges in distribution to $\mathcal{N}(0, 1)$, the standard normal distribution (see e.g. [4], Section 9.1).

52) Kendall's statistic is denoted in Table A21 as K.

H Completing the censored data

In the present appendix we sketch some ideas related to the challenge to complete the uncensored data set, which then is used to estimate a regression line using OLS. It is likely that a successive estimate requires an iteration of these steps to obtain a satisfactory estimate of the regression line.

In this appendix we describe this estimation problem in a general, distribution-free context. In Section 8, however, an application of this method is described using a parametric model for the residues in the data, in particular the data in the censored data set. The method used here to compute an acceptable regression line is iterative. In fact, because a parametric model is used, the mode of computation is somewhat different from what is described here.

The basis idea of the method described here is as follows. If a candidate regression line is known for the censored data, we can apply a kind of mirroring of these data in the regression line to obtain a sensible imputation for the missing data situated above this line.

So let the candidate regression line be given by

$$y = a + bx. \tag{H.1}$$

If (x_i^w, y_i^w) is a point in the censored data set (hence situated below the regression line) then its mirror point above the regression line is $(x_i^w, 2a + 2bx_i^w - y_i^w)$. Note that the original point and its mirror point have the same distance to the regression line (H.1). So formally we have the mapping $\mathfrak{m} : \mathcal{C}^w \to \mathbb{R}^2$, where $\mathcal{C}^w \subset \mathbb{R}^2$ are the censored data, with

$$\mathfrak{m} : (x_i^w, y_i^w) \mapsto (x_i^w, 2a + 2bx_i^w - y_i^w). \tag{H.2}$$

Then $\mathfrak{m}(\mathcal{C}^w)$ could be viewed as an imputation for the missing part of the data cloud, i.e $\mathcal{C}^c$, albeit a rather primitive one: the mirror image would be too perfect, as it is without any uncertainty. But it should be viewed as a first attempt to produce a mirror distribution. More realistic ones can be constructed fairly easily, as we will show.

The next step follows by noting that the 'observed' residues

$$\rho_i = a + bx_i - y_i, \tag{H.3}$$

for $i = 1, \ldots, n$ are quantities (ideally) independent of the x-values, as we assume homoskedasticity for the original regression line, and also for the reconstructed regression lines.

So, if true, we could sample[53)] from the set of residues $\{\rho_1, \dots, \rho_n\}$ and assign, to the point (x_i^w, y_i^w), as its mirror image the point $(x_i^w, a + bx_i^w + \rho_j)$, where ρ_j is supposed to be the randomly selected residue for the image of point (x_i^w, y_i^w). In this way we add more variability to the imputed missing data points above the regression line.

Remark In fact, we could even go further and take x-values that are not used in the observed data points. Assuming that the range of the x-values is $[x_{(1)}^w, x_{(n)}^w]$, where $x_{(1)}^w$ is the smallest x value in the censored data and $x_{(n)}^w$ the largest one, we can randomly sample n values from this range, say $\xi_1, \dots, \xi_n$ and compute $\eta_i = a + b\xi_i$. To these values we add randomly chosen (srswr)[54)] residues ρ_j so that we obtain $\theta_i = \eta_i + \rho_{j(i)}$, for $i, j(i) = 1, \dots, n$, where $j(i)$ is intended to express that j belongs to i, not that j depends on i. So we now have the set of mirror image points (ξ_i, θ_i), for $i = 1, \dots, n$. A variant for such an approach (using a parametric model for the residues) is actually applied in Section 8. ☐

In practice we do not know the regression line, of course. (We are trying to find one!) So we have to work with estimates $y = \hat{a} + \hat{b}x$, instead, obtained from methods in the main text using the censored data directly. Repeating the above method we obtain a completed synthetic data set from which we can estimate a regression line, adjust it a bit if necessary and use it to mirror the uncensored data as an imputation method. This alternating process of estimating the regression line and using it in turn to mirror the uncensored data to obtain complete synthetic data set can be repeated many times.[55)] Ideally this process should converge, in which case one has both a completed (partly) synthetic data set as well as a corresponding regression line to match.

Remark When the data set has been completed by a mirroring step in the above procedure, it needs to be checked if the regression line used fited well enough. If not we might see, for instance, a widening gap in the data with increasing x, or, with decreasing x. This would indicate that the slope needs adjustment. Or we see a rather wide (or reversely, a rather narrow) gap between the uncensored data and their mirrored counterpart. This would indicate that the intercept is not right. And adjusting of any of these shortcomings is obvious. To track them one should consider the residues $\varepsilon_i = a + bx_i - y_i$, for both the censored data and the imputed data. Ideally they should be independent of the indices i (assuming homoskedaticity). So the order of the positive residues $\varepsilon_{(n+1)} \leq \cdots \leq \varepsilon_{(2n)}$ should be independent of the indices i. Also the gaps $\varepsilon_{(n+1)} - \varepsilon_{(n)}$ and $\varepsilon_{(n)} - \varepsilon_{(n-1)}$ should be comparable in size. How to make adjustments on the basis of this information we leave as an open question to be researched. Here we only wanted to point out a direction in which possibly to proceed. ☐

We complete this appendix with three final remarks.

Remark Another possibility to judge the suitability of a regression line before a mirroring step is to consider the roof of the uncensored data and its counterpart for the imputed data, which we call the floor. One needs to consider these two sets in order to judge whether the regression line was ok, in particular its slope. Also a very wide (or very narrow) gap over the entire x domain might indicate an intercept that is too big (or too small, respectively). Using roof and floor might be more suitable for making these judgements than using the entire data sets, of censored and imputed data, because these sets are smaller in size. ☐

53) Using, say, simple random sampling with replacement (srswr), which means that we draw these residues with equal probability and independently of each other, in each step of a bootstrap procedure.

54) srswr = simple random sampling with replacement.

55) It resembles the EM algorithm as is noted in Section 8.

Remark Suppose that by some method the censured points have been imputed. This yields an imputed data cloud hovering on top of the observed data, with the floor at the bottom. A feasible regression line should separate the imputed and the observed data clouds. This means it should lie between the floor of the imputed data and the roof of the observed data. One could look at all the lines that accomplish this and average them. This can be used as an estimator of the regression line. Instead of considering al these lines one could sample them and consider the average from this sample as an estimate. If both ‘half clouds’ are close, there is not much margin for such regression lines, and the estimates should be pretty ‘sharp’. □

Remark In Section 8 the adjustment of the regression line is controlled by the points of the censored data set that lie above the estimated regression line: they are excluded in the next estimation step.[56)] This tends to push the next regression line upwards and/or tilts it. □

56) Because these points are by definition below the original regression line, and also should be below the reconstructed regression line.

Colophon

Publisher
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

Prepress
Statistics Netherlands, Grafimedia

Design
Edenspiekermann

Information
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contact form: www.cbs.nl/information