Title
Linear regression with an observation distribution model

Authors
D D Lichti
Corresponding Author
Department of Geomatics Engineering
The University of Calgary
Calgary, Alberta, Canada
ddlichti@ucalgary.ca
ORCID 0000-0001-7038-073X

T O Chan
Guangdong Provincial Key Laboratory of Urbanization and Geo-simulation
School of Geography and Planning
Sun Yat-sen University
Guangzhou, China
chantingon@mail.sysu.edu.cn

D Belton
School of Earth and Planetary Sciences
Curtin University
Perth, Western Australia, Australia
D.Belton@curtin.edu.au

Abstract
Despite the high complexity of the real world, linear regression still plays an important role in estimating parameters to model a physical relationship between at least two variables. The precision of the estimated parameters, which can usually be considered as an indicator of the solution quality, is conventionally obtained from the inverse of the normal equations matrix for which intensive computation is required when the number of observations is large. In addition, the impacts of the distribution of the observations on parameter precision are rarely reported in the literature. In this paper, we propose a new methodology to model the distribution of observations for linear regression in order to predict the parameter precision prior to actual data collection and performing the regression. The precision analysis can be readily performed given a hypothesized data distribution. The methodology has been verified with several simulated and real datasets. The results show that the empirical and model-predicted precisions match very well, with discrepancies of up to 6% and 3.4% for simulated and real datasets, respectively. Simulations demonstrate that these differences are simply due to finite sample size. In addition, simulation also demonstrates the relative insensitivity of the method to noise in the independent regression variables that causes deviations from the data distribution function. The proposed methodology allows straightforward prediction of the parameter precision based on the distribution of the observations related to their numerical limits and geometry, which greatly simplify design procedures for various experimental setups commonly involved in geodetic surveying such as LiDAR data collection.

Keywords
regression, least-squares, estimation, observation distribution, normal equations

Funding
This work is supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2018-03775)

Conflicts of interest/Competing interests
None

Availability of data and material
Not available

Code availability
Not available

Authors' contributions
DDL designed the research; DDL, TOC and DL performed the research; DDL, TOC and DB analyzed data; DDL and TOC wrote the paper

1. Introduction

Linear regression is a least-squares estimation methodology commonly used in many science and engineering disciplines to infer the physical relationship between variables (Abraham and Ledoliter 2006). It finds use in geodesy for a range of applications including, but certainly not limited to, geodetic instrument calibration (Lichti et al. 2010; Jazireeyan and Ardalan 2017), coordinate transformation estimation (Schaffrin and Felus 2008; Mahboub 2012; Ruffhead 2018), seasonal glacier mass balance determination (Pelto et al. 2019) and gravimetric geoid fitting (Featherstone and Lichti 2009). Much literature exists on regression for robust estimators (Yang 1999), differently-distributed data types (homoscedastic and heteroscedastic; Hekimoglu and Berbe 2003), multiple variables, errors-in-variables (Schaffrin and Wieser 2008), etc. However, to the authors' best knowledge, the role of the distribution of the independent-variable observations is not addressed. The impact of the observation distribution on the precision of parameter estimates determined by linear regression is investigated here.

1.1. Linear regression

Assume the following linear relationship between two variables, x and y, which is a function of two unknown parameters, the slope m and y-axis intercept, b

$$y + e = f(x) = mx + b \tag{1}$$

Unlike in the error-in-variables approach, the x observation is considered to be error-free. The y observation is affected by random error, e, which is modelled in terms of the first two moments of its symmetric probability density function

$$E\{e\} = 0 \tag{2}$$

$$E\{e^2\} = \sigma^2 \tag{3}$$

For a set of n observations, the relationship can be expressed as follows

$$\mathbf{y} + \mathbf{e} = \mathbf{A}\mathbf{x}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} \tag{4}$$

where $\mathbf{y}$ and $\mathbf{e}$ are n×1 vectors of observations and random errors, respectively, $\mathbf{x}$ is the u×1 unknown parameter vector, and $\mathbf{A}$ is n×u the design matrix. The stochastic model of the random errors is specified by

$$E\{\mathbf{e}\} = \mathbf{0} \tag{5}$$

$$E\{\mathbf{e}\mathbf{e}^T\} = \mathbf{C}_y \tag{6}$$

Here, the observation errors are assumed to be uncorrelated and drawn from distributions having equal variance, so the n×n positive-definite covariance matrix $\mathbf{C_y}$ is a scalar matrix

$$\mathbf{C_y} = \sigma^2 \mathbf{I} \tag{7}$$

The weight matrix, **P**, is generally defined as follows

$$\mathbf{P} = \frac{1}{\sigma_0^2} \mathbf{C_y^{-1}} \tag{8}$$

where the a priori variance factor is usually assumed to be unity, i.e. $\sigma_0^2 = 1$.

The least-squares normal equations are obtained from the minimization of the quadratic form of the errors, $\mathbf{e}^T\mathbf{Pe}$:

$$\begin{aligned} \mathbf{A}^T\mathbf{PAx} &= \mathbf{A}^T\mathbf{Py} \\ \mathbf{Nx} &= \mathbf{U} \end{aligned} \tag{9}$$

Under the assumptions stated above, the weight matrix is given by

$$\mathbf{P} = \frac{1}{\sigma^2} \mathbf{I} \tag{10}$$

Thus, for the linear regression problem of Equation 1, the analytical form of the normal equations is given by

$$\frac{1}{\sigma^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} 1 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \frac{1}{\sigma^2} \begin{pmatrix} \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{pmatrix} \tag{11}$$

If the mean value is subtracted from the x coordinates, then the off-diagonal term of the normal equations matrix **N** equals zero, leaving

$$\frac{1}{\sigma^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & 0 \\ 0 & \sum_{i=1}^{n} 1 \end{pmatrix} \begin{pmatrix} m \\ b \end{pmatrix} = \frac{1}{\sigma^2} \begin{pmatrix} \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{pmatrix} \tag{12}$$

Note that no notational distinction is made between original and reduced x coordinates.)

1.2 Data imbalance and data distribution

It is hypothesized that the distribution of the x coordinate observations used to solve Equation 12 can have a significant impact on the estimated parameters and their precision. In the design of an experiment, prior knowledge of exactly how the sampling of the function f will impact the regression results is desirable. This is tantamount to first order network design (Grafarend 1974) where one seeks a geometric network configuration to meet some target quality criteria. This is a familiar problem (e.g., Grafarend and Sansò 1985) for which general principles governing the preferred values of the observations are known (e.g. Berné et al. 2004). However, the focus of the design process is in parameter space (point placement) rather than in observation space.

One might intuitively strive for a uniform sampling of the function f. This is the aim of electronic distance measurement baseline design (Rüeger 1990). Ideally, the configuration of instrument stations/pillars should be such that there is an equal distribution of all measured distances without any repeats. However, this cannot be achieved for five or more stations because a gracefully-numbered ruler must have four or fewer marks in order to measure all integral distances, from zero to its length, only once (Rüeger 1990). In another context, sampling theory has been employed for third-order design in high-precision levelling networks used for subsidence monitoring (Holst et al 2013). The optimal number of stations and their spacing are determined by analyzing the observability of sinusoidal terms in the estimated deformation surface and redundancy numbers.

In reality, the desired sampling regime may not be realized due to a number of factors including experimental conditions, available instrumentation and gross errors such as lost data due to accidents (Rawlings et al. 1989). The result is known as data imbalance. Searle (1986) describes techniques to handle unbalanced data in the context of multiple regression. It is currently a topic of great interest for handling big datasets (Leevy et al. 2018) and machine learning (Haixiang et al. 2017). The focus of this work is not, however, multivariable regression for which model classes have been observed unequally. Instead, the aim is to quantify the impact of sampling distribution on the quality of linear regression estimates.

Simulation approaches developed in recent years attempt to solve first order network design for indoor 3D mapping (e.g., Mozaffar and Varshosaz 2016; Soudarissanane 2016; Jia and Lichti 2019) where the aim is to create a model representation of a complex environment. Broadly speaking, these methods generate a set of hypothesized instrument locations within an environment that meet certain criteria such as maximum range and incidence angle. The impact of the observation distribution on the model quality could be investigated rather easily by simulation of synthetic observations. For example, Holst et al. (2014) investigate the impact of scanning geometry and sampling density on estimated surface models for deformation monitoring by numerical simulation. Whilst these approaches are sound, the simulation process adds extra computational load to a task that is already computationally intensive. Thus, a less burdensome but also more broadly applicable methodology is desired. Such a method has been developed herein and is demonstrated on both simulated and real datasets.

2. Linear Regression

2.1 Uniform Symmetric Data Distribution

The methodology development commences with the assumption that the x-coordinates are uniformly distributed on a finite interval [-a, a]. Both sides of Equation 12 are multiplied by the sampling interval, Δx, where

$$\Delta x = \frac{2a}{n} \quad -a \le x_i \le a \tag{13}$$

so the normal equations become

$$\frac{\Delta x}{\sigma^2}\begin{pmatrix} \sum_{i=1}^{n} x_i^2 & 0 \\ 0 & \sum_{i=1}^{n} 1 \end{pmatrix}\begin{pmatrix} m \\ b \end{pmatrix} = \frac{\Delta x}{\sigma^2}\begin{pmatrix} \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{pmatrix} \tag{14}$$

The x data limits [-a, a] are symmetric due to the uniform sampling and the subtraction of the mean.

Noting that all terms within the summations are continuous on [-a, a], the definition of the definite integral can be used. As the number of observations becomes large, that is as $n \to \infty$, the normal equations become

$$\mathbf{N} = \lim_{n\to\infty} \frac{\Delta x}{\sigma^2}\begin{pmatrix} \sum_{i=1}^{n} x_i^2 & 0 \\ 0 & \sum_{i=1}^{n} 1 \end{pmatrix} = \frac{1}{\sigma^2}\begin{pmatrix} \int_{-a}^{a} x^2 dx & 0 \\ 0 & \int_{-a}^{a} dx \end{pmatrix} \tag{15}$$

and

$$\mathbf{U} = \lim_{n\to\infty} \frac{\Delta x}{\sigma^2}\begin{pmatrix} \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{pmatrix} = \frac{1}{\sigma^2}\begin{pmatrix} \int_{-a}^{a} x \cdot y \cdot dx \\ \int_{-a}^{a} y \cdot dx \end{pmatrix} \tag{16}$$

2.2 General Symmetric Data Distribution

It has been assumed in the formulation above that the distribution of the x coordinates on [-a, a] is uniform, which is not necessarily always true. A more general expression can be developed by redefining the weight matrix with diagonal elements that quantify the sampling distribution, $p(x_i)$, as follows

$$\mathbf{P} = \frac{1}{\sigma^2}\begin{pmatrix} p(x_1) & 0 & \cdots & 0 \\ 0 & p(x_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p(x_n) \end{pmatrix} \tag{17}$$

Taking the limit of the resulting normal results in

$$\mathbf{N} = \lim_{n\to\infty}\frac{\Delta x}{\sigma^2}\begin{pmatrix} \sum_{i=1}^{n} p(x_i)x_i^2 & 0 \\ 0 & \sum_{i=1}^{n} p(x_i) \end{pmatrix} = \frac{1}{\sigma^2}\begin{pmatrix} \int_{-a}^{a} x^2 p(x)\,dx & 0 \\ 0 & \int_{-a}^{a} p(x)\,dx \end{pmatrix} \tag{18}$$

and

$$\mathbf{U} = \lim_{n\to\infty}\frac{\Delta x}{\sigma^2}\begin{pmatrix} \sum_{i=1}^{n} p(x_i)x_i y_i \\ \sum_{i=1}^{n} p(x_i)y_i \end{pmatrix} = \frac{1}{\sigma^2}\begin{pmatrix} \int_{-a}^{a} x\cdot y\cdot p(x)\,dx \\ \int_{-a}^{a} y\cdot p(x)\,dx \end{pmatrix} \tag{19}$$

The function p(x) is defined such that it satisfies the following property

$$\int_{-\infty}^{\infty} p(x)\,dx = 1 \tag{20}$$

The function p(x) represents the distribution of the observations in x on the interval [-a, a] according to the sampling process. This function should be chosen such that the integrands of Equations 18 and 19 are continuous. Moreover, the condition p(x)>0 must be satisfied so that the weight matrix is positive definite. Although p(x) possesses some properties similar to those of a probability density function, it is not interpreted as such.

The solution for **x** is determined in the usual way by multiplying **U** by the inverse of **N**. Proof that the estimated line parameters are unbiased is given in Appendix A. The covariance matrix of the parameters, which quantifies parameter precision, is given by the inverse of the full-rank normal equations matrix

$$\mathbf{C_x} = \mathbf{N}^{-1} = \begin{pmatrix} \sigma_m^2 & \sigma_{mb} \\ \sigma_{mb} & \sigma_b^2 \end{pmatrix} \tag{21}$$

The matrix $\mathbf{C_x}$ has the following form due to the translation of the x data to the centroid and the condition given by Equation 20.

$$\mathbf{C_x} = \begin{pmatrix} \sigma_m^2 & 0 \\ 0 & \sigma_b^2 \end{pmatrix} = \begin{pmatrix} \dfrac{\sigma^2}{\int_{-a}^{a} x^2 p(x)\,dx} & 0 \\ 0 & \sigma^2 \end{pmatrix} \tag{22}$$

Equation 22 reveals that the precision of the intercept parameter b depends only on the observation precision, σ^2. It is independent of the distribution of the observations. The precision of the slope parameter is also a function of observation precision but is strongly dependent on the observation distribution because it is inversely proportional to the variance of the x coordinates. Numerical estimates of the parameter variances obtained from the inverse of the normal equations matrix (Equation 11) can be directly compared with those of Equation 18 following multiplication by the sample size, n. The estimates agree with the sample size is large, as will be demonstrated.

3. Specific Observation Distributions

Here, the impact of specific observation distributions on the precision of the linear regression parameters is examined. All functions are symmetric and centred at x=0.

3.1 Uniform distribution

A typical example of data following this distribution is a uniformly-sampled time series.

$$p(x) = \begin{cases} \dfrac{1}{2a} & -a \le x \le a \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

$$\mathbf{C_x} = \begin{pmatrix} \dfrac{3\sigma^2}{a^2} & 0 \\ 0 & \sigma^2 \end{pmatrix} \tag{24}$$

3.2 Triangular (tent function) distribution

In this case, the observations are highly concentrated near the mean x coordinate value and their density decays linearly.

$$p(x) = \begin{cases} \dfrac{1}{a}\left(1 - \left|\dfrac{x}{a}\right|\right) & -a < x < a \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

$$\mathbf{C}_{\mathbf{x}} = \begin{pmatrix} \frac{6\sigma^2}{a^2} & 0 \\ 0 & \sigma^2 \end{pmatrix} \tag{26}$$

3.3 Raised cosine distribution

This function is a bounded approximation to the Gaussian function. The observations are more heavily concentrated near the mean x coordinate value than in the triangular case.

$$p(x) = \begin{cases} \frac{1}{2a}\left(1+\cos\left(\frac{\pi x}{a}\right)\right) & -a < x < a \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

$$\mathbf{C}_{\mathbf{x}} = \begin{pmatrix} \frac{3\pi^2\sigma^2}{a^2\left(\pi^2-6\right)} & 0 \\ 0 & \sigma^2 \end{pmatrix} \approx \begin{pmatrix} \frac{7.65\sigma^2}{a^2} & 0 \\ 0 & \sigma^2 \end{pmatrix} \tag{28}$$

3.4 Simulations

Analyses of Equations 24, 26 and 28 illustrate the strong influence of data distribution on the slope parameter precision. The variance of m for the triangular data distribution is a factor of two worse than for the uniform distribution case. Moreover, the raised cosine slope variance slope is more than 2.5 times higher than the uniform case. These results are depicted graphically in Fig. 1, which shows a line fit and confidence region determined from the covariance matrix of parameters for the uniform, triangular and the raised cosine distributions. The confidence region of the regression line is defined by the standard deviation of the y coordinate, σ_y. Since the off-diagonal terms of $\mathbf{C}_{\mathbf{x}}$ are zeroes, σ_y is determined from the variance propagation law as

$$\sigma_{\hat{y}} = \pm\sqrt{x^2\sigma_m^2 + \sigma_b^2} \tag{29}$$

Each example case comprises 5000 samples in x drawn from the respective distributions. The x coordinates are error free. Data for the raised cosine case were drawn from a truncated Gaussian function since the capability to do so was readily available in MATLAB, which was not true for the raised cosine. As the results show, the model fits the data distribution sufficiently well for the sake of the demonstration. The y coordinates contain additive Gaussian random errors with standard deviation of 0.5. The confidence regions are shown with exaggerated scale to ensure the differences among the three cases are clearly visible.

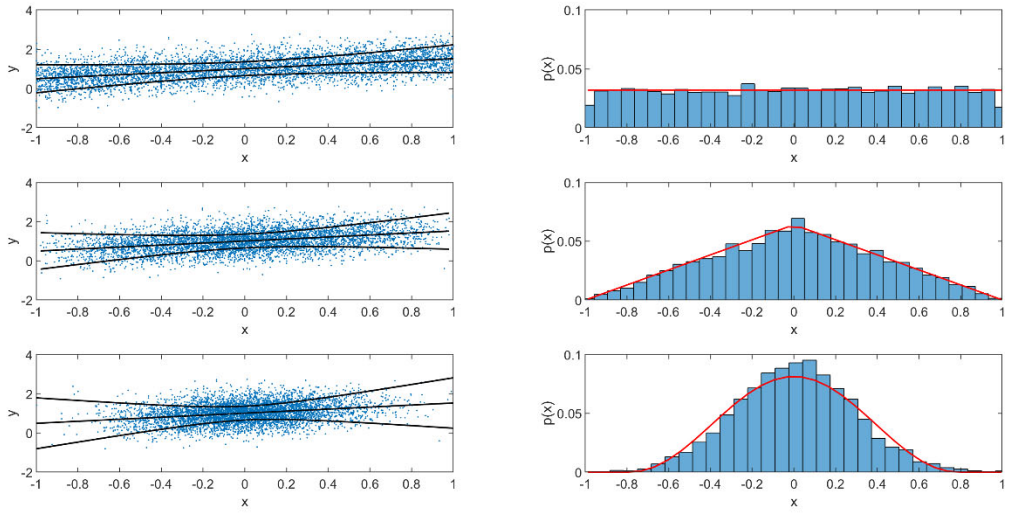

Fig. 1. Simulated data linear regression examples. Left: line fits and scaled confidence regions. Right: empirical histograms and overlain distribution functions. Top: uniform observation distribution. Middle: triangular observation distribution. Bottom: truncated Gaussian distribution.

3.5 Numerical vs. model predictions

A simulation was conducted to compare the numerical and model standard deviations of the slope parameter as a function of sample size. The simulations were performed for the symmetric uniform and triangular distributions with a=1 in both cases. The sample size of x observations was varied from 10^2 to 10^6. The results (Fig. 2) show that the numerical estimates for slope standard deviation, obtained from inversion of the normal equations matrix (Equation 11), closely match the model predictions computed from Equations 20 and 22. In both cases, the differences between the estimates are small, less than 6% of the standard deviation for small sample sizes and asymptotically become zero.

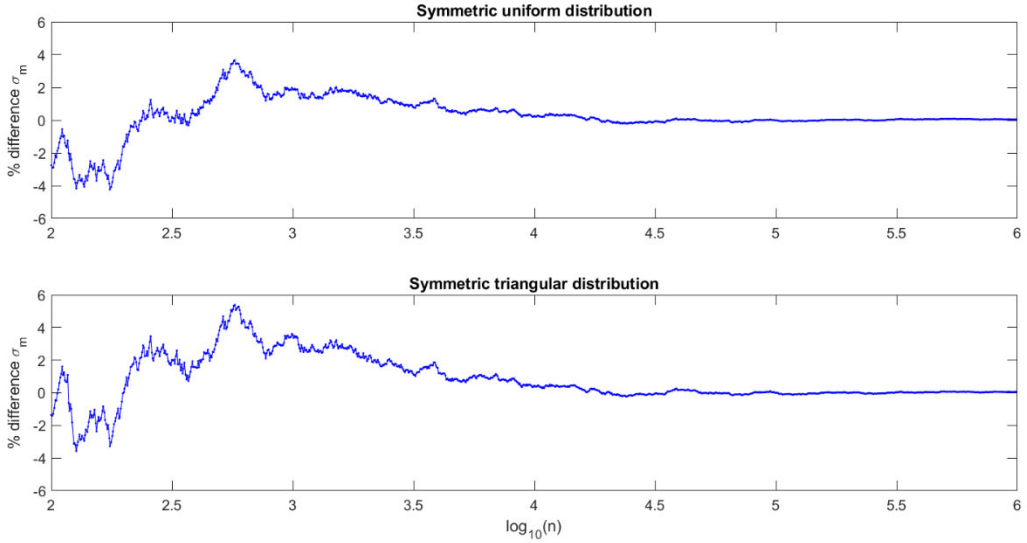

Fig. 2. Per cent difference between model and numerical slope standard deviations, σ_m, as a function of sample size, n, for two x-observation distributions.

4. Linear Regression—Asymmetric Data Distribution

Often in reality, the sampling distribution of the x-observations is asymmetric. To develop more general expressions for this case, the integration limits $[a_1, a_2]$ are unequal ($a_1 < a_2$) and the distribution has been translated such that the mean is at x=0

$$\mathbf{N} = \frac{1}{\sigma^2}\begin{pmatrix} \int_{a_1}^{a_2} x^2 p(x)\,dx & 0 \\ 0 & \int_{a_1}^{a_2} p(x)\,dx \end{pmatrix} \tag{30}$$

and

$$\mathbf{U} = \frac{1}{\sigma^2}\begin{pmatrix} \int_{a_1}^{a_2} x \cdot y \cdot p(x)\,dx \\ \int_{a_1}^{a_2} y \cdot p(x)\,dx \end{pmatrix} \tag{31}$$

As shown in Appendix A, the asymmetry of p(x) does not affect the unbiasedness of the line parameters as long as it is centred at the mean.

It is instructive to analyze a specific case to demonstrate how the slope parameter precision is influenced by the x-coordinate distribution asymmetry. The example chosen is the general form of the triangular distribution, which is a function of three parameters: the lower limit, a; the upper limit b; and the peak location, c (Evans et al. 2000).

$$p(x) = \begin{cases} \dfrac{2(x-a)}{(b-a)(c-a)} & a \le x \le c \\ \dfrac{2(b-x)}{(b-a)(b-c)} & c \le x \le b \\ 0 & \text{otherwise} \end{cases} \tag{32}$$

where $a < b$ and $a \le c \le b$. The covariance matrix of parameters in this case is given by

$$\mathbf{C}_x = \begin{pmatrix} \dfrac{18\sigma^2}{a^2+b^2+c^2-ab-ac-bc} & 0 \\ 0 & \sigma^2 \end{pmatrix} \tag{33}$$

Slope precision is strongly affected by the symmetry of the observation distribution. It is poorest for the symmetric case, a=-b and c=(a+b)/2, which is reflected by Equation 26. It is greatest where the peak location, c, coincides with one of the limits where the variance of the x observations is highest. Examples are shown for simulated data in Fig. 3.

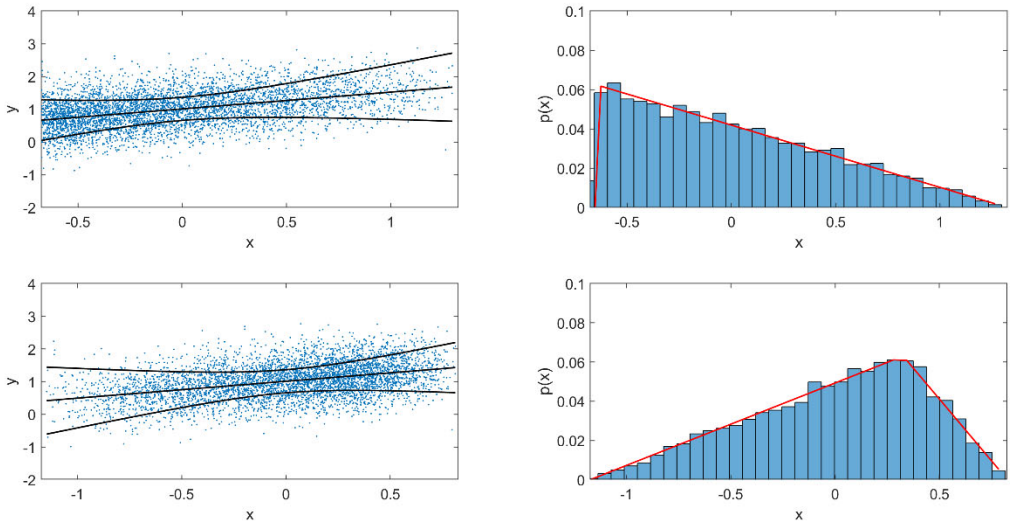

Fig. 3. Simulated linear regression examples with x-observations drawn from asymmetric triangular distributions. Note the data are centred at the mean x coordinate in each case.

5. Multidimensional Linear Regression

The methodology is extended to regression as a function of two independent variables where the 2D observation distribution function, p(x,y), is assumed to be separable:

$$p(x,y) = p(x)p(y) \tag{34}$$

Consider the following model of a plane, which is a function of two slope parameters, a and b, and an offset parameter, c

$$z + e = f(x,y) = ax + by + c \tag{35}$$

The independent variables x and y are assumed to be error free and the z coordinate observations are affected by random error. As with the line fitting case, uniform sampling is initially assumed. The number of samples, sampling interval and observation range in x and y are (n_x, n_y), (Δx, Δy) and (x_0, y_0), respectively. They are related as follows:

$$\Delta x = \frac{2x_0}{n_x} \tag{36}$$

$$\Delta y = \frac{2y_0}{n_y} \tag{37}$$

The following system of equations can be written for the n_y y-coordinates and x-coordinate i, where $i \in n_x$

$$\mathbf{y}_i + \mathbf{e}_i = \mathbf{A}_i \mathbf{x}$$

$$\begin{pmatrix} z_{i1} \\ z_{i2} \\ \vdots \\ z_{in_y} \end{pmatrix} + \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{in_y} \end{pmatrix} = \begin{pmatrix} x_i & y_1 & 1 \\ x_i & y_2 & 1 \\ \vdots & \vdots & \vdots \\ x_i & y_{n_y} & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} \tag{38}$$

The normal equations comprising the contributions from all observations can be formed using the summation of normals method (Mikhail 1976)

$$\mathbf{N} = \sum_{i=1}^{n_x} \mathbf{A}_i^T \mathbf{P}_i \mathbf{A}_i \tag{39}$$

$$\mathbf{U} = \sum_{i=1}^{n_x} \mathbf{A}_i^T \mathbf{P}_i \mathbf{y}_i \tag{40}$$

Similar to the line fit model development, the weight matrix is assumed to be a scalar matrix (Equation 10). The analytical form of the normal equations is given by double summations over the samples in x and y

$$\mathbf{N} = \frac{1}{\sigma^2} \begin{pmatrix} \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} x_i^2 & \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} x_i y_i & \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} x_i \\ & \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} y_i^2 & \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} y_i \\ \text{sym} & & \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} 1 \end{pmatrix} \tag{41}$$

$$\mathbf{U} = \frac{1}{\sigma^2} \begin{pmatrix} \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} x_i z_{ij} \\ \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} y_i z_{ij} \\ \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} z_{ij} \end{pmatrix} \tag{42}$$

Next, the weight matrix is defined in a manner similar to Equation 17

$$\mathbf{P} = \frac{1}{\sigma^2}\begin{pmatrix} p(x_1)p(y_1) & 0 & \cdots & 0 & \cdots & 0 \\ 0 & p(x_1)p(y_2) & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 & \cdots & 0 \\ 0 & 0 & 0 & p(x_2)p(y_1) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & p(x_{n_x})p(y_{n_y}) \end{pmatrix} \tag{43}$$

The normal equations are then multiplied by $\Delta x \Delta y$ and the limit is taken as the numbers of observations, n_x and n_y, tend to infinity

$$\begin{aligned} \mathbf{N} &= \lim_{\substack{n_x \to \infty \\ n_y \to \infty}} \frac{\Delta x \Delta y}{\sigma^2} \begin{pmatrix} \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} p(x_i,y_j)x_i^2 & \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} p(x_i,y_j)x_i y_i & \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} p(x_i,y_j)x_i \\ & \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} p(x_i,y_j)y_i^2 & \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} p(x_i,y_j)y_i \\ \text{sym} & & \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} p(x_i,y_j) \end{pmatrix} \\ &= \frac{1}{\sigma^2} \begin{pmatrix} \int_{-x_0}^{x_0}\int_{-y_0}^{y_0} x^2 p(x,y)\,dxdy & \int_{-x_0}^{x_0}\int_{-y_0}^{y_0} x\cdot y\cdot p(x,y)\,dxdy & \int_{-x_0}^{x_0}\int_{-y_0}^{y_0} x\cdot p(x,y)\,dxdy \\ & \int_{-x_0}^{x_0}\int_{-y_0}^{y_0} y^2 p(x,y)\,dxdy & \int_{-x_0}^{x_0}\int_{-y_0}^{y_0} y\cdot p(x,y)\,dxdy \\ \text{sym} & & \int_{-x_0}^{x_0}\int_{-y_0}^{y_0} p(x,y)\,dxdy \end{pmatrix} \end{aligned} \tag{44}$$

$$\mathbf{U} = \lim_{\substack{n_x \to \infty \\ n_y \to \infty}} \frac{\Delta x \Delta y}{\sigma^2} \begin{pmatrix} \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} p(x_i,y_j)x_i z_{ij} \\ \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} p(x_i,y_j)y_i z_{ij} \\ \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} p(x_i,y_j)z_{ij} \end{pmatrix} = \frac{1}{\sigma^2} \begin{pmatrix} \int_{-x_0}^{x_0}\int_{-y_0}^{y_0} x\cdot z\cdot p(x,y)\,dxdy \\ \int_{-x_0}^{x_0}\int_{-y_0}^{y_0} y\cdot z\cdot p(x,y)\,dxdy \\ \int_{-x_0}^{x_0}\int_{-y_0}^{y_0} z\cdot p(x,y)\,dxdy \end{pmatrix} \tag{45}$$

where p(x,y) is the function jointly representing the distributions of observations in x and y according to the sampling process and has the following properties

$$\int_{-x_0}^{x_0}\int_{-y_0}^{y_0} p(x,y)\,dxdy = 1 \tag{46}$$

$$p(x,y) > 0 \tag{47}$$

If the x and y coordinates are reduced by their respective means, then elements (1,3) and (2,3) of **N** (Equation 44) are equal to zero. Moreover, if $p(x)$ and/or $p(y)$ are symmetric, then element (1,2) is also equal to zero. Under these conditions, **N** becomes

$$\mathbf{N} = \frac{1}{\sigma^2}\begin{pmatrix} \int_{-x_0}^{x_0}\int_{-y_0}^{y_0} x^2 p(x,y)\,dxdy & 0 & 0 \\ 0 & \int_{-x_0}^{x_0}\int_{-y_0}^{y_0} y^2 p(x,y)\,dxdy & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{48}$$

The covariance matrix of parameters is obtained by inversion of Equation 48 and has the following form

$$\mathbf{C_x} = \begin{pmatrix} \sigma_a^2 & 0 & 0 \\ 0 & \sigma_b^2 & 0 \\ 0 & 0 & \sigma_c^2 \end{pmatrix} \tag{49}$$

Since this matrix is diagonal, it can be readily deduced that the slope parameters of the plane, a and b, are dependent on the dispersion of the observations in the x and y directions, respectively.

6. Real Data Examples

6.1 Example 1: line fitting of laser rangefinder data

The indoor built environment can be mapped with integrated mobile systems comprising inertial navigation units and laser scanners such as the Velodyne VLP-16 sensor (Chan et al. 2019). The operator moves through the environment to build up a point cloud of 3D point samples from the integrated system. The Velodyne collects laser range measurements, ρ, throughout at 360° horizontal field-of-view by means of rotation about its vertical axis. The resolution of the horizontal direction measurements, θ, varies from 0.1° to 0.4°, depending on the user-defined rotation rate. The collection of data from multiple sensor revolutions at a static location results in the collection of datasets with greater density. The 16 lasers of the VLP-16 are nominally spaced in uniform increments to provide a 30° vertical field-of-view.

The desired end product of the mapping exercise is a model of the environment reconstructed from the point cloud. This may be a full 3D model or a 2D floor plan. To construct the latter, groups of points belonging to linear structures are identified and extracted from a 2D generalization of the

point cloud, either extracted horizontal profiles or the projection of points onto a horizontal plane. Line fitting is performed to obtain the model representation of each extracted segment.

In this example, line fitting is performed for Velodyne VLP-16 data of a 12 m long wall segment captured from a normal distance, d, of 6.5 m. The geometry is depicted in Fig. 4. The approximate angular scanning limits, θ_{min} and θ_{max}, -1.9° and 61.6°, respectively. The dataset comprises n=2526 samples from four adjacent laser rangefinders. The standard deviation of the observation errors was $\sigma=\pm 0.02$ m.

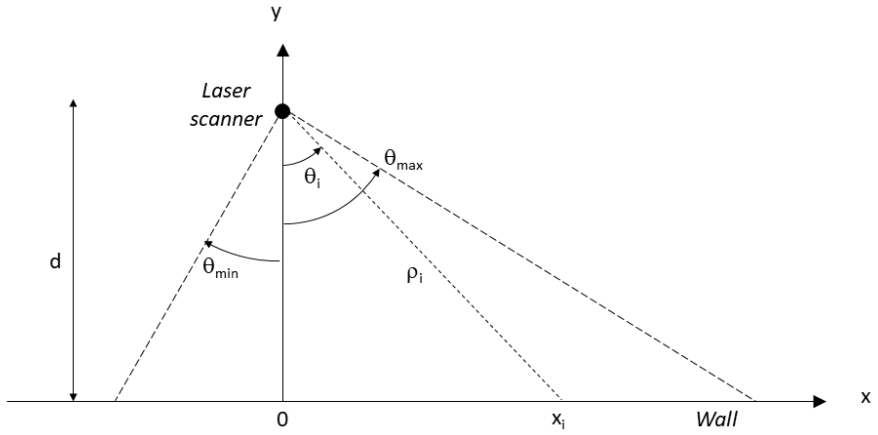

Fig. 4. 2D scanning geometry for Example 1.

If the distribution of horizontal direction observations is uniform on $[\theta_{min}, \theta_{max}]$, then the distribution of the x observations can be derived from the geometric relationship between θ and x

$$\theta = \arctan\left(\frac{x}{d}\right) \tag{50}$$

It can be shown that the distribution of x coordinates follows the functional form of a truncated Cauchy density function

$$p(x) = \begin{cases} \dfrac{1}{(\theta_{max} - \theta_{min})} \dfrac{d}{(d^2 + x^2)} & a_1 < x < a_2 \\ 0 & \text{otherwise} \end{cases} \tag{51}$$

where

$$a_1 = d \tan \theta_{min} \tag{52}$$

and

$$a_2 = d \tan \theta_{max} \tag{53}$$

As with previous examples, it is assumed that the x coordinates are reduced to the mean and are error free. The latter is a very common assumption in such problems.

Staneski (1990) derives the moments of the truncated Cauchy distribution, which can be computed unlike those of the non-truncated Cauchy distribution (Evans et al. 2000). From Staneski's equations, the covariance matrix of line parameters is given by

$$\mathbf{C_x} = \begin{pmatrix} \dfrac{\sigma^2}{(t_1 - t_2)} & 0 \\ 0 & \sigma^2 \end{pmatrix} \tag{54}$$

where

$$t_1 = d\left(\frac{(b-a)}{(\theta_{max} - \theta_{min})} - d \right) \tag{55}$$

and

$$t_2 = \left\{ \frac{d}{2(\theta_{max} - \theta_{min})} \left(\ln\left(1 + \left(\frac{b}{d}\right)^2\right) - \ln\left(1 + \left(\frac{a}{d}\right)^2\right) \right) \right\}^2 \tag{56}$$

The results of the line fitting as well as the empirical and theoretical observation distributions are shown in Fig. 5. The estimated standard deviation of the slope obtained from the inverse of the normal equations matrix (Equation 11) is ±0.00617 (unitless). The corresponding truncated Cauchy model prediction for the standard deviation, computed from Equation 54, is ±0.00616. These figures agree quite well; the difference between them is only 0.16%.

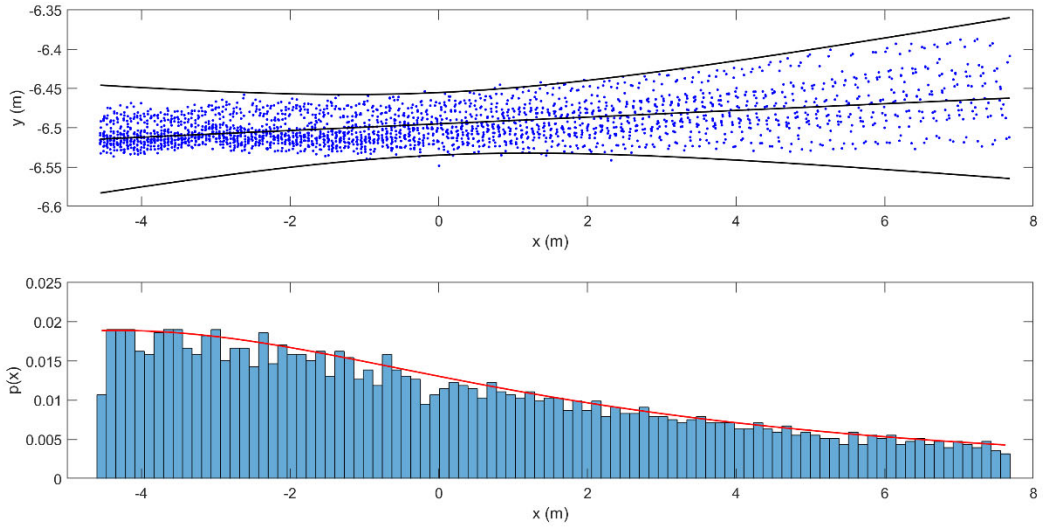

Fig. 5. Line fitting from laser rangefinder results for Example 1. Top: line fit and confidence region. Bottom: empirical histogram and model distribution.

6.2 Example 2: terrain modelling

The second example is the modelling of a small patch of terrain comprising 2700 (n_x=90, n_y=30) samples (Fig. 6). The area in question is half of a football field comprising an 89 m × 29 m (easting × northing) parcel. The terrain slopes gently in both cardinal directions, approximately 1 m in the easting direction and 0.3 m in northing. The height data were derived from airborne LiDAR and have a uniform sample spacing in each dimension of 1 m. The standard deviation of the height (z) errors was assumed to be ±0.03 m. Modelling this dataset with a plane is a direct application of Equation 48 with a two-dimensional uniform distribution function having limits of x_0=44.5 m and y_0=14.5 m. Using the separability property of Equation 34, the covariance matrix of plane parameters is given by

$$\mathbf{C}_{\mathbf{x}} = \begin{pmatrix} \frac{3\sigma^2}{x_0^2} & 0 & 0 \\ 0 & \frac{3\sigma^2}{y_0^2} & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix} \tag{57}$$

The respective numerical and model standard deviations for the (unitless) slope parameters (a and b), σ_a and σ_b, are (±0.001168, ±0.00358) and (±0.001155, ±0.003466), which agree quite well. The differences between estimates are only 1.1% and 3.4% and can be attributed to finite sample size as suggested by the results in Figure 2.

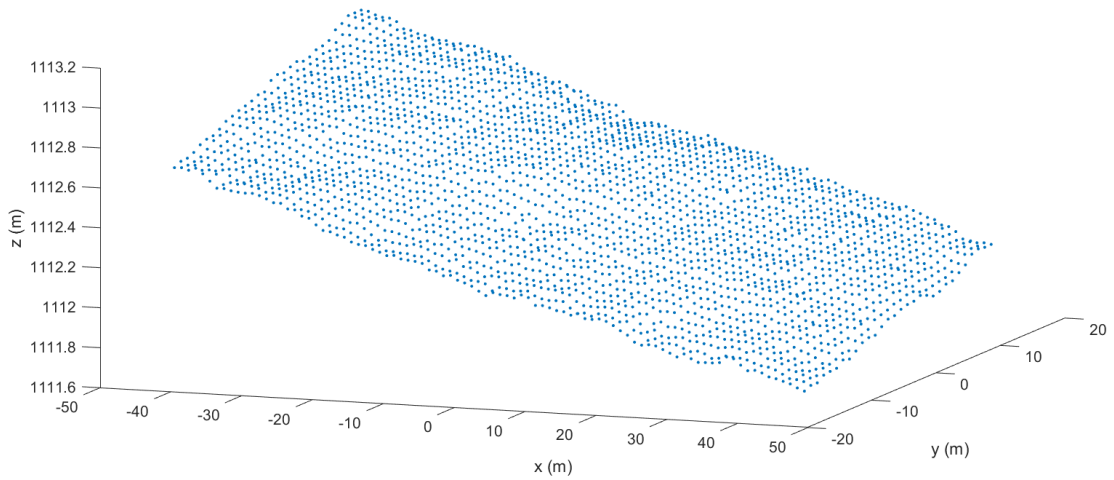

Fig. 6. Digital elevation model data for Example 2.

6.3 Example 3: indoor plane fit

The final example is a plane fit to data also captured with a Velodyne VLP-16 sensor. Pictured in Fig. 7, the dataset comprises 8857 point measurements collected with 11 of the instrument's 16

laser rangefinders. The data cover a 2.7 m × 2.7 m patch of an indoor wall located 10.0 m from the scanner. The aforementioned disparity between the horizontal (x) and vertical (y) sampling increments is clearly evident. Some gaps in the data due to missing points are also evident. The setup was such that instrument's z-axis was aligned with the local gravity vector, but the axes have been permutated so as to follow the parameterization of Equation 35. Note that doing so does not change the geometry of the plane fitting problem.

To apply the proposed modelling scheme, an additional angular variable, α, is introduced and defined similarly to θ

$$\alpha = \arctan\left(\frac{y}{d}\right) \tag{58}$$

The function, p(x,y), is once again assumed to be separable. Both p(x) and p(y) are of the same form as Equation 51. The standard deviation of the observation (z) errors was set to ±0.02 m.

The estimated slope parameter standard deviations also match quite well in this example. The numerical estimates of σ_a and σ_b are ±0.026623 and ±0.025341, respectively. The corresponding model predictions are ±0.025845 and ±0.025573. The differences between estimates are 3.0% and 0.9%.

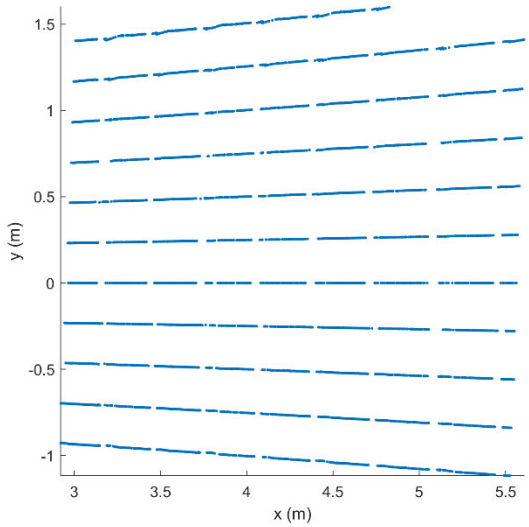

Fig. 7. Indoor Velodyne plane fit dataset for Example 3.

6.4 The influence of noise

The proposed model for predicting regression parameter precision presumes an idealized sampling structure. The x coordinates (and y in the 2D case) are assumed to be error free. Noise in the independent variable(s) can cause the actual data to deviate from the idealized sampling structure given by p(x). Therefore, a simulation study was performed to quantify the effect of noise in the independent variable on estimates of slope precision.

Conditions similar to those of the first two examples were simulated. Linear regression with uniform sampling was first considered. One dimension of the terrain modelling example was used for this test: 90 samples with 1 m uniform sample spacing (Δx) were generated. Random Gaussian noise was added to the x coordinates. The standard deviation of the added noise, σ_x, was varied in ten increments as a proportion of the sample spacing from 10% to 100%, i.e. $0.1\Delta x$ to Δx. One hundred trials were performed for each value of σ_x. The covariance matrix of the parameters was computed by inverting **N** of Equation 11. Fig 8 shows the results in terms of the difference between the numerical estimates of σ_m and the predicted model value. Under these experimental conditions, the differences are less than 2.5%, even when the noise standard deviation is equal to the sampling interval, i.e. $\Delta x=\sigma_x$. It should be noted that the sample size is small in this case in order to match real-data experiment conditions. The differences are expected to be much smaller if the sample size were larger, as per Fig. 2

A similar test was performed for line fitting to replicate the laser rangefinding conditions of Example 2: 2500 uniformly-spaced angle (θ) observations were simulated on the interval $-2.5° \leq \theta \leq 60°$ such that the nominal sample spacing, $\Delta\theta$, was 0.025°. For this test, a greater proportion of random error was added to the angle observations to simulate very noisy data: the increments of the standard deviation of added noise, σ_θ, started at $10\Delta\theta$ and were increased up to $100\Delta\theta$. The standard deviation of the corresponding errors in x was 0.52 m for the final testing increment. One hundred trials were performed for each value of σ_θ and the x coordinates were computed from the noisy θ observations. As can be seen in Fig. 8, the differences in variance estimates are less than 3.5% for the chosen conditions despite the proportionally higher introduced noise levels. The results suggest that accurate slope precision prediction is obtained from the proposed model even with large amounts of additive noise in the independent variables.

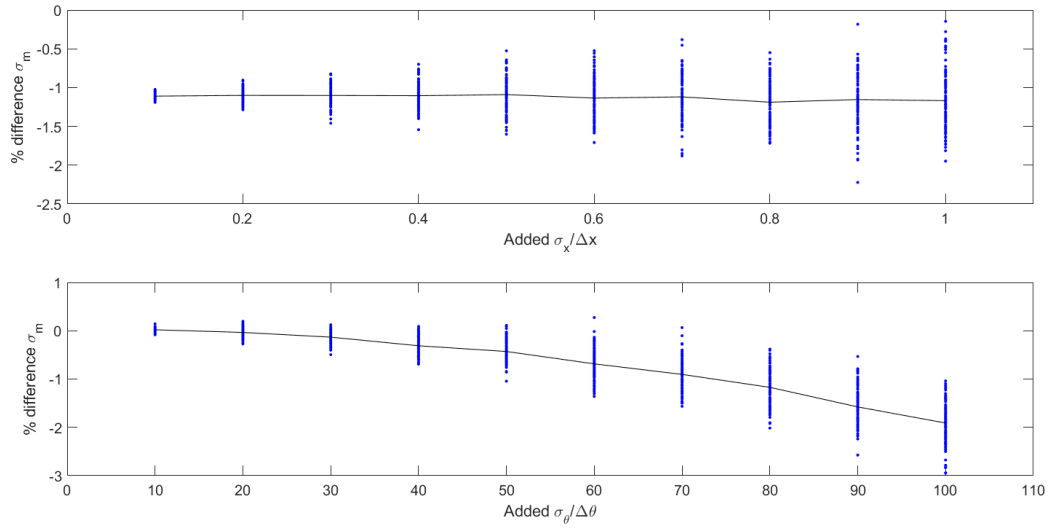

Fig. 8. Per cent difference between model and numerical slope standard deviations, σ_m, due to the presence of additive random noise in the independent variable x. Top: results of the line fit simulation of the terrain modelling example. Bottom: results of the line fit simulation of the laser rangefinder example. The trend lines pass through the mean values.

7. Conclusions

The role of the observation distribution on linear regression estimates has been investigated in this work. A methodology has been developed to model the distribution of the observations and incorporate it into the least-squares solution such that parameter quality can be analytically predicted for a given hypothesized distribution. Experimental results demonstrate the value of the proposed methodology as both a planning and analysis tool. Its greatest advantage is elimination of the numerical computation of the normal equations matrix required to obtain model parameter precision, thus avoiding the need to perform intensive simulations to achieve the same results. Thus, geodetic network quality can be obtained without having to generate synthetic observations of the environment. Only the instrument locations and the sampling pattern are required.

Our results show that the empirical and model precisions are consistent, with differences only up to 6% and 3.4% for the simulated and real datasets, respectively. Simulation demonstrated that these differences are due to the finite sample size. Moreover, simulation demonstrated relative insensitivity to noise in the independent regression variables that causes deviations from the hypothesized data distribution function. Our results also indicate that the slope precision can be strongly affected by the shape and the symmetry of the observation distribution but can be modelled in terms of the variables deduced from the numerical limits and distribution geometry. The proposed methodology is also shown to be extendable to multivariable linear regression, which therefore supports many real world applications involving more than two variables.

Thanks to advanced develop ment of sensors and big data strategies, much more data are becoming available that will require linear regression. The proposed methodology shows that instant and accurate precision estimation for large datasets becomes possible. Not only can the method benefit the geodetic applications as shown in the paper, but it also could serve as an efficient tool for data mining and forecasting applications.

Many avenues exist for future research to develop this methodology further. Some of these stem from the assumptions made herein. The assumption that only the dependent regression variable is affected by random errors allows use of the Gauss-Markov model. The influence of noise in the x coordinates on regression precision estimates has been studied by simulation. However, further investigation into the more general case of all coordinates being affected by random error and, therefore, the use of the Gauss-Helmert model or the errors-in-variables model is warranted. We also assumed the variance of random errors to be common to all observations, which is a widespread practice but is not always realistic. A logical advancement would be to consider the more general case of homoscedasticity. Furthermore, only linear functional models were investigated. The extension to non-linear functions encountered in other applications of geodesy, photogrammetry and laser scanning should also be studied.

Finally, a key assumption in the two-dimensional case is the ability to model the sampling geometry in the x and y coordinates as separate functions. Whilst theoretically sound for uniform sampling in Cartesian coordinates, deeper investigation is needed into the validity of this assumption for the spherical geometry of laser scanners. The close agreement between the numerical and theoretical standard deviations of Example 3 may be partly due to the relatively narrow extents of the data compared to the standoff distance of the sensor from the plane. Any y-coordinate-dependence of the x coordinate sampling function, and vice-versa, may be significant over a larger field-of-view.

Acknowledgements
This work is supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2018-03775). Data for example 2 have been made available by the City of Calgary.

The authors wish to thank the anonymous reviewers who provided very constructive comments to help improve our paper.

Appendix A
In this appendix, the unbiasedness of the estimates of the slope parameter, m, and the y-axis intercept parameter, b, estimated by the linear least-squares model incorporating the x-observation distribution function, p(x), is proven. The limits of integration, $[a_1, a_2]$ are unequal ($a_1 < a_2$) to keep the proof general. The distribution function is assumed to be centred at x=0.

The expectation of the estimated slope stems from the solution to the least-squares normal equations (Equation 9) and substituting the specific forms given by Equations 30 and 31

$$\begin{aligned}\mathbf{x} &= \mathbf{N}^{-1}\mathbf{U} \\ &= \begin{pmatrix} \int_{a_1}^{a_2} x^2 p(x)\,dx & 0 \\ 0 & \int_{a_1}^{a_2} p(x)\,dx \end{pmatrix}^{-1} \begin{pmatrix} \int_{a_1}^{a_2} x \cdot y \cdot p(x)\,dx \\ \int_{a_1}^{a_2} y \cdot p(x)\,dx \end{pmatrix}\end{aligned} \tag{A1}$$

and results in

$$E\{\hat{m}\} = E\left\{ \frac{\int_{a_1}^{a_2} x \cdot y \cdot p(x)\,dx}{\int_{a_1}^{a_2} x^2 p(x)\,dx} \right\} \tag{A2}$$

Since x has been assumed to be error free, this expression reduces to

$$E\{\hat{m}\} = \frac{E\left\{ \int_{a_1}^{a_2} x \cdot y \cdot p(x)\,dx \right\}}{\int_{a_1}^{a_2} x^2 p(x)\,dx} \tag{A3}$$

The numerator is analyzed by substituting the model of Equation 1

$$
\begin{aligned}
E\left\{\int_{a_1}^{a_2} x\cdot y\cdot p(x)\,dx\right\} &= E\left\{\int_{a_1}^{a_2} x\left(m\cdot x+b-e\right)p(x)\,dx\right\} \\
&= E\left\{\int_{a_1}^{a_2} m\cdot x^2 p(x)\,dx\right\}+E\left\{\int_{a_1}^{a_2} b\cdot x\cdot p(x)\,dx\right\}-E\left\{\int_{a_1}^{a_2} x\cdot e\cdot p(x)\,dx\right\} \\
&= mE\left\{\int_{a_1}^{a_2} x^2 p(x)\,dx\right\}+bE\left\{\int_{a_1}^{a_2} x\cdot p(x)\,dx\right\}-E\left\{e\int_{a_1}^{a_2} x\cdot p(x)\,dx\right\}
\end{aligned}
\tag{A4}
$$

Under the stated assumptions, the numerator reduces to

$$
\begin{aligned}
E\left\{\int_{a_1}^{a_2} x\cdot y\cdot p(x)\,dx\right\} &= m\int_{a_1}^{a_2} x^2 p(x)\,dx+b(0)-E\{e(0)\} \\
&= m\int_{a_1}^{a_2} x^2 p(x)\,dx
\end{aligned}
\tag{A5}
$$

Division of Equation A5 by the denominator of Equation A3 results in the following

$$
E\{\hat{m}\}=\frac{m\int_{a_1}^{a_2} x^2 p(x)\,dx}{\int_{a_1}^{a_2} x^2 p(x)\,dx}=m \tag{A6}
$$

Therefore, the estimated slope parameter is unbiased.

The expected value of the intercept parameter is given by

$$
E\{\hat{b}\}=E\left\{\frac{\int_{a_1}^{a_2} y\cdot p(x)\,dx}{\int_{a_1}^{a_2} p(x)\,dx}\right\}=\frac{E\left\{\int_{a_1}^{a_2} y\cdot p(x)\,dx\right\}}{\int_{a_1}^{a_2} p(x)\,dx} \tag{A7}
$$

Since the denominator is unity by definition (Equation 20), it is sufficient to analyze only the numerator

$$
\begin{aligned}
E\left\{\int_{a_1}^{a_2} y \cdot p(x)\,dx\right\} &= E\left\{\int_{a_1}^{a_2} (m \cdot x + b - e)\,p(x)\,dx\right\} \\
&= m\int_{a_1}^{a_2} x \cdot p(x)\,dx + b\int_{a_1}^{a_2} p(x)\,dx - E\left\{e\int_{a_1}^{a_2} p(x)\,dx\right\} \\
&= m(0) + b(1) - E\{e(1)\} = b
\end{aligned}
\tag{A8}
$$

Therefore, the estimated y-axis intercept parameter is also unbiased.

References

Abraham, B, Ledoliter, J (2006). Introduction to Regression Modelling. Thompson Brooks/Cole. Belmont, CA.

Berné, J, Baselga (2004). First-order design of geodetic networks using the simulated annealing method. Journal of Geodesy 78: 47-54.

Chan, TO, Lichti, D, Roesler, G, Cosandier, D, Al-Durgham, K (2019) Range scale-factor calibration of the Velodyne VLP-16 LiDAR system for position tracking applications. In Proceedings of the 11th International Conference on Mobile Mapping. Shenzhen, China, 6-8 May. 70-77. 350-355.

Evans, M, Hastings, N, Peacock, B (2010). Statistical Distributions, 3rd Ed. John Wiley & Sons, New York.

Featherstone, WE, Lichti, DD (2009) Fitting gravimetric geoid models to vertical deflections. Journal of Geodesy 83: 583–589.

Grafarend, E (1974). Optimization of geodetic networks. Bolletino di Geodesia a Science Affini 33 (4): 351-406.

Grafarend, WE, Sansò, F (Eds.) (1985). Optimization and design of geodetic networks. Springer, Berlin.

Haixiang, G, Yijing, L, Shang, J, Mingyun, G, Yuanyue, H, Bing, G (2017) Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications, 73: 220-239.

Hekimoglu, S, Berbe, M (2003). Effectiveness of robust methods in heterogeneous linear models. Journal of Geodesy 76: 706-713.

Holst, C, Eling, C, Kuhlmann, H (2013). Automatic optimization of height network configurations for detection of surface deformations. Journal of Applied Geodesy 7(2): 103-113

Holst, C, Artz, T, Kuhlmann, H (2014). Biased and unbiased estimates based on laser scans of surfaces with unknown deformations. Journal of Applied Geodesy 8(3): 169-184.

Jazireeyan, I, Ardalan, AA (2017) Absolute calibration of satellite altimetry using linear regression and harmonic analysis. Geodesy and Cartography 43(3): 83-91.

Jia, F, DD Lichti (2019) A model-based design system for terrestrial laser scanning networks in complex sites. Remote Sensing 11: 1749.

Leevy, JL, Khoshgoftaar, TM, Bauder, RA, Seliya, N (2018) A survey on addressing high-class imbalance in big data. Journal of Big Data, 5: 42.

Lichti, DD, O'Keefe, K, Jamtsho, S (2010) Propagation of an unmodeled additive constant in range sensor observations. ASCE Journal of Surveying Engineering 136 (3): 111-119.

Mahboub, V (2012) On weighted total least-squares for geodetic transformations. Journal of Geodesy 86: 359-367.

Mikhail, E.M (1979) Observations and Least Squares. IEP, New York.

Mozaffar, M, Varshosaz, M. (2016) Optimal placement of a terrestrial laser scanner with an emphasis on reducing occlusions. Photogrammetric Record 31(156): 374–393

Pelto, BM, Menounos, B, Marshall, SJ (2019) Multi-year evaluation of airborne geodetic surveys to estimate seasonal mass balance, Columbia and Rocky Mountains, Canada. The Cryosphere 13: 1709-1727.

Rawlings, JO, Pantula, SG, Dickey, DA (1989). Applied Regression Analysis: a Research Tool. 2nd Ed. Springer-Verlag New York.

Rüeger, JM (1990). Electronic distance measurement: An introduction, 3rd Ed., Springer, Heidelberg, Germany.

Ruffhead, A (2018) Introduction to multiple regression equations in datum transformations and their reversibility. Survey Review 50(358): 82-90.

Schaffrin, B, Wieser, A (2008) On weighted total least-squares adjustment for linear regression. Journal of Geodesy 82: 415-421.

Schaffrin, B, Felus, YA (2008) On the multivariate total least-squares approach to empirical coordinate transformations. Three algorithms. Journal of Geodesy 82: 373-383.

Searle. SR (1986) Linear Models for Unbalanced Data. Wiley, New York.

Soudarissanane, S. (2016) The Geometry of Terrestrial Laser Scanning: Identification of Errors, Modeling and Mitigation of Scanning Geometry. Dissertation, Delft University of Technology.

Staneski, PG (1990). The Truncated Cauchy Distribution: Estimation of Parameters and Application to Stock Returns. Dissertation, Old Dominion University.

Yang, Y (1999) Robust estimation of geodetic datum transformation. Journal of Geodesy 73: 268-274.