SCHULICH School of Engineering
UNIVERSITY OF CALGARY
DEPARTMENT OF GEOMATICS ENGINEERING

ENGO 361 - Least Squares Estimation

Winter 2018

Lab #1 - Multivariate statistics, mathematical models, and error propagation

Course instructor: Dr. Ivan Detchev

Teaching assistant: Ahmed Youssef

Date of lab: January 19, 2018

Date of submission: February 8, 2018

Student name: Erica Lemieux

Certificate of work:

"I, Erica Lemieux, certify that this is my own work which has been done expressly for this course, and where appropriate I have acknowledged the work of others. Further, I have read and understood the section in the university calendar on plagiarism/cheating/other academic misconduct and I am aware of the implications thereof."

Contents

Part I – Univariate statistical analysis

a. Field measurements and error computations

Table 1. Distance measurements by two observers (A and B), in metres

Meas No.	Observer		*Meas No.*	Observer		*Meas No.*	Observer	
	A	B		A	B		A	B
1	153	149.6	*18*	151.7	148.4	*35*	155.1	150.6
2	152.9	151.4	*19*	154.5	150.3	*36*	150.3	151.6
3	149	149	*20*	155	149	*37*	146.1	148.7
4	149.2	150.2	*21*	144.3	149.2	*38*	150.4	151
5	145.5	15.2*	*22*	149.8	151.1	*39*	152.1	149.4
6	147.3	150	*23*	152.7	151	*40*	149.6	149.5
7	147.1	149.4	*24*	151.2	151	*41*	157.9	150.7
8	147.7	151.3	*25*	151.6	150.2	*42*	146.9	149
9	153.3	150.5	*26*	151.7	148.2	*43*	142.7	150.4
10	149.2	150.7	*27*	142.5	147.9	*44*	152.9	150.5
11	147.7	149.7	*28*	149.4	151.1	*45*	147.3	151.8
12	152.2	150.2	*29*	149.8	150	*46*	148.7	149.8
13	150	150.4	*30*	151.5	151.3	*47*	147.5	148.7
14	155.6	148.7	*31*	148.1	148.9	*48*	147.8	151.2
15	149.7	152.5	*32*	150.2	149.8	*49*	146.8	151.3
16	143.9	151.7	*33*	149.4	150.6	*50*	151.6	150
17	144.5	149.6	*34*	154.4	150.6			

**Gross error (blunder)*

For the following error calculations, the gross error indicated above has been removed (A: *n = 50*, B: *n = 49*).

Sample mean ($\bar{x}$)

The best estimate of a data set, or mean, were calculated using the equation:

$$\bar{x} = \frac{1}{n} * \sum_{i=1}^{n} l_i \qquad (1)$$

where n is the number of observations and l_i is the i^{th} sample. Using this equation, the mean values for Datasets A and B are shown below:

Mean ($\bar{x}$, metres)	
A	B
149.79	150.16

Residuals (v)

The residuals of the data were calculated using the equation:

$$v_i = \bar{x} - l_i \qquad (2)$$

and the units are metres (Table 2).

Table 2. Measurement residuals of the two observer data sets, in metres

Meas No.	Residuals		*Meas No.*	Residuals		*Meas No.*	Residuals	
	A	B		A	B		A	B
1	-3.21	0.56	*18*	-1.91	-0.14	*35*	-5.31	-1.44
2	-3.11	-1.24	*19*	-4.71	1.16	*36*	-0.51	1.46
3	0.79	1.16	*20*	-5.21	0.96	*37*	3.69	-0.84
4	0.59	-0.04	*21*	5.49	-0.94	*38*	-0.61	0.76
5	4.29	0.16	*22*	-0.01	-0.84	*39*	-2.31	0.66
6	2.49	0.76	*23*	-2.91	-0.84	*40*	0.19	-0.54
7	2.69	-1.14	*24*	-1.41	-0.04	*41*	-8.11	1.16
8	2.09	-0.34	*25*	-1.81	1.96	*42*	2.89	-0.24
9	-3.51	-0.54	*26*	-1.91	2.26	*43*	7.09	-0.34
10	0.59	0.46	*27*	7.29	-0.94	*44*	-3.11	-1.64
11	2.09	-0.04	*28*	0.39	0.16	*45*	2.49	0.36
12	-2.41	-0.24	*29*	-0.01	-1.14	*46*	1.09	1.46
13	-0.21	1.46	*30*	-1.71	1.26	*47*	2.29	-1.04
14	-5.81	-2.34	*31*	1.69	0.36	*48*	1.99	-1.14
15	0.09	-1.54	*32*	-0.41	-0.44	*49*	2.99	0.16
16	5.89	0.56	*33*	0.39	-0.44	*50*	-1.81	
17	5.29	1.76	*34*	-4.61	-0.44			

Standard deviation of a single observation (σ):

$$\sigma = \sqrt{\frac{1}{n-1} * \sum_{i=1}^{n} {v_i}^2} \qquad (3)$$

where n is the number of observations and v_i is the i^{th} residual. The *n* values for Dataset A and B are 50 and 49, respectively.

Standard deviation of the mean ($\sigma_{\bar{x}}$):

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad (4)$$

The standard deviation, standard deviation of the mean, average and probable error are summarized in Table 3.

Table 3. Summary of the standard deviation and error computations of the two observer data sets

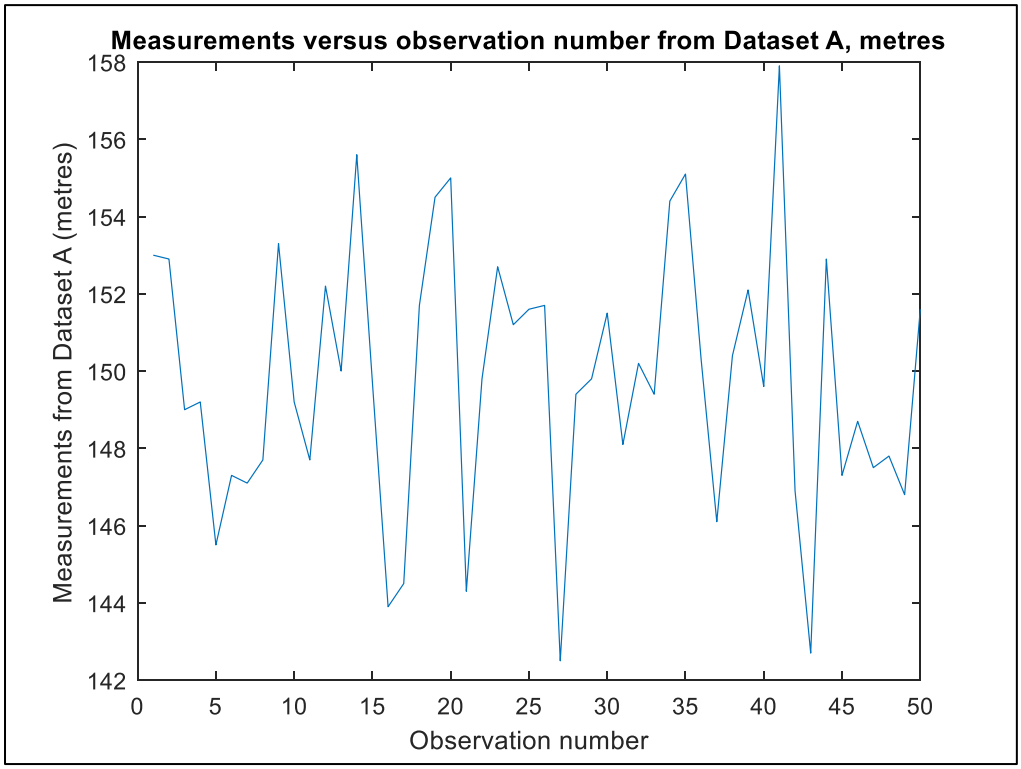
Computation parameter	Observer	
	A	B
Standard deviation of a single observation (σ, metres)	3.41	1.04
Standard deviation of the mean ($\sigma_{\bar{x}}$, metres)	0.48	0.15
Average Error (a_e, metres)	2.72	0.87
Probable Error (p_e, metres)	2.30	0.84

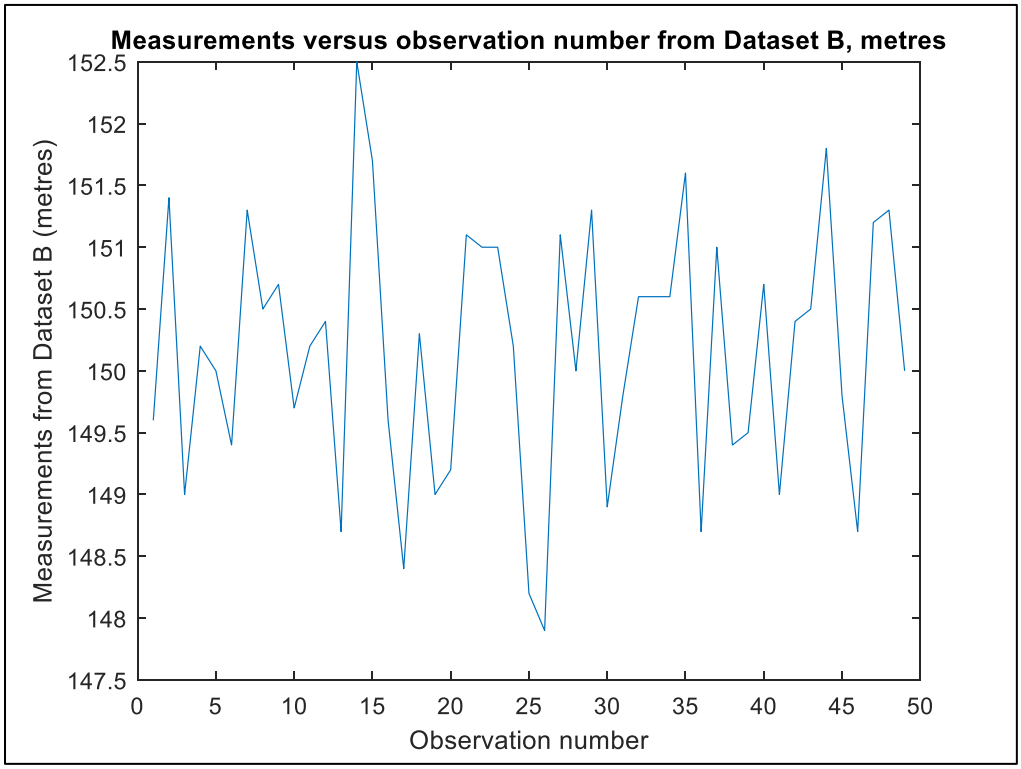
b. Best estimates

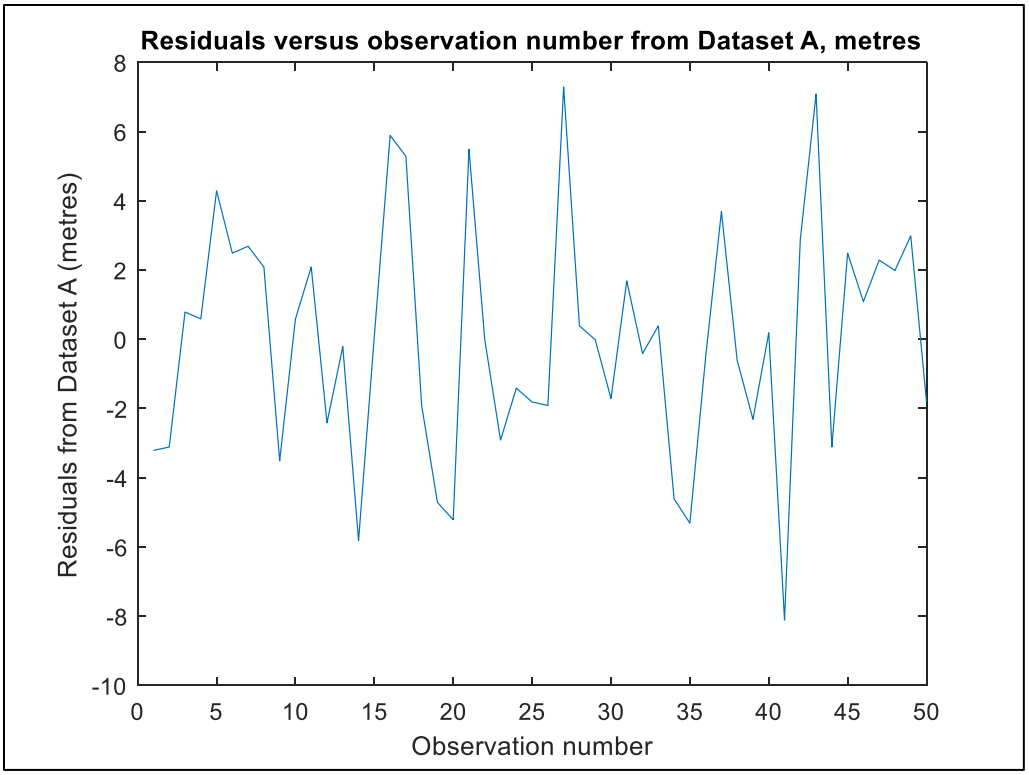
For a 95% confidence interval, a tolerance of 1.96 standard deviations (σ) from the true mean is used:

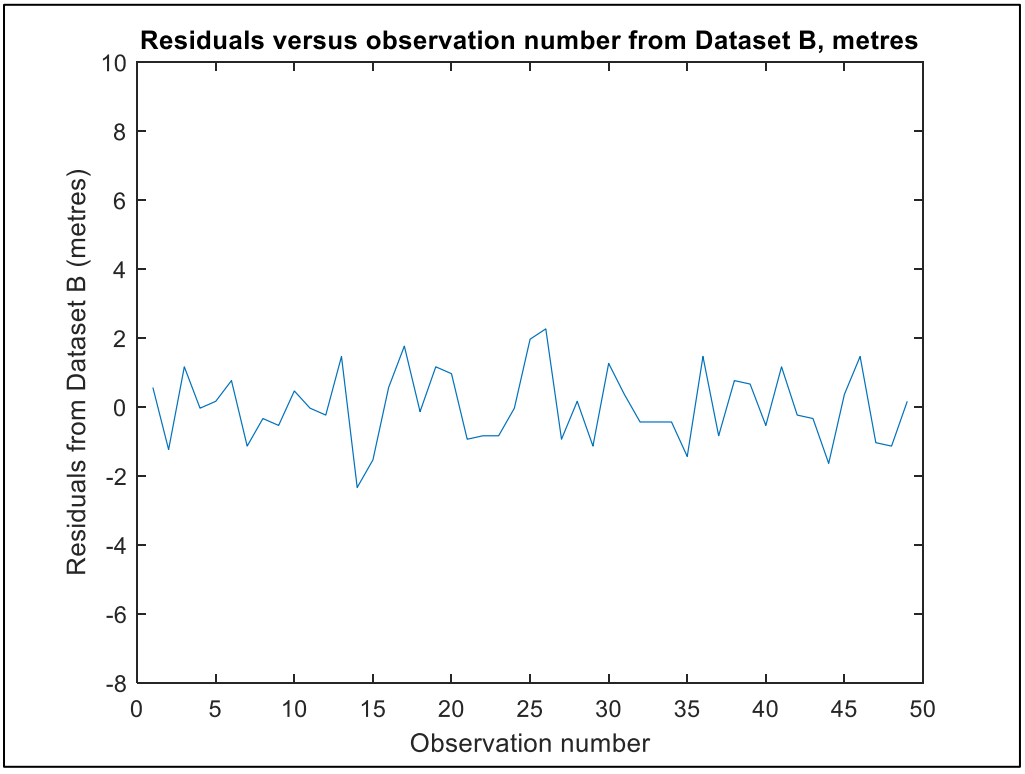
Best estimate at a 95% confidence level	
A	B
$149.79 \pm 1.96(3.41) = [143.11, 156.47]$	$150.16 \pm 1.96(1.04) = [148.12, 152.20]$

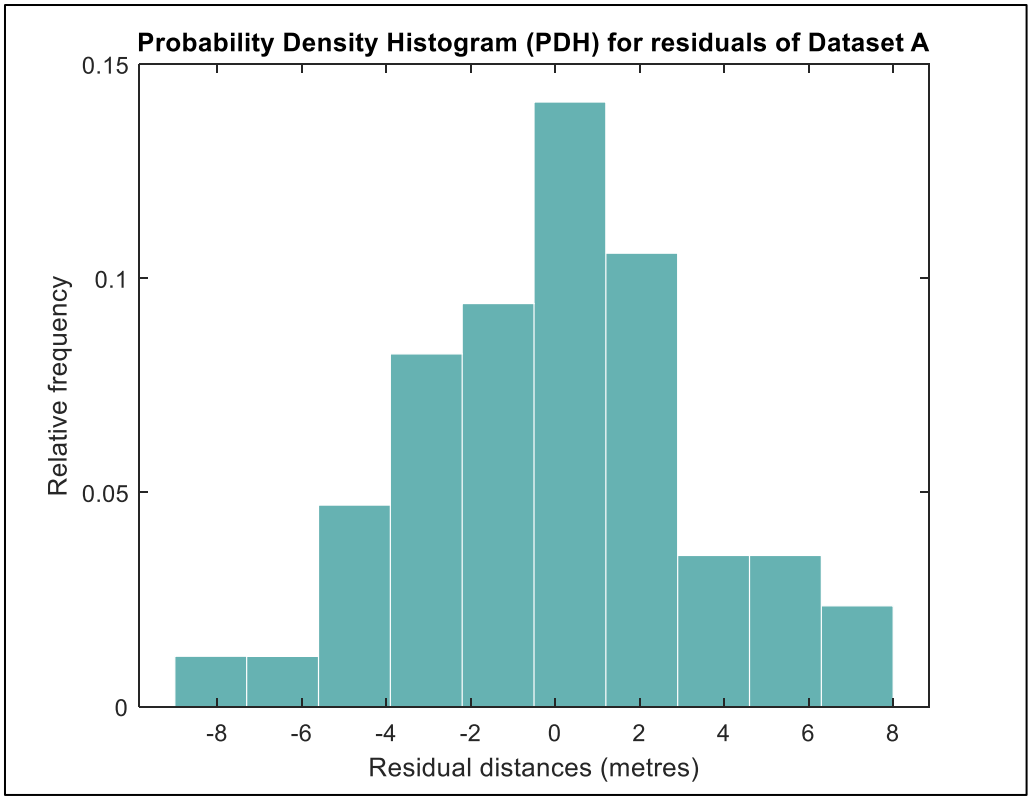
c. Graphical representations

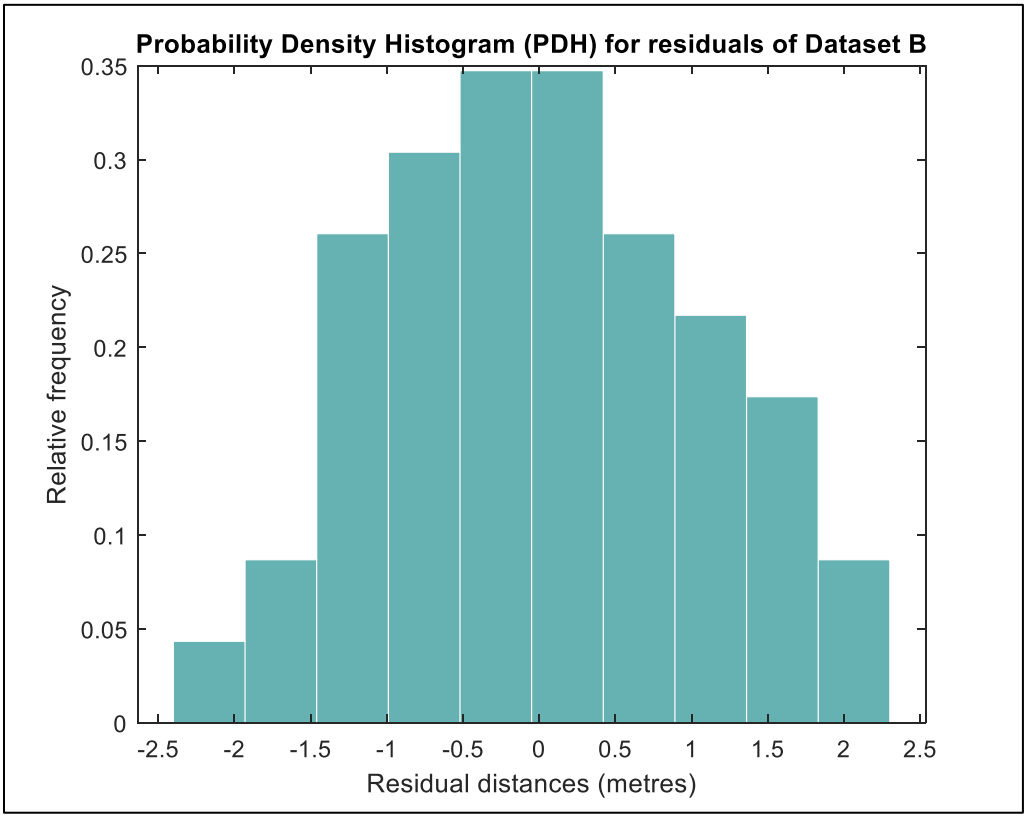





Residuals versus observation number from Dataset A, metres
Residuals from Dataset A (metres)
Observation number
8
6
4
2
0
-2
-4
-6
-8
-10
0
5
10
15
20
25
30
35
40
45
50


Residuals versus observation number from Dataset B, metres
Residuals from Dataset B (metres)
Observation number
10
8
6
4
2
0
-2
-4
-6
-8
0
5
10
15
20
25
30
35
40
45
50


Probability Density Histogram (PDH) for residuals of Dataset A
Relative frequency
0.15
0.1
0.05
0
-8
-6
-4
-2
0
2
4
6
8
Residual distances (metres)


Probability Density Histogram (PDH) for residuals of Dataset B
Relative frequency
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
Residual distances (metres)

d. Precision comparison

The standard deviation, average error, and probable error from the data obtained by observer B were all lower in magnitude compared to the data of observer A (see Table 3). While Dataset B had a blunder that needed to be removed from the analysis, the rest of the observations had a higher level of precision than those of Dataset A. All three precision parameters being considered (standard deviation, average error, and probable error) support the conclusion.

Standard deviation from a Gaussian distribution is the preferred method of measuring precision for two major reasons. It has a physical representation called a Probability Distribution Function (PDF), which uses confidence intervals to detect and remove outliers from the data. It also has the ability to account for large errors because of the squaring process of the residuals. These characteristics make it the most common and reliable choice for reporting on error and precision.

e. Weighted mean

When comparing means, a measure is used to account for the differences in precision between datasets: a weighted mean. To calculate the weighted mean (x_{WM}), the weight (P) of each dataset was determined using the inverse of the squared standard deviation from the mean:

$$P = \frac{1}{\sigma_{\bar{x}}^2} \qquad (5)$$

Using Equation 5, the weights of each dataset were found to be $P_A = 4.31$ and $P_B = 44.97$. The dataset with higher probability, Dataset B, indicates that it will be much closer to the weighted mean than the other. They were then plugged into the weighted mean equation given by:

$$\bar{x}_{WM} = \frac{P_A \bar{x}_A + P_B \bar{x}_B}{P_A + P_B} \qquad (6)$$

Using Equation 6, the weighted mean for Datasets A and B was determined to be:

$$\bar{x}_{WM} = \frac{(4.31)(149.79) + (44.97)(150.16)}{4.31 + 44.97}\ [m]$$

$$\underline{\bar{x}_{WM} = 150.13\ m}$$

The weighted standard deviation (σ_{WM}) equation is given by the square root of the following:

$$\sigma_{WM}^2 = \frac{1}{N-1}\left[\frac{P_A * \sum_{i=1}^{n}(v_i{}^A)^2 + P_B * \sum_{i=1}^{n}(v_i{}^B)^2}{P_A + P_B}\right] \qquad (7)$$

where N is the number of sample sets. The weighted standard deviation for Datasets A and B ($N = 2$) was determined to be:

$$\sigma_{WM}^2 = \left[\frac{(4.31)(568.32) + (44.97)(52.30)}{4.31 + 44.97}\right]\ [m]^2$$

$$\sigma_{WM}^2 = 97.4308\ m^2$$

$$\underline{\sigma_{WM} = \pm 9.87\ m}$$

Part II – Multivariate statistics

a. Mean, variance, and standard deviation

The multivariate set of samples, N, is given by:

$$N = (L_1, L_2, L_3, L_4)$$

where:

$$L_1 = (73, 77, 71, 68, 82, 73, 80, 82, 71, 85)\ [kg]$$

$$L_2 = (1.80, 1.70, 1.86, 1.70, 1.78, 1.71, 1.70, 1.90, 1.65, 1.79)\ [m]$$

$$L_3 = (7.2, 7.2, 8.0, 7.3, 5.0, 7.4, 5.1, 5.2, 7.1, 5.9)\ [m/s]$$

$$L_4 = (13, 7, 19, 10, 7, 14, 8, 8, 12, 6)\ [unitless]$$

Using Equations 1, 2 and 3, the mean, variance, and standard deviation of each set of observations was calculated (See Table 4).

Table 4. The mean, variance, and standard deviation of the random sample of football players' weight, height, speed, and number of goals scored

Analysis parameter	Weight (w)	Height (h)	Speed (s)	Goals scored (g)
Mean ($\bar{x}$)	$76.20\ kg$	$1.76\ m$	$6.54\ m/s$	10.40
Variance (σ^2)	$33.51\ kg^2$	$0.0064\ m^2$	$1.25\ (m/s)^2$	16.71
Standard deviation (σ)	$5.79\ kg$	$0.08\ m$	$1.12\ m/s$	4.09

b. Variance-covariance matrix

Covariance measures the level of correlation between any two components of a multivariate dataset. It is determined by the summation of dot products of the residuals of two components, divided by $n - 1$:

$$\sigma_{ij} = \frac{1}{n-1} * \sum_{i=1}^{n} v_i \cdot v_j \qquad [i * j\ units] \qquad (8)$$

where σ_{ij} is the covariance between the i^{th} and j^{th} observation type and n for each dataset must be equal. In this case, $n_{L_1} = n_{L_2} = n_{L_3} = n_{L_4} = 10$.

The variance-covariance matrix (C_N) for dataset N is given as:

$$C_N = \begin{bmatrix} \sigma_w^2 & \sigma_{wh} & \sigma_{ws} & \sigma_{wg} \\ \sigma_{hw} & \sigma_h^2 & \sigma_{hs} & \sigma_{hg} \\ \sigma_{sw} & \sigma_{sh} & \sigma_s^2 & \sigma_{sg} \\ \sigma_{gw} & \sigma_{gh} & \sigma_{gs} & \sigma_g^2 \end{bmatrix} \qquad (9)$$

Using Equations 8 and 9, the variance-covariance matrix for Datasets A and B was found to be:

$$C_N = \begin{bmatrix} 33.5111\ kg^2 & 0.1658\ kg*m & -5.4311\ kg*m/s & -17.0889\ kg*goal \\ 0.1658\ kg*m & 0.0064\ m^2 & -0.0187\ m*m/s & 0.0493\ m*goal \\ -5.4311\ kg*m/s & -0.0187\ m*m/s & 1.2538\ (m/s)^2 & 3.3822\ (m/s)*goals \\ -17.0889\ kg*goal & 0.0493\ m*goal & 3.3822\ (m/s)*goals & 16.7111\ goals^2 \end{bmatrix}$$

c. Correlation matrix

Correlation is a statistical computation which measures how closely the quantities are related to one another, given by the coefficient ρ. The correlation matrix (ρ_N) for the dataset N, is given as:

$$\rho_N = \begin{bmatrix} 1 & \rho_{wh} & \rho_{ws} & \rho_{wg} \\ \rho_{hw} & 1 & \rho_{hs} & \rho_{hg} \\ \rho_{sw} & \rho_{sh} & 1 & \rho_{sg} \\ \rho_{gw} & \rho_{gh} & \rho_{gs} & 1 \end{bmatrix} \qquad (10)$$

where, (ρ_{ij}) is the correlation between (i) and (j) observations, and by definition, $\rho_{ii} = 1$.

Table 5 summarizes the level of correlation between each two variables in the sample set, where the magnitude of correlation is categorized as: $0 < |\rho_{ij}| \leq 0.35$ is a weak correlation, $0.35 < |\rho_{ij}| \leq 0.75$ is a significant correlation, and $0.75 < |\rho_{ij}| \leq 1.0$ is a strong correlation. The sign of the correlation value reflects the positive or inverse nature of the relationship. As shown in Equation 10, the relationship between any variable and itself (ρ_{ii}) is a complete positive correlation, equal to 1.

Table 5. Correlation relationships for the football player sample set

Matrix Element Indices	Correlation Coefficients Values	Discussion
$\rho_N(1,1), \rho_N(2,2), \rho_N(3,3), and\ \rho_N(4,4)$	$\rho_{ww} = \rho_{hh} = \rho_{ss} = \rho_{gg} = 1.0$	The relationship between any variable and itself is a complete positive correlation, equal to 1.0
$\rho_N(2,1)$, and $\rho_N(1,2)$	$\rho_{hw} = \rho_{wh} = 0.3571$	Weight and height have a borderline weak to significant positive correlation, which means it would have a fairly strong solution.
$\rho_N(3,1)$, and $\rho_N(1,3)$	$\rho_{sw} = \rho_{ws} = -0.838$	Weight and speed have a strong negative correlation – the strongest of the sample set - meaning it would have a weak solution.
$\rho_N(3,3)$, and $\rho_N(4,4)$	$\rho_{gw} = \rho_{wg} = -0.722$	Weight and number of goals scored have a significant negative correlation
$\rho_N(3,3)$, and $\rho_N(4,4)$	$\rho_{sh} = \rho_{hs} = 0.209$	Speed and height have a weak positive correlation, which means it would have a strong solution
$\rho_N(3,3)$, and $\rho_N(4,4)$	$\rho_{gh} = \rho_{hg} = 0.151$	Height and number of goals scored have a weak positive correlation – the weakest of the sample set - meaning it would have a strong solution.
$\rho_N(3,3)$, and $\rho_N(4,4)$	$\rho_{gs} = \rho_{sg} = 0.739$	Speed and number of goals scored have a significant positive correlation

Part III – Methods of mathematical models

Table 6. A summary of the constants, unknowns, observations, and functions of a four-station levelling network

Constants (c)	Unknowns (x)	Observations (l)
$c = [H_A\ H_B]^T$	$x = [H_C\ H_D]^T$	$l = [\Delta h_{AC}, \Delta h_{AD}, \Delta h_{BC}, \Delta h_{CD}, \Delta h_{DB}]^T$
	$u = 2$	$n = 5$

a. Math model by definition

(i) Direct Model
The direct model has one equation per parameter ($x = f(l)$). It is possible to solve for the elevations of C and D directly. One such possibility is:

$$H_C = H_A + \Delta h_{AC}$$
$$H_D = H_A + \Delta h_{AD}$$

(ii) Indirect (Parametric) Model
The direct model has one equation per observation ($l = f(x)$). It is possible to solve for the elevations of C and D indirectly:

$$\Delta h_{AC} = H_C - H_A$$
$$\Delta h_{AD} = H_D - H_A$$
$$\Delta h_{BC} = H_C - H_B$$
$$\Delta h_{CD} = H_D - H_C$$
$$\Delta h_{DB} = H_B - H_D$$

(iii) Conditional Model
The conditional direct model expresses no parameters per equation ($0 = f(l)$). It is possible to solve for the elevations of C and D with the conditional model:

$$\Delta h_{CD} + \Delta h_{DB} + \Delta h_{BC} = 0$$
$$\Delta h_{AC} + \Delta h_{CD} - \Delta h_{AD} = 0$$
$$\Delta h_{CD} + \Delta h_{DB} - \Delta h_{AD} = 0$$

where we know that $H_C = H_A + \Delta h_{AC}$ and $H_D = H_A + \Delta h_{AD}$.

b. Math model in matrix form

(i) Direct Model
In the matrix form of the direct math model (below), the number of the equations representing the relation between the unknowns and the measurements is $m = u = 2$. The functions with respect to observations yields a matrix of constants, signifying that the model is linear.

$$\begin{bmatrix} H_C \\ H_D \end{bmatrix}_{2\times1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}_{2\times5} \begin{bmatrix} \Delta h_{AC} \\ \Delta h_{AD} \\ \Delta h_{BC} \\ \Delta h_{CD} \\ \Delta h_{DB} \end{bmatrix}_{5\times1} + \begin{bmatrix} H_A \\ H_A \end{bmatrix}_{2\times1}$$

(ii) Indirect Model

In the matrix form of the indirect math model (below), the number of the equations representing the relation between the unknowns and the measurements is $m = n = 5$. The functions with respect to unknowns yields a matrix of constants, signifying that the model is linear.

$$\begin{bmatrix} \Delta h_{AC} \\ \Delta h_{AD} \\ \Delta h_{BC} \\ \Delta h_{CD} \\ \Delta h_{DB} \end{bmatrix}_{5\times1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix}_{5\times2} \begin{bmatrix} H_C \\ H_D \end{bmatrix}_{2\times1} + \begin{bmatrix} -H_A \\ -H_A \\ -H_B \\ 0 \\ H_B \end{bmatrix}_{5\times1}$$

(iii) Conditional Model

In the matrix form of the conditional math model (below), the functions with respect to observations yields a matrix of constants. The model is therefore linear. The number of equations is given by the number of independent conditions, which is $m = n - u = 3$.

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & -1 & 0 & 1 & 0 \end{bmatrix}_{3\times5} \begin{bmatrix} \Delta h_{AC} \\ \Delta h_{AD} \\ \Delta h_{BC} \\ \Delta h_{CD} \\ \Delta h_{DB} \end{bmatrix}_{5\times1} + \begin{bmatrix} H_A - H_B \\ H_A - H_B \\ 0 \end{bmatrix}_{3\times1} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}_{3\times1}$$

c. List of all possible solutions

There are 12 possible equations that can be used to find the elevations for C and D, assuming a unique formulation.

$H_C = H_A + \Delta h_{AC}$
$H_C = H_A + \Delta h_{AD} - \Delta h_{CD}$
$H_C = H_A + \Delta h_{AD} + \Delta h_{DB} + \Delta h_{BC}$
$H_C = H_B + \Delta h_{BC}$
$H_C = H_B - \Delta h_{DB} - \Delta h_{CD}$
$H_C = H_B - \Delta h_{DB} - \Delta h_{AD} + \Delta h_{AC}$

$H_D = H_A + \Delta h_{AD}$
$H_D = H_A + \Delta h_{AC} + \Delta h_{CD}$
$H_D = H_A + \Delta h_{AC} - \Delta h_{BC} - \Delta h_{DB}$
$H_D = H_B - \Delta h_{DB}$
$H_D = H_B + \Delta h_{BC} + \Delta h_{CD}$
$H_D = H_B + \Delta h_{BC} - \Delta h_{AC} + \Delta h_{AD}$

Part IV – Error propagation and pre-analysis of survey measurements

$L = 1607.25 \pm 6\ cm$
$W = 1227.67 \pm 7\ cm$

a. Area

$$A = A_{rectangle} + 2A_{semicircle}$$
$$A = L * W + \frac{\pi W^2}{4}$$

b. Length

Length of the track (P):

$$P = 2L + 2\pi W$$

c. Standard deviation in perimeter

Law of the propagation of variances:

$$\sigma_P^2 = \left(\frac{\partial P}{\partial L}\right)^2 \sigma_L^2 + \left(\frac{\partial P}{\partial W}\right)^2 \sigma_W^2$$
$$\frac{\partial P}{\partial L} = 2$$
$$\frac{\partial P}{\partial W} = 2\pi$$

d. Standard deviation in area

Law of the propagation of variances:

$$\sigma_A^2 = \left(\frac{\partial A}{\partial L}\right)^2 \sigma_L^2 + \left(\frac{\partial A}{\partial W}\right)^2 \sigma_W^2$$
$$\frac{\partial A}{\partial L} = W$$
$$\frac{\partial P}{\partial W} = L + \frac{\pi W}{2}$$

Part V