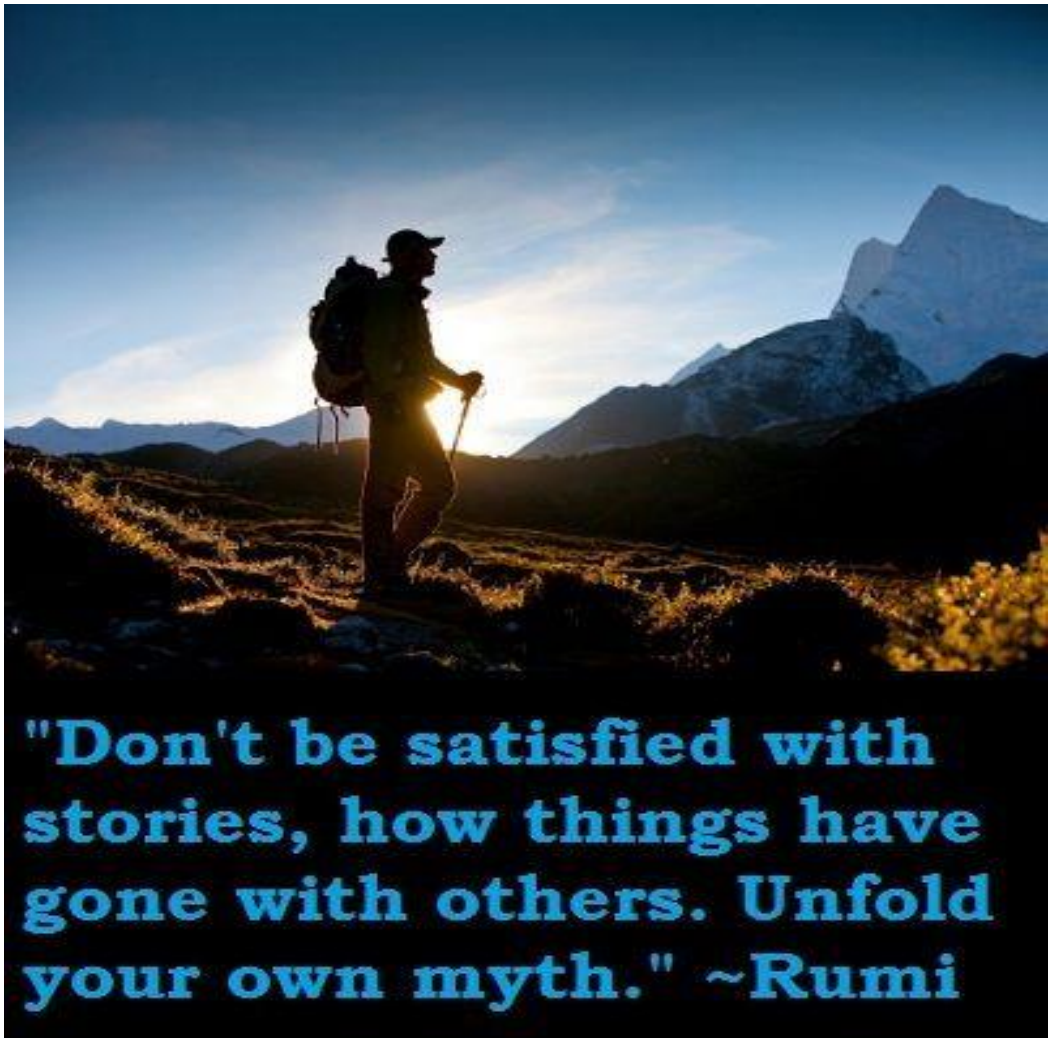
بسم الله الرحمن الرحيم

In the name of Allah the most Beneficial ever merciful



Artificial Intelligence (AI) in Software Engineering

Linear Regression

Department of Computer Science , Univeristy of Karachi (DCS-UBIT)

3rd February 2020

Agenda

0- Introduction to AI in SE

1- Introduction to SIL

2- Linear Regression

3- Linear Regression Examples

4- Linear Regression Derivation

1 Evolution of Artificial Intelligence

Artificial Intelligence

Any tool that accepts the inputs of prior knowledge and then creates an output that implements an action.

Machine Learning

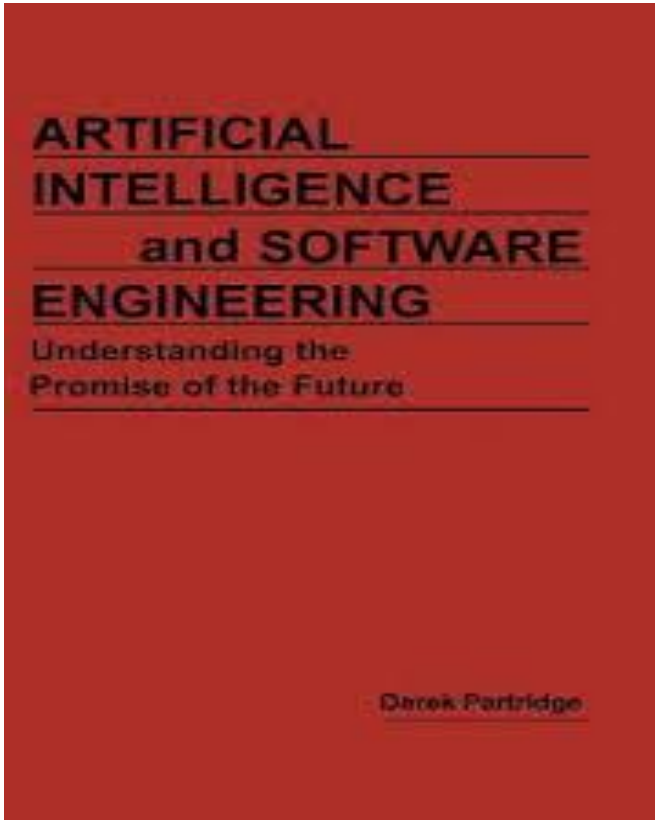
The computer receives inputs and features to create its own program that produces the desired output.

Deep Learning

The computer receives inputs and independently identifies features to generate the desired output.

Two AI Aspects in SE

Intelligent Softwares



Improved software development

Two AI Aspects in SE

Computational Linguistic (NLP)

Autonomous Vehicles

Internet of things (IOT)

Robotics and industrial automation

Military systems

Medical devices and healthcare related applications

Space-based applications

Safety Critical System

Safety Critical Systems

Safety-Critical Software: 15 things every developer should know - Small Business Programming

- ✓ Safety-critical systems are those systems whose failure could result in loss of life, significant property damage, or damage to the environment.
- ✓ Safety-critical software systems are often embedded, distributed systems.
- ✓ Safety Integrity Level (SIL) Approaches Are Common
- ✓ The IEC 61508 was the first international standard to quantify the safety performance of an electrical control system and introduce the concept of lifecycle. The main goal of this standard was to minimize the failures in all electrical/electronic/programmable electronic (typically shortened to E/E/PE) safety-related systems, irrespective of where and how they are used.

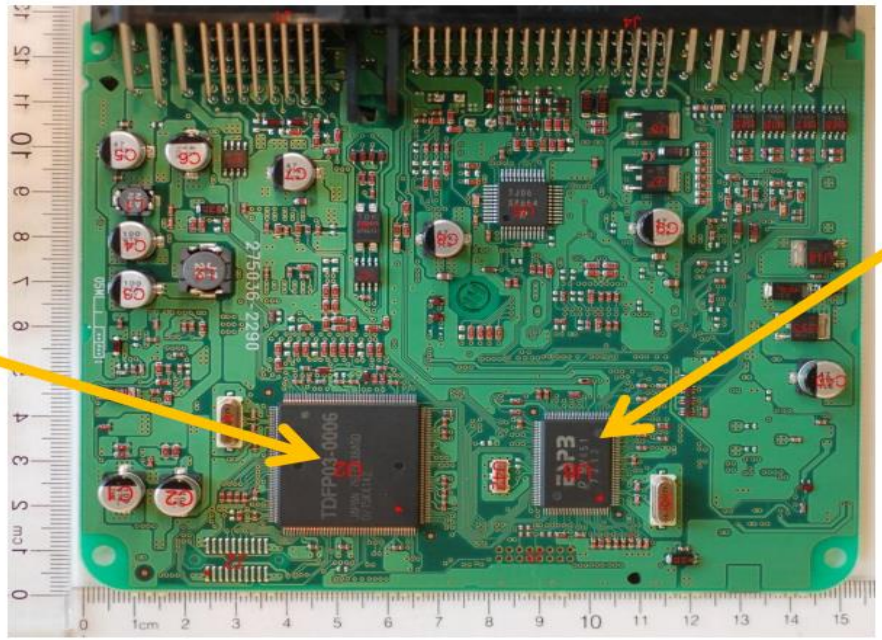
UNIVERSITY OF KARACHI

Embedded and Real Time Systems

Electronic Throttle Control System (ETCS)

Controls air + fuel + spark ➔ engine power

Toyota 2008 ETCS – Two CPUs



http://m.eet.com/media/1201063/Toyota_ECM.jpg

A Case Study of Toyota Unintended Acceleration and Software Safety

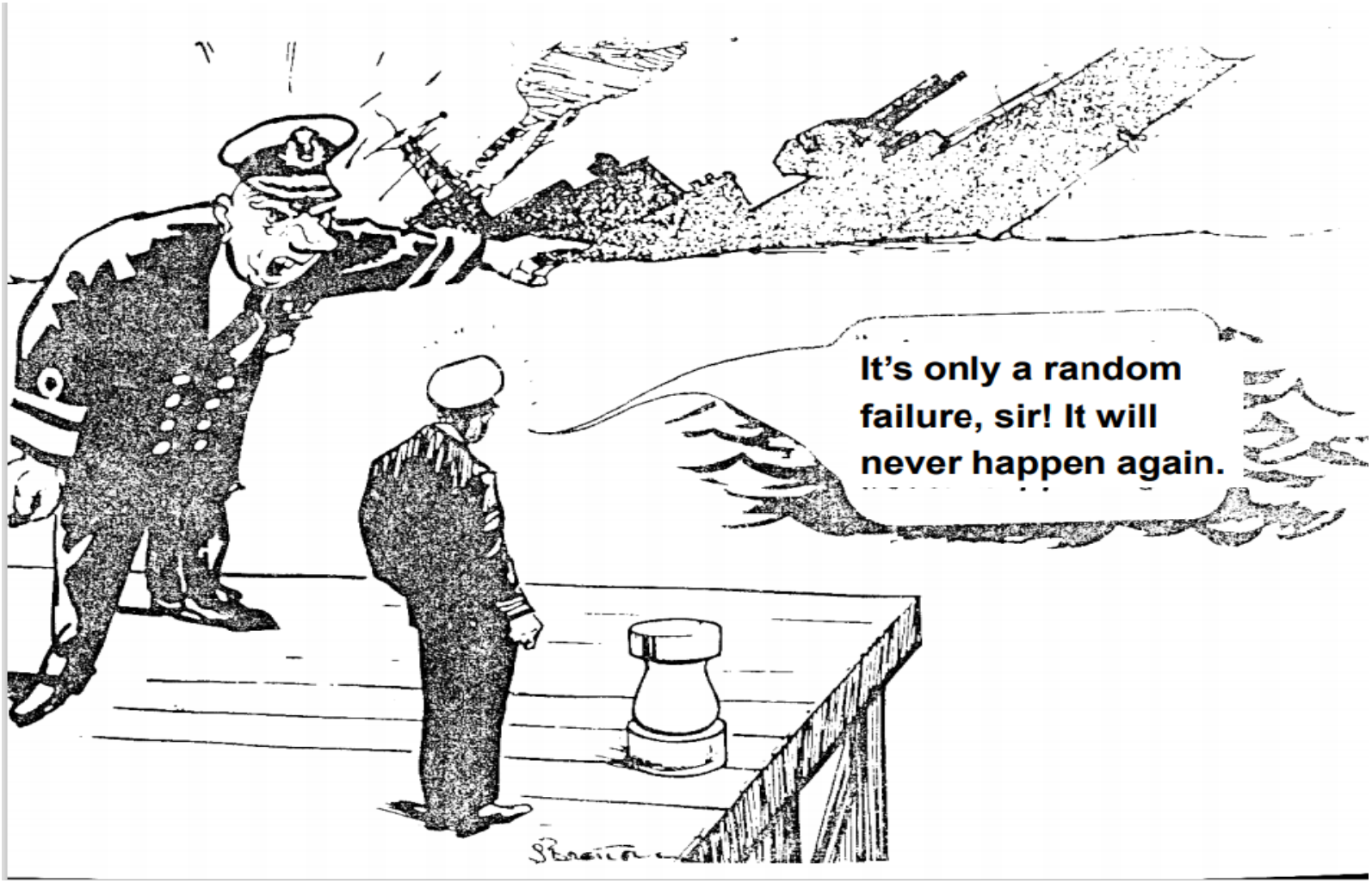
May 25, 2010

Toyota "Unintended Acceleration" Has Killed 89

A 2005 Toyota Prius, which was in an accident, is seen at a police station in Harrison, New York, Wednesday, March 10, 2010. The driver of the Toyota Prius told police that the car accelerated on its own, then lurched down a driveway, across a road and into a stone wall. (AP Photo/Seth Wenig) / AP PHOTO/SETH WENIG

Software does not "fail"

General Purpose Machine + Software = Special Purpose Machine



Safety Integrated Level (SIL)

Terminology Explained: What is Safety Integrity Level (SIL)? - DNV GL - Software

Putting simple: safety Integrity level is a measure of performance required from a safety instrumented system to maintain or achieve the safety state

There are two basic elements associated with this measure:

- Hardware safety integrity: which is typically based upon random hardware failures can normally be estimated to a reasonable level of accuracy via probability of failure on demand (PFD).

- Systematic safety integrity: systematic integrity tends to be harder to quantify. This is due to the diversity of causes of failures; systematic failures may be introduced during the specification, design,implementation, operational and modification phase and may affect hardware as well as software

How to measure SIL ?

- ✓ Risk Matrix
- ✓ Risk graph and Layer of Protection Analysis (LOPA)

A discrete level (one out of a possible four) for specifying the safety requirements of the safety functions which must be allocated to the system. This is the main benefit of SIL as it allows a high-level understanding of each level is typically all that is necessary to convey SIL at management levels.

Safety Integrity level	Probability of Failure on Demand	Risk Reduction Factor
SIL 4	$10^{-5} \geq PofD < 10^{-4}$	100,000 to 10,000
SIL 3	$10^{-4} \geq PofD < 10^{-3}$	10,000 to 1,000
SIL 2	$10^{-3} \geq PofD < 10^{-2}$	1,000 to 100
SIL 1	$10^{-2} \geq PofD < 10^{-1}$	100 to 10

The higher the level of safety integrity, the lower the probability that the safety-related system will fail to carry out the required safety functions.

Terminology Explained: What is Safety Integrity Level (SIL)? - DNV GL - Software

Linear Regression

1

Associating years of professional experience with remuneration.

Predict salary — the dependent variable — based on years of experience

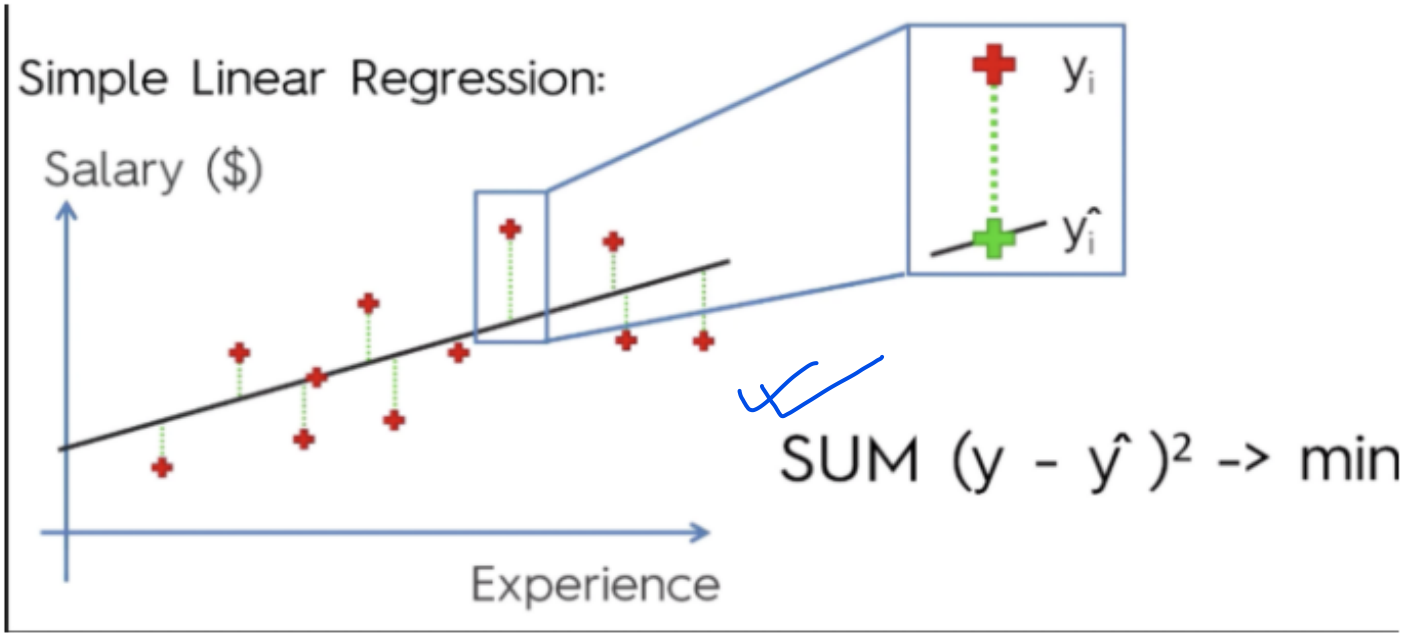
"Sum of Squared Errors" (SSE) is a simple, straightforward method to fit intercept lines between points — and compare those lines to find out the best fit through error reduction. The errors are the sum difference between actual value and predicted value.

$$SSE = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Standard Deviation

y_i = Dependent Variables (Salary)

$\bar{y}$ = Average of Dependent Variables



UNIVERSITY OF KARACHI

Recapitulation and Assumptions

Let's recap from last time. The simple linear regression model is a statistical model for two variables, X and Y. We use X — the **predictor** variable — to try to predict Y, the **target** or **response**[1]. The assumptions of the model are:

1. The distribution of X is arbitrary (and perhaps X is even non-random).
2. If $X = x$, then $Y = \beta_0 + \beta_1 x + \epsilon$, for some constants ("coefficients", "parameters") β_0 and β_1, and some random noise variable ϵ.

[1] Older terms would be "independent" and "dependent" variables, respectively. These import an unwarranted suggestion of causality or even deliberate manipulation on the part of X, so I will try to avoid them.

Y is linear combination of X and weights or Parameters

lecture-05.pdf (cmu.edu)

3 Regression → Ordinary Least Square Error (OLS)

In a typical situation, we also possess observations $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$, which we presume are a realization of the model. Our goals are to estimate the parameters of the model, and to use those parameters to make predictions.

In the notes for the last lecture, we saw that we could estimate the parameters by the **method of least squares**: that is, of minimizing the in-sample mean squared error:

$$\widehat{MSE}(b_0, b_1) \equiv \frac{1}{n} \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2 \qquad (1)$$

Regression is a powerful analysis that can analyze multiple variables simultaneously to answer complex research questions. However, if you don't satisfy the OLS assumptions, you might not be able to trust the results.

UNIVERSITY OF KARACHI

Classical OLS Assumptions

OLS Assumption 1: The regression model is linear in the coefficients and the error term

OLS Assumption 2: The error term has a population mean of zero

Statisticians refer to systematic error like this as bias, and it signifies that our model is inadequate because it is not correct on average.

OLS Assumption 3: All independent variables are uncorrelated with the error term

3. $\mathbb{E}[\epsilon|X = x] = 0$ (no matter what x is), $\text{Var}[\epsilon|X = x] = \sigma^2$ (no matter what x is).

4. ϵ is uncorrelated across observations.

Error simply means standard deviation

In Sample MSE vs. True MSE

The true regression coefficients minimize the true MSE, which is (under the simple linear regression model):

$$(\beta_0, \beta_1) = \underset{(b_0, b_1)}{\operatorname{argmin}} \mathbb{E}\left[(Y - (b_0 + b_1 X))^2\right]$$

What we minimize instead is the mean squared error on the data:

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{(b_0, b_1)}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

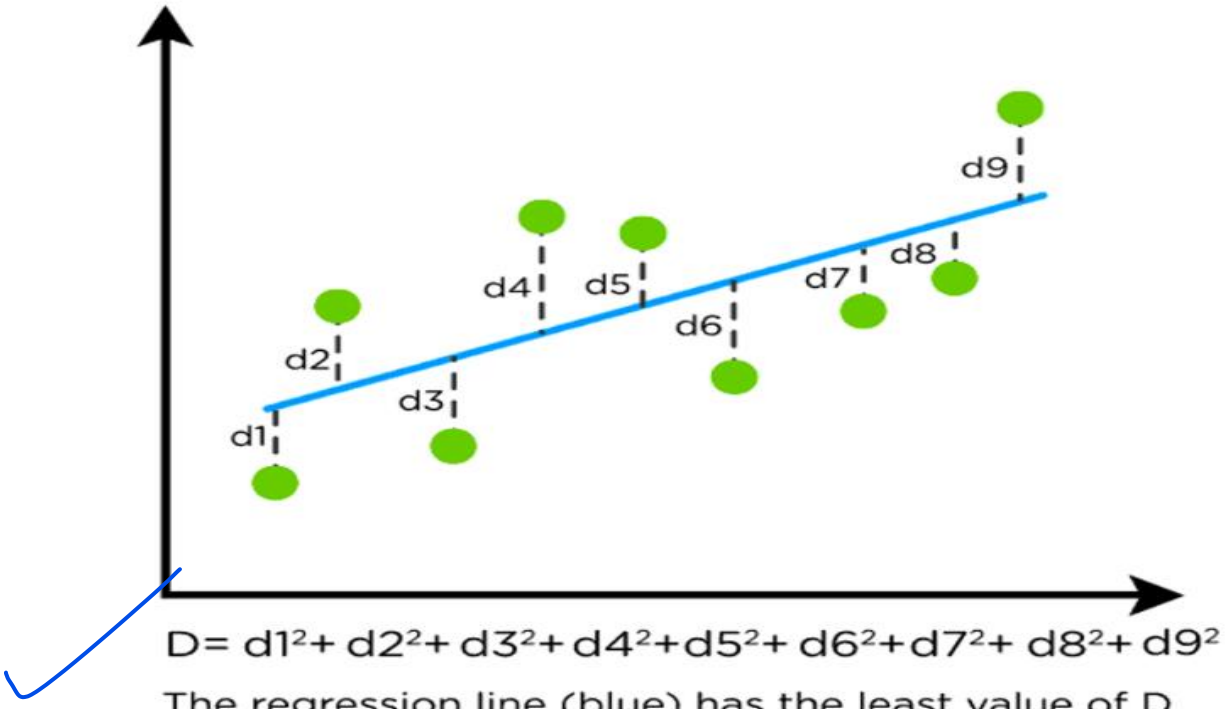
This is the **in-sample** or **empirical** version of the MSE. It's clear that it's a sample average, so for any *fixed* parameters b_0, b_1, when the law of large numbers applies, we should have

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2 \to \mathbb{E}\left[(Y - (b_0 + b_1 X))^2\right]$$

Convergence and minima

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2 \to \mathbb{E}\left[(Y - (b_0 + b_1 X))^2\right]$$

as $n \to \infty$. This should make it *plausible* that the minimum of the function of the left is going to converge on the minimum of the function on the right, but there can be tricky situations, with more complex models, where this convergence doesn't happen.



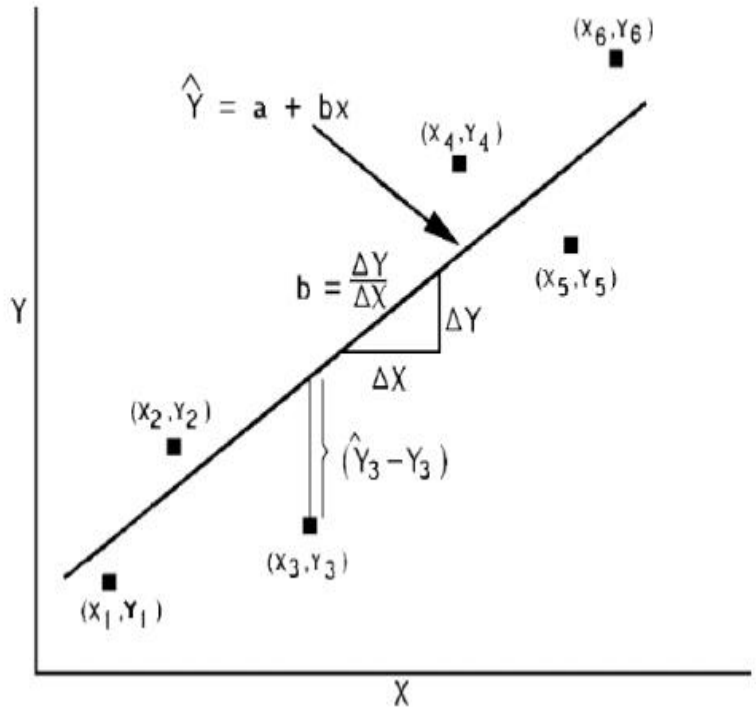
Linear Regression(stats) → Gradient Descent (AI/ML)

- ✓ we can think of linear regression as the task of fitting a straight line (or, in the case of multiple linear regression, a "hyperplane") through a set of points.
- ✓ You can take the leftmost point and the rightmost point and draw a line between them
- ✓ You could compute the slopes of the lines connecting each pair of points and calculate the average slope, drawing a line with this slope that passes through the point at the average of the "x values" and the average of the "y values"
- ✓ You could find the line for which there are an equal number of points above the line and below the line.
- ✓ You could draw a line, and then for each of the data points, measure the vertical distance between the point and the line, and add these up; the fitted line would be the one where this sum of distances is as small as possible
- ✓ you could draw a line, and then for each of the data points, measure the vertical distance between the point and the line, square it, and add these up; the fitted line would be the one where this sum of distances is as small as possible. This is OLS or SSE.

Linear Regression (Example 1)

1 Scatter Plot to Illustrate Linear Relationship

To illustrate the principle, we will use the artificial data presented as a scatter diagram in Figure 10-1.



Because of the existence of experimental errors, the observations (Y) made for a given set of independent values (X) will not permit the calculation of a single straight line that will go through all the points.

The least squares line is the line that goes through the points so that the sum of the squares of the vertical deviations of the points from the line is minimal.

Those with a knowledge of calculus should recognize that this is a problem of finding the minimum value of a function. T

UNIVERSITY OF KARACHI

Estimates of Parameters a,b ; b0, b1 ; θ_0, θ_1

That is, set the first derivatives of the regression equation with respect to a and b to zero and solve for a and b. This procedure yields the following formulas for a and b based on k pairs of X and Y: If X is not a random variable, the coefficients so obtained are the best linear unbiased estimates of the true parameters.

$$b = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{\Sigma(X - \overline{X})^2} = \frac{\Sigma XY - (\Sigma X \Sigma Y)/k}{\Sigma X^2 - (\Sigma X)^2/k}$$

$$a = \frac{(\Sigma X^2)\overline{Y} - \overline{X}(\Sigma XY)}{\Sigma X^2 - (\Sigma X)^2/k} = \overline{Y} - b\overline{X}$$

Fixed Design Points

we assume that a linear response was appropriate to describe the effect of N fertilizer on the sucrose content of beet roots. Note that the N rates were specifically chosen by the experimenter and, therefore, are considered fixed design points

Table 1. Elements necessary to compute the least squares regression for changes in % sucrose associated with changes in N-fertilizer.

X lbs N (acre)	Y mean % (sucrose)	X^2	XY	$\hat{Y}$ predicted (% sucrose)	$\hat{Y}$ - Y
0	16.16	0	0	16.22	-0.06
50	15.74	2,500	787	15.78	-0.04
100	15.29	10,000	1,529	15.35	-0.06
150	15.29	22,500	2,293.5	14.92	0.39
200	14.36	40,000	2,872	14.48	-0.12
250	13.94	62,500	3,485	14.05	-0.11
(Total) 750	90.78	137,500	10,966		
(Mean) 125	15.13	22,916.67			

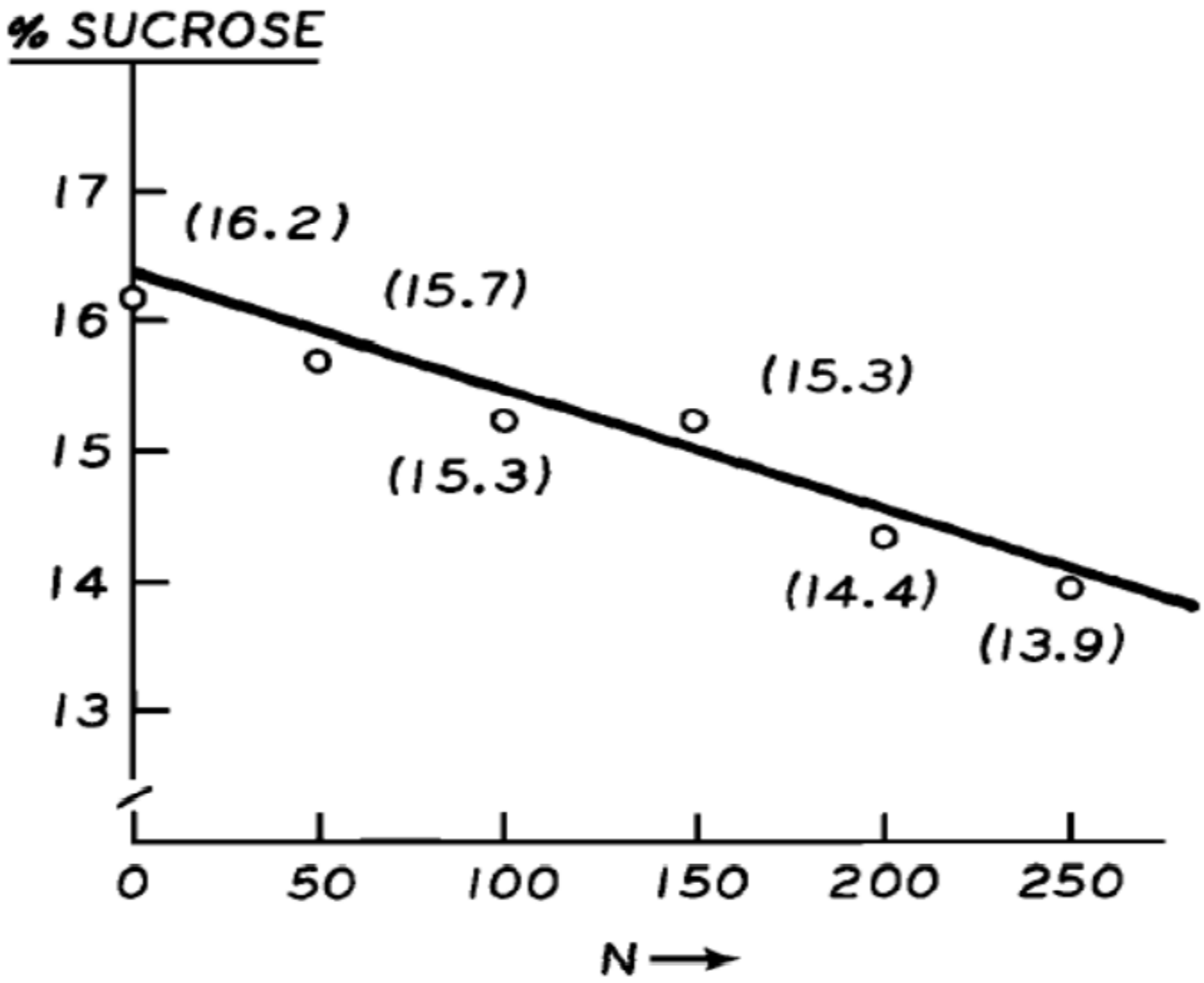
Result : Equation of Straight Line

$$b = \frac{\Sigma XY - (\Sigma X \bullet \Sigma Y)/k}{\Sigma X^2 - (\Sigma X)^2/k} = \frac{10966.5 - (750)(90.78)/6}{137500 - (750)^2/6}$$

$$= \frac{-381.0}{43750} = -0.0087$$

$$a = \overline{Y} - b\overline{X} = 15.13 - (-0.0087)(125) = 16.22$$

The relationship between % sucrose and levels of N-fertilizer



Linear Regression (Example 2)

1 Independence variable - measurement with error

Sometimes researchers are interested in estimating a quantity that is difficult to measure directly.

It is desirable to be able to predict this quantity from another variable that is easier to measure.

For example, to predict leaf area from the length and width of leaves, sugar content from percent total solids, or rate of gain from initial body weight

For a case study we will use data collected to see if it is possible to predict the weight of the livers of mice from their body weights. The data are given in Table and the calculation of the regression line is shown below the table.

UNIVERSITY OF KARACHI

Data Points and Predictions

Table · Mice body and liver weights (grams) and predicted liver weights from a linear regression of Y on X.

X body wts (x10g)	X^2	Y liver wt.	Y^2	XY	$\hat{Y}$ predicted liver wt.	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
16.4	268.96	2.67	7.13	43.79	2.37	0.30	0.09
17.2	295.84	2.75	7.56	47.30	2.95	-0.20	0.04
17.6	309.76	2.99	8.94	52.62	3.24	-0.25	0.06
18.0	324.00	3.14	9.86	56.52	3.53	-0.39	0.15
18.2	331.24	3.88	15.05	70.62	3.68	-0.20	0.04
18.5	342.25	4.23	17.89	78.25	3.89	0.34	0.12
(Total) 105.9	1,872.05	19.66	66.44	389.10		0.00	0.50
(Mean) 17.65		3.28					

UNIVERSITY OF KARACHI

Calculations

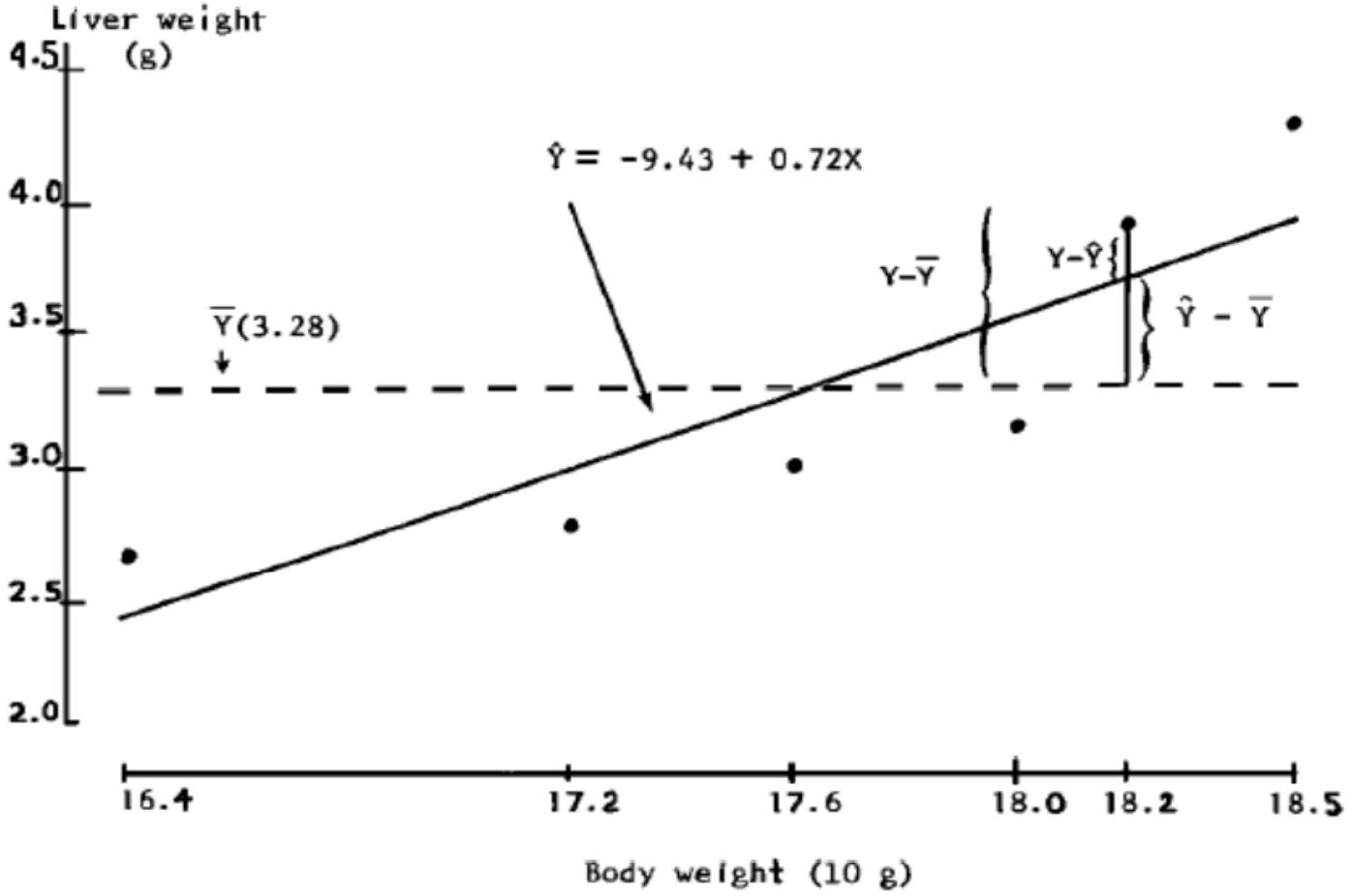
$$b = \frac{\Sigma XY - (\Sigma X \bullet \Sigma Y)/k}{\Sigma X^2 - (\Sigma X)^2/k} = \frac{349.10 - (105.9)(19.66)/6}{1872.05 - (105.9)^2/6}$$

$$a = \overline{Y} - b\overline{X} = 3.28 - 0.72(17.65) = -9.43$$

$$\hat{Y} = -9.43 + 0.72X$$

4 Illustration and Relationship

The relation between body and liver weights and the regression line are plotted in Figure



UNIVERSITY OF KARACHI

References

Safety-Critical Software: 15 things every developer should know - Small Business Programming

Microsoft PowerPoint - 2014_phil_public_toyota_talk_version7_handouts.pptx (cmu.edu)

Engineering a Safer and More Secure World (mit.edu)

Terminology Explained: What is Safety Integrity Level (SIL)? - DNV GL - Software

Department of Compute Science (UBIT Building), Karachi, Pakistan.

1200 Acres (5.2 Km sq.)
53 Departments
19 Institutes
25000 Students

UNIVERSITY OF KARACHI