Linear Regression

- Linear regression and correlation can help you one of the target variable. eg: if an auto mechanic's salary is related to his work experience.
- Professionals often want to know how two or more numeric variables are
  - For example, is there a relationship between the grade on the second math exam a student takes and the grade on the final exam? If there is a relationship, what is the relationship and how strong is it?
- In another example, your income may be determined by your education, your profession, your years of experience, and your ability. The amount you pay a repair person for labor is often determined by an initial amount plus an hourly fee.
- The type of data described in the examples is **bivariate** data — "bi" for two variables. In reality, statisticians use **multivariate** data, meaning many variables.
- Simplest form of regression is "linear regression" with one independent variable (x). This involves data that fits a line in two dimensions.

- Linear regression for two variables is based on a linear equation with one independent variable. The equation has the form:

- $y=a+bx$
- where a and b are constant numbers.

- The variable x is the independent variable, and y is the dependent variable. Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

- The following examples are linear equations.
- $y=3+2x$
- $y=-0.01+1.2x$

- Data rarely fit a straight line exactly.
- Typically, you have a set of data whose scatter plot appears to “fit” a straight line. This is called a Line of Best Fit or Least-Squares Line.
- eg: The exam score, x, is the independent variable and the final exam score, y, is the dependent variable.
  - We can plot a regression line that best “fits” the data.
  - If each of you were to fit a line , we can use what is called a **least-squares** regression line to obtain the **best fit line**.

Error

- Data points not on the best fit line is called the "error" or residual. It is not an error in the sense of a mistake.
- The absolute value of a residual measures the vertical distance between the actual value of y and the estimated value of y.
- In other words, it measures the vertical distance between the actual data point and the predicted point on the line.

- If the observed data point lies above the line, the residual is positive.
- If the observed data point lies below the line, the residual is negative.

- For each data point, you can calculate the residuals or errors,
  - $y_i - \hat{y}_i = \epsilon_i$ for i = 1, 2, 3, …, 11
  - Each $|\varepsilon|$ is a vertical distance.

- For the example about the third exam scores and the final exam scores for the 11 statistics students, there are 11 data points.
- Therefore, there are 11 ε values. If you square each ε and add this is called the Sum of Squared Errors (SSE).

- When you make the SSE a minimum, you determine the points that are on the line of best fit.

Least Squares Criteria for Best Fit

- The process of fitting the best-fit line is called linear regression.
- The idea behind finding the best-fit line is based on the assumption that the data are scattered about a straight line.
- The criteria for the best fit line is that the sum of the squared errors (SSE) is minimized, that is, made as small as possible.
- Any other line you might choose would have a higher SSE than the best fit line.
- This best fit line is called the **least-squares regression line.**

- The slope of the line, b, describes how changes in the variables are related.
- It is important to interpret the slope of the line in the context of the situation represented by the data.
- Interpretation of the Slope: The slope of the best-fit line tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average.

- eg:
- Third Exam vs Final Exam Example: Slope: The slope of the line is b = 4.83.
- Interpretation: For a one-point increase in the score on the third exam, the final exam score increases by 4.83 points, on average.

The Correlation Coefficient r

- The correlation coefficient, r, developed by Karl Pearson in the early 1900s, is numerical and provides a measure of strength and direction of the linear association between the independent variable x and the dependent variable y.

- The correlation coefficient is calculated as r

Correlation Coefficient Formula

$$\mathbf{r} = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]}}$$

- If you suspect a linear relationship between x and y, then r can measure how strong the linear relationship is.

- What the VALUE of r tells us: The value of r is always between –1 and +1: $-1 \le r \le 1$. The size of the correlation rindicates the strength of the linear relationship between x and y. Values of r **close** to **–1 or to +1** indicate a stronger linear relationship between x and y.
- If $r = 0$ there is absolutely no linear relationship between x and y (no linear correlation).
- If $r = 1$, there is perfect positive correlation.
- If $r = -1$, there is perfect negativecorrelation.

- What the **SIGN** of r tells us: A positive value of r means that when x increases, y tends to increase and when x decreases, y tends to decrease (positive correlation). A negative value of r means that when x increases, y tends to decrease and when x decreases, y tends to increase (negative correlation). The sign of r is the same as the sign of the slope,b, of the best-fit line.

- Note
- Strong correlation does not suggest that x causes y or y causes x. We say "correlation does not imply causation."