REGRESSION

Module 3

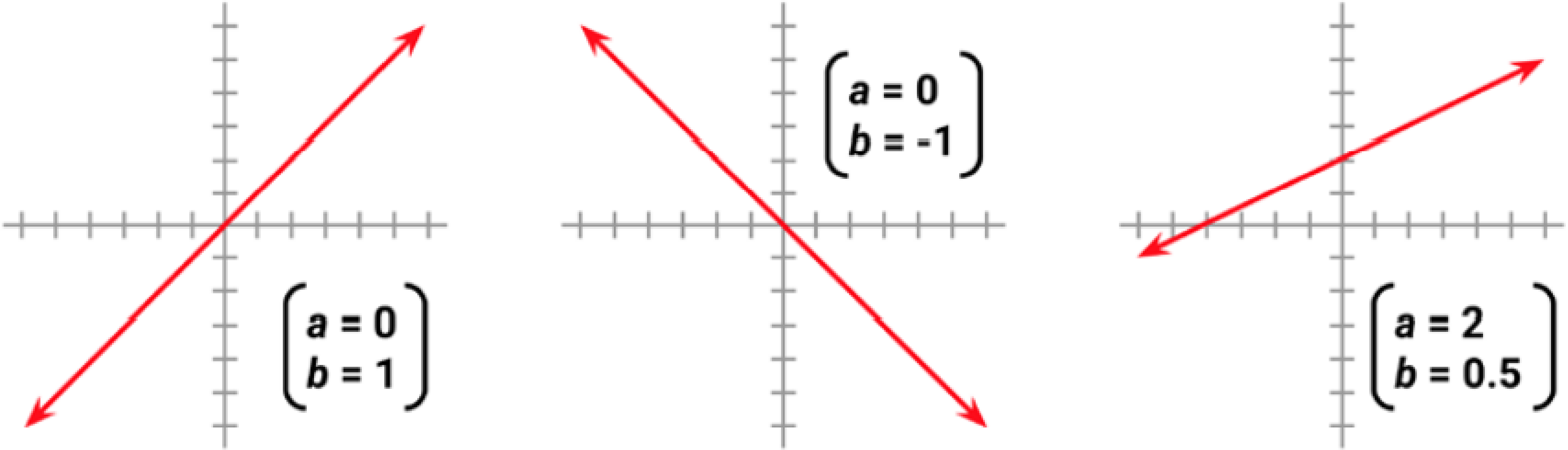
Understanding Regression

Regression is concerned with specifying the relationship between a single numeric dependent variable (the value to be predicted) and one or more numeric independent variables (the predictors). The dependent variable depends upon the value of the independent variable or variables. $Y= X_1^2$

We'll begin by assuming that the relationship between the independent and dependent variables follows a straight line.

- slope-intercept form $y = a + bx$
  - where y is the dependent variable and x is the independent variable. In this formula, the slope ‘b’ indicates how much the line rises for each increase in x.
  - Variable ‘a’ is known as the intercept because it specifies where the line crosses the vertical axis.

- slope-intercept form y = a + bx



- The machine's job is to identify values of a and b such that the specified line is best able to relate the supplied x values to the values of y.
- Regression methods are also used for **hypothesis testing**, which involves determining whether data indicate that a presupposition is more likely to be true or false.

- Regression analysis is commonly used for modeling complex relationships among data elements, estimating the impact of a treatment on an outcome, and extrapolating into the future.
- Some specific use cases include:
  - Examining how populations and individuals vary by their measured characteristics, for use in scientific research across economics, sociology, psychology, physics, and ecology
  - Quantifying the causal relationship between an event and the response, such as those in clinical drug trials, engineering safety tests, or marketing research.
  - Identifying patterns that can be used to forecast future behavior criteria, such as predicting insurance claims, natural disaster damage, election results, and crime rates.

Regression Models

- If there is only a single independent variable, this is known as simple linear regression, otherwise it is known as multiple regression.
- Both of these models assume that the dependent variable is continuous.

Regression can also be used for other types of dependent variables and even for some classification tasks.

- For instance, Logistic regression can be used to model a binary categorical outcome;
- while Poisson regression—named after the French mathematician Siméon Poisson—models integer count data.
- The method known as Multinomial logistic regression models a categorical outcome; thus, it can be used for classification.

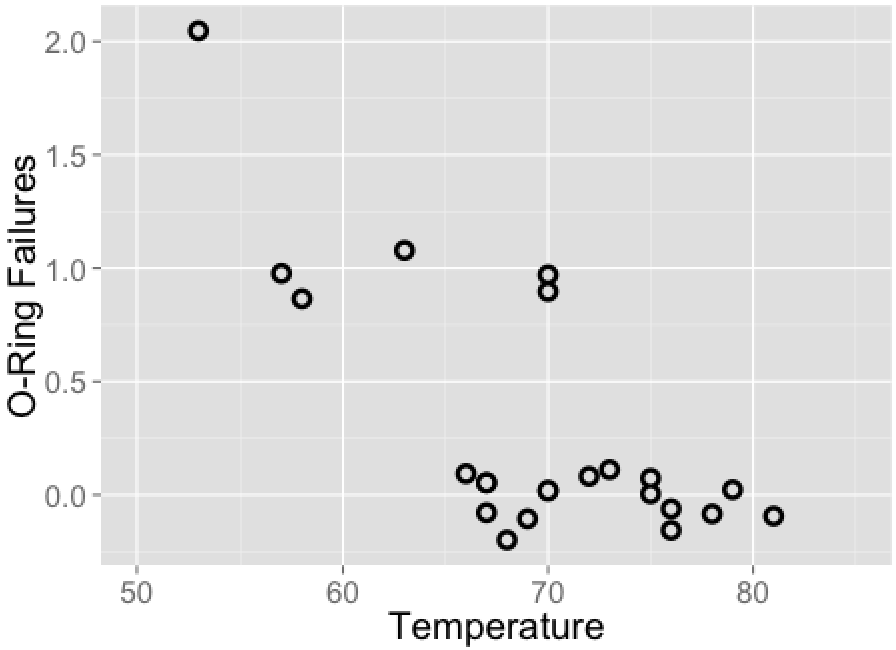
Simple linear regression- Example

- On January 28, 1986, seven crewmembers of the United States space shuttle *Challenger* were killed when a rocket booster failed, causing a catastrophic disintegration. In the aftermath, experts focused on the launch temperature as a potential culprit. The rubber O-rings responsible for sealing the rocket joints had never been tested below 40ºF (4ºC) and the weather on the launch day was unusually cold and below freezing when O-rings responsible for sealing the joints of the rocket booster failed and caused a catastrophic explosion.

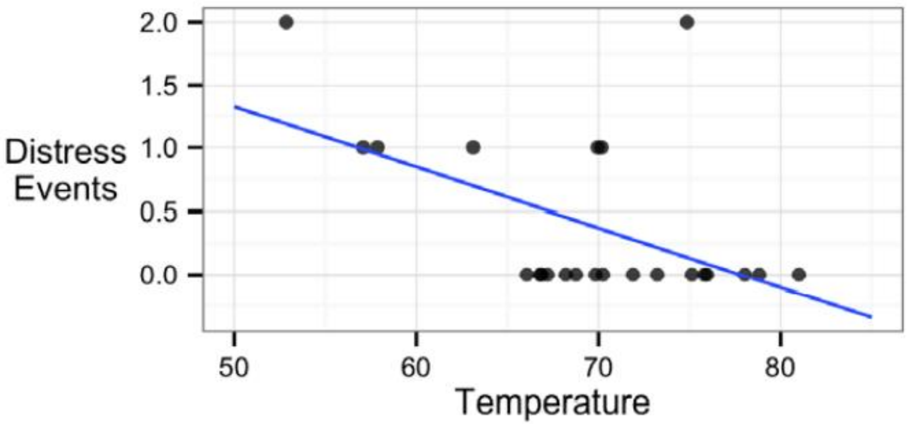
- The rocket engineers believed that cold temperatures could make the components more brittle and less able to seal properly, which would result in a higher chance of a dangerous fuel leak.

- *A regression model that demonstrated a link between temperature and O-ring failure, and could forecast the chance of failure given the expected temperature at launch, might be very helpful.*

- The scientists' discussion turned to data from 23 previous successful shuttle launches which recorded the number of O-ring failures versus the launch temperature. Launches occurring at higher temperatures tend to have fewer O-ring failures. Additionally, the coldest launch (58 degrees F) had two rings fail, the most of any launch. To answer this, we can turn to simple linear regression.



- Suppose we know that the estimated regression parameters in the equation for the shuttle launch data are: a = 3.70 and b = -0.048.
- Hence, the full linear equation is y = 3.70 – 0.048x. we can plot the line on the scatterplot like this:



As the line shows, at 60 degrees Fahrenheit, we predict just under one O-ring distress. At 70 degrees Fahrenheit, we expect around 0.3 failures. If we extrapolate our model, all the way to 31 degrees—the forecasted temperature for the Challenger launch—we would expect about 3.70 - 0.048 * 31 = 2.21 O-ring distress events.

Regression Problem-1

The sales of a company (in million dollars) for each year are shown in the table below.

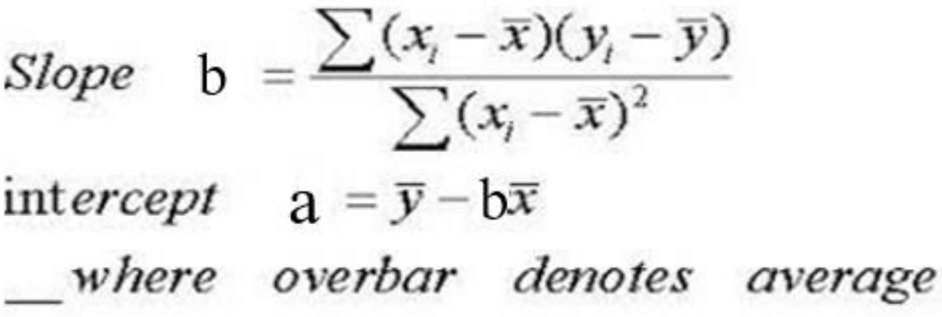
x (year)	2005	2006	2007	2008	2009
y (sales)	12	19	29	37	45

a) Find the least square regression line y = a x + b.

b) Use the least squares regression line as a model to estimate the sales of the company in 2012.

Regression Problem-1

$$y = a + bx$$



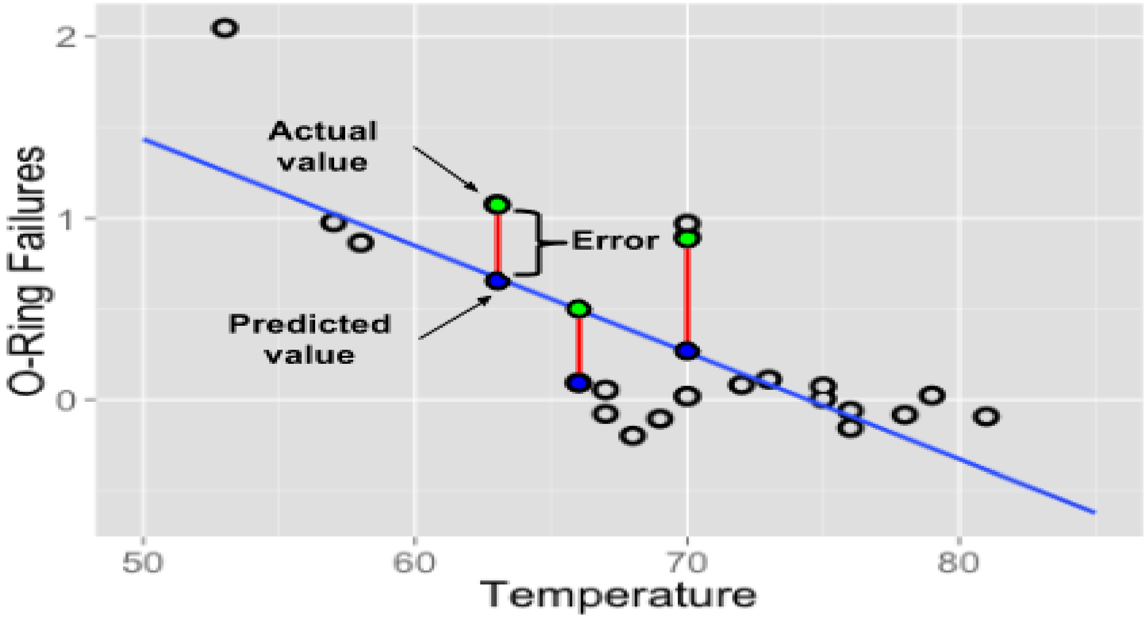
Slope $b = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$

int*ercept* $a = \bar{y} - b\bar{x}$

__where overbar denotes average

Ordinary Least Squares

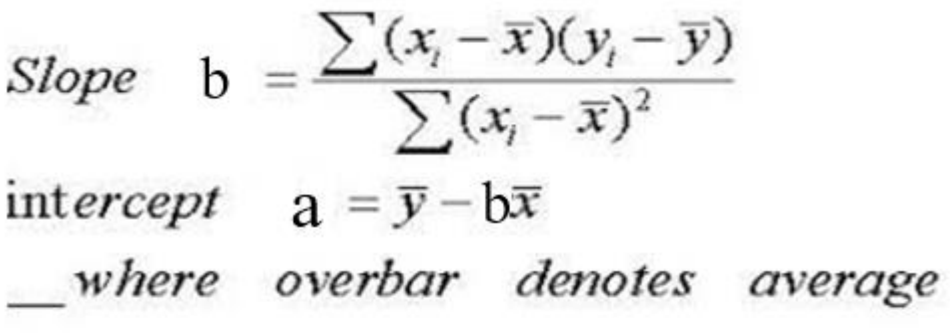
- In order to determine the optimal estimates of α (a) and β (b), an estimation method known as **Ordinary Least Squares** (OLS) was used.
- In OLS regression, the slope and intercept are chosen such that they minimize the sum of the squared errors, that is, the vertical distance between the predicted y value and the actual y value. These errors are known as residuals.



In mathematical terms, the goal of OLS regression can be expressed as the task of minimizing the following equation:

$$\sum(y_i - \hat{y}_i)^2 = \sum e_i^2$$

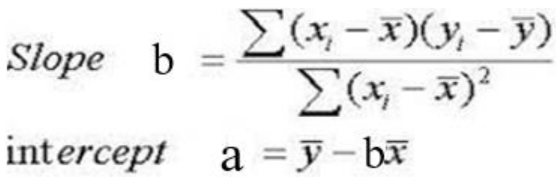
- In plain language, this equation defines ***e*** (the error) as the difference between the actual y value and the predicted y value. The error values are squared and summed across all points in the data.



$$\textit{Slope} \quad \mathrm{b} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\text{int}\textit{ercept} \quad \mathrm{a} = \bar{y} - \mathrm{b}\bar{x}$$

__*where overbar denotes average*



$$\textit{Slope} \quad b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\textit{intercept} \quad a = \bar{y} - b\bar{x}$$

- Numerator in slope formula involves taking the sum of each data point's deviation from the mean x value multiplied by that point's deviation away from the mean y value.

- This is known as the covariance of x and y, denoted as Cov(x, y). With this in mind, we can re-write the formula for ‘b’ as:

  - **b= cov(x,y)/var(x)**

Correlations

- The correlation between two variables is a number that indicates how closely their relationship follows a straight line.
- The correlation ranges between -1 and +1. The extreme values indicate a perfectly linear relationship, while a correlation close to zero indicates the absence of a linear relationship.
- Correlation typically refers to Pearson's correlation coefficient, which was developed by the 20th century mathematician Karl Pearson.

Correlations…

- The following formula defines Pearson's correlation:
  - $\mathrm{Corr}\ (x, y) = \mathrm{Cov}\ (x, y)/\sigma\ x\ \sigma\ y$

- For the above example problem, the correlation between the temperature and the number of distressed O-rings is -0.51.
- The negative correlation implies that increases in temperature are related to decreases in the number of distressed O-rings. To the NASA engineers studying the O-ring data, this would have been a very clear indicator that a low temperature launch could be problematic.

Correlations…

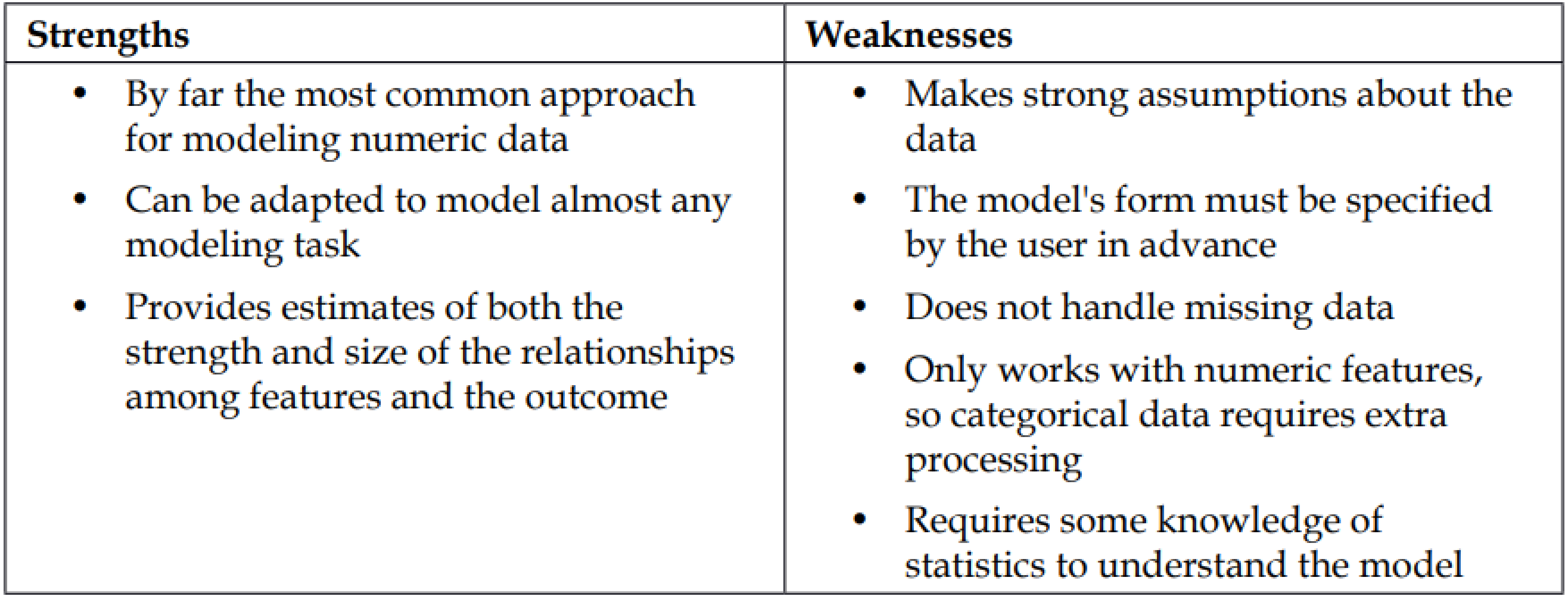
- There are various rules used to interpret correlation strength. One method assigns a status of "weak" to values between 0.1 and 0.3, "moderate" to the range of 0.3 to 0.5, and "strong" to values above 0.5
- These also apply to similar ranges of negative correlations.
- A weak correlation to values between 0.1 and 0.3, moderate for 0.3 to 0.5, and strong for values above 0.5

Multiple Linear Regression

- Most real-world analyses have more than one independent variable.
- In this context, multiple linear regression is used for the numeric prediction task.

Multiple Linear Regression…..

- The strengths and weaknesses of multiple linear regression are shown in the following table:



Strengths	**Weaknesses**
• By far the most common approach for modeling numeric data • Can be adapted to model almost any modeling task • Provides estimates of both the strength and size of the relationships among features and the outcome	• Makes strong assumptions about the data • The model's form must be specified by the user in advance • Does not handle missing data • Only works with numeric features, so categorical data requires extra processing • Requires some knowledge of statistics to understand the model

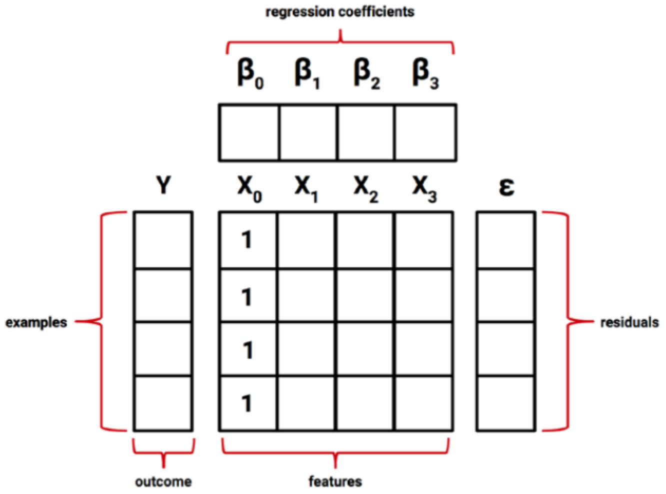
Multiple Linear Regression…..

- Multiple regression is an extension of simple linear regression. The goal in both cases is similar, and the key difference is that there are additional terms for additional independent variables.

- Multiple regression generally follow the form:
  - $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \varepsilon$

- The dependent variable ‘y’ is specified as the sum of an intercept term plus the product of the estimated β value and the x value for each of *i* features. An error term (epsilon) has been added here as a reminder that the predictions are not perfect.

- $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \varepsilon$
  - Since the intercept term α is really no different than any other regression parameter, it is also denoted as β0 as shown in the following equation:
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \varepsilon$
  - Using Vector notation, it becomes
    - $y = \beta x + \varepsilon$



Multiple Linear Regression…..

- Now the goal is to solve for the vector β that minimizes the sum of the squared errors between the predicted and actual y values.
- The best estimate of the vector β can be computed as:
  - $\beta\,\hat{} = (X^T X)^{-1} X^T Y$
- This solution uses a pair of matrix operations, T indicates the transpose of matrix X, while the negative exponent indicates the matrix inverse.