Linear Regression

- Is the equation of the linear relationship between an explanatory variable (x) and response variable (y)
- It's used to predict the average response for all subjects with a given value of the explanatory variable
- Some relationships are such that the points of a scatterplot tend to fall along a straight line – a linear relationship
- A **regression line** is a straight line that describes how a response variable y changes as an explanatory variable x changes

$$\mathbf{y = a + bx}$$

- b = slope (if x goes up by 1, there is a constant incremental change in y by b), a = intercept, the value of y when x=0.
- Used to predict the value of y for a given value of x

Predicted y-value: $\hat{y} = \hat{a} + \hat{b}x$

Residuals

- We want the regression line to be as close as possible to the data points in the vertical (y) direction (since that is what we are trying to predict)
- We want the residuals to be as small as possible overall
- A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line, for each x value in the data

$$\boldsymbol{\varepsilon} = \textbf{residual = observed y – predicted y = y} - \hat{y}$$

eg. Regression line: y = 30.1 x – 159
Predicted value of y for x=22.1℃ is 506.21
The observed ice cream sales was $552
The residual for 22.1 ℃ = 552 – 506.21
= 45.79

- Residuals can be negative or positive, or they can cancel out
- To avoid them cancelling each other out, we use the line that minimizes the sum of the squares of the residuals (minimises the overall residuals)
- Because residuals can be positive and negative, we don't want them to cancel out when we add them so we square them

$$SSR = \varepsilon_1^2 + \varepsilon_2^2 + \varepsilon_3^2 + \ldots + \varepsilon_n^2$$

- The best fitted line is the one with slope and intercept such that the SSR is the smallest value for this data (**method of least squares**)
eg. Which of the two possible lines of best fit: $\hat{y}1 = 6 - 1.5x$ or $\hat{y}2 = 6.5 - 2x$ is better, according to the method of least squares?

For $\hat{y}1$ = 6 – 1.5x, substitute x = 0 to find the predicted value
When x = 0, $\hat{y}1$ = 6
Then subtract the predicted values from the observed values in the y column to find the residual
6 – 6 = 0
$\varepsilon^2 = 0^2 = 0$
Find all of the residuals and sum the squares of the residuals.
Repeat this for the second line of best fit.
SSR ($\hat{y}1$ = 6 – 1.5x) = 7.5
SSR ($\hat{y}2$ = 6.5 – 2x) = 9
$\hat{y}1$ = 6 – 1.5x is a better line of best fit.
The smaller the sum of squares of residuals (SSR), the better the line of best fit.
- You can also plot the residuals by placing a dot where the residual is when x = 0 etc

Regression Line Equation

$$\hat{\boldsymbol{y}} = \hat{\boldsymbol{a}} + \hat{\boldsymbol{b}}\boldsymbol{x}$$

- Where slope = $\hat{\boldsymbol{b}} = \boldsymbol{r}\frac{S_y}{S_x}$, intercept = $\hat{\boldsymbol{a}} = \overline{\boldsymbol{y}} - \hat{\boldsymbol{b}}\overline{\boldsymbol{x}}$
- S_x and S_y are the standard deviations of the two variables (always > 0)
- r is the correlation co-efficient
- $\bar{x}$ and $\bar{y}$ are the averages of the x and y data
- If r > 0, $\hat{b}$ > 0 and vice versa

R^2 (R Squared)

- R^2 = fraction of variation in the values of y that is explained by the regression line
- Square root R^2 to find r (correlation) and determine positive/negative association to determine if r is positive or negative
- Measures accuracy of the linear regression prediction
- Not a perfect measure because the correlation coefficient can be ambiguous

  eg. r = 1 → R^2 = 1 regression line explains all (100%) of the variation in y (perfect correlation)

  eg. r = 0.6 or -0.6 → R^2 = 0.36 regression line explains 36% of the variation in y

  Residual Plot

  - To determine if data is following a linear trend, we look at the residual plot
  - The horizontal axis in a residual plot corresponds to the regression line in the scatterplot
    - To plot the residual plot, place the points above and below the regression line like how they are on the scatterplot
  - It is linear if the residual plot shows no pattern
    - If all points on residual plot are zero = perfect linear relation
  - It is not linear if the residual plot does show a pattern
    - Patterns include a u or n shape
    - Note: data may look linear until you plot the residuals

  Steps to determine if data is really following a linear trend:

  - Identify which is explanatory and which is response variable

- Always have a scatterplot first
- Obtain the regression line (LINEAR trendline)
- Calculate the residual at each point
- Draw a residual plot, a scatterplot of these residuals (on y-axis) against the variable of x
- Look for a random pattern around zero

Things to look out for:

- Only used for linear relationships (plot data before interpreting)
- To determine linearity, use the **RANDOM RESIDUAL PLOT** (NOT R^2)
  - Because both r and the regression equation are affected by outliers
- Be careful of **extrapolation** (predicting outside of the range of x using the equation of the line of best fit), because we don't know the trend past the last point and the line of best fit may be different with all the other data that we can't see included
- Correlation does not imply causation, even if r →1
  - There may be other variables (other than the explanatory) that are affecting correlation because they have an important effect on the variables in a study, but are not included in the study

Tests of Significance

- Tests of significance about a mean when σ is unknown
- Two main areas of statistical inference are **estimation** and **test of significance**
- The test of significance is to assess the evidence provided by data about some claim concerning a population parameter
- Is the hypothesized population value correct or not, with some probability?
- We take a random sample to test the claim about a population
- The obvious question is "does this sample give evidence for the claim?"
- The answer will be either "the sample is an evidence for the claim" or "the sample is not an evidence for the claim"

Steps

- **1. State the null and alternative hypotheses**

  #### Hypothesis Testing

  - Hypothesis testing asks whether the parameter differs from a specific null expectation, it doesn't ask by how much it differs
  - **Hypothesis testing** compares data to what we would expect to see if a specific null hypothesis were true
  - If the data differs, compared to what we would expect to see if the null hypothesis were true, then the **null hypothesis is rejected**
  - The alternative hypothesis includes every other possibility except the one stated in the null hypothesis
  - One of the hypotheses is true, the other is false
  - They are not scientific hypotheses, which are statements about the existence and possible causes of natural phenomena

    ##### Null Hypothesis (H_0)

    - The **null hypothesis** is a **specific** statement about a population parameter
    - Often, the null hypothesis is that the population parameter of interest is zero

- Generally thought of as the status quo, or no effect, no relationship, or no difference
- Usually the researcher hopes to disprove or reject the null hypothesis
- When performing a hypothesis test, we assume that the null hypothesis is true until we have sufficient evidence against it

  eg. The teller machine seller suggests that a software upgrade will not change the mean time for a customer to complete a transaction
- If the scientific hypothesis predicted a beneficial effect of medication X, but the null hypothesis, which predicted that the medication X would have no effect, was rejected, then this supports the scientific hypothesis
- The null hypothesis is the only statement being tested with the data
- If the data is consistent with the null hypothesis, then we have failed to reject it (we never accept it)
- If the data is inconsistent with the null hypothesis, we reject it and say the data supports the alternative hypothesis
- Rejecting a hypothesis by itself reveals nothing about the magnitude of the population parameter, **estimation is used to provide magnitudes**

Alternative Hypothesis (H_A)

- The **alternative hypothesis** includes all other feasible values for the population parameter besides the value stated in the null hypothesis
- It is a statement of "there is an effect" or there is a relationship, or that there is a difference
- The alternative hypothesis is often the statement that the researcher hopes is true

  eg. The teller machine seller suggests that a software upgrade will reduce the mean time for a customer to complete a transaction

Hypotheses for a Mean

- Null hypothesis:
  - $H_0 : \mu = \mu_0$, where μ_0 is an actual value and is called the null value
- Alternative hypothesis:
  - Two-sided H_a: $\mu \neq \mu_0$ (there is a difference, or it has changed, in either direction, increase or decrease)
  - One-sided
    - H_a: $\mu < \mu_0$ (it has decreased, or it is now less)
    - H_a: $\mu > \mu_0$ (it has increased, or it is now more)

  eg. Does the sample support the claim that the software upgrade reduces the mean transaction time?

  Let μ be the mean transaction time in population

  Alternative hypothesis: statement we want to prove H_a: $\mu < 270$ (mean transaction time is reduced)

  Null hypothesis: statement of current status H_o: $\mu=270$ (there's no reduction) We call 270 the null value and we assume the null hypothesis is true until otherwise proven.

- **2. (a) Check conditions and then (b) Calculate the test statistic**

  Conditions:

  - The value of μ is unknown whereas the value of the population standard deviation σ is known

- Must be a random sample from the population of interest
- The population distribution of a variable of interest has a Normal distribution N(μ, σ) OR, if the population distribution is non normal, n is large enough (eg n>30) (central limit theorem)

Calculating Test Statistic

- We assume the null hypothesis is true until the sample data conclusively demonstrates otherwise
- If the null hypothesis is true about the population, what is the probability of observing sample data like that observed? → P-value
- How extreme the sample value is from the hypothesized value (null value), if it were true
  - How many standard deviations is the sample value from the null value?

$$\textbf{test statistic} = \frac{\textbf{sample value} - \textbf{null value}}{\textbf{standard deviation of sampling distribution}}$$

- Test statistic for testing a single mean when population SD,σ, is known:

$$\mathbf{z - statistic} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- When H_0 is true, the test statistic z has a standard normal distribution
  eg. How far is the sample mean (258 seconds) from the hypothesized mean (270 seconds)? → find the z-statistic
  The sampling distribution of sample mean is Normal because the population is normal
  The standard deviation of sampling distribution of sample mean = $\frac{24}{\sqrt{38}} = 3.89$
  The mean of sampling distribution of sample mean = 270
  So the z-statistic of $\bar{x}$ = 258 is: $\frac{258-270}{3.89} = -3.08$
  The sample mean $\bar{x}$ = 258 falls **below** 3 standard deviation of the sampling distribution from the null value of 270 → very extreme sample mean

- **3. Find the P-value using the appropriate distribution**
  - **P-value** is the probability of observing a test statistic as extreme as what we did, or something more extreme, when the null hypothesis is true
  - P-value measures the strength of the sample evidence
  - The smaller p-value, the stronger sample evidence
    eg. P-value is the probability of getting z as extreme as -3.08, or more extreme than -3.08
    = area under the z-curve to the left of -3.08
    Find z = -3.08 on the z-table to find the probability
    P-value = only 0.001 (Very low)
    If H_0 is true, there is only a 0.001 (0.1%) chance of this sample result or more extreme.
    An observed value this small would rarely occur just by chance if the mean transaction time was 270
    The mean transaction time after the upgrade is not 270, but maybe somewhere near our sample value

- **4. Make a decision and state your conclusion in the context of the specific setting of the test**
    - Choose the level of significance (α)
    - If **p-value ≤ α** → **Reject** the null hypothesis
        - The sample result is **statistically significant** and we have convincing evidence that H_a is true
    - If **p-value > α** → **Fail to reject** the null hypothesis
        - The sample result is **not statistically significant** and we do not have convincing evidence that H_a is true
    - Common significance levels are $\alpha=0.05$, $\alpha=0.10$, or $\alpha=0.01$
    - If a significance level is not specified, use **$\alpha=0.05$**

      eg.

      Decision: As the p-value is less than 0.05, we reject the null hypothesis

      Conclusion: The sample is an evidence at 5% level of significance that the software upgrade reduces the mean transaction time

      eg. For test statistic $z = 3.23$ and $H_a : \mu > 0$, P-value = $P(Z > 3.23) = 1 – 0.9994 = 0.0006$

      → If H_0 (no loss) is true, there is only a 0.0006 (0.06%) chance of this sample result or more extreme

      → The sample was unlikely to occur under H_0

      → This sample is evidence against the null hypothesis, H_0, and is in favour of the alternative hypothesis, H_a

  Interpreting the Level of Significance

    - If we choose $\alpha = 0.05$, we are requiring that the data give evidence against H_0 so strong that it would occur no more than 5% of the time when H_0 is true
    - If we choose $\alpha = 0.01$, we are insisting stronger evidence against H_0, evidence so strong that it would occur only 1% of the time when H_0 is true
    - If we have a smaller level of significance, we require stronger evidence against the null hypothesis
    - So, smaller α you choose
    - → smaller p-value is required to reject the null hypothesis
    - → more extreme sample mean is required to reject H_0
    - → you are insisting stronger evidence to reject H_0