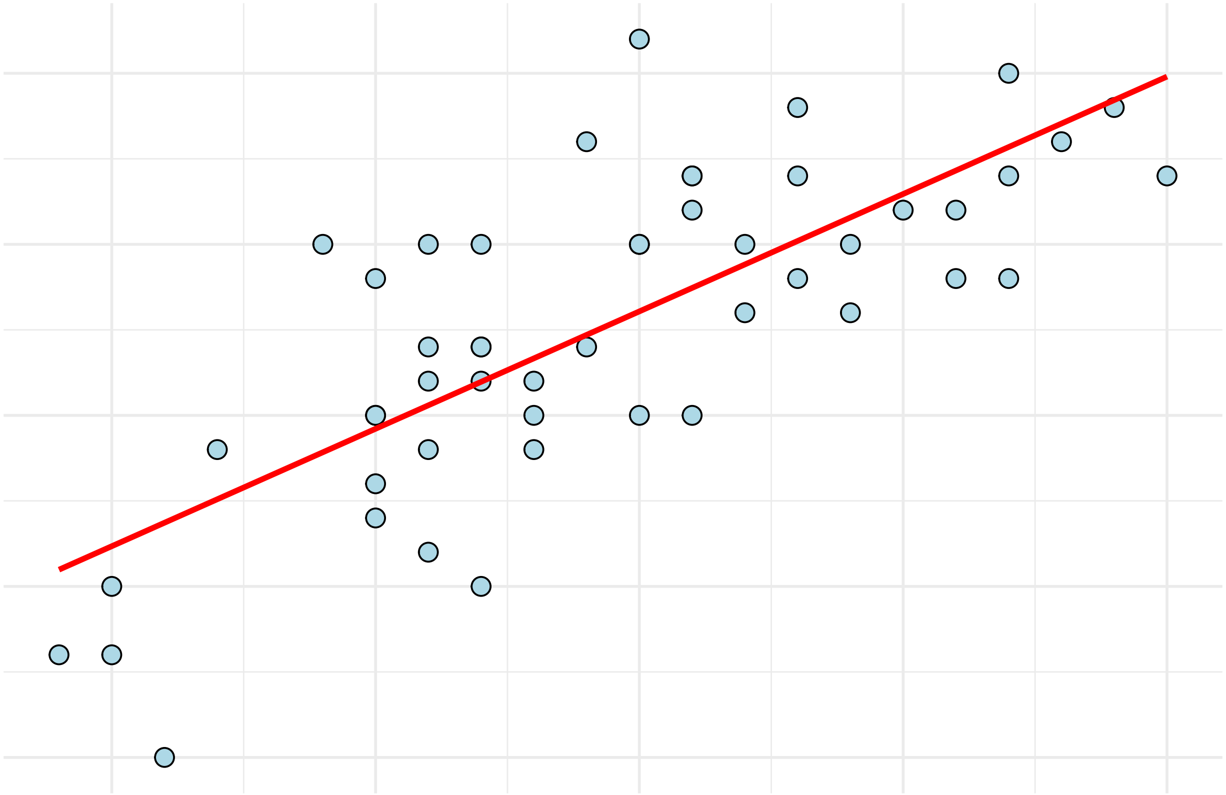
Linear Regression

What is Linear Regression

Linear regression is a statistical technique used to model and quantify the relationship between a continuous outcome variable (dependent continuous variable) and one or more predictor variables (independent continuous variables).

Unlike correlation, which describes the strength and direction of a relationship, linear regression allows us to predict values, estimate how much one variable changes with another, and test whether that change is statistically meaningful.

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



Can we predict our Y variable with our X variable?

When to use linear regression?

When you are in a situation where you have two **continuous** variables and you think the value of one can be predicted by the value of the other.

For example; using height to predict body weight, predicting DMFT score from fluoride exposure or saliva pH from food consumption.

Assumptions of linear regression

The assumptions of a linear regression that must be met are as follows:

1) Linearity
2) Independence of observations
3) Equal variance as one value increases (Homoscedastic)
4) Normality of residuals (residuals should follow a normal distribution)
5) No multicolinnearity when doing multiple regression (predictor variables can't be correlated)

Different types of linear regression

Type of test: Simple linear regression

Example: Can plaque index be used to predict probing depth?

Type of test: Multiple linear regression

Example: How do both sugar consumption and brushing habits together influence caries development?

How to implement in RStudio

```r
# Simple linear regression
model <- lm(outcome ~ predictor, data = your_data)
summary(model)

# Multiple linear regression
model <- lm(outcome ~ predictor1 + predictor2 + predictor3, data = your_data)
summary(model)

# Check diagnostics of the model (important for the assumptions)
plot(model)
```

Interpreting results

After fitting a regression model, pay attention to:

The p-value

The p-value tells whether a predictor has a statistically significant association with the outcome after controlling for other variables (if multiple linear regression). So a low p-value means that it is very unlikely to see your data when there is no real difference. A threshold of 0.05 is commonly used in research and this means there is a 5% chance the results you observe are due to chance and not another reason.

Coefficients (Gradient and Intercept)

In linear regression, the effect size is represented by the regression coefficients rather than a correlation coefficient. The intercept is the outcome value when predictor values are zero. The gradient is the increase in the outcome variable for every 1-unit increase in predictor variable (If 0.7 then for every 1-unit increase in plaque index the probing depth increases by 0.7 mm)

R-squared

Represents how much of the variability in the outcome is explained by the model. Generally speaking; 0.1 = small effect, 0.3 = moderate effect, 0.5+ = strong effect.