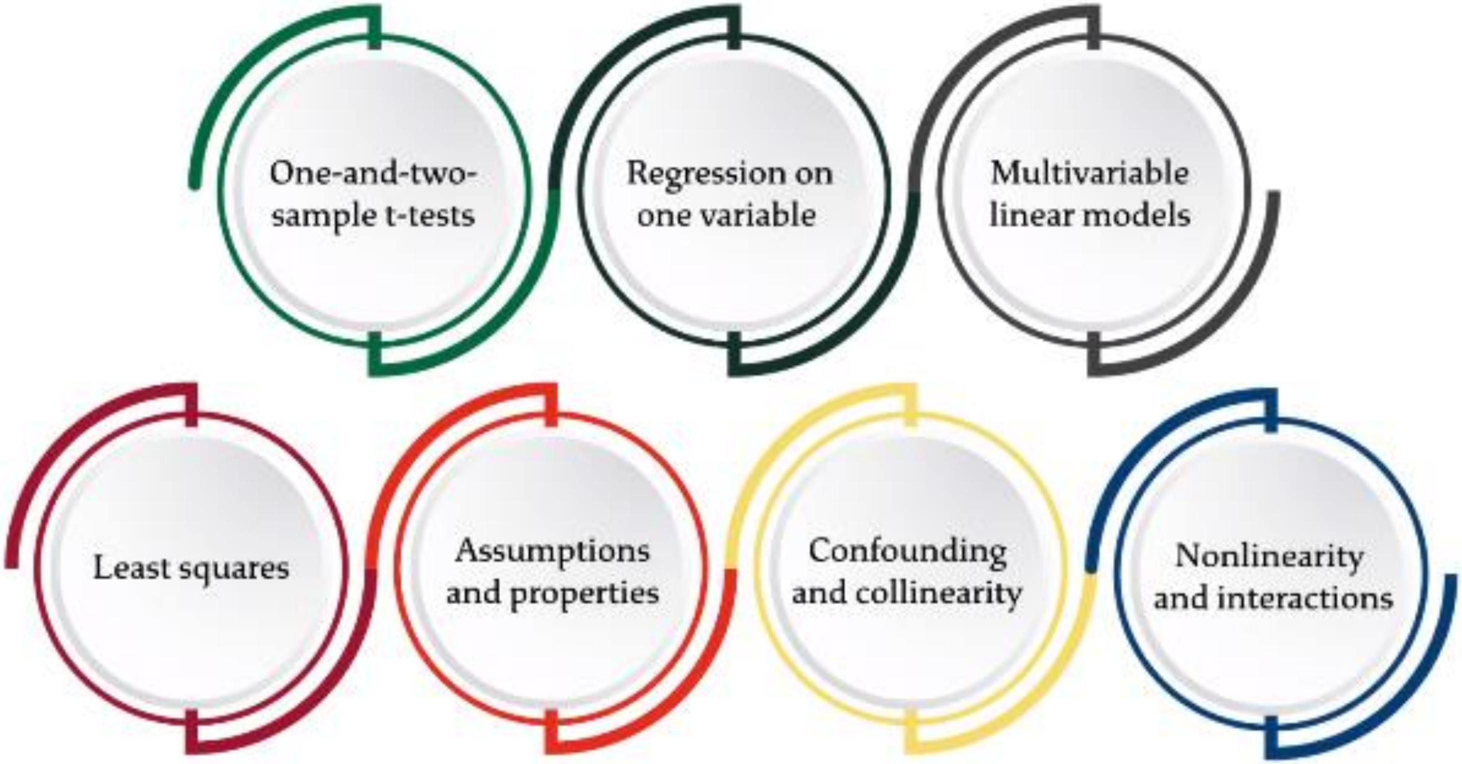
Module 4:
Linear Regression

Introduction to Linear Models



Dartmouth GEISEL SCHOOL OF MEDICINE | QBS Program in Quantitative Biomedical Sciences

What is Statistic?

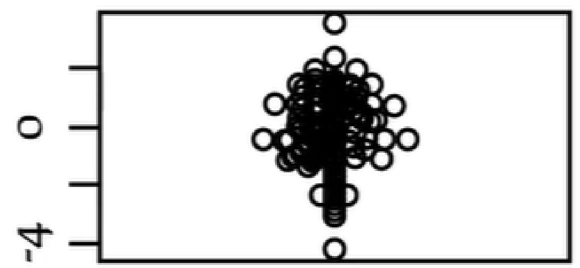
Statistics is the study of random variables, such as:

- Weight
- Body mass index (BMI)
- Eyesight

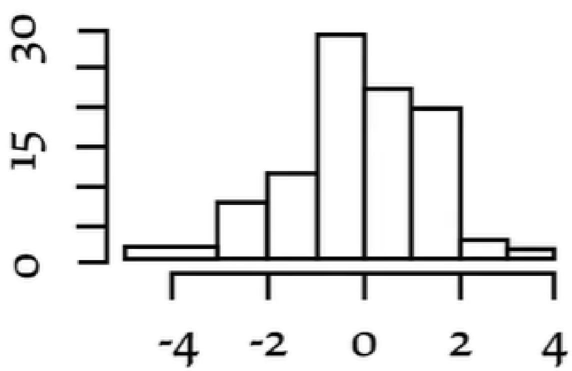
Dartmouth GEISEL SCHOOL OF MEDICINE | QBS Program in Quantitative Biomedical Sciences

Visualizing a Distribution

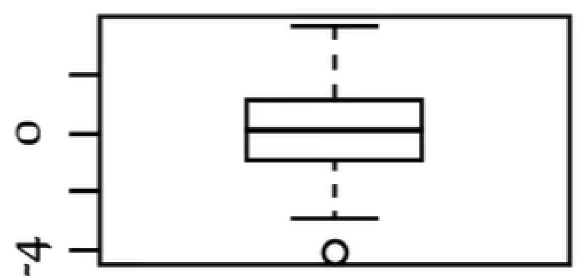
Violin plot



Histogram



Boxplot



Dartmouth
GEISEL SCHOOL OF MEDICINE
QBS
Program in Quantitative Biomedical Sciences

Simple Statistical Model

Where,

Y = Random variable

μ = Average or central tendency

ε = Error or distribution around the central tendency

$E[Y] = \mu$

OR

$Y = \mu + \varepsilon$

?

What can you do with this simple model?

It can be used to test if μ is equal to a particular value, such as $\mu = 0$

Dartmouth GEISEL SCHOOL OF MEDICINE | QBS Program in Quantitative Biomedical Sciences

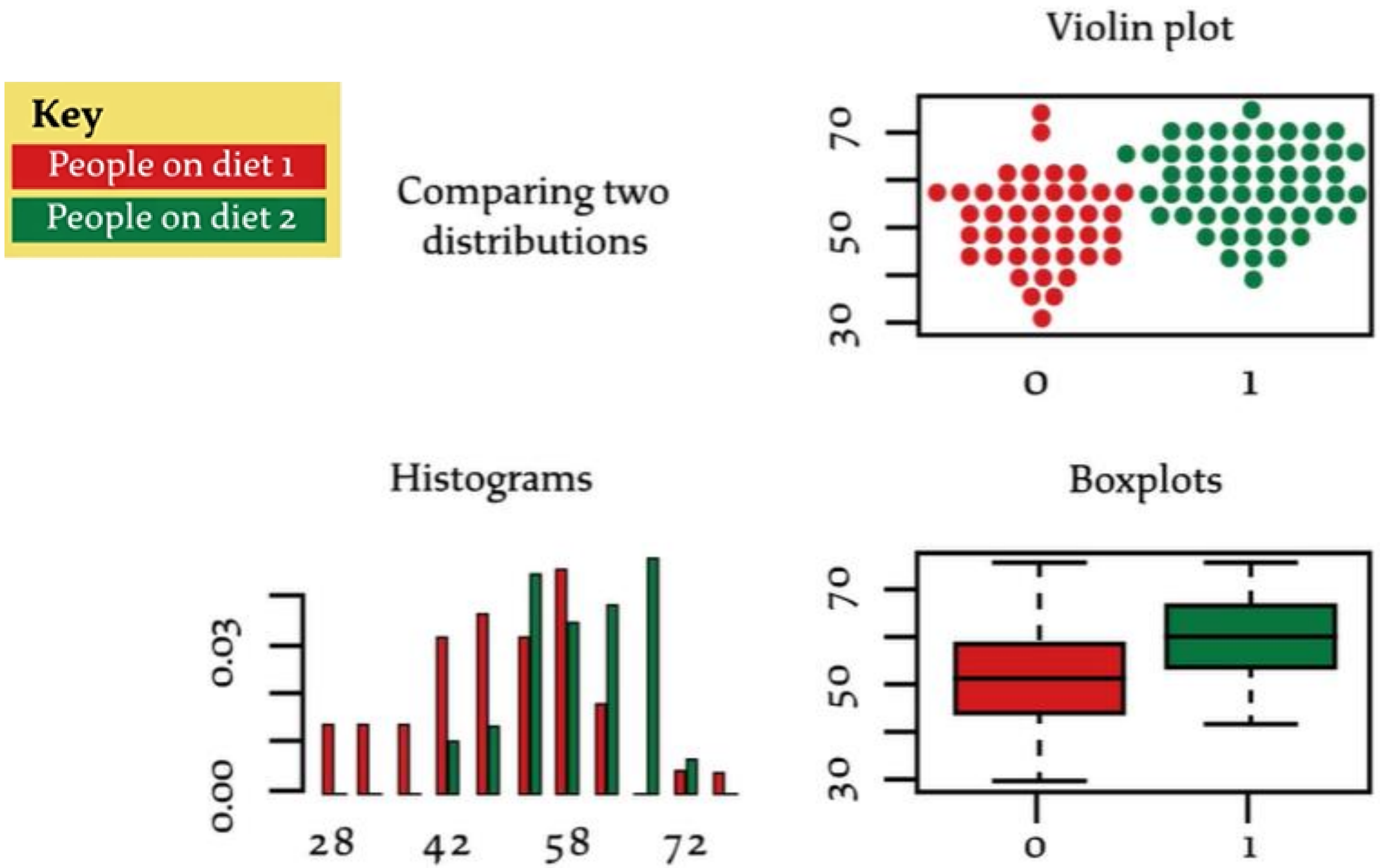
Paired or One-Sample t-Test

Assumption: Observations are independent

$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma/\sqrt{n})$$

- In large samples, the z-test and t-test are identical
- In small samples, the t-test is preferred

Paired or One-Sample t-Test



Dartmouth GEISEL SCHOOL OF MEDICINE | QBS Program in Quantitative Biomedical Sciences

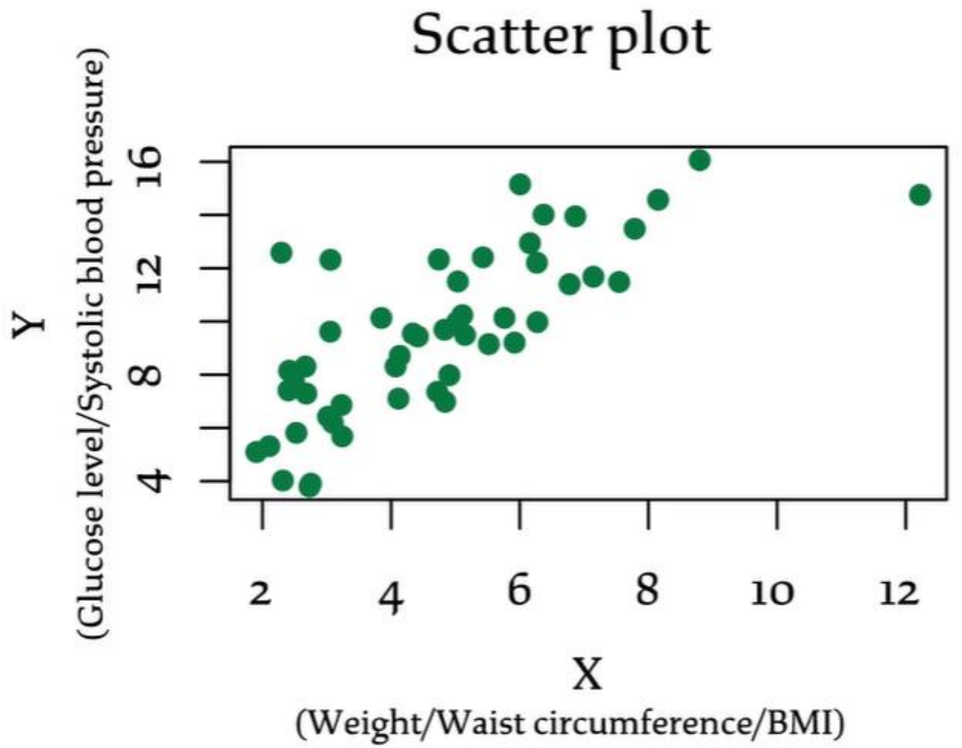
Two-Sample t-Test

$$\frac{\frac{1}{n_1}\sum_{i=1}^{n} X_{1i} - \frac{1}{n_0}\sum_{i=1}^{n} X_{0i}}{\sqrt{\hat{\sigma}_0^2/n_0 + \hat{\sigma}_1^2/n_1}}$$

Here are the observations of a two-sample t-test:

- If the value is close to zero, the null hypothesis is true
- If the value is far away from zero, the null hypothesis should be rejected

Model Relating Two Variables



Linear model is the simplest way to relate two variables

Expressing a Linear Model

How does the average of Y depend on X?

$$E[Y|X] = a + bX$$

$$Y = a + bX + \epsilon$$

Where,

E = Average
Y = Dependent variable
X = Independent variable
b = Slope/Coefficient

Estimation by Least squares

Using the least squares approach to estimate the slope (b)

Where,

b = Parameter/Coefficient

y = Dependent variable

x = Independent variable

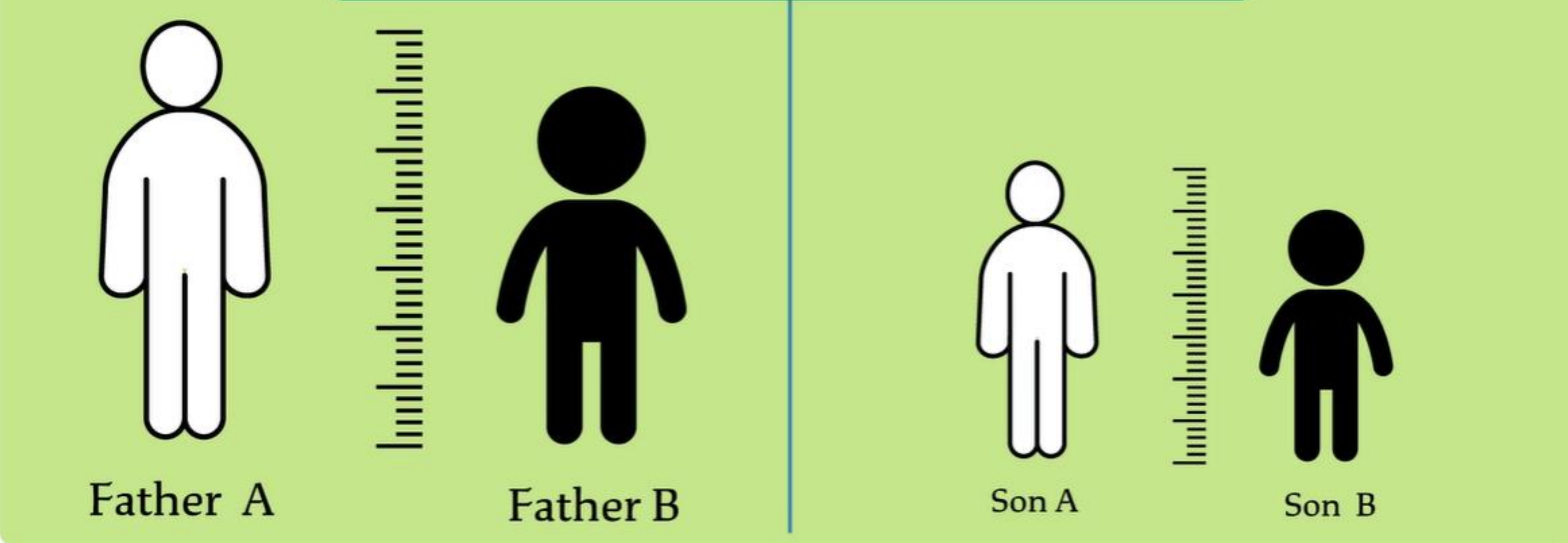
$$\sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

Minimize the sum of the squared distances between the observed values

The least squares approach is always preferred over the least absolute deviations approach

Galton's Application of Least Squares

Observation: By applying the least squares approach, it was found that there was no strong relationship between the heights of the two different age groups



Regression to the mean, an important concept in epidemiology and biostatistics, means that things come back to average

Estimation by Least Squares

Where,

b = Parameter/Coefficient/Slope

Y = Dependent variable

X = Independent variable

$$\sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

$$= \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Taking a derivative

$$\hat{b} = \frac{1}{\hat{\sigma}_x^2} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})/n$$

$$= \frac{1}{\hat{\sigma}_x^2} \sum_{i=1}^{n} (x_i - \bar{x}) y_i / n$$

Parameter $\hat{a} = \bar{y} - \hat{b}\bar{x}$ (of no use)

Standard Error of the Estimate

Where,

$\hat{\sigma}$ = Variation of the dependent variable

$\hat{\sigma}_x$ = Variation of the independent variable

n = Sample size

$$\sigma(\hat{b}) = \frac{\hat{\sigma}}{n\hat{\sigma}_x}$$

Dartmouth GEISEL SCHOOL OF MEDICINE | QBS Program in Quantitative Biomedical Sciences

Key Points in Statistic

- Sample size has an inverse relationship to the standard error
- Confidence intervals are the intervals around the standard error or margin of error
- The value of confidence intervals is usually 95%

Equivalence of Regression

Pearson correlation

Pearson correlation, a statistical concept, is a number ranging between -1 and +1, depicting correlation

-1 = Perfect negative correlation

+1 = Perfect positive correlation

Two-sample t-tests

- The equivalence of regression involves a dependent variable, such as glucose value regressed on a binary variable

Multivariable Model

Where, Y = Dependent variable (mortality rate) X = Independent variable (government spend on healthcare/welfare)	$E[Y \mid X_1, \ldots, X_k] = b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k$

A linear model is the simplest way to connect one or more variables to a dependent variable

Design Matrix

Healthcare spend (per person)	Welfare spend (per person)	Spend on policy	Average income

A design matrix is an abstract yet efficient model

Testing

- Key concept in statistics
- Often referred to as the Wald test
- Uses slope estimations and divides it by the standard error calculation

$$\hat{b}_j^2 / \sigma^2(\hat{b})_{jj}$$

Wald Test

Versions of the Wald test

- Z-test
- F-test/t-test

Wald test for multiple parameters:

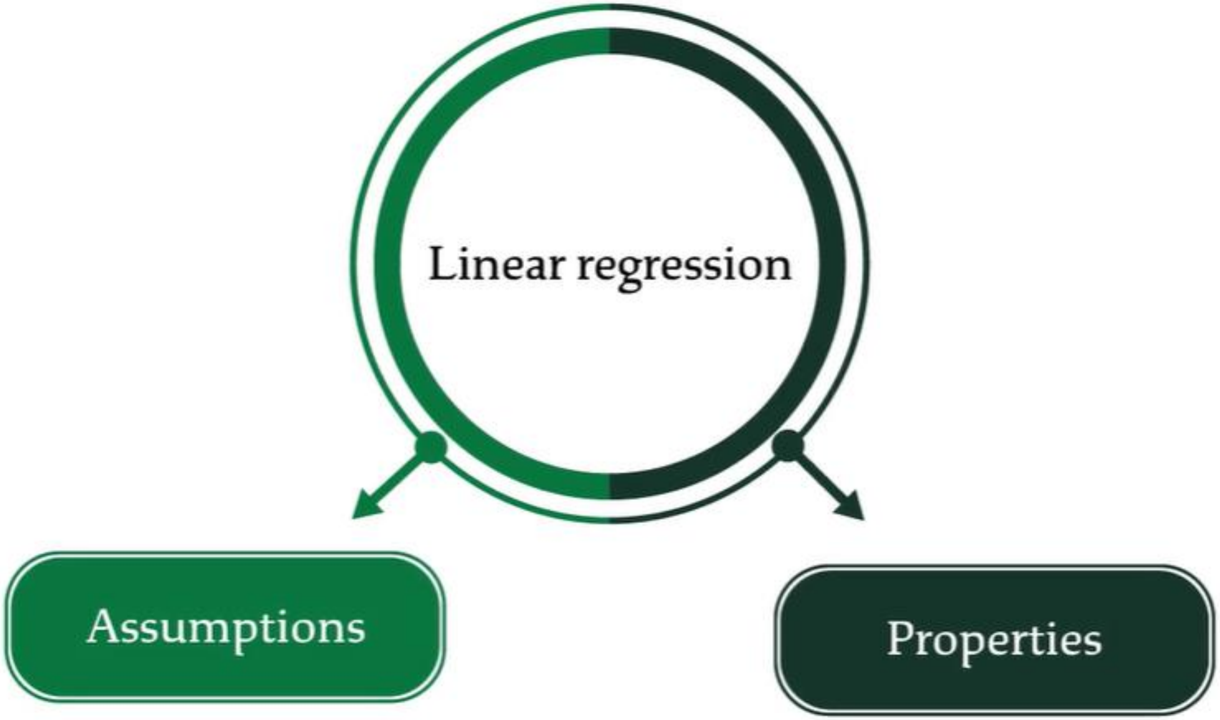
$$(L\hat{b})^T L^t [\mathrm{Cov}(\hat{b})L]^{-1}(L\hat{b})$$

For large samples, both can be used

For small samples, t-test is preferred

Dartmouth GEISEL SCHOOL OF MEDICINE | QBS Program in Quantitative Biomedical Sciences

Properties and Assumptions of Linear regression



Dartmouth GEISEL SCHOOL OF MEDICINE | QBS Program in Quantitative Biomedical Sciences

Properties and Assumptions of Linear regression

Assumptions

- Model is right
- Independence
- Exogeneity
- Homogeneity of variance
- Normality/ Large sample

Dartmouth GEISEL SCHOOL OF MEDICINE | QBS Program in Quantitative Biomedical Sciences

Properties and Assumptions of Linear regression

Assumptions

- The true phenomenon is approximately linear
- Approximated models are easier to interpret
- Complicated models are harder to interpret

- The errors are independent of each other
- The lack of independence creates problems when come up with standard errors
- Violating independence assumption results in wrong standard errors

- Also known as "confounding" in statistics and epidemiology
- The error term factor in what cannot be explained with independent variables

- The variation does not change from observation to observation
- For smaller data, the precision of dependent variables across observations lack homogeneity

- Depicted through a bell-shaped distribution, also referred to as Gaussian
- Applied to small samples and not suitable for large sample sizes
- For a large sample, the central limit theorem is used

Properties and Assumptions of Linear regression

Properties

- Unbiased estimators
- Unbiased confidence intervals
- Unbiased tests
- Efficiency

Properties and Assumptions of Linear regression

Properties

- Estimators are on target if all assumptions are met
- The derived confidence intervals will capture the true value on target

- The probability that the 95% confidence interval contains the true value 95% of the time
- Unless the assumptions are met, the unbiasedness of the estimators and 95% confidence interval may not hold true

- When creating a test, the usual cut off of the p-value is 0.05
- If assumptions are met, 5% of the time the p-value would be wrong
- If assumptions are not met, more than 5% of the time the p-value would be wrong

Categorical variables as Indicators

Dependent variable

Y

Systolic blood pressure

Categorical variable

X

Race
- Whites
- Blacks
- Asians
- Native Americans/Pacific Islanders

Create indicator variables to capture the information in a categorical variable

To perform linear regression and least squares, a linear model creates an indicator variable for each group except the referent group

Dartmouth GEISEL SCHOOL OF MEDICINE | QBS Program in Quantitative Biomedical Sciences

Categorical Variables as Indicators

Dependent variable

Y

Systolic blood pressure

Categorical variable

X

Race
- Whites
- Blacks
- Asians
- Native Americans/Pacific Islanders

For each indicator variable, R creates a variable with a value of 1 or 0 in the background

K(variable) = 0 or 1

Person is not Black

Person is Black

In a linear model, indicator variables split up the categorical variables when considering the referent group

ANOVA Features

Generalization of two-sample t-tests to more than two samples

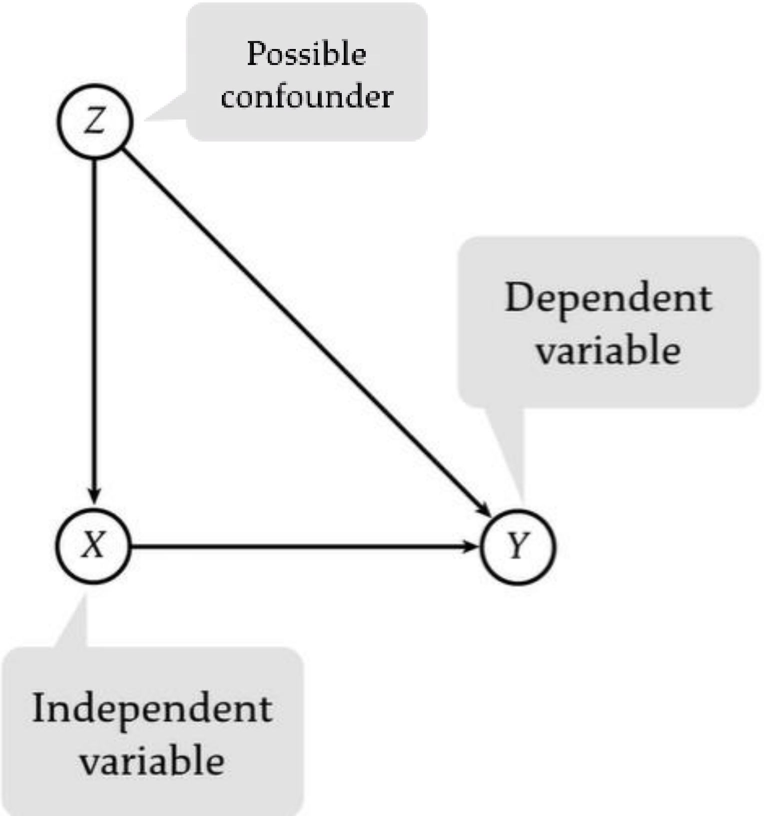
Can be replaced with the LM command in R

Provides a single p-value as opposed to doing multiple two-sample t-tests

Analysis of Covariance

- Is a generalization of ANOVA
- Requires one more covariant to be added
- Is a linear model with a single categorical variable and a single continuous variable

Directed Acyclic Graphs (DAG)



D—Directed = Arrows go in **one direction**

A—Acyclic = **No loops**

This DAG helps to understand confounding

Key question: How does X affect Y?

A confounder is a variable that affects both the exposure of interest X and the dependent variable Y

Cofounded Results

- Refer to the output of the related study
- Have significant limitations when interpreted
- Are difficult to share with other researchers and media

Dartmouth GEISEL SCHOOL OF MEDICINE | QBS Program in Quantitative Biomedical Sciences

Collinear Variables: Example

	Temperature
Collinear	Celsius
	Fahrenheit

- Variables are said to be collinear if they are highly correlated
- Collinearity provides the third variable given the other two

Collinearity

It is a mathematical problem

It can be epidemiological or statistical problem

Dartmouth GEISEL SCHOOL OF MEDICINE | QBS Program in Quantitative Biomedical Sciences

Collinearity

To be modeled: Impact of weight on health outcomes

To be measured: Whether people will have a stroke in the next ten years

Data set (as independent variables):

- Body mass index (BMI)
- Weight
- Circumference

Highly correlated variables

Collinearity

To be modeled: Impact of education on health outcomes

To be measured: Whether people will have a stroke in the next ten years

Data set:

- Years of education
- Highest degree

Coefficients of these will be difficult to interpret

In order to interpret a variable in a multi-variable model, all other variables are held constant

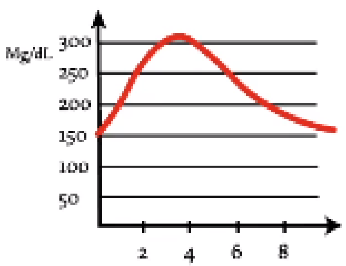
Interpreting Coefficients from a Multi-Variable Model

- Coefficient corresponds to how much you can expect a dependent variable to change if you change this independent variable by one unit, keeping all the other variables constant
- Interpretations are difficult with highly collinear variables

Nonlinearity

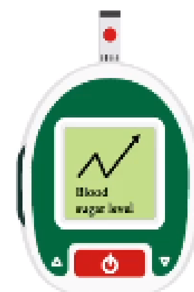
Nonlinearity refers to creating an interaction between two variables.

Does being female affect how your glucose control today affects your glucose control three months from now?



Glucose levels at the baseline

Predict



Glucose levels after three months

Gender (male/female)

Nonlinear Effects

Where,

Y = Dependent variable

a_0 = Intercept term

X_1 = First variable

X_2 = Second variable

Interaction term

$$Y = a_0 + a_1X_1 + a_2X_2 + \boxed{a_{12}X_1X_2} + \epsilon$$

In a linear model, a nonlinear effect can also be accommodated

Interactions

Variable/Term	To measure	Interaction
Gender/Sex	Glucose control at baseline Glucose control at follow up	Does gender/sex modify the effect of glucose control?
Race		Does race interact with glucose control?

Effect Modification

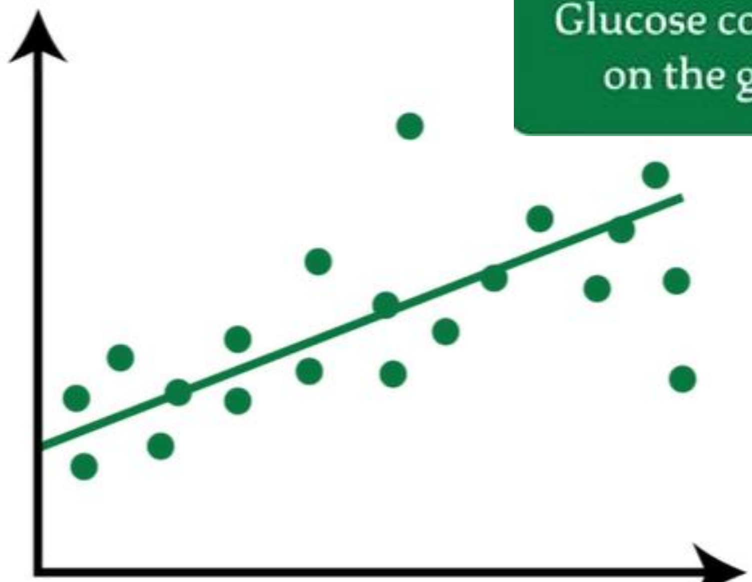
01. Calculates whether a categorical variable/continuous variable modifies the effect of another variable

02. Interprets interactions

Dartmouth GEISEL SCHOOL OF MEDICINE | QBS Program in Quantitative Biomedical Sciences

Nonlinear Effects

Glucose control at baseline has an effect on the glucose control at follow up



$$Y = a_0 + a_1 X_1 + a_2 \boxed{X_1^2} + \epsilon$$

Coefficient ≠ Non-zero or significantly different from non-zero = Quadratic effect/Nonlinearity

Dartmouth GEISEL SCHOOL OF MEDICINE | QBS Program in Quantitative Biomedical Sciences

Transformation: Log Transform

Dependent variable has zero (0) → Cannot take a log, as it has negative infinity

Dependent variable doesn't have zero (0) → Consider taking a log

When a log is transformed, the interpretation of coefficients comes down to a percentage scale

Dartmouth
GEISEL SCHOOL OF MEDICINE

QBS
Program in Quantitative Biomedical Sciences