Linear models

Katie Pollard

BMI 206

In this unit we will learn…

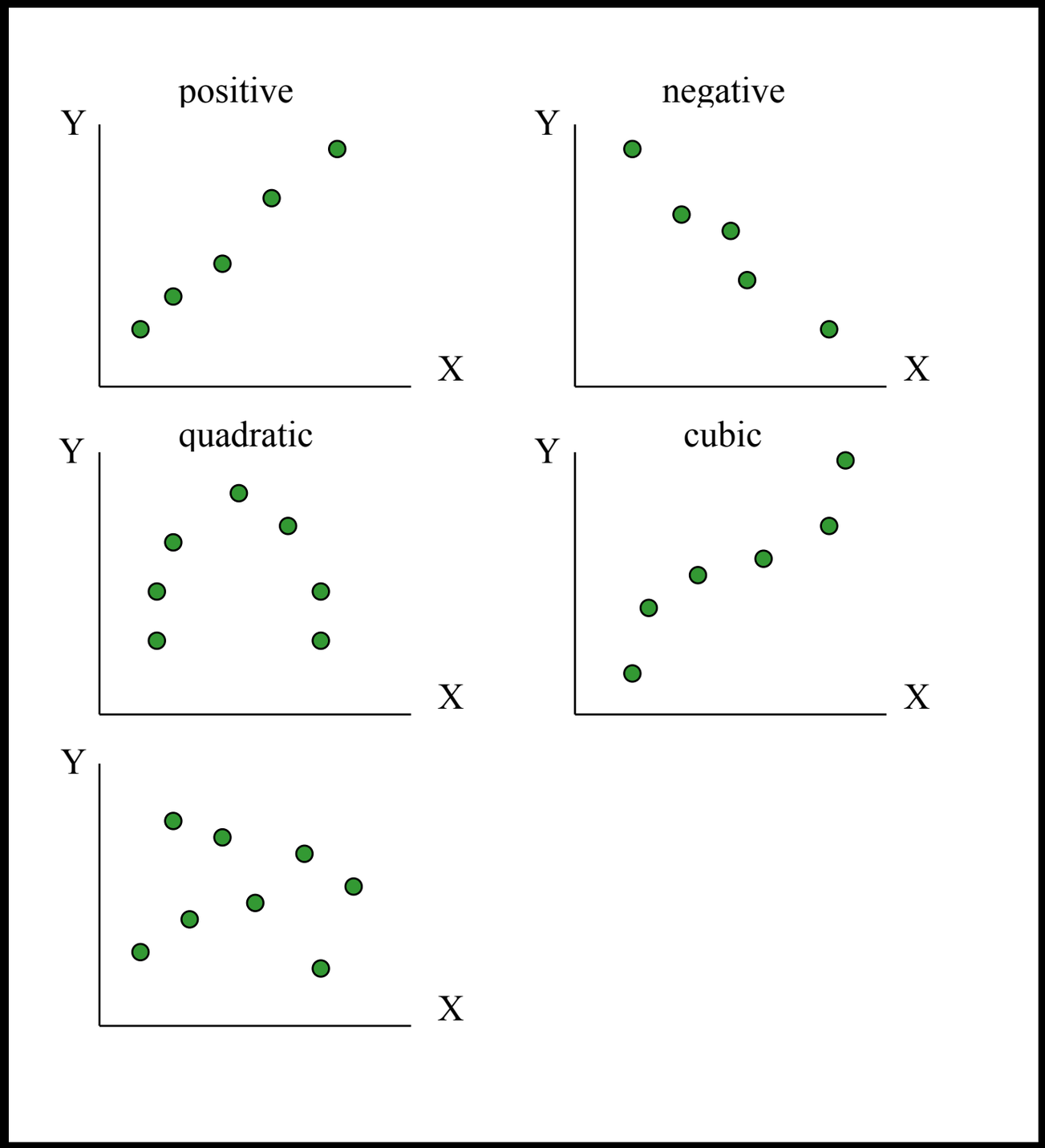
- The meaning of a linear relationship
- How to formulate linear models (LMs) with outcomes that are Normally distributed
- Least squares regression to fit LMs
- The main components of LMs
- Interpretation of parameters in LMs
- Formulation and interpretation of parameters in multiple regression LMs

Relating Different Data Types

Covariate (independent variable)

Outcome (dependent variable)	Continuous or Both	Categorical
Continuous	Linear Regression / ANCOVA	ANOVA
Categorical	Generalized Linear Model Regression	Contingency Tables / Log-linear Model Regression

Relating Continuous Variables

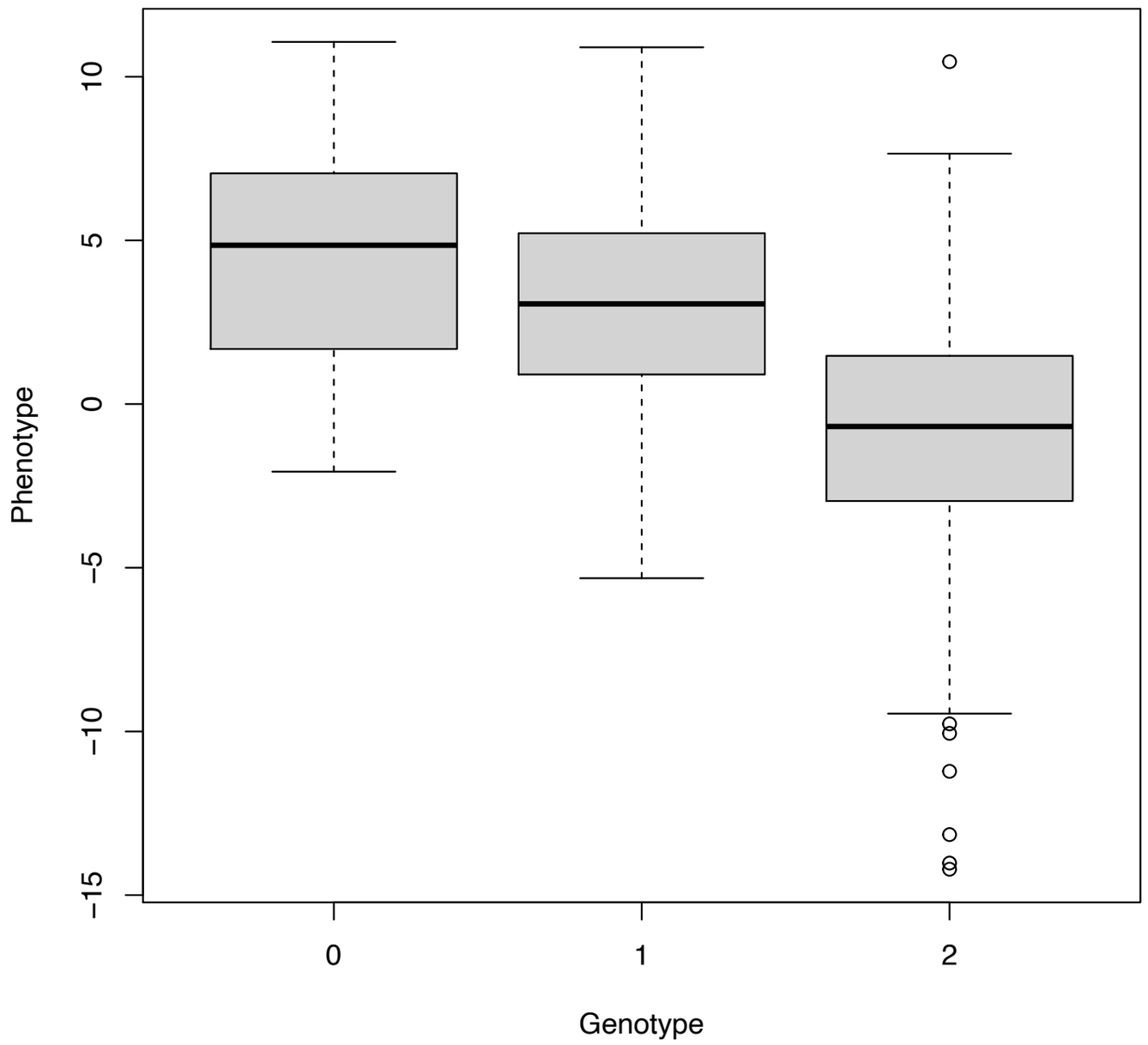


Linear relationship

Non-linear relationship

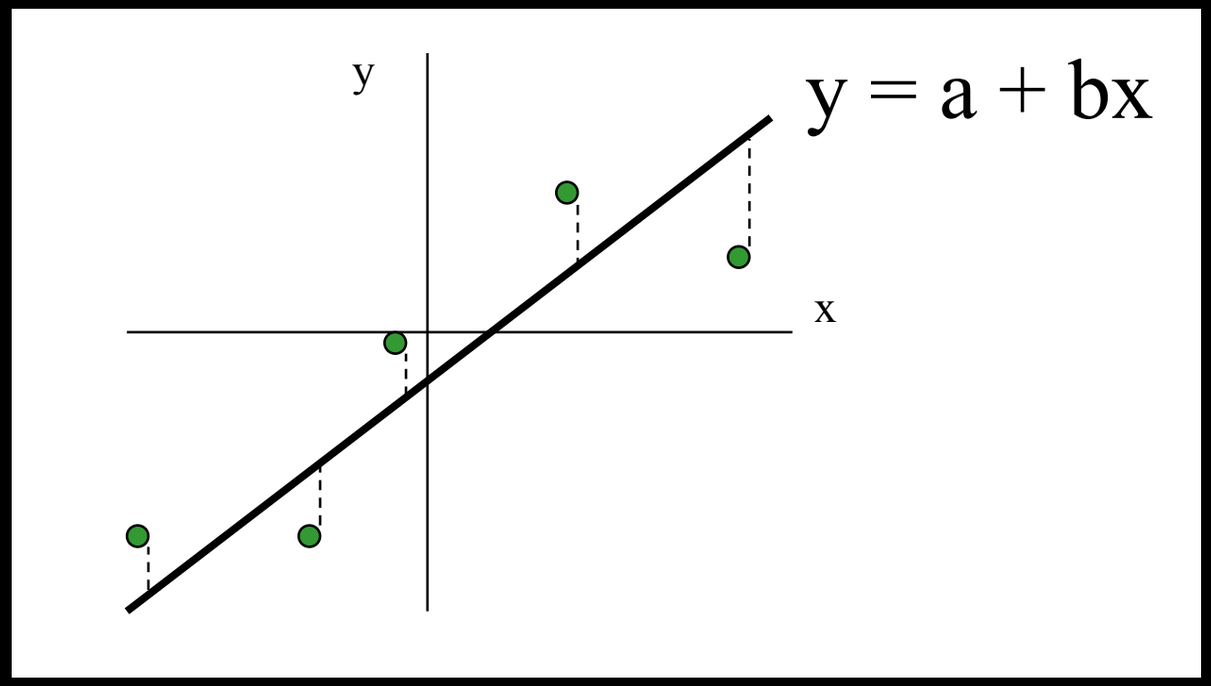
No obvious relationship

Relating Continuous and Categorical Variables



Different means

Linear Model



a is the intercept
b is the slope

Seek the line that minimizes sum of squared residuals.

- Substituting estimates of (a,b) provides a prediction of the outcome (y) for any x value. This is the line.
- Residual is observed minus predicted value for each observed x. This is the vertical distance to the line.

Least Squares Regression

The solution to the least squares problem is:

$$a = \bar{y} - b\bar{x} \qquad b = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

- What does the intercept (a) represent?
- What is the meaning of the slope (b)?

Assumptions:

- The linear model assumes the outcome (y) is normally distributed.
- This means the residuals are normally distributed around the least squares line (mean zero, same variance as y).
- The covariate x can have any distribution, even discrete.

Multiple Regression

$$E[Y] = a + b\,X + c\,W + d\,X^{*}W$$

- One outcome variable (Y), 2+ covariates (X and W)
  - Covariates can be continuous or categorical
  - May include powers or other transformations of covariates
  - May include interactions between covariates (X*W)
- Coefficient estimates and their standard errors can be used to test for association with Y. This is called a Wald test.
- Predicted values can be computed from the fitted model for different covariate combinations.

Conditional Associations in Multiple Regression Models

$$E[Y] = a + b\,X + c\,W + d\,X^{*}W$$

- Coefficients represent expected change in Y per unit increase in that covariate, while holding the other covariates constant (i.e., adjusted for them).
- X and Y can be conditionally associated, but marginally independent.
- X and Y can be conditionally independent, but marginally associated…. and other combinations.
- Can we still think about a line in multiple regression?