General linear models:

Continuous response + discrete or continuous covariates + normally distributed error

Outline

- ANCOVA
  - ANOVA + regression
  - Interactions with continuous predictors
- General linear models
  - Linking together linear regression, t-tests, ANOVAs, and ANCOVAs
- Degrees of freedom

ANCOVA

- ANCOVA = Analysis of Covariance

- ANCOVA is really "ANOVA with covariates" or, more simply, a combination of ANOVA and regression

- When do you use this model?
  - You have some categorical factors and some quantitative (continuous) predictors.
  - Continuous variables are typically referred to as *covariates*

ANCOVA

- Historically: the idea is that concomitant variables are not necessarily of primary interest, but their inclusion in the model will explain more of the response, and hence reduce the error variance.

- For example, want to know which of three types of cancer treatments let's patients live longest??
  - Must also include the stage of cancer as a covariate for an accurate assessment (as stage of cancer directly correlates with time to death)

- In some situations, failure to include an important covariate can yield misleading results.

ANCOVA

- Modern statistical EEBB: we may be interested in both the difference in group means AND the effects of a particular covariate
- For example, want to know if population density varies for amphibians by state?
  - Need to include the amount of wetland habitat available in each state.
- Thus the continuous covariate could be a nuisance parameter or it could be an important quantity of interest.

Standard ANCOVA

- What is the effect of region (North, Central, South) on mass?
- However, we know that elevation can play a key role…

$$mass_i = \alpha + \beta_{j(i)} * region_i + \delta * elevation_i + \varepsilon_i$$

$$\varepsilon_i \sim Norm(0, \sigma^2)$$

- Snake mass is composed of three components:
  1. A constant (alpha)
  2. The product of another constant (beta) with the value of the indicator for region where the individual was caught
  3. The product of a constant (gamma) with the elevation where the individual was caught
  4. The error (ε_i) that is specific to snake

Standard ANCOVA

Question: Is mass different in each region? How does elevation influence mass?

Example with 10 data points:

Individual	Location	Elevation	Mass
1	North	54	6
2	North	25	8
3	North	75	5
4	North	44	7
5	Central	99	5
6	Central	96	4
7	Central	85	6
8	South	39	8
9	South	32	8
10	South	10	9

Standard ANCOVA

Translates into a set of equations:

$$6 = \alpha * 1 + \beta_1 * 0 + \beta_2 * 0 + \delta * 54 + \varepsilon_1$$
$$8 = \alpha * 1 + \beta_1 * 0 + \beta_2 * 0 + \delta * 25 + \varepsilon_2$$
$$5 = \alpha * 1 + \beta_1 * 0 + \beta_2 * 0 + \delta * 75 + \varepsilon_3$$
$$7 = \alpha * 1 + \beta_1 * 0 + \beta_2 * 0 + \delta * 44 + \varepsilon_4$$
$$5 = \alpha * 1 + \beta_1 * 1 + \beta_2 * 0 + \delta * 99 + \varepsilon_5$$
$$4 = \alpha * 1 + \beta_1 * 1 + \beta_2 * 0 + \delta * 96 + \varepsilon_6$$
$$6 = \alpha * 1 + \beta_1 * 1 + \beta_2 * 0 + \delta * 85 + \varepsilon_7$$
$$8 = \alpha * 1 + \beta_1 * 0 + \beta_2 * 1 + \delta * 39 + \varepsilon_8$$
$$8 = \alpha * 1 + \beta_1 * 0 + \beta_2 * 1 + \delta * 32 + \varepsilon_9$$
$$9 = \alpha * 1 + \beta_1 * 0 + \beta_2 * 1 + \delta * 10 + \varepsilon_{10}$$

Or in matrix notation:

$$\begin{pmatrix} 6 \\ 8 \\ 5 \\ 7 \\ 5 \\ 4 \\ 6 \\ 8 \\ 8 \\ 9 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 54 \\ 1 & 0 & 0 & 25 \\ 1 & 0 & 0 & 75 \\ 1 & 0 & 0 & 44 \\ 1 & 1 & 0 & 99 \\ 1 & 1 & 0 & 96 \\ 1 & 1 & 0 & 85 \\ 1 & 0 & 1 & 39 \\ 1 & 0 & 1 & 32 \\ 1 & 0 & 1 & 10 \end{pmatrix} * \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \delta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \end{pmatrix}$$

$$\varepsilon_i \sim Norm(0, \sigma^2)$$

ANCOVA

Assumptions

- The continuous covariate is not related to the treatment variables (factors).
- The covariate is linearly related to the response and the relationship will be the same for all levels of the factor (no interaction between covariate and factor).
  - Of course, this assumption can be modified if we need to account for such an interaction.

ANCOVA with interactions

What is an interaction?

- Talked about this last time with discrete variables -> different mean values in the dependent variable that change with each factor combination (e.g., mean snake mass in each region/habitat combination)

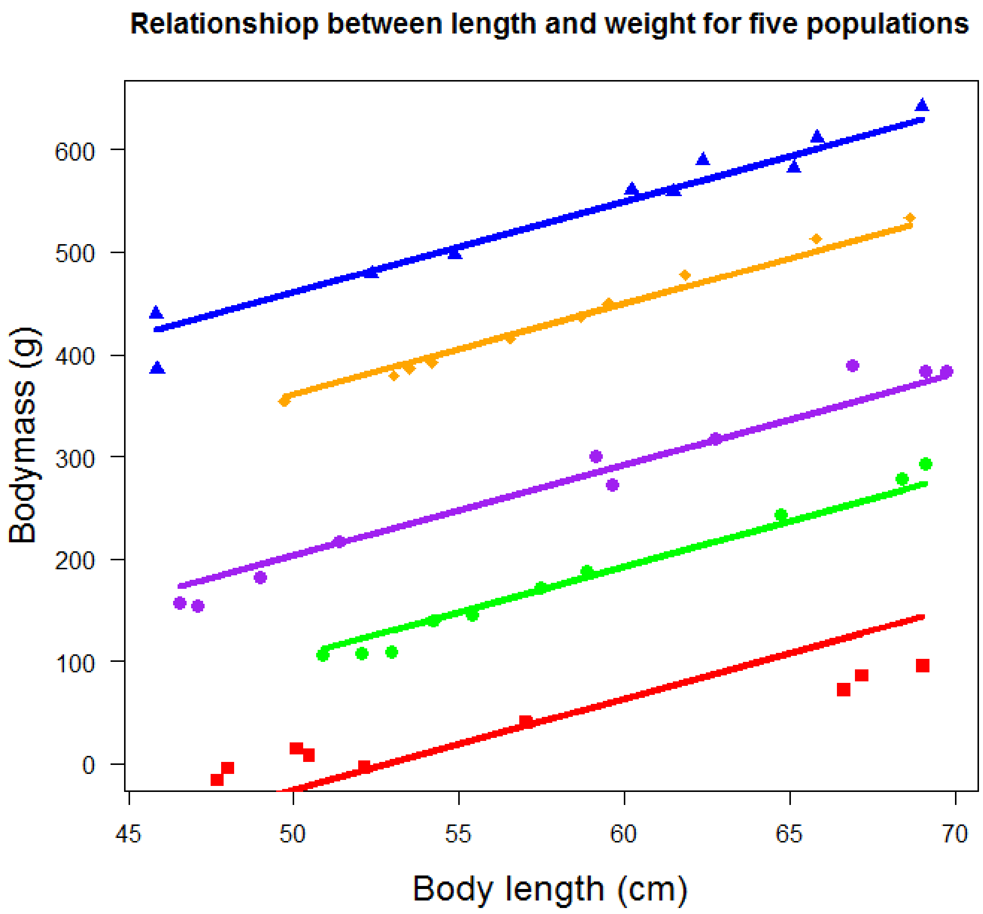
- What about interaction with discrete and continuous variables? Two continuous variables?

ANCOVA with interactions

What is an interaction?

Examine the length-weight relationship in five populations

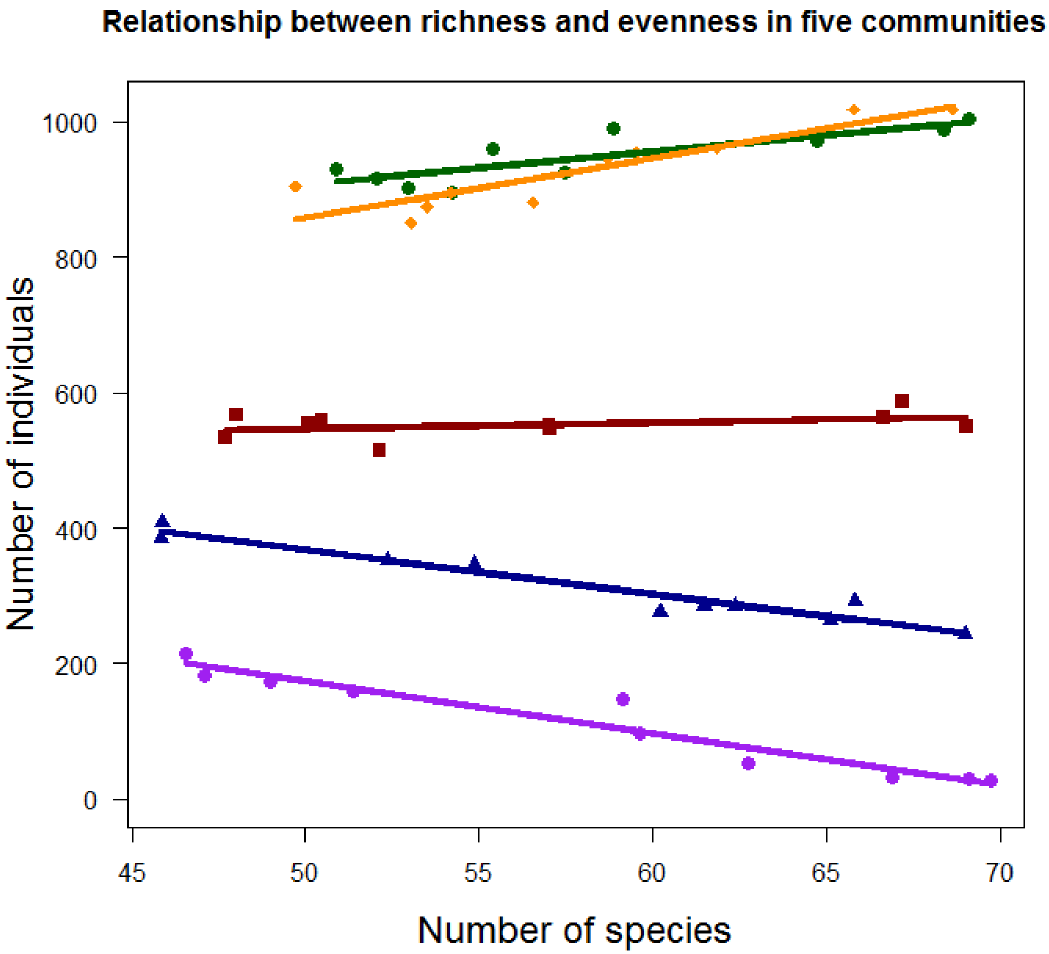
What do we see about the slope of the relationship?



ANCOVA with interactions

In some instances, we may want the relationship of x to y to vary among factors.

We do this by including an interaction



ANCOVA with interactions

- What is the effect of region (North, Central, South) on mass?
- Where elevation can differentially influence mass?

$$mass_i = \alpha + \beta_{j(i)} * region_i + \delta_{k(i)} * elevation_i + \varepsilon_i$$

$$\varepsilon_i \sim Norm(0, \sigma^2)$$

ANCOVA with interactions

- What is the effect of region (North, Central, South) on mass?
- Where elevation can differentially influence mass?
- What about the scenario where we hypothesize that the relationship between mass and elevation may be nonlinear? (i.e., peaking at some intermediate value)

$$mass_i = \alpha + \beta_{j(i)} * region_i + \delta_{k(i)} * elevation_i + \gamma_{k(i)} * elevation^2{}_i + \varepsilon_i$$

$$\varepsilon_i \sim Norm(0, \sigma^2)$$

General Linear Models

Linear Regression:

Continuous response, one continuous explanatory variable

T-test:

Continuous response, one discrete explanatory variable with only two categories

One-way ANOVA (Analysis of Variance):

Continuous response, one discrete explanatory variable with more than two categories

Two-way ANOVA:

Continuous response, two discrete explanatory variable

ANCOVA (Analysis of Covariance):

Continuous response, one discrete explanatory variable and one continuous explanatory variable

General linear models

- All of these models (model of the mean, t-test, linear regression, ANOVA) are special cases of the ***general linear model***

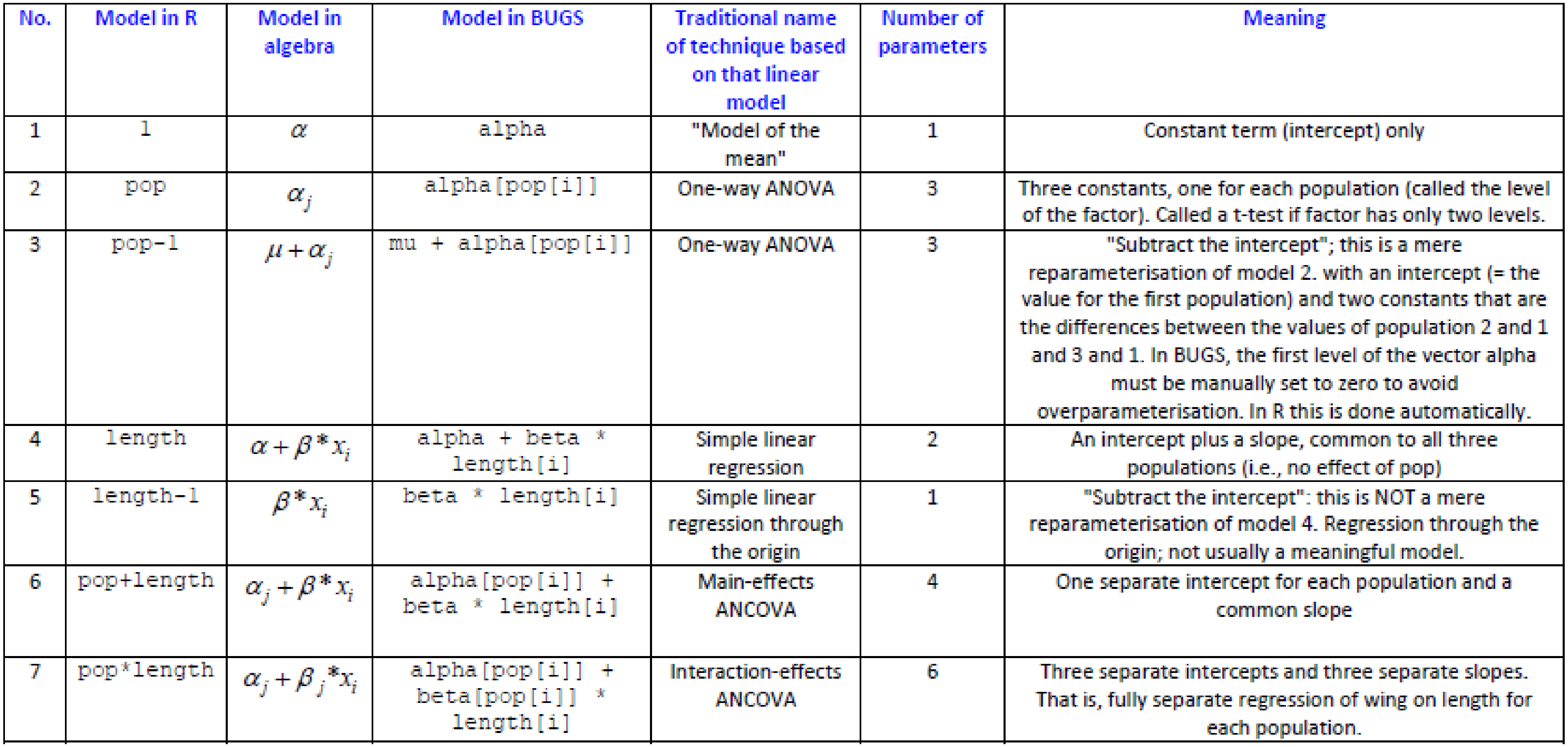
- General linear models express a continuous response as a linear combination of the effects of discrete (categorical; factors) and/or continuous explanatory variables (covariates)

  Plus a single random contribution from a Normal distribution whose variance is estimated along with the coefficients of all – discrete and continuous – covariates.

General linear models

- People used to make a sharp and artificial distinctions between linear models that contain categorical explanatory variables and those that contain continuous covariates
- Nowadays, in most practical applications we have several explanatory variables of both types (discrete, continuous) and may want to fit both the main effects of these covariates as well as some or all of their pair-wise or even higher-order interactions.

Linear models



No.	Model in R	Model in algebra	Model in BUGS	Traditional name of technique based on that linear model	Number of parameters	Meaning
1	`1`	α	`alpha`	"Model of the mean"	1	Constant term (intercept) only
2	`pop`	α_j	`alpha[pop[i]]`	One-way ANOVA	3	Three constants, one for each population (called the level of the factor). Called a t-test if factor has only two levels.
3	`pop-1`	$\mu + \alpha_j$	`mu + alpha[pop[i]]`	One-way ANOVA	3	"Subtract the intercept"; this is a mere reparameterisation of model 2. with an intercept (= the value for the first population) and two constants that are the differences between the values of population 2 and 1 and 3 and 1. In BUGS, the first level of the vector alpha must be manually set to zero to avoid overparameterisation. In R this is done automatically.
4	`length`	$\alpha + \beta * x_i$	`alpha + beta * length[i]`	Simple linear regression	2	An intercept plus a slope, common to all three populations (i.e., no effect of pop)
5	`length-1`	$\beta * x_i$	`beta * length[i]`	Simple linear regression through the origin	1	"Subtract the intercept": this is NOT a mere reparameterisation of model 4. Regression through the origin; not usually a meaningful model.
6	`pop+length`	$\alpha_j + \beta * x_i$	`alpha[pop[i]] + beta * length[i]`	Main-effects ANCOVA	4	One separate intercept for each population and a common slope
7	`pop*length`	$\alpha_j + \beta_j * x_i$	`alpha[pop[i]] + beta[pop[i]] * length[i]`	Interaction-effects ANCOVA	6	Three separate intercepts and three separate slopes. That is, fully separate regression of wing on length for each population.

Degrees of freedom

- The number of values that are free to vary

- Say you have a function: x + y + z =10
  - You can change x and y at random but then you have no choice (i.e., freedom) about the value of z. Therefore, there are 2 DF for this model.

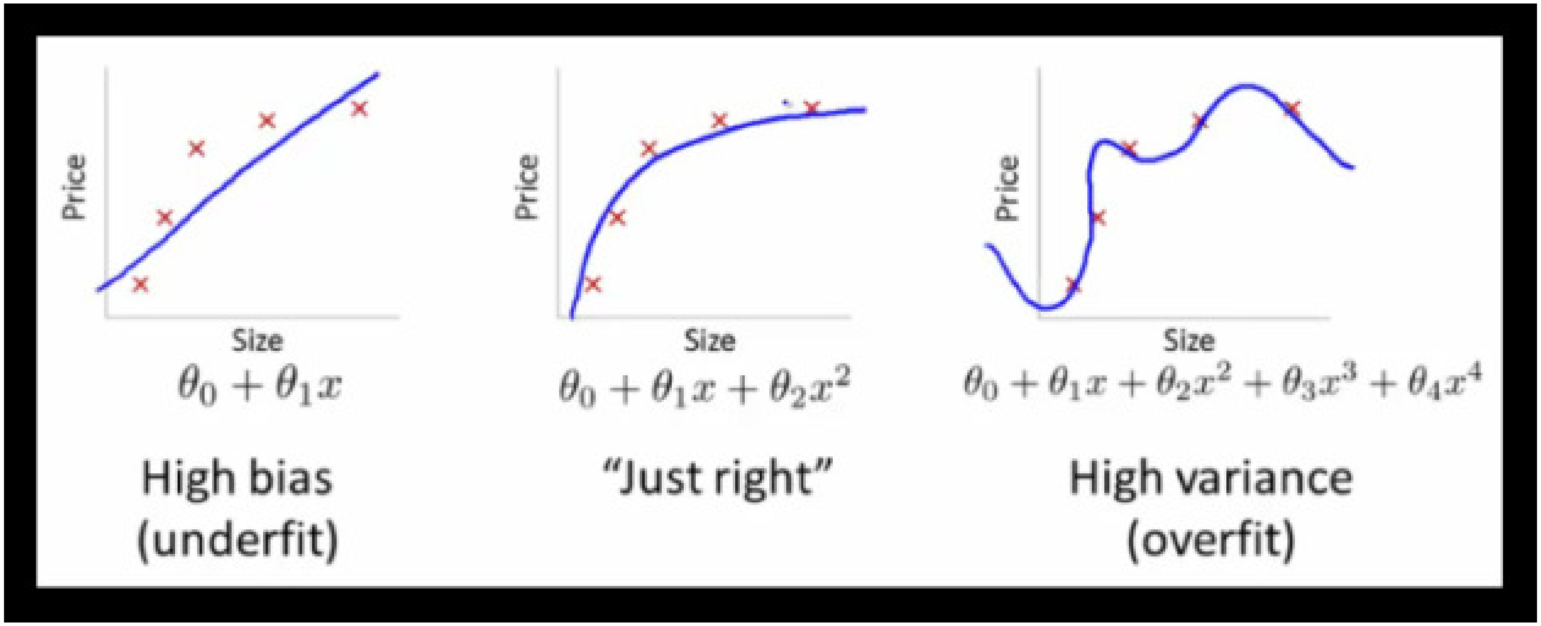
Degrees of freedom

- DF for an analysis depends on:
  - number of independent observations in a sample
  - number of population parameters to be estimated from the sample data

- Why is it important to understand DF in your model?
  - Gives a sense for how complicated you can reasonable make your model and for whether you are overfitting

Degrees of freedom

Too many parameters in a model lead to overfitting and lack of generalizability.



Lab

ANCOVA:

-Work in small groups on the ANCOVA homework file