ECO 602
Analysis of Environmental Data

FALL 2019 – UNIVERSITY OF MASSACHUSETTS

DR. MICHAEL NELSON

Today's Agenda

Recap of dual-model thinking and Frequentist inference.

Methods snapshot group activity/quiz

Constellation of methods: Group 1 – linear models

Group 1 coefficient interpretation

Question set 3 time

Landscape of Statistical Methods…

The Landscape

The basic statistical model:

Y = deterministic part + stochastic part

- Univariate
- Multivariate

- Linear
- Nonlinear
- Smoothed

- Distribution
- Heterogeneity
- Auto-correlation
- Nested data (random effects)
- Random noise

Model thinking: dual models

Predictors, responses

Expected values, model of the means

Errors: 3 sources

Constellation of methods

Group 1: Linear methods

Group 2: Extended linear methods

Group 3: Multivariate methods

Univariate methods snapshot

Univariate Methods

We'll be focusing on about regression-type methods:

- Predictors used to explain patterns in a response

Our models will only consider a **single response**, but may accommodate **multiple predictors**

Group 1 : Linear Models

Key assumptions:

1. Single response variable: **Univariate**
2. **Independent** observations
3. Linear* predictor/response relationships
4. Normally-distributed errors
5. Constant variance: no Heteroscedasticity

*Linear in the parameters

Group 1: terms and coefficients

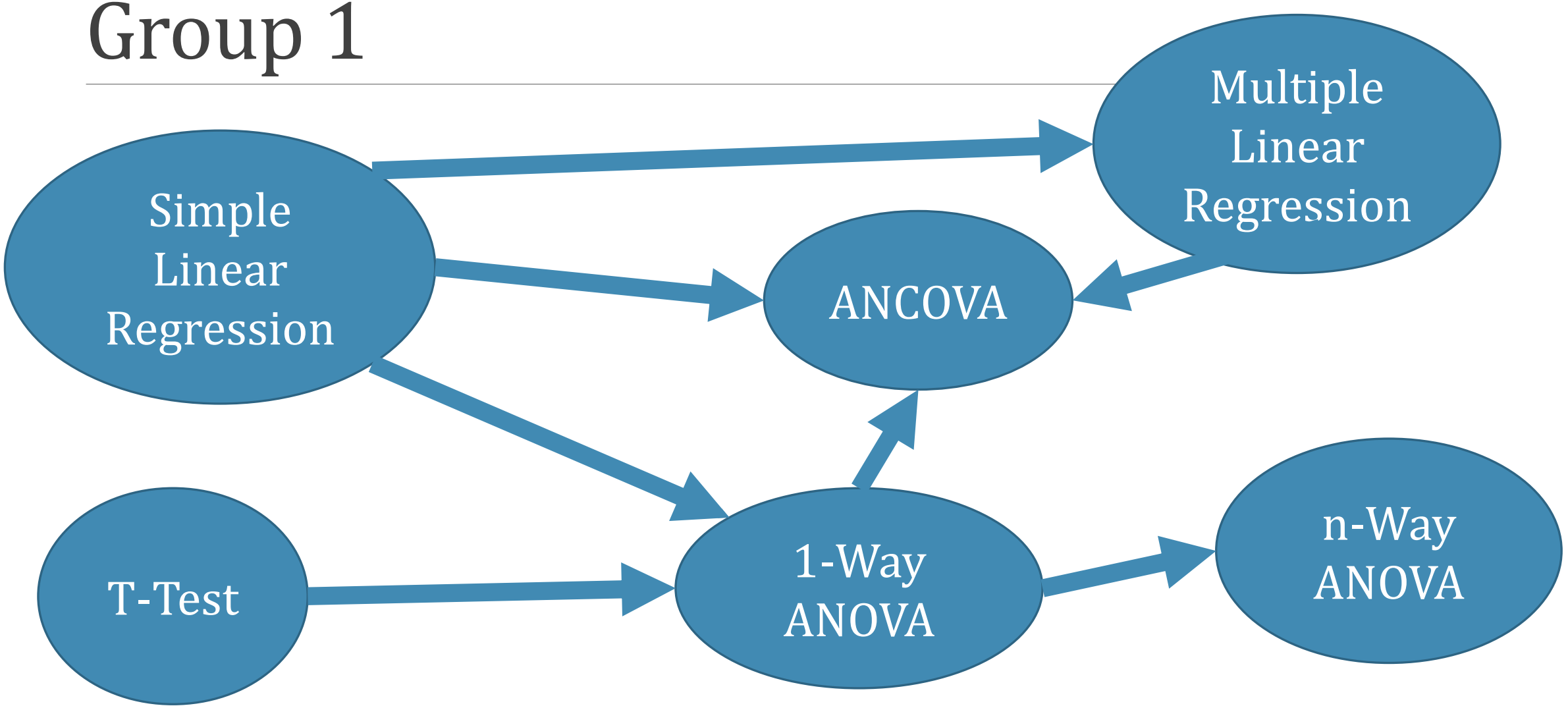
Response: Y

Predictor(s): X

Intercept(s): alpha

Slope(s): beta

Group 1


Simple Linear Regression
Multiple Linear Regression
ANCOVA
T-Test
1-Way ANOVA
n-Way ANOVA

Group 1: general equation format

Element-by-element form:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_n x_{ni} + \epsilon$$

Vector form:

$$\boldsymbol{Y} = \alpha + \beta_1 \boldsymbol{X_1} + \beta_2 \boldsymbol{X_2} + \cdots + \beta_n \boldsymbol{X_n} + \epsilon$$

Linearity in the parameters

Landscape of Statistical Methods...

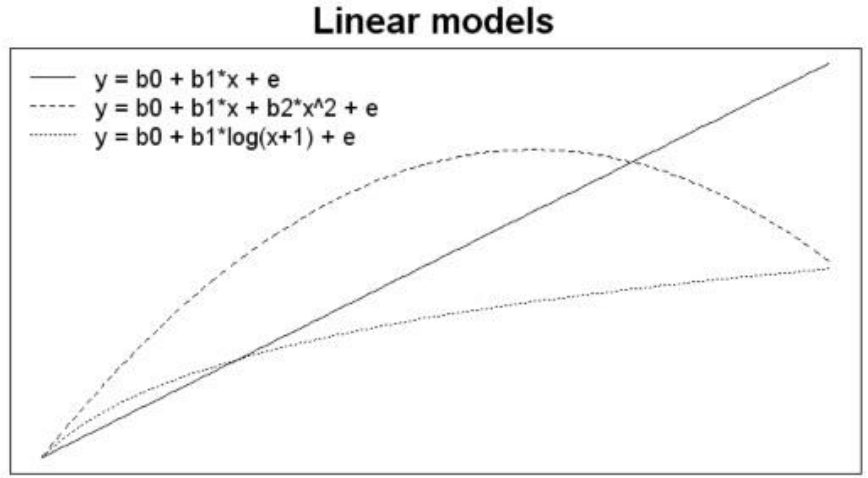
General linear models

Example linear models:

$$Y \sim Normal(b_0 + b_1 x, \sigma^2)$$

$$Y \sim Normal(b_0 + b_1 x + b_2 x^2, \sigma^2)$$

$$Y \sim Normal(b_0 + b_1 \log(x), \sigma^2)$$

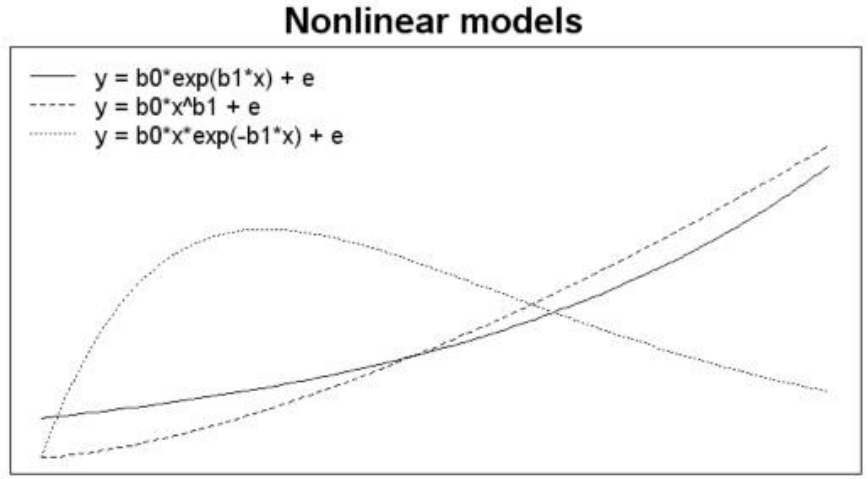


Example nonlinear models:

$$Y \sim Normal(b_0 e^{b_1 x}, \sigma^2)$$

$$Y \sim Normal(b_0 x^{b_1}, \sigma^2)$$

$$Y \sim Normal(b_0 x e^{-b_1 x}, \sigma^2)$$



Thunder Basin antelope

Response: spring fawn count

Predictors:

1. Adult population (previous year)
2. Annual precipitation
3. Winter severity index (1 – 5)

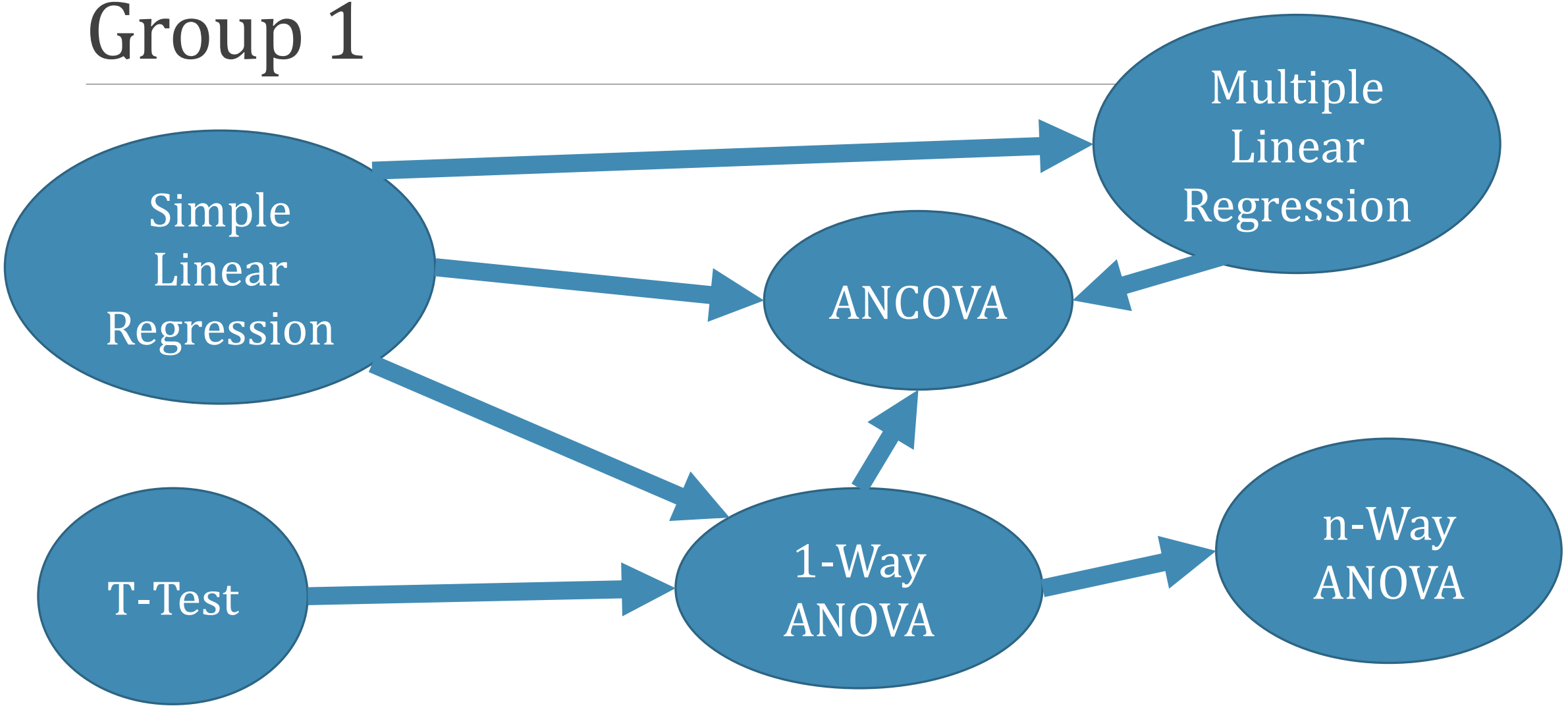
Thunder Basin antelope: what are the data types?

Response: spring fawn count

Predictors:

1. Adult population (previous year)
2. Annual precipitation
3. Winter severity index (1 – 5)

Group 1


Simple Linear Regression
Multiple Linear Regression
ANCOVA
T-Test
1-Way ANOVA
n-Way ANOVA

T-test

One **categorical** predictor with 1 or 2 levels

One **continuous** response

T-test

Analyzes the following questions:

1. Is the **mean of one group** different from **a fixed value**?
2. Are the means of **two groups** different from **each other**?

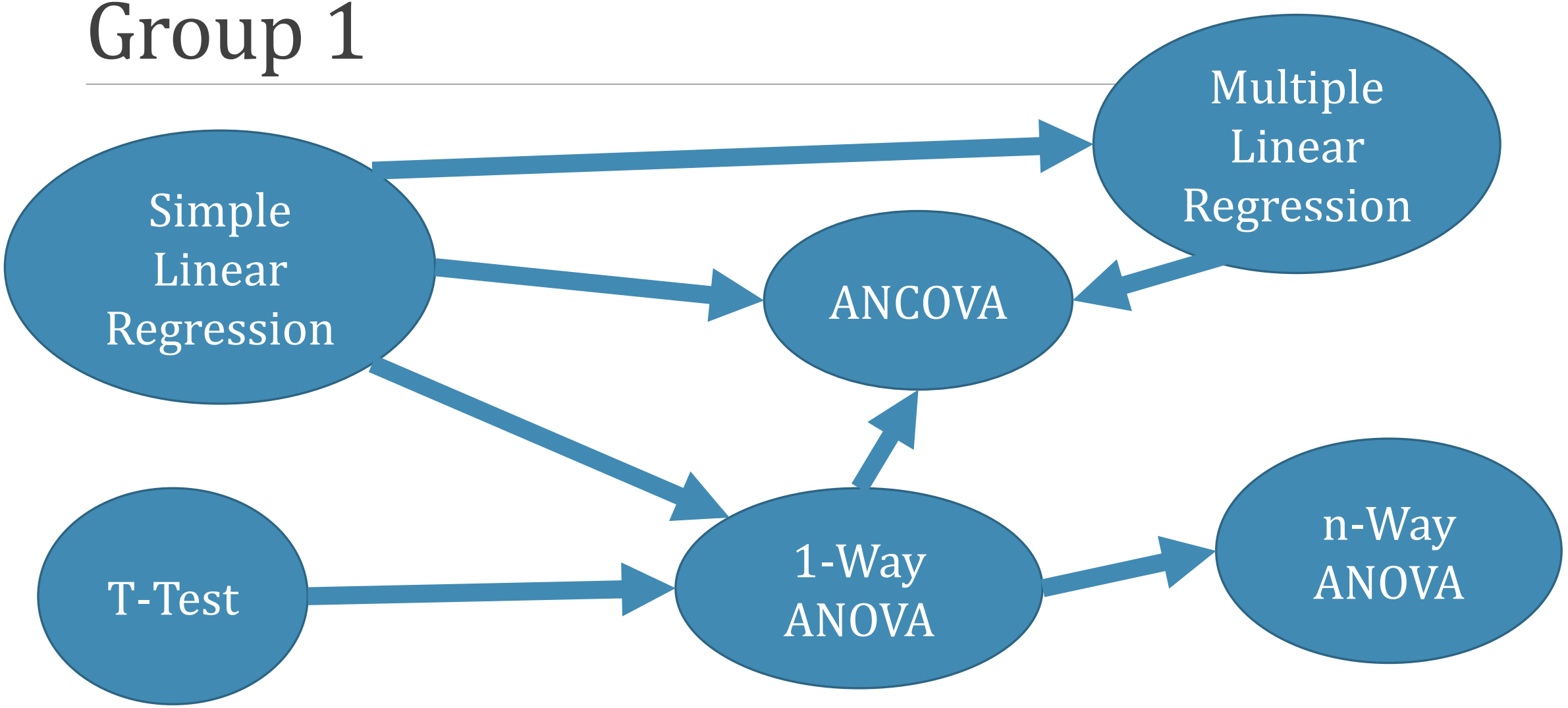
T-test elaboration

1. 1-way ANOVA extends t-test to 3 or more groups.

What antelope questions could we address with t-tests?

Hint: What are the categorical predictors?

Group 1


Simple Linear Regression
Multiple Linear Regression
ANCOVA
T-Test
1-Way ANOVA
n-Way ANOVA

Simple Linear Regression

One **continuous** response

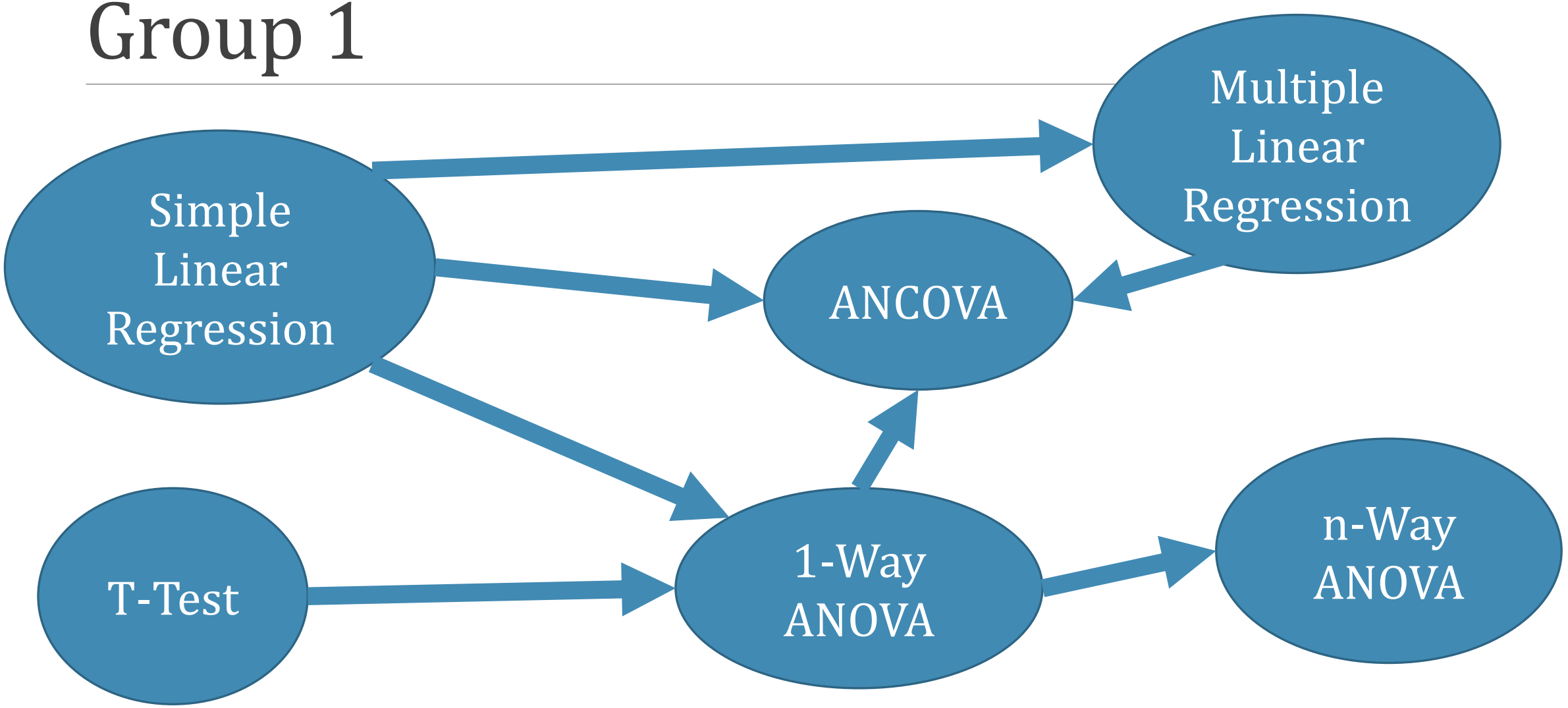
One **continuous** predictor

What questions could we address in the antelope data?

Simple Linear Regression elaborations

1. Multiple linear regression: More than one **continuous** predictors
2. ANOVA: One **categorical** predictor (instead of continuous)
3. ANCOVA: **Mixture** of categorical and continuous predictors

Group 1


Simple Linear Regression
Multiple Linear Regression
ANCOVA
T-Test
1-Way ANOVA
n-Way ANOVA

Multiple Linear Regression

One **continuous** response

Two or more **continuous** predictors

Multiple Linear Regression

Attempts to quantify the **pairwise relationships** between each **predictor** and the **response**.

Can consider **interactions**:

- **combined effect of 2 or more predictors** on the response

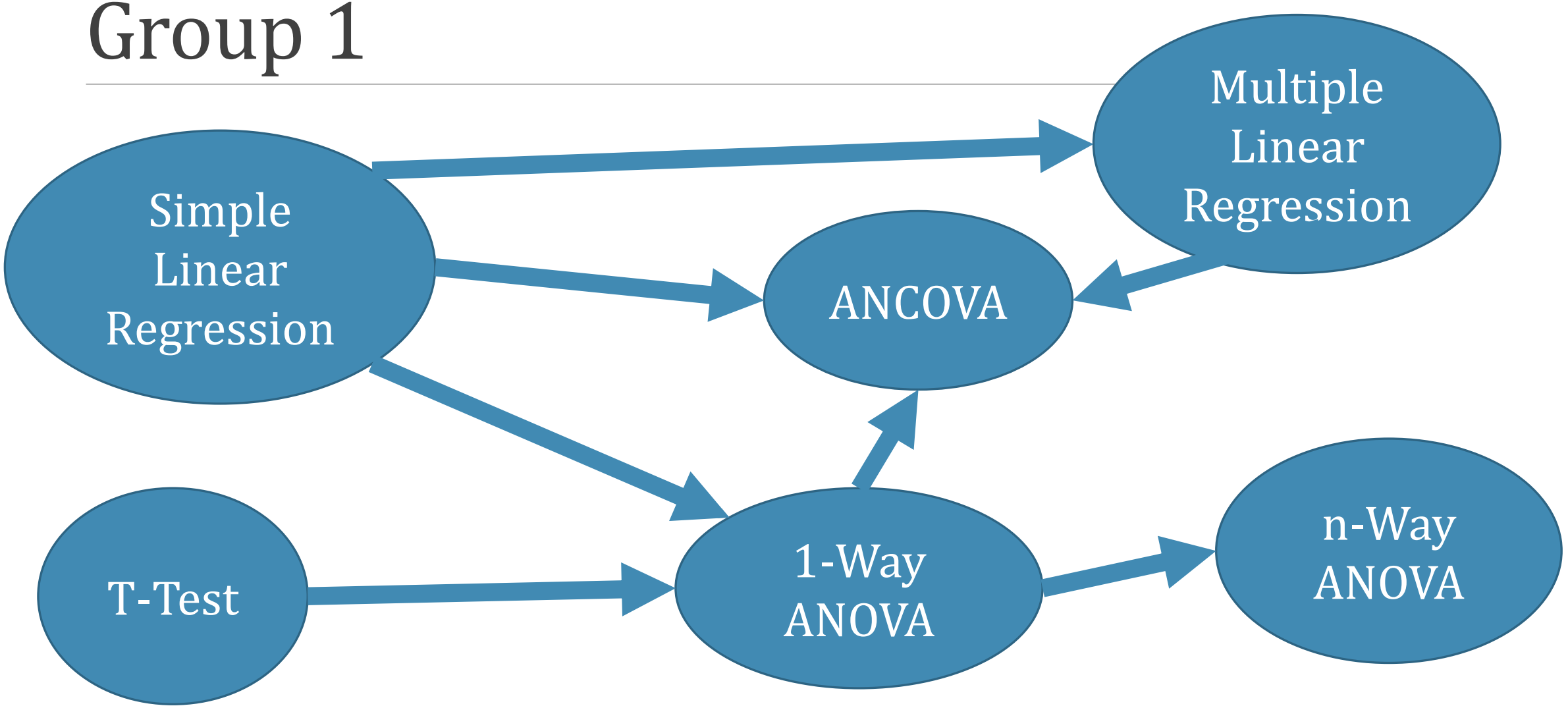
Multiple regression can fail with **highly correlated predictors**.

Multiple Linear Regression elaboration

Mixture of categorical and continuous predictors: Analysis of Covariance (ANCOVA)

Multiple Linear Regression: what questions could we address in the antelope data?

Group 1


Simple Linear Regression
Multiple Linear Regression
ANCOVA
T-Test
1-Way ANOVA
n-Way ANOVA

Analysis of Variance - ANOVA

Categorical predictor, 3 or more levels

Continuous response

Analyzes the following questions:

1. Are the group means different from one another?

Note: ANOVA **does not** specify which pairs of groups are different from one another.

ANOVA: What could we ask with the antelope data?

ANOVA elaborations

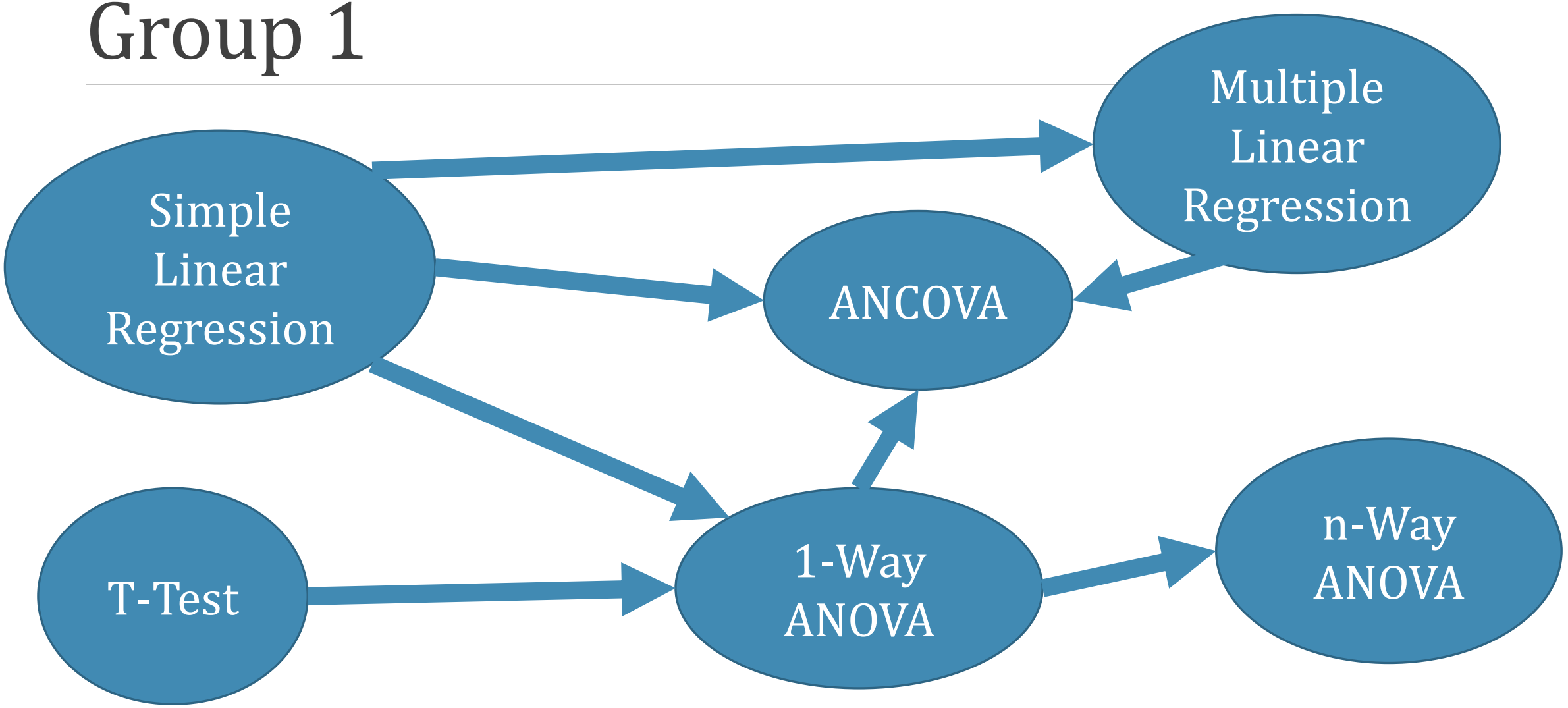
Two or more categorical predictors: multi-way ANOVA

Categorical and continuous predictors: Analysis of Covariance (ANCOVA)

Post-ANOVA analysis: which groups are different from one another?

- For example: Multiple testing corrections, Tukey Honest Significant Difference (HSD) test, Bonferroni correction, etc

Group 1


Simple Linear Regression
Multiple Linear Regression
ANCOVA
T-Test
1-Way ANOVA
n-Way ANOVA

Multi-way ANOVA

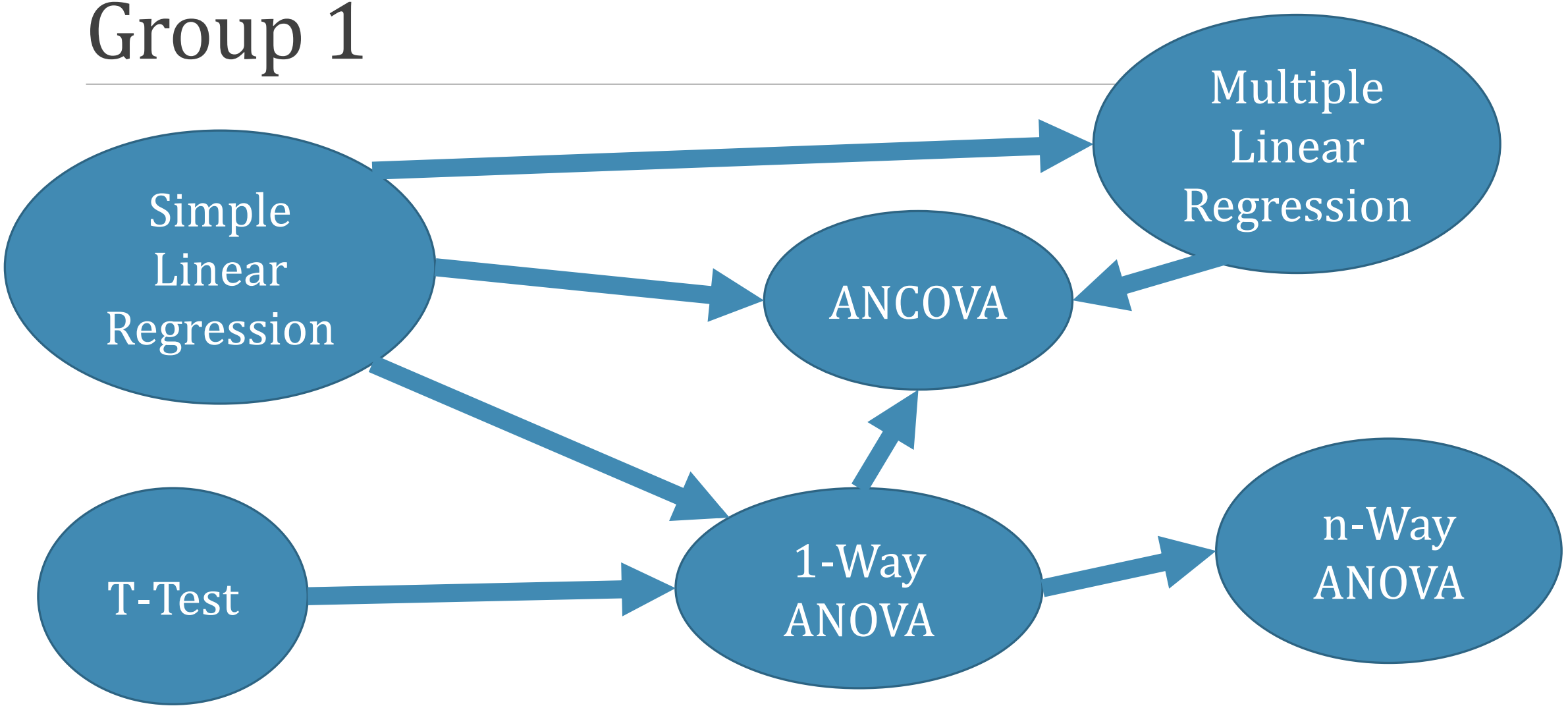
Categorical analogue of multiple regression

Main effects, interactions

Is multi-way ANOVA applicable to the antelope data?

Elaboration: Mix of categorical and continuous variables – Analysis of Covariance (ANCOVA)

Group 1


Simple Linear Regression
Multiple Linear Regression
ANCOVA
T-Test
1-Way ANOVA
n-Way ANOVA

Analysis of Covariance ANCOVA

Categorical and continuous predictors

Continuous response

What about the antelope data?

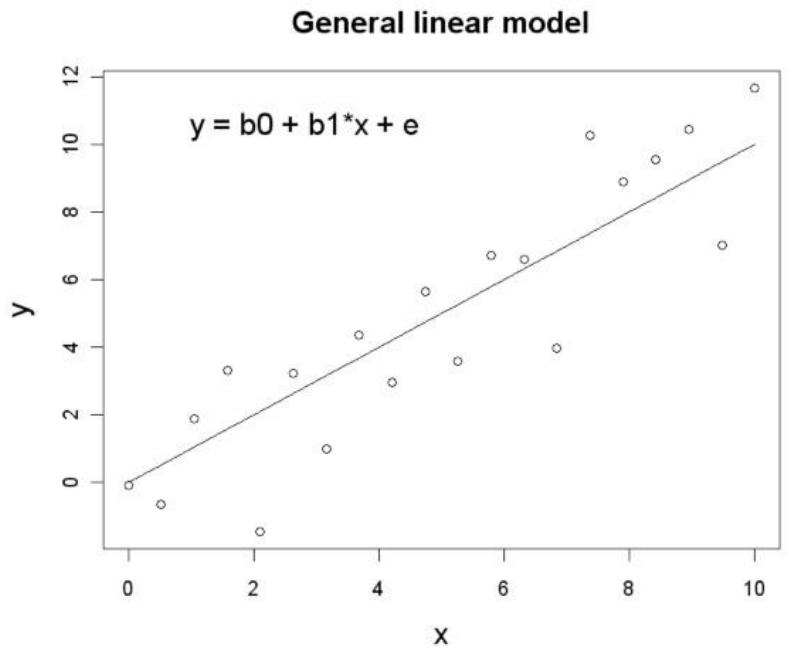
Landscape of Statistical Methods...

General linear models

Models that are linear functions *of the parameters*, not necessarily of the independent variables

$$Y \sim Normal\left(b_0 + b_1 z, \sigma^2\right)$$

- Y is *continuous*
- All observed values are *independent* and *normally* distributed with a *constant variance* (homoscedastic); any continuous predictor variables (covariates) are measured without error
- Method: *ordinary least squares*



Group 1 model summary presentations

1. Table of model coefficients – model summary, strength of effects of predictors, overall model significance test
2. ANOVA table – model variability attributed to each factor, factor-specific significance tests

Group 1 model summary interpretation: coefficients

1. Intercept: What is the value of the response when the predictor has value zero?
2. Slope: What is the change in the response with each unit change in the predictor?
3. Standard Errors: shape of sampling distribution
4. F-test: overall model significance test

Simple Linear Regression: Iris petal width predicted by petal length

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.363076   0.039762  -9.131  4.7e-16 ***
Petal.Length  0.415755   0.009582  43.387  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2065 on 148 degrees of freedom
Multiple R-squared:  0.9271, Adjusted R-squared:  0.9266
F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```

Multiple regression: petal width predicted by petal length, sepal length

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.008996   0.182097  -0.049   0.9607
Petal.Length  0.449376   0.019365  23.205   <2e-16 ***
Sepal.Length -0.082218   0.041283  -1.992   0.0483 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2044 on 147 degrees of freedom
Multiple R-squared:  0.929, Adjusted R-squared:  0.9281
F-statistic: 962.1 on 2 and 147 DF,  p-value: < 2.2e-16
```

ANOVA: Species and petal width

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          0.24600    0.02894    8.50 1.96e-14 ***
I(Species)versicolor 1.08000    0.04093   26.39  < 2e-16 ***
I(Species)virginica  1.78000    0.04093   43.49  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2047 on 147 degrees of freedom
Multiple R-squared:  0.9289, Adjusted R-squared:  0.9279
F-statistic:   960 on 2 and 147 DF,  p-value: < 2.2e-16
```

ANCOVA: species and petal length

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          -0.09083    0.05639  -1.611    0.109
Petal.Length          0.23039    0.03443   6.691 4.41e-10 ***
I(Species)versicolor  0.43537    0.10282   4.234 4.04e-05 ***
I(Species)virginica   0.83771    0.14533   5.764 4.71e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1796 on 146 degrees of freedom
Multiple R-squared:  0.9456, Adjusted R-squared:  0.9445
F-statistic: 845.5 on 3 and 146 DF,  p-value: < 2.2e-16
```

Group 1 ANOVA table interpretation

1. Degrees of freedom: Reflects the number of samples, number of factor levels, number of individuals per factor level etc.
2. Sum of squares: Reflects the total squared deviation from the mean explained by a source.
3. Mean squares: Mean SS due to a source (per DF)
4. F-tests: Test for ratio of variability explained by a source

1-way ANOVA

```
Analysis of Variance Table

Response: Petal.Width
             Df Sum Sq Mean Sq F value    Pr(>F)
I(Species)    2 80.413  40.207  960.01 < 2.2e-16 ***
Residuals   147  6.157   0.042
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA table for an ANCOVA

```
Analysis of Variance Table

Response: Petal.Width
                Df Sum Sq Mean Sq  F value    Pr(>F)
Petal.Length     1 80.260  80.260 2487.017 < 2.2e-16 ***
I(Species)       2  1.598   0.799   24.766 5.482e-10 ***
Residuals      146  4.712   0.032
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA table vs. model coefficient table

Model coefficient table tells you:

1. Intercept and slope coefficients
2. Overall model significance test, correlation test

ANOVA table tells you:

1. Variability explained by each factor in the model
2. Significance tests for each factor separately

When do group 1 methods start to fail?

Recall some of the key group 1 assumptions.

Independent observations

Linear relationships

Normal errors

Constant variance

Nonlinear relationships

Try adding terms: polynomial, power, interaction

Try a data transformation

Other options:

1. Additive models
2. Nonlinear least squares

Non-normal errors

Data transformation

Generalized linear models

Non constant variance

Data transformation

GLS

Non independent observations

Autocorrelation

Spatial dependence

Repeated measurements

Correlated error structures

Complicated experimental scenarios

Pseudoreplication

Too many zeroes

Blocking, random effects, nesting, hierarchical models

Categorical responses

For next time:

Select group paper for critical paper review

Group 2: extended linear techniques.