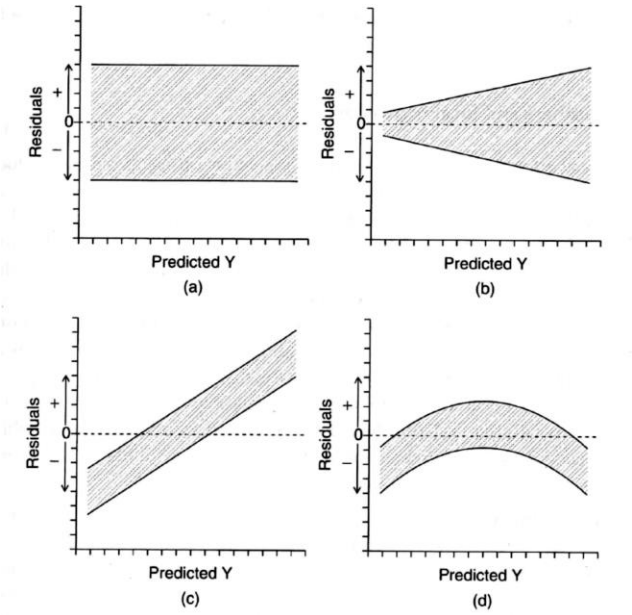
(370 pts total):

## Section 1 – Short answer

1. (15 pts) Let's say an ecologist wants to model the presence or absence (binary 0/1) of a species as a function of Elevation and Rainfall. Write the appropriate equation(s) to completely describe this model.

2. (15 pts) In Week #8, we discussed several different coding schemes for linear models (dummy coding, effect coding, Helmert contrast coding, etc.). In 1-2 sentences, explain why a modeler may choose to select a particular coding scheme for an analysis. (In other words, why have so many different ways of writing down the same model?)

3. (16 pts) Below are four plots depicting the residuals of a linear model plotted as a function of $\hat{Y}$.




For each panel, state whether the model violates any of the assumptions of linear regression and, if yes, which assumption(s) of linear regression are violated.

Panel a)

Panel b)

Panel c)

Panel d)

4. (15 pts) Ordinary least squares regression assumes that the residuals of a model are independent and identically distributed (i.i.d.), i.e. $Y = \beta_0 + \beta_1 X + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$. Name three common scenarios in which the residuals are <u>not</u> i.i.d.

5. (9 pts) Shmeuli makes the case that predictive modelling requires a larger sample size than explanatory modeling. Why?

6. (10 pts) When and why do we use Fisher's z transformation of the correlation coefficient r?

7. (15 pts) Fill in the three empty boxes.

| Test | Hypothesis (assuming two-tailed tests) | Test statistic T | $f(T\|H_0)$ (Distribution of T under $H_0$) | Assumptions |
|---|---|---|---|---|
| **Two sample paired t-test** | $H_0: \mu_A = \mu_B$<br>$H_A: \mu_A \neq \mu_B$ | | | |

8. (15 pts)

a. What is the difference between AIC and likelihood (both mathematically and in terms of how they assess model fit)?

b. How do you use AIC to calculate model weights (include the appropriate equation)?

9. (10 pts) What is the difference between a data point with high leverage and a data point that has high influence?

10. (10 pts) Let's say you have a two-factor crossed ANOVA

$$Y_{ijk} \sim \mu + A_i + B_j + AB_{ij} + \varepsilon_{ijk}$$

and the ANOVA finds that the interaction term is statistically significant. In words, how would you interpret a statistically significant interaction between the two factors $A$ and $B$?

## Section 2 – Long answer

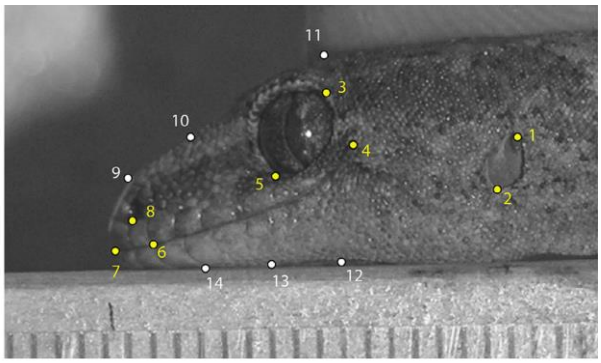
11. (85 pts)

Sexual size and shape dimorphism and allometric scaling patterns in head traits in the New Zealand common gecko *Woodworthia maculatus*

Clint D. Kelly [a,b,*]

[a] *Department of Ecology, Evolution and Organismal Biology, Iowa State University, Ames, IA 50011, USA*
[b] *Département des Sciences Biologiques, Université du Québec à Montréal, CP-8888 succursale centre-ville, Montréal, QC, Canada H3C 3P8*

Sexual dimorphism (the existence of different sizes or shapes for male vs. female members of the same species) is studied in the New Zealand common gecko by measuring particular points along the head (see Figure).




Let's say the authors were interested in studying head width (HW) as a function of body size (BS). One model that they might use would be the following equation:

$$HW = \alpha \times BS^{\beta}$$

We will call this the "Allometric model".

Another approach would be to take the log of both sides of this equation and fit this equation:

$$\log(HW) = \log(\alpha) + \beta\log(BS)$$

We will call this the "Log-linear model".

Part 1 (15 pts): Assuming both models are fit using ordinary least-squares regression (OLS), what is the statistical difference between these two models? Would they yield exactly the same answer? Why or why not?

Part 2a (5 pts): What function would you use to fit the “Allometric model” in R?

Part 2b (5 pts): What function would you use to fit the “Log-linear model” in R?

Let’s say that instead of modeling the effect of body size, the authors were interested in the relationship between head width (HW) and Sex (male vs. female), and they designed a study in which 10 female and 10 male geckos were measured, and the measurements of HW on each individual were replicated 4 times.

Part 3 (5 pts each for a-f; 30 for ANOVA table):

a) Would ‘Sex’ be considered a fixed or a random effect and why?

b) State the null hypothesis being tested with regards to the effect of ‘Sex’.

c) Would ‘Individual be considered a fixed or a random effect and why?

d) State the null hypothesis being tested with regards to the effect of ‘Individual’.

e) Why might the researchers want to replicate the measurements on each gecko?

f) Why would it be wrong to analyze all 40 male and 40 female head width measurements together using a simple t-test?

g) Fill in the gray squares to complete the appropriate ANOVA table for the analysis of HW ~ Sex + Individual. (Since you don't have data, just fill in the appropriate equations in the ANOVA table.) For the p-value column, use the indicated space below the table to fill in the R code you would need to calculate the associated p-value.

| Source | Degrees of freedom | SS | MS (leave as ratio) | F-ratio (leave as ratio) | p-value |
|---|---|---|---|---|---|
| | | | | | write answer below |
| | | | | | write answer below |
| | | | N/A | N/A | N/A |
| Total | | | N/A | N/A | N/A |

p-value for the first row of the ANOVA table =

p-value for the second row of the ANOVA table =

12. (75 pts)

Consider the following model for the relationship between the *Sex Ratio* ("SR"=response) of eggs in a clutch, *Temperature* and *Maternal Age* ("T" and "MA" = covariates). You examined 50 clutches of eggs in the field.

In all cases

$$SR_i \sim Binomial(n_i, p_i)$$

where $i$ represents the $ith$ clutch and $n_i$ is assumed to be known (fixed value, counted in the field).

Model 1: $logit(p) \sim \beta_0 + \beta_1 * T + \beta_2 * MA$; Log-likelihood (Model 1) = -92

Model 2: $logit(p) \sim \beta_0 + \beta_1 * T$; Log-likelihood (Model 2) = -95

Model 3: $ogit(p) \sim \beta_0 + \beta_2 * MA$ ; Log-likelihood (Model 3) = -99

Model 4: $ogit(p) \sim \beta_0$ ; Log-likelihood (Model 4) = -100

A) (15 pts) Which combinations of models can be compared by a likelihood ratio test?

B) (15 pts) What is the test statistic and distribution under the null hypothesis for the likelihood ratio test?

C) (24 pts) Fill in the AIC model selection table below (for the last column, feel free to leave the answer as a mathematical expression).

| Model | # parameters | AIC | AICc |
|---|---|---|---|
| 1 | | | |
| 2 | | | |
| 3 | | | |
| 4 | | | |

D) (21 pts) Discuss briefly the pros and cons of using the likelihood ratio test vs. an Information Theoretic approach such as AIC.

13. (40 points) In a linear regression, we model

$$Y_i = \beta_0 + \beta_1 * X_i + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

a. Derive the maximum likelihood estimate of $\widehat{\beta_1}$. (Note: Substantial credit will be awarded for correctly setting up the solution.)

(TOP PORTION OF PAGE LEFT BLANK AS EXTRA SPACE FOR SHOWING YOUR WORK.)

b. List two R functions that could be used to estimate $\widehat{\beta_0}$ and $\widehat{\beta_1}$. Be sure to include the entire R command, including any parameters that must be included as inputs to the R function.

## Section 3 – Essay (40 pts)

(Each question can and should be answered succinctly in 4-5 sentences max.)

14a. In Hurlbert (1984), Cox suggests that one possible method of overcoming segregated designs after randomization is to "rerandomize" until some pre-specified amount of interspersion has been achieved. Why would this suggestion be so problematic for Fisher?

14b. What is the difference between the *pre-layout* and *layout specific* alpha? Which is of greater interest to experimenters and why?