NAME: _____

Section 1 – Short answer

1. (13 pts) Estimate the Pearson product-moment correlation coefficients for the following scatterplots (write answer next to scatterplot, there are 21 panels in total). Note that for some panels, only rough estimates will be possible; make sure that the relative magnitude and signs are correct among the plots. There is one panel for which the Pearson product-moment correlation coefficient is Undefined. Please mark it accordingly.

2. (10 pts) In 2-3 sentences, explain the motivation for using and basic idea behind robust regression. Draw a sketch if needed.

3. (15 pts) The following data (Figure 1) violates an assumption of linear regression. What assumption is violated? How would you test whether the assumption is violated? How would you analyze the data appropriately?

4. (12 pts) There are four experimental designs represented here. Circle the correct design type:

5.(20 pts)

A) What kind of response variables are modeled by Poisson regression?

B) Write down the Poisson regression model (Hint: Need two equations to completely describe the model.).
(2.5 pts for each equation)

C) What is meant by overdispersion when discussing Poisson regression?

D) What is the difference between doing a Poisson regression and doing a normal linear regression using log-transformed responses?

6. (30 points) Derive the expression for the slope parameter β_1 in a linear regression model in two ways.

A) Method #1:

B) Method #2:

7. (80 pts)

A) (5 pts) Why do you think the authors transformed the response variables?

B) (44 pts) Write out the three-way ANOVA table for ANPP assuming all factors are treated as fixed effects. ($Y_{ijkl} =$
$ANPP\ for\ the\ l^{th}\ sample\ in\ sapling\ category\ i, elevation\ class\ j, and\ soil\ category\ k$).

Source	Degrees of freedom	SS	MS (leave as ratio)	F-ratio (leave as ratio)	p-value
Sapling density					Leave blank- See Part 3 below
Elevation					Leave blank
Soil category					Leave blank
Sapling density x Elevation					Leave blank
Sapling density x Soil category					Leave blank
Within groups (residual)				N/A	N/A
Total			N/A	N/A	N/A

C) (5 pts) Write the R command needed to generate the p-value associated with the main effect of Sapling density.

D) (5 pts) State the null hypothesis being tested with regards to the main effect of “Sapling density”.

E) (5 pts) What is the implicit alternative hypothesis with regards to the main effect of “Sapling density”.

F) (5 pts) State the null hypothesis being tested with regards to the interaction of “Sapling density x Elevation”.

G) (5 pts) Why do you think the authors chose not to include the interaction between Elevation and Soil Category in their analysis?

8. (50 pts)

A) (10 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) (12 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) (13 pts) Discuss briefly the pros and cons of using the likelihood ratio test vs. an Information Theoretic approach such as AIC.

9. (80 pts)

A) (10 pts): Given that

$$\lambda(t) = -\frac{d}{dt} log S(t)$$

what is the survival function $S(t)$ if we assume a constant hazard function $\lambda(t) = \lambda$?

B) (20 pts): Prove that, under a constant hazard function, the log-likelihood associated with these data on the age of death across the population is given by

$$\log(L) = klog\lambda - \lambda \sum_{i=1}^{n} t_i$$

C) (10 pts): What is the maximum likelihood estimator for the hazard λ?

D) (20 pts): Prove that the log-likelihood for censored exponential data (what you just derived above in Part B) coincides exactly (**except for constants**) with the log-likelihood that would be obtained by treating k as a Poisson random variable with expected value equal to $\lambda \sum_{i=1}^{n} t_i$. In other words,

$$Probability\ of\ dying \sim Pois(\lambda \sum_{i=1}^{n} t_i)$$

E) (20 pts): There are a large number of statistical models for survivorship, many of which focus on modeling the hazard function. Cox (1972) introduced a class of models called *proportional hazards models*, in which the hazard function takes on the following form

$$\lambda_i(t|X) = \lambda_0(t)e^{X\beta}$$

where $X\beta$ is just the model matrix representing the covariates and their coefficients influencing the hazard function.

Notice that we can rewrite this as

$$\log\big(\lambda_i(t|X)\big) = \log\big(\lambda_0(t)\big) + X\beta$$

where it is easy to see that we are using a log "link" function to model the relationship between the hazard rate at time t . The covariates represent an additive offset to the baseline hazard rate.

To apply this to a concrete example, let's say a researcher studying bird survivorship wants to know whether chlorophyll-a (a year-specific continuous covariate used as a measure of biological productivity in the ocean) has an effect on the hazard function (which is, as a reminder, the instantaneous death rate at age t). Other covariates of interest include colony size and average sea surface temperature (SST).

The two models to be compared are then

$$\log\big(\lambda_i(t|ChlA_i, Size_i, SST_i)\big) = \log\big(\lambda_0(t)\big) + \beta_0 * ChlA_i + \beta_1 * Size_i + \beta_2 * SST_i$$

and

$$\log\big(\lambda_i(t|ChlA_i, Size_i, SST_i)\big) = \log\big(\lambda_0(t)\big) + \beta_1 * Size_i + \beta_2 * SST_i$$

Identify two methods of determining whether chlorophyll-a has a **statistically significant** influence on the hazard function.

Section 3 – Essay (50 pts)

10. Shmueli (2010) makes the distinction between explanatory modeling and predictive modeling. Briefly describe the difference between these two types of modeling in terms of their goals and differing approaches to model selection. Be sure to include in your answer some discussion of the bias-variance trade-off in the expected prediction error. What is predictive modeling most concerned with? What is explanatory modeling most concerned with?