**PSC 204B – Winter 2018**

**Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

The final is due on Thursday, March 22nd by 12:00 PM (noon), uploaded to Canvas.

Instructions:

* This exam includes 8 questions summing up to 120 points. 100 points are needed for a perfect score. This exam includes a total of 20 bonus points, which may be used toward the exam portion of your final grade.
* Answers to essay and short answer questions should be typed.
* Answers from questions requiring computer work should include **edited, relevant output** (unedited or irrelevant output will result in deducted points).
* This exam gives you the opportunity to work independently. You may consult with Kristine, Jared or Philippe concerning the final, but not with anyone else. However, you can consult your book and class notes and search the web.
* Please, show all your work on calculation problems. Answers should be expressed as clearly and concisely as possible.
* Address every question – we can’t give you points for skipped questions. Good luck!

***Final Exam***

**1. Causality (10 points)**

Assume that a linear regression is appropriate for the regression of an outcome, *y*, on treatment indicator, *T* , and a single confounding covariate, *x*. Sketch hypothetical data (plotting *y* versus *x*, with treated and control units indicated by circles and dots, respectively) and regression lines (for treatment and control group) that represent each of the following situations:

1. No treatment effect *(3 points)*,
2. Constant treatment effect *(3 points)*,
3. Treatment effect increasing with x *(4 points)*.

**2. Information Criteria (10 pts)**

We talked about Akaike’s information criteria. Another is the Bayesian Information Criterion (BIC) which is defined as BIC = Dtrain + ln(*n*)\**k*

From this definition, try to explain what the idea of BIC is and how it is different or similar to AIC.

**3. Logistic Regression (10 points)**

Say you run a logistic regression with predictor variable X1 and outcome variable Y. You fit the following model:

a) What is the interpretation of B0 in logits? What is the interpretation of B1 in logits? *(2 points)*

b) How would you transform B1 so that it would be expressed in odds? How would this change your interpretations of B0 and B1? *(3 points)*

**After running the analysis you remember a covariate that you believe is related to Y, but is not substantively of interest.**

c) What are the benefits of including the covariate in the model? Include two benefits and explain in them in detail. *(3 points)*

**Finally, you include the covariate within the analysis and fit the following model:**

d) What is the interpretation of each of the coefficients in this model? *(2 points)*

**4. Interactions (10 points)**

It is not uncommon for researchers to incorporate interaction terms in regression models. For the following questions, please consider the following regression equation:

y = B0 + B1\*x + B2\*d + B3\*x\*d

a. Generally, what do the coefficients (B0, B1, B2, & B3) above represent? How would you interpret them? *(2 points)*

b. Imagine that *d* is a dummy coded variable indicating whether a participant is exposed to an experimental condition (1) or a control condition (0). Assume that x and y represents continuous scores. Generally, indicate what the intercept and slope are for the control condition and for the experimental condition. *(2 points)*

c. What are the advantages and disadvantages to z-scoring variables before conducting regression analysis? Be sure to comment on the effects of z-scoring when including interaction terms. *(1 point)*

d. If the interaction term in the model above was not statistically significant but the main effects were, would you be able to interpret them? If all terms were significant, would you be able to interpret main effects? Explain your answers. *(2 points)*

e. What are simple main effects and why do we use them? If *d* had also been a continuous score how would you conduct simple main effects using the above regression equation? Indicate values you would use to demonstrate how trajectories of the x and y relation are influenced by values of *d*. Demonstrate predicted values for at least three trajectories as a function of *d* with at least two points between x and y. Do not use actual values, simply indicate values you’d use formulaically (e.g., B2\*mean(x), B2\*(mean(x)+2SD(x))). *(3 points)*

**5. Emotion (20 points):**

You are investigating positive and negative affect in participants who provided daily measures of affect. These data come from a partnership satisfaction study and they also include a measure on general satisfaction with their partnership.

For the current analyses, the outcome is positive affect (pa\_gral) and the main predictor is negative affect (na\_gral). Both variables have been measured repeatedly over a period of 30 to 91 days within 165 individuals.

The data are in the *emotion.csv* dataset, where you’ll find following variables: ID\_INDIV = person ID; GENDER = male or female; DAY = day of measurement indexing the time-variable; pa\_gral = positive affect; na\_gral = negative affect; V1RelSat = general satisfaction with their relationship.

Affect scores range from 1 to 5; you can assume that these variables are continuous.

a. Plot the data (spaghetti plots) and comment. You are then going to predict positive affect with negative affect and you will also include gender and relationship satisfaction as explanatory variables. *(3 points)*

b. Comment on whether you need to detrend or center (or both) your data. Go ahead and follow your recommendations and model your data. Settle on a model with which you are comfortable. *(10 points)*

c. Illustrate your model(s) with different plots of the predicted scores and compare models via an information criterion. *(5 points)*

d. Describe your final model and give an indication on how confident you are with your results. *(2 points)*

**6. Please use the *pigeon.csv* data set to answer the following questions (20 points).**

Dr. Cher Ami is conducting a mega-analysis of 500 pigeons under similar fixed-interval reinforcement schedules. Pigeons were placed in a Skinner box where reinforcement (food) was administered after the pigeon pecked a target button following the fixed interval period. These trials lasted ~20 min.

a. Plot the pigeon data. Make sure the plot is properly labeled. *(3 points)*

b. Test at least three regression models for this data. At least one model should be a nonlinear regression model. (Hint: The data generating model yields an AIC = 1277.38 and BIC = 1294.24 these should be treated as benchmarks for your models) *(10 points)*

c. Report fit indices for your models. *(2 points)*

d. Plot the predicted trajectories of your models against the empirical data. Your plot trajectories should be labeled. If all the trajectories cannot be clearly seen in a single plot then please make additional plots as needed. *(3 points)*

e. Summarize your results: (a) indicate which model you have selected and why; and (b) interpret your model to explain the effect of fixed-interval reinforcement on pigeon pecking behavior over a 20 min trial. *(2 points)*

**7**. **Linear Mixed-Effects Models (20 points).**

Using the following link, please load your data into your statistical program: <https://www.hsph.harvard.edu/fitzmaur/ala/lead.txt>. The link provides both the data and the description of the data. The data is a longitudinal study examining lead exposure in children over time. Please answer the following questions:

a. The data given to you are in wide format. Please convert the data from wide to long and provide the first few lines of your data to show that it is indeed in long format (through the “head” function in R, for example). *(3 points)*

b. Fit a linear-mixed effects model with a random intercept. Interpret the fixed intercept and fixed slope in the context of the problem. Depending on the statistics program or R library you use, you will get a standard deviation/variance in the random intercept. Please report the intraclass correlation (ICC) (please remember the ICC uses VARIANCE, not SD). What does a small intraclass correlation tell us? A large ICC? Assume that the standard deviation/variance in the random intercept is significant. What does this mean in the context of the problem? *(5 points)*

c. Now fit a linear mixed-effects model with a random slope. Interpret the fixed intercept and fixed slope in the context of the problem. Again, assume that both the standard deviation/variance of the random intercept/slope are significant. What does this mean in the context of the problem? *(5 points)*

d. Which model is better: the random intercept only model or the model with a random intercept and a random slope? Provide three methods of evidence to back your assertion as to which model is better. *(3 points)*

e. Please explain why we are using a linear mixed-effects model as opposed to a traditional linear regression model for this data. What are the assumptions of a traditional linear regression model and what are the assumptions of a linear mixed-effects model? How does fitting a traditional linear regression model to this data violate the assumptions and how does a linear mixed-effects model remedy these violations? Even though fitting a linear mixed-effects model is the right choice for this data, what, if any, are the disadvantages to fitting such a model (in general)? *(4 points)*

**8. Car accidents (20 points)**

This is an actual question I obtained two years ago from a former schoolmate. He works in an engineering business and they resurfaced a section on a highway between two cities in Switzerland. Here’s the email I got (translated):

Dear Philippe,

….

I have been working in an engineering office for a few years now and mainly deal with risk analysis. Currently I'm working on a security report, which I then have to represent in court, and regarding a statistics question, I am a little uncertain. Since I do not want to embarrass myself in court, I am now looking for your advice.

I’m interested in three questions:

1. Is the accident frequency of sample 2 significantly smaller than that of sample 1?
2. Is the accident frequency on section S1 [a defined section of the road with a given length of some miles] of sample 2 significantly smaller than that section S1 of sample 1?
3. Is the accident frequency in sections S2 and S5 and S6 and S8 of sample 2 significantly lower than the corresponding sections of sample 1?

The tests can be on a standard alpha level of .05.

What statistical tests would you use and what results do you obtain?

Very best,

Mänu

The file (highway.csv) contains all the information I got from Mänu. The month denotes the observation time.

Months 1-61 denote the observations before they renewed the surface. Months ranging from 91-105 denote the observations after they renewed the surface.

S1 to S8 are given sections of the highway, each number denotes a crash (0=no accident, 1=one accident, 2=two accidents...)

The variable “Sample” indicates the pre (coded as 1) and post (coded as 2) intervention – which can be derived from the month’s as well.

Please answer the three points in Mänu’s email, with your conclusion and your explanation on what you did and describe your results (hint: the outcome here is **count** of accidents--this is NOT a continuous variable--so you should take this into account when choosing a model). *(Question 1: 6 points; Question 2: 6 points; Question 3: 8 points)*