This assignment has 2 questions, each with multiple parts.

Please put your name only on the back of the last page and make sure all pages are securely attached. I want to grade anonymously.

General notes:
1) You can write log likelihood functions or use glm() or glm.nb() for all parts, but some parts will be easier with a log likelihood function. Using glm() for those requires knowing some things I haven't talked about.
2) Use the asymptotic normal confidence interval formula (the one in the week 1 equation notes) to compute all confidence intervals.
3) Do not log transform the counts for any parts. Some parts ask you to use log lambda as the parameter. That's very different from log transforming the response.
4) You do not need to include R code with your answers, but I encourage you to do so. That way I have more information if you give me an answer I didn't expect.

1. This problem is intended to give you practice working with likelihood and distributions. The data comes from a study of white oaks in two woodlands. One woodland has dry soils; the other has moister soils. Ecology students counted the number of white oak trees in 10m x 10m plots in both types of woodland. There were 22 dry and 18 moist woodland plots. The data are provided in two forms. oakD.csv and oakM.csv have the counts in the dry and moist woodlands, respectively. bothoak.csv has one row for each of the 40 plots. The two variables are the type of woodland (dry, moist) and the number of white oaks. Throughout we will assume that the number of white oak trees in a plot has a **Negative Binomial distribution**.

   (a) Use the log-likelihood function lnlNB0() defined in the lnlNB.r code on the class website to fit a Negative Binomial distribution to the counts in **dry** woodland plots. Start the numerical optimization at $\lambda = 20$, $r = 5$, use method='BFGS', and turn on trace=T to see the path.
   Does optim() venture "outside" the valid parameter ranges ($\lambda \geq 0$, $r > 0$) then recover? If so, give the values of μ and r where optim() evaluates lnL but recovers.
   Does optim() get stuck "outside" the valid parameter ranges?
   Do you trust the reported mle's and lnL?

   (b) What would be a more reasonable starting value for λ? (Many answeres are possible, some more reasonable than others). Use optim() with that starting value. If optim() still gives outrageous values, try different choices of r. Report the estimated λ (not log λ) and the NegBin overdispersion parameter.

   (c) Calculate the standard error of $\hat{\lambda}$ and the asymptotic 95% confidence interval for the λ. For this confidence interval, use the estimated mean count, λ. and the asymptotic confidence interval equation given in the note summaries.

   (d) Find the mle's when the parameters are $\log \lambda$ and r. Is the log likelihood for this fit the same as when you use λ as the parameter?

   (e) Estimate the log scale mean number of oaks in a **dry** woodland plot and the standard error of the log-scale estimate. Use the log mean and the asymptotic confidence interval formula to compute a 95% confidence interval for the log mean. Backtransform the log scale confidence interval to provide a 95% confidence interval for λ.
   Notes: 1) If you use glm() or glm.nb(), do not use the R confint() function for this question. confint() uses a different approach, profile likelihood, for glm() parameters.

We'll talk about profile likelihood in a few weeks.
2) You should get the same estimated mean but different confidence intervals for problems 1c and 1e.
3) Stats students (and perhaps others) should be able to explain why the two intervals are different. Explanation not required.

(f) Use a likelihood ratio test to test the null hypothesis of no difference in mean counts between the dry and the moist woodland. Continue to assume that the counts are approximately Negative Binomial and assume that the two woodlands have the same overdispersion parameter. Report the test statistic (twice the change in log likelihood) and the p-value.

(g) Estimate the log ratio, as log (mean # oaks in dry / mean # oaks in moist woodland). Use the asymptotic confidence interval formula to compute a 95% confidence interval for that log ratio.

(h) Convert the estimates in question 1g into an estimate of the ratio, mean # oaks in dry / mean # oaks in moist woodland, and a 95% confidence interval for that ratio.

2. The data in worms.csv are from an ecotoxicology study evaluating the effect of a soil fungicide on earthworms. Soil fungicides are designed to kill pathogenic fungi in soils, but they may also kill earthworms, which are beneficial creatures. The study followed a before-after-control-impact (BACI) design. BACI designs are commonly used to evaluate environmental impacts. This study used the simplest BACI design, a 2 x 2 factorial. The treatment factor denotes whether the fungicide (D) or a water control (A) was applied to the plot. The time factor denotes whether a plot ws measured before (pre) or after (post) application of the treatment. There are 9 species of earthworms found in the study area. The response is the number of each species in each plot. The 9 variables (Lt, Lr, ..., Acc) are names by the first letters of the genus and species. The full names are not important. Because the measurement method is destructive, each plot was measured only once. Hence the four combinations of time and treatment are randomly assigned to plots. (Many BACI studies measure the same plot pre- and post-treatment. That wasn't possible here; that means there are no repeated measures to worry about). There are data for 60 plots: 30 each for fungicide and water control, with 10 measured pre-treatment and 20 measured post-treatment.

   The time x treatment interaction is the important quantity in a BACI analysis. If the fungicide has an effect, you expect the mean change (post - pre) in the treatment plots to differ from that in the control polots.

   The group variable identifies each unique combination of treatment and time. This variable has 4 levels (A/post, A/pre, D/post, and D/pre).

   (a) Plot the mean-variance relationship for each combination of species and group. Using this plot, choose the more appropriate distribution, Poisson or Negative Binomial, to model the data. Your answer is the plot, your choice, and explanation for your choice.

   (b) Fit both the Poisson and Negative Binomial models to the data. Use AIC or lnL (your choice) to assess which distribution is more appropriate. Remember that manyglm() requires one distribution for all species, so you can't choose different distributions for each species. Your answer is your choice of distribution and an explanation, supported by numbers, for your choice.

   (c) Do the data provide evidence of a time x treatment interaction? Your choice of inference method, but remember that likelihood-based inference (AIC or LRT) is preferred to Wald inference. Report your answer, including supporting numeric values.

(d) Think about the time effect in this model. Don't forget to include the interaction in your thinking, even if it is small. Does it make sense to evaluate whether there is an effect of time? In other words, does the time effect answer a biologically relevant question? Explain why or why not.

(e) Identify which species show evidence of a change over time. Include appropriate numbers and your explanation of how you used those numbers to identify those species.

(f) Think about the treatment effect in this model. Don't forget to include the interaction in your thinking, even if it is small. Does it make sense to evaluate whether there is an effect of treatment? In other words, does the treatment effect answer a biologically relevant question? Explain why or why not.