Lab 9 (3 points)

ENVIRON 710

Introduction

In this lab, we will fit a series of logistic and Poisson regression models to understand associations between forest health, species diversity, and various different covariates. A description of the variables in the data set `forest_health.csv` are as follows:

agb: Above Ground Biomass in Megagrams/hectare (Mg/ha)
growth: A categorical variable with two values: "Old Growth" and "New Growth"
state: A categorical variable with the following values: "CA","CO","ME","NC","WA"
basal_area: A numeric variable measured in meters squared (m^2)
temp_anomaly: A numeric variable measured in degrees C, representing long-term changes in the mean temperature from 1900-1950. nspecies: A numeric variable that measures the number of unique plant species within each sampled area area: the area of each forest sample (in km^2) disease: A binary variable that measures whether or not a specific disease was detected in the forest (coded as 1) or if the forest does not have the disease (coded as 0)

Fitting a generalized linear model follows the same general form as if fitting a linear model in R. Specifically, you can fit a logistic regression model in R using one of the two following lines of code:

```
fit1 <- glm(y~x1 + x2 + x3, family="binomial",data=df)

# OR

fit2 <- glm(y~x1 + x2 + x3, family="binomial",weights=ntrials,data=df)
```

The first model (fit1) corresponds to a logistic regression in which the outcome is binary (0 or 1). Hence, each row of the outcome variable corresponds to the result of a Bernoulli trial. The second model (fit2) corresponds to a series of Bernoulli trials that were conducted in each row of data. For example, suppose that the outcome were the number of diseased trees in a given forest stand (row of data), and that you determined that y_i trees were diseased out of N_i trials, where the index i denotes the row of the data. In this latter case, you are fitting a logistic regression in which you are accounting for the number of trials in each sampled forest.

In this lab, we will work with the code corresponding to fit1. In this case, we are only interested in whether or not we found the disease in the sampled forest or not.

For the following questions, be sure to provide code documenting how you arrived at the answer you did.

Question 1 (0.3 points)

Start by fitting an intercept-only model: **disease** $= \beta_0$ Call this model 'fit_1'.

a. Report the estimated value of β_0. What does it correspond to?

b. Given the estimate of the intercept, what is the predicted probability (i.e. $\hat{p}$) of finding the disease in a forest?

c. What are the odds of finding the disease in a forest?

Question 2 (0.4 points)

Now suppose you hypothesize that disease status differs between old-growth and new-growth forests. Create a binary indicator variable from the growth variable. Code this variable as 1 if the sample is from an old-growth forest, and 0 otherwise. Next, expand on the model above by including this new indicator variable, as follows: $\textbf{disease} = \beta_0 + old.growth\beta_1$. Call this model 'fit_2'.

a. Based on your estimates from the model, what is the predicted probability of finding the disease in an old-growth forest?

b. Based on your estimates from the model, What is the predicted probability of finding the disease in a new-growth forest?

c. Calculate the ratio of the probability of disease in new-growth forests versus the probability of disease in old-growth forests.

d. Now calculate the Odds Ratio for old-growth forest and interpret the effect in terms of the odds of finding the disease in old- versus new-growth forest. Is this the same as your answer for c.? Why or why not?

Question 3 (0.4 points)

Now suppose you are interested in building on your model to understand whether or not basal area (centered on its mean) is associated with disease status, independent of old-growth status. Now fit a logistic regression of the following form: $\textbf{disease} = \beta_0 + +old.growth\beta_1 + basal.area.centered\beta_2$. Call this model 'fit_3'

a. Calculate the Odds Ratio for the effect of basal area (centered) on disease status and interpret its meaning (not in terms of a hypothesis test, but in terms of expected changes in the odds of disease).

b. Make a plot of the above regression model. The y-axis should consist of the log odds of disease. The x-axis should consist of the centered basal area. Be sure to color each regression line by old- versus new-growth forest. Comment on this plot. Are the two regression lines parallel? If so, what is the magnitude of the distance between the two parallel lines?

c. Repeat the above plotting exercise, but instead of having the y-axis be the log odds of disease, the y-axis should be the odds of disease. Again, the x-axis should consist of the centered basal area, and you should use a different color for old- versus new-growth forest. Comment on this plot. What is the magnitude of the distance between the two lines?

d. Repeat the above plotting exercise, but instead of having the y-axis be the odds of disease, the y-axis should be the probability of disease. Again, the x-axis should consist of the centered basal area, and you should use a different color for old- versus new-growth forest. Comment on this plot. Does the change in probability of disease change constantly as a function of basal area? Why or why not?

Question 4 (0.2 points)

Now suppose you wish to better understand whether temperature anomalies are associated with the disease status of forests. To do this, you fit the following regression model: $\textbf{disease} = \beta_0 + old.growth\beta_1 + basal.area.centered\beta_2 + temp.anomaly\beta_3$. Call this model 'fit_4'.

a. Calculate the Odds Ratio for the effect of temperature anomalies on disease status and interpret its meaning as in prior questions.

b. Now add to this model by creating an interaction term between old-growth forest and the temperature anomaly covariate. Fit a new model of the form: $\textbf{disease} = \beta_0 + old.growth\beta_1 + basal.area.centered\beta_2 +$

$temp.anomaly\beta_3 + old.temp.anomaly\beta_4$. Call this model 'fit_4b'. Calculate the Odds Ratio for the effects of temperature anomalies on odds of disease for both new-growth forests and old-growth forests. Interpret your findings.

Question 5 (0.2 points)

Now you are interested in understanding the level of biodiversity in the forests that you sampled. During your survey, you recorded the number of unique plant species (nspecies) identified. You also recorded the area (in square kilometers) of the forest area that you sampled.

To understand biodiversity, you will fit Poisson regression models to understand how different covariates influence the number of species (nspecies) per sampled area. You can fit a Poisson regression model (with two covariates: x1 and x2) in R using the following code:

```
fit <- glm(nspecies~x1 + x2,data=df,family="poisson",offset=log(area))
```

Recall from lecture that we include an offset term to account for the different areas sampled (intuitively this should make sense: the more area you sample, the more species you would expect to find, so you need to account for that). Start by fitting a model with just an intercept, as follows: $y = \beta_0$. Call this model 'fit_5'.

a. What is the estimated value of the intercept? What does it correspond to?

b. Given the estimate of the intercept, what is the expected number of species per square kilometer (i.e. $\hat{\lambda}$) across all the forests sampled? How does this compare to the mean number of species per forest area in your data set?

Question 6 (0.4 points)

Now suppose you are interested in understanding the effect of old- versus new-growth forest on species diversity. Now fit a model of the following general form: $y = \beta_0 + old\beta_1$. Call this model 'fit_6'.

a. What is the estimated value of the intercept $\hat{\beta}_0$ now? What does it correspond to?

b. What is the estimated value of $\hat{\beta}_1$? What does it correspond to? Interpret the effect of old-growth versus new-growth forest on the number of species per square kilometer.

c. Given the estimate of $\hat{\beta}_0$ and $\hat{\beta}_1$, what is the expected number of species per square kilometer for old- and new-growth forests?

d. Make a plot with the expected number of species per square kilometer on the y-axis, and old-growth and new-growth on the x-axis.

Question 7 (0.5 points)

Now you wish to extend your analysis by investigating the association between basal area and species diversity.

a. Make a scatterplot with basal area on the x-axis and the number of species on the y-axis. Do you see any clear patterns?

b. The above plot did not take into account the different sizes of the forest areas sampled. Now, remake this plot, but replace the number of species on the y-axis with the number of species per square kilometer. How did the plot change? Do you see any clear patterns now? Describe them.

c. Fit the following model: $y = \beta_0 + old\beta_1 + basal.area.centered\beta_2$. Call this model 'fit_7'. Interpret the effect of basal area on species diversity. Does diversity increase with increasing basal area, or does it decrease, and by how much?

d. Make a scatterplot with the log of the number of species per square kilometer from the model on the y-axis and the observed basal area on the x-axis. Include the regression lines for both old- and new-growth forest. Be sure to use separate colors to indicate old- versus new-growth forest.

e. Now make a scatterplot with the number of species per square kilometer from the model on the y-axis and the observed basal area on the x-axis. Include regression lines for both old- and new-growth forest. Be sure to use separate colors to indicate old- versus new-growth forest.

Question 8 (0.6 points)

Now you wish to include the possibility that temperature anomalies have an impact on species diversity.

a. Make a scatterplot with the number of species per square kilometer on the y-axis and temperature anomalies on the x-axis. Be sure to use separate colors to indicate old- versus new-growth forest. Comment on any patterns you observe. Does species diversity appear to increase or decrease with increasing temperature anomalies?

b. Fit a Poisson regression model of the following form: $y = \beta_0 + old\beta_1 + basal.area.centered\beta_2 + temp.anomaly\beta_3$. Call this model 'fit_8'. Interpret the effect of a 1-degree increase in temperature on species diversity. Does species diversity increase or decrease as a function of temperature?

c. Make a scatterplot with the number of species per square kilometer from the model on the y-axis and temperature anomalies on the x-axis. Include regression lines for old-growth versus new-growth forest, and be sure to use separate colors to indicate old- versus new-growth forest. Comment on any patterns you observe. Does species diversity appear to increase or decrease with increasing temperature anomalies? Is this consistent with what you found in the plots of the raw data?