Lab 8

ENVIRON 710

Introduction

In this lab, we will fit a series of logistic regression models to understand associations between forest health 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 deviations from the 20-year mean.

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:

```r
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 (0.5 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?

Question 2 (1 point)

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: **disease** $= \beta_0 + old.growth\beta_1$. Call this model 'fit_2'.

a. Report the estimate of β_0. What does it represent, and how does it compare to the estimate from question 1?
b. Based on your estimates from the model, what is the predicted probability of finding the disease in an old-growth forest?
c. Based on your estimates from the model, What is the predicted probability of finding the disease in a new-growth forest?
d. Calculate the Odds Ratio for old-growth forest and interpret the effect of old- versus new-growth forest on the odds of finding the disease.
e. Conduct a hypothesis test in which the null hypothesis is that there is no difference in disease status between new-growth and old-growth forests.

Question 3 (1 point)

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: **disease** $= \beta_0 + +old.growth\beta_1 + basal.area\beta_2$. 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 substantively).
b. Conduct a hypothesis test for the effect of basal area (centered) on disease status under the null hypothesis of no effect.

Question 4 (0.5 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: **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. Based on the model you just fit, what factors would you conclude are associated with increased odds of disease, and what factors would you include are associated with decreased odds of disease? Why? Hint: do not simply note which variables were "statistically significant", since statistical significance is not a measure of scientific importance.