Project

Acacia, Kate, Julia

3/23/2021

Scenario

Genes A and B have both been shown to be related to lifetime chance of cancer. We want to know if an organism with gene A has the same chance of cancer as an organism with gene B, and how that differs when an organism has both genes A and B, or neither gene. We need to include the covariate of the organism's mass, because we know this also affects the lifetime chance of cancer.

We sample N organisms of each of the four genotypes, and we record their body mass and whether or not cancer is present in that organism.

Story

Your lab is studying cancer-related genes using feline mammary carcinomas (FMCs) to both validate these tumors as models for human breast cancer (HBC) studies and to improve small animal veterinary practice. FMCs have been emerging as valuable models for human breast cancer, and the domestic cat is highly affected by spontaneous mammary tumors (Ferreira et al., 2019).

The cancer-related genes A and B are conserved between cat and human. Both genes have been shown to be related to lifetime incidence of HBC, but they have not been studied in cats. You want to know if a cat with gene A has the same lifetime chance of developing a FMC as a cat with gene B, and how that differs when an organism has both genes A and B, or neither gene (wild type).

A cat's breed, sex, age, and whether or not it is intact all greatly influence their lifetime risk of FMCs (Ohio State University Veterinary Medical Center, 2021). To control for these factors, you use only intact female domestic shorthair cats that are 10 years and older (which show the highest incidence of FMCs) as study subjects. A cat's mass affects their lifetime chance of cancer as well.

Working with the MSU Small Animal Clinic, you recruit 100 participants for the study: 25 with gene A, 25 with gene B, 25 with both genes A and B, and 25 with neither gene. For each cat, you record its genotype, mass at time of death and whether or not it had a FMC in its lifetime. Your null hypothesis is that the lifetime chance of developing a FMC is equal across all four genotypes. Your alternative hypothesis is that the lifetime chance of developing a FMC is different between genotypes.

Ferreira D, Martins B, Soares M, Correia J, Adega F, et al. (2019) Gene expression association study in feline mammary carcinomas. PLOS ONE 14(8): e0221776. https://doi.org/10.1371/journal.pone.0221776

Ohio State University Veterinary Medical Center (2021) Feline Mammary Tumors. Retrieved from https://vet.osu.edu/vmc/companion/our-services/oncology-and-hematology/common-tumor-types/feline-mammary-tumors

Hypothesis

Null Hypothesis: The probability of cancer is equal across all genotypes.

Alternative Hypothesis 1 (no interaction effect): A, B, and AB all show a higher probability of cancer than WT, where A+B = AB.

Alternative Hypothesis 2 (interaction effect increases probability): A, B, and AB all show a higher probability of cancer than WT, where A + B > AB.

Alternative Hypothesis 3 (interaction effect decreases probability): A, B, and AB all show a higher probability of cancer than WT, where A + B < AB.

Variables

Genotypes: WT (neither A nor B), A, B, and AB. A and B each have an incidence rate for cancer.

Mass: a continuous variable between 7 and 15 lbs

Genotype A x Genotype B: Interaction effects

Response variable: Presence or absence of cancer in the organism (0/1)

Predictor variables: Genotype A, Genotype B, interaction between Genotypes A & B, and organism mass

Modeling and Justification

Binomial Distribution

$y \sim Bin(p, N)$

- p is probability of cancer
- N is total number of individuals

Justification: We measure cancer as either present/ absent in each organism. The data simulation evolves N multicellular individuals. We have fixed N number of trials that we can repeat *ad infinitum*. This is a frequentists' approach. Hence, we chose the Binomial distribution.

Deterministic Function (model) & Joint Probability (likelihood)

$counts = \alpha + \beta_1 * genotype_A + \beta_2 * genotype_B + \delta * genotype_A * genotype_B + \gamma * mass + \epsilon$

- $counts$ = expected number of organisms with cancer present
- α = (intercept) – expected incidence rate of cancer in wildtype reference
- β_1 = constant for genotype_A (how much gene A influences incidence of cancer)
- β_2 = constant for genotype_B (how much gene B influences incidence of cancer)
- δ = slope term indicating the interaction between genes A and B (how much genes A and B combined influence incidence of cancer)
- γ = constant for mass (how much organism mass influences incidence of cancer)
- ϵ = randomness pulled from binomial distribution
  - $\epsilon \sim Bin(p, N)$

Simulate the Data

```
#### Number of Datapoints #### Assume a balanced design 2 genes - A and B Arbitrary
#### number of replicates performed
n.reps <- 80  #50   #changed to 80 so that 4 divides evenly into it

# Total number of data points
n <- n.reps

# For a binomial distribution, we need to know the number of organisms in each
# replicate In our work we usually use the same number of organsims in each
# replicate
n.orgs <- rep(100, n)
```

```r
#### Data Simulation ####

# Specify a categorical variable which indicates genotype
genA <- factor(rep(c(0, 1), each = n/2))  # produces 0000...1111....
genB <- factor(rep(rep(c(0, 1), each = n/4), 2))  # produces 00...11...00...11...

# We also need a vector to indicate mass for each organism Assume this is a
# continuous variable between 7 and 15 pounds
mass <- round(runif(n, min = 7, max = 15), digits = 2)

# Chose the values for the parameters (logit transformed) Labeled to make it
# easier to think about Each element in beta.vec.names indicates the effect of
# that variable or interaction of variables
beta.vec.names <- c("WT", "genA", "genB", "genA:genB", "mass")
beta.vec <- c(0, 0.03, 0.01, 0.09, 0.02)  # we can freely edit these
names(beta.vec) <- beta.vec.names


#### Model Matrix Creation #### Build the design matrix of the interactive
#### combination of genotype and mass
Xmat = model.matrix(~genA * genB + mass)

#### Create Stochastic Data #### Generate the linear predictor (mutliple Xmat by the
#### beta.vec)
lin.pred = Xmat %*% beta.vec

# Transform the data with an inverse logit to get the expected proportion of
# cancerous samples
exp.p <- exp(lin.pred)/(1 + exp(lin.pred))

# Add binomial noise
cancer.counts <- rbinom(n = n, size = n.orgs, prob = exp.p)

#### Combine Data #### Combine type data, mass data, and cancer counts
df <- data.frame(genA, genB, mass, cancer.counts)

#### Export #### Export the data to a csv file
filename <- "cancer_data.csv"
write.csv(df, file = filename, row.names = FALSE)
```

Statistical Test

Plots

Analysis