Lesson Plan

- A **generalized linear model (GLM)** is a category of linear regressions models that allows for response variables that have error distributions other than the normal distribution, and response variables that are both numerical and non-numerical. The books lists five components of a GLM. What are they? (p185–Chapter 11 page)
  - A data vector
  - a linear predictor
  - a link function g
  - a data distribution
  - other parameters such as variances, overdispersions, cutpoints, etc.
- Hand out summaries, fill in
- For each example, say which model it is best suited for:
  1. A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, affect admission into graduate school.
     (a) Ans: Logit or Probit
  2. And if my data for this example had a few extreme observations, such as a kid that aced the GRE, went to an Ivy league with a 4.0 GPA but decided to go to find a job, or the kid that somehow networked his way into grad school despite dismal GRE and GPA scores.
     (a) Ans: Robit
  3. Suppose I want to model the number of flying-bomb hits in London during WWII, where the average number of hits to a particular neighborhood in London is very small but there are nearly years and years worth of events where each neighborhood could have been bombed. .
     (a) Ans: Poisson. I can divide London into X number of km squared areas, and count the number of bombs for each area over the course of the war, which is measured in N days. For each trial, or day, each area has a small probability of being hit, the probability that the same area will be hit twice in the same day is also very small, and whether one particular area was hit is independent of what happens in neighboring areas.

4. A teacher has decided to offer her students a gift at the end of the year. Each student can choose between an action game, a puzzle game, or a sports game, and she wants to know how each student's average grade in math, reading, and writing affected their choice of games. What model should she use?

   (a) Ans: Multinomial. The response variable is an unordered categorical variable, with three options.

- **Latent data**

  - What is latent data? Latent data is data that is not measured directly and cataloged in a study, but something that can inferred, and then modeled. And then latent variable models are models that attempt to explain the observed data in terms of the inferred variable.

    * So, for example, let's say I've sent a survey out to customers of a brand, and the questions are things like "On a scale of 1 to 10, would you repurchase the same product?" "Are you likely to purchase a different product from this brand?" "How positively do you regard this brand?" "How likely are you to recommend it to a friend?" These variables, in 1 to 10 scale, represent specific questions, but together I'm trying to use them to model something I can't directly measure: customer loyalty. How loyal are these customers to this brand?
    * It's also kind of similar to Principle Component Analysis. This example was brought up in another class, I think. If you are trying to relate finch diet to beak shape, beak shape is really kind of hard to measure and categorize on its own. So what do you do? You measure beak length, beak curvature, the ratio in height of the top part to the bottom part, and you throw all this into a PCA and you can infer "beak shape" through the conflagration of all these other variables.

- **Multinomial Example:**

  - Let's say I've been collecting data about whether students in a highschool will go into a general program, a vocational program, or an academic program based on attributes such as social status, channel type, awards, gender, economic status, and grades in reading, writing, math.
  - Start with a basic model: $program \sim ses(socio-economicstatus) + write$, where socio-economic status has three levels (low, medium, and high) and write is just a continuous grade.
  - My outcome looks like this:

Coefficients:	(Intercept)	sesmiddle	seshigh	write
general	2.852	-0.5334	-1.163	-0.05793
vocational	5.218	0.2914	-0.9827	-0.1136

  - So how do we interpret this?
    - The first step is to realize that everything here is a comparison to a baseline. In terms of the categorical y variables, the academic program is the baseline. In terms of socioeconomic status, it is low.
    - So it we're interpreting the baseline of low economic status and you're failing writing, you have an increase in log odds of 2.852 of choosing general program over academic, or 94.54% increase in the odds. Likewise with vocational, you have a 5.218 increase in the log odds, or a 99.46% increase in the odds that you'll choose vocational over academic.
    - A 1-unit increase in write will result in a 0.0214 decrease in log odds of going into the general program over the academic program. What is this in regular percentage change?
      - $\frac{e^{-0.0214}}{1+e^{-0.0214}} = -0.042$ or a 49.47% decrease in the odds that you'll choose general over academic for just one extra writing grade unit.
    - Doing the same for vocational, we can say that a 1 unit increase will decrease the log odds by 0.1136 or by -47.16%
    - Now let's jump back to socio-economic status. This is a little tricky to interpret, because the ses value low is also not there. You can interpret the coefficient for sesmiddle as, if I changed from a low socio-economic status to a middle status, how would the odds change for choosing general over academic? And the answer is that you would see a decrease in the odds of choosing the general program but an increase in choosing the vocational. Going from low to high status would result in students choosing academic over both general and vocational more often.

- **Poisson Example:**
  - Let's say we've collected data on horseshoe crabs. Specifically, in the study, each female horseshoe crab had a male crab attached to her in the nest, and looked at if there were any satellite males handing nearby. They measured female color, weight, and carapace width as predictors, and looked at 173 females.
  - So the model is $numSatellites \sim 1 + carapaceWidth + math$
  - My outcome looks like this:

Coefficients:	Estimate	Std. Error
(Intercept)	-3.305	0.5422
Carapace Width	0.1641	0.019973

* For each unit increase in carapace width, with see an 0.16 increase in the log odds, or a $e^(0.1641) = 1.18$, or 118% increase in the number of satellite males per female.

Chapter 10 Questions

- **What is the equation that corresponds to the inverse logit function?**
  - Ans p156: $logit^{-1}(x) = \frac{e^x}{1+e^x}$
- **What are the first two tricks when interpreting a logit fit summary table?**
  - Ans: p169
  - evaluating the intercept at the mean of the data
    - Follow-Up:**Why the mean and not the zero point?**
      - Ans p158: Zero may not be interesting or in the case of categorical variables, not even in the model
  - The "Divide by 4" rule to get an approximate predictive difference in the probability for each unit change in predictor variables
    - Follow-Up:**What is the equation for the actual upper bound predictive difference?**
      - Ans p158: $\frac{\beta e^x}{(1+e^x)^2}$
  - A bonus third tip is to evaluate the model using centered inputs. Thus, the coefficients at the zero points already represent the interpretation at the mean.
- **What are the primary requirements for running a logistic regression?**
  - Ans p155: That the response variable be binary and that there are no outliers in the data
    - Exception: logit and probit can be expanded for ordered and unordered categorical data
- **In logistic regression, what can cause a parameter to be nonidentified?**
  - Ans p177: If it is co-linear with another predictor or if the predictor is completely aligned with the outcome (y=1 for x>T, y=0 for x<T)

Chapter 11 Questions

- **What are the similarities and differences between the binomial and poissonian models?**
  - Ans p189: They both measure the number of certain random events within a certain time/space frame. However, binomial is based on discrete events, while Poisson is based on continuous events.
  - Poisson models can be thought of as a Binomial with a very large n (number of attempts) and very small p (probability of success)
- **What is overdispersion and how do you account for it?**
  - Ans p187: Overdispersion of a Poisson model occurs when there is greater variation in the observed data than is expected. You can account for overdispersion by multiplying the standard errors of the coefficients by the square-root of the overdispersion parameter ϖ.
  - Follow-Up: **What is an easy calculation to check dispersion?**
    - Ans p187: Using the ratio of the variance to the mean:
      - $\frac{\sigma}{\mu} = 1$, there is no dispersion
      - $\frac{\sigma}{\mu} > 1$, there is overdispersion
      - $\frac{\sigma}{\mu} < 1$, there is underdispersion
- **How is robit different from logit?**
  - Ans p200: It uses the Student-t distribution of errors rather than the normal distribution, so it is better qualified to deal with outliers in the data. This is due to the variance parameter incorporated into the distribution–as the user, you can widen it to make the 95% CI include the outliers.