Intro to Bayesian Statistics

We are normally taught the frequentist approach to statistics, in which we assume that our population parameters are fixed but unknown quantities. We can estimate these parameters using samples from a population, but different samples give us different estimates

* The distribution of these different estimates is called the sampling distribution
* The sampling distribution quantifies the uncertainty of the parameter estimate, but the parameter is still considered fixed

In contrast, Bayes assumes that population parameters are random variables, rather than fixed quantities. These can be described using a probability distribution, where probability is our degree of belief

Bayes’ Law for Discrete Hypotheses

* This mathematical expression about the belief of our parameter is called the prior distribution
* In other words, the prior distribution is the probability distribution of what we think the parameter is, which we estimate before we actually see the data
* Bayesian statistics allows us to update our belief about the population parameter based on the results of our experiment, resulting in the posterior distribution
  + The likelihood is what you sample and actually collect data on
  + If you use uninformative priors, the posterior distribution will be identical to the likelihood distribution. Larger sample sizes will result in the likelihood distribution having greater and greater effect on the posterior

Posterior = Prior x Likelihood

P(θ|y) = P(θ)P(y|θ), where θ is the population parameter. The posterior and prior distribution will always be of the same distribution family

In Bayesian modeling, you want to test the probability of your data given the model you’ve created. You are asking what is the probability of the covariate coefficients that you’ve estimated. Assumptions are

* Whatever assumptions we made for the frequentist model, since those are the assumptions about the likelihood function
* That the likelihood function we’ve chosen is a reasonable representation of the data
* We do not make assumptions about our priors, since our priors themselves are *a priori* assumptions we’ve made about the parameters

Meaning of a 95% confidence interval

* In frequentist statistics, 95% CI means that 95 out of 100 replications of exactly the same experiment capture the fixed but unknown regression coefficient
* In Bayesian statistics, 95% CI means that there is a 95% chance that the regression coefficient resides between the upper and lower limit

Intro to Logistic Regression

Logistic regression is a regression analysis used to predict the probability of a binary outcome, such as win/lose or attractive/unattractive (streams). It fits an S-shaped curve which ranges from 0-1 on the y-axis

* Instead of comparing simple models to more complex models like you would in linear regression, you test to see if the variable’s effect on the prediction is significantly different from 0 using *Wald’s Test*
* Logistic regression does not have the same concept of residuals as in linear regression, so you can’t evaluate the fit using R2. Instead, you use something called *maximum likelihood* 
  + The goal of maximum likelihood is to find the optimal way to fit a distribution to the data by finding an estimate of the parameters that maximizes the likelihood of observing the data. You basically plot a bunch of values for the parameter you are estimating from different placements of your chosen distribution on your data. Then, you compare the likelihood of all of these parameter estimates given the data and choose the maximum likelihood, or the estimate and associated distribution placement that maximizes the probability of observing your parameter estimate given the data
  + Maximum likelihood is good for comparing models
* In the logit model, the log odds of the outcome is modeled as a linear combination of the predictor variables
* Logit models assume that observations are independent of one another (no matching or repeated measurements), that there is no collinearity, and that the independent variables are linearly related to the log odds

Random vs fixed effects

Fixed effects are what you are interested in and random effects are what you are controlling for. When comparing blood pressure between smokers and nonsmokers for example, smoking would be a fixed effect, but participant ID and the different doctors taking the blood pressure measurements would be random effects. This is because different study participants will naturally have lower blood pressure than others independent of whether or not they smoke, and different doctors may have variable ability to measure blood pressure accurately. You are not interested in the doctors or the participants’ natural blood pressure variability, so these are accounted for with random effects

* In your study, *year* is a random effect because different years may be slightly different than others for reasons not accounted for in your model, and you are not interested in year effect
* *Stream* could also be a random effect because the sites in your study were selected out of a pool of potential sites at random. We did not expect the attractive sites like Sawmill Creek to be highly attractive *a priori*
  + Update 1/18/21: Year is my random effect, not stream. I cannot have both due to limited degrees of freedom and year makes more sense because I want to look at trends in attractiveness over time, not within years. To deal with the stream issue (not including it as a random or fixed effect), I will evaluate my results at the stream level. I.e., summarize results/take averages by stream, thereby assuming that there are stream (site) specific effects
* Random effects are centered around 0 and have a normal distribution. More information on random intercepts vs effects is included in the model code for my thesis project (*Model\_fitting2.R*)

Summary of Curry’s Nov. 23, 2021 email

* Curry recommends building my model as a mixed-effects model first to see if it is estimable given the unbalanced predictions. I may run out of degrees of freedom rather quickly given the number of covariates I am proposing
  + To build this as a Bayesian model, there are two options. 1) is to use restrictive (informative) priors that penalize your covariates towards zero (no effect) unless there is strong evidence from the data. 2) is to use rather uninformative (hyper)priors for the mean and SD (hyperparameters) of your prior distribution for covariate effects
  + Either way, does that mean that the aforementioned mixed model that I should build is the likelihood function that goes into the Bayesian modeling setup? Or is it entirely separate and Curry is saying that I shouldn’t use Bayes at all?
* My next step should be to write out my model structure, which should include answering the following questions:
  + What are the groups within my data? What is my unit of replication I will use to define my statistical relationships (stream or year)?
  + What information do I want to share across model groups? Why do I want to share information across model groups?
* Example: Stream flow is observed across streams (s) and years (y), so a hierarchical structure would be advantageous in which there is a separate relationship between flow and stray proportion (p\_s,y) by location and year
  + p\_s,y = intercept\_s + effect\_s \* flow\_s,y + e
* Now in this context, both location (stream) specific intercept (Intercept\_s) and effect of flow on straying (Effect\_s), differ by stream s
* In the hierarchical structure we could assume that the stream-specific flow effects arise from a common normal distribution: **Effect\_s ~ Normal(mu\_flow, sd\_flow),** wherein we estimate the hyperparameters (mu\_flow and sd\_flow) which describe the mean and variation in the effect of flow across streams
  + Why can we assume a normal distribution? The size of your effect is coming from a population of potential effect sizes for flow which are approx. normally distributed
* We would then need to define hyperpriors for mu\_flow and sd\_flow.
* The hyperprior for average flow should be centered at zero (null hypothesis of no effect), and have a standard deviation equal to something: **mu\_flow ~ Normal(0, sd\_something)**
  + Hyperpriors are not site-specific, these are the overall parameters for all streams that you are estimating
* if sd\_something is very low, you are penalizing everything toward zero (no effect), or you could have the a fairly broad/uninformative prior on sd\_something and let the data dictate its value
  + So, I use my data to dictate the sd for a prior?
* sd\_something ~ Uniform(0,10), or even better a broad normal distribution truncated at zero: sd\_something ~ Normal(0,10)

Defining the model structure (preliminary version of model, not final)

*Py,s* = probability of straying into stream *s* in year *y*

*β0,y* = randomly varying intercept

*Zy* = random effect of year (see notes for why only year is random effect)

*βs,y* = stream (categorical)

*βd,y* = effect of stream distance to the nearest release site

*βh,y* = effect of stream distance to the nearest hatchery

*βn,y* = effect of number of fish released within 40km of stream

*βca,y* = effect of conspecific abundance (recipient population size) in stream

*βcd,y* = effect of conspecific density in stream

*βpa,y* = effect of Pink Salmon abundance in stream

*βpd,y* = effect of Pink Salmon density in stream

*βa,y* = effect of average seasonal stream discharge

*βc,y* = effect of the coefficient of variation of seasonal stream discharge

*βsh,y* = effect of stream hydrology (categorical)

*βl,y* = effect of % watershed made up of lakes and wetland

*βf,y* = effect of fishing pressure on stream straying index

#except for stream and stream hydrology (*βs,y*), all covariates are continuous and will be z-scored (within each year if you have year as your random effect)

Hierarchy

#varies by year and stream

#varies by stream, not by year

Some years there are more strays than others and/or control for issue of non-independence of years

Hierarchical models

First, what does it mean to say that a model is hierarchical?

* In a hierarchical model, lower levels are sorted under a hierarchy of successively higher-level units. In other words, data is grouped into clusters at one or more levels, and the influence of the clusters on the data points in each cluster is taken into account in statistical analyses (usually by including random effects).
  + Example: Children are more likely to share worldviews with their brothers and sisters than they would with children from other families. So, in the analysis, children would be grouped into clusters of their family
  + In your study, fish in the same year are probably going to be more similar to one another than fish in other years, so fish will be grouped into year clusters

What are the advantages of hierarchical modeling?

* **You avoid statistical errors associated with assuming observations are independent when in reality they aren’t.** Fish sampled in the same year may not be independent from one another, children growing up in the same house may not have totally independent worldviews, etc.. You need to account for this non-independence by creating hierarchies. Also
  + Pool information across small sample sizes 🡪 You may need to share information amongst streams for streams which are lacking in data, particularly those that were only sampled once. The way this would work is those streams would end up with responses that revert back to the overall mean
    - Example: You are almost always going to have a greater range of data in a classroom with more test scores (more high and low outliers) because of the law of large numbers. So, if you treat classroom as a random effect, you can pool shared information about the means across classrooms
  + If you want to predict out of your system to other regions in which you don’t have data (like you’re going to do in chapter 2), you will have a better understanding of existing relationships and we can preserve some degrees of freedom because we don’t have to estimate a specific effect for each stream for each covariate

What is my random effect? Pros and cons of stream vs year:

**Objective 1 Objective 2**

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Stream 1 in green

Stream 2 in blue

|  |  |
| --- | --- |
| * Year as random effect (more useful for making generalizations about the covariate effects over time and then predicting out to places where you don’t have data, i.e., Chapter 2) * Points show stream data in a single year * Understand stray-flow relationship across streams and how it varies among years * You need to have multiple streams observed in each year, which you do! * **Z-score within streams among years** * *Logit(p) ~ Flow + (Flow|Year)* OR *logit(p) ~ Year + Year:Flow* | * Stream as random effect (more useful for making statements about specific streams for which you already have eco. data) * Points show stream data for all years * Understand how stray-flow relationship across time varies among streams * You need to have multiple years of observation for each stream, which you don’t for a few streams * *Logit(p) ~ Stream + Stream:Flow* OR *logit(p) ~ Flow + (Flow|Stream)* |

**Conclusion:** *Year* will be the random effect, with the unit of replication at the stream level. By doing this, I can look across the range of streams I have and compare how these streams rank relative to one another amongst years. Making year a random effect will also account for the variation in surveying among years. This means I need to z-score within streams across years, except for the covariates that don’t vary over time (flow), you can scale these within a year to reduce their SD

Notes from committee meeting (12/7/20):

* We can replicate a hierarchical model using a mixed effects structure. Curry recommends trying to fit and run mixed effects model first and then trying to implement them in a Bayesian context
* Year as a random effect (RE) can mean 2 things: When you talk about year as a RE, you are saying that your intercept is varying randomly as a function of year. We can also assume that the effect of one of our covariates varies randomly as a function of year
  + Across which axis of observation do we assume the effect of covariate X (say, flow, for example) is drawn? Implicit in our hierarchy, we have some distribution of effects for flow. We are either quantifying the distribution of that flow effect across streams or across years. If we do it across years, we are assuming an underlying relationship between observations at the stream level and attractiveness as a function of the stream-specific flow
* Due to your difference in sampling pressure, you need to take the average, rather than the total number of strays, for calculating straying index for each stream
  + So, for a given stream, calculate the average number of strays and divide it by the sum total of average calculated for each stream (not the total number of strays in SEAK). Done. This column is now included in the dataset that I am using to model. For each stream, it is the Avg\_number\_strays, which is calculated as the number of strays detected divided by the number of surveys. I’ve summed this column within each subregion for the Sum\_avg\_strays\_by\_subregion column. Straying index is calculated as Avg\_number\_strays divided by Sum\_avg\_strays\_by\_subregion
  + Possibly include a plot in supplementary material showing timing vs number of strays detected

Notes from actually trying to fit the mixed effects model (December 2020)

On scaling

* I will scale all of my continuous covariates because they are measured on different scales and this could affect the results (? 🡪 considerable uncertainty on this topic though, read below)
  + Variables measured on different scales do not give equal contribution to the analysis. For instance, a variable that ranges between $100 and $10000 will carry more weight than a variable that ranges between $0 and $30. This is definitely a problem in your dataset (e.g., CV\_flow ranges between 0.4 and 0.66 while Pink\_density ranges between 0 and 105973.668!)
  + By scaling, your variables will have equal range and/or variance (source: <https://www.listendata.com/2017/04/how-to-standardize-variable-in-regression.html>)
  + But there seems to be a lack of consensus on this, that is, scaling may not be necessary (but it also can’t hurt). Scaling can speed up model convergence, but it makes interpretation more challenging and isn’t necessary because the betas are estimated by converting the units of each explanatory variable into the units of the response variable. See this stack exchange forum: <https://stats.stackexchange.com/questions/29781/when-conducting-multiple-regression-when-should-you-center-your-predictor-varia>
  + As Curry mentioned earlier, I will scale within years across streams because year is my random effect
  + Also, when I tried to fit the model WITHOUT scaling, I got a lot of warnings
* Scale within or across years?
  + Only z-score within a year if you are treating each year as truly independent, which you decide based on your unit of replication -> what is your unit of replication, stream or year? It is stream! Year is RE
  + For time-independent predictors (hydrology type, stream width, etc), z-score across streams
  + For time-varying predictors, z-score across all observations (years and streams)
  + I think the reason for this is that the time-independent predictors only vary across streams and are the same across all years. So, to make things simpler, you can just set a value (z-score) for each stream in one year and then copy and paste those values into all subsequent years. Doing this within one year instead of across all years also makes the SD slightly smaller, which is advantageous because that means there is less variance around your z-scores for the time-independent predictors (as there should be because they are always constant within a stream)
  + See Curry email 1/29/21 (not sure how helpful it is in explaining but at least shows the source of this information^^)

Update (5/30/21): The problem with scaling within a year for the time-invariant predictors (mean\_flow and CV\_flow) as opposed to across years like I did for the other covariates is that it makes streams comparable within a year, but not necessarily across years. I found when I did this that it made the extreme values more extreme, causing out-of-whack model predictions. For instance, take a stream with a high CV\_flow like Sawmill Creek (CV\_flow = 0.63). This high value results in it having significantly above average z-scores in general as it should, but especially so in a year where the other streams have lower CV\_flow on average. This was the case in 2011, when there were fewer streams in the dataset, and it gave CV\_flow for Sawmill an extremely large z-score of 2.54, resulting in a model-predicted 420 strays later on. The photo below shows the z-scores for Sawmill Creek where mean\_flow and CV\_flow were scaled within years.

Thus, if I z-score within years and not across years like I did for the time-varying predictors, then my values for those predictors are dependent on what other streams are included in the model in that particular year, and I don’t think this is correct. I want to make predictions based on data across ALL years. So, despite Curry’s recommendation to z-score within years for time-invariant covariates, I will instead z-score across years.

Graphical user interface, application

Description automatically generated



Interpretation of model covariates

*#Note that I refer to “SI” as my response variable below, which was my old model response variable (straying index = # strays in stream/ total # strays in region). I am no longer using this response variable but the interpretation of the following covariates below should be the same with my new response variable, the average number of strays*

Hydrology\_Type

* If one stream type has a stronger relationship with SI than others, then that stream type is more attractive to H strays than others
  + Snowmelt
    - Snowmelt-fed streams have more predictable and less variable flow patterns than rainfall-fed streams (Gordon et al. 2004 stream ecology book)
    - Colder water temperatures and less thermal sensitivity to the air (Lisi et al. 2015)
  + Rain
    - Generally more flow overall than snowmelt streams (Berghuijs et al. 2014)
  + Rain-snow
    - Rain-snow streams are cooler than rain-fed streams, though I don’t think quite as cold as snowmelt streams (Leach and Moore 2014)

Fishery\_harvest

* If fishery harvest is inversely related to SI, then fishery removal of potential strays is most likely occurring and influencing straying rates
* If fishery harvest is directly related to SI, then fishery removal of potential strays is probably not affecting straying indices

Cons\_Abundance

* If conspecific abundance is directly related to SI, then presence of other Chum salmon in the stream *is* attractive to hatchery strays

Pink\_Abundance

* If conspecific abundance is directly related to SI, then presence of Pink salmon in the stream *is* attractive to hatchery strays

WMA\_Releases\_by\_Yr

* If Chum salmon release volume is directly related to SI, then larger hatchery Chum salmon releases *near* streams results in more strays

Dist\_nearest\_R

* If distance to the nearest hatchery is directly related to SI, then being close to a release site means that a stream is more attractive to potential Chum salmon strays

percent\_lake\_wetL\_A

* If the percentage of a stream watershed made up of lakes and wetlands is directly related to SI, then a more stable stream hydrograph is more attractive to strays because lakes and wetlands serve to regulate streamflow

mean\_flow

* If average stream flow is directly related to SI, then larger streams (measured by discharge) are more attractive to hatchery strays

CV\_flow

* If the CV of stream flow is directly related to SI, then a less stable stream hydrograph is more attractive to hatchery strays #seems like this is pretty much the same thing as percent\_lake\_wetL\_A, maybe you can toss out percent\_lake\_wetL\_A for this reason. Toss out percent\_lake\_wetL\_A over CV\_flow because it is slightly more confusing to interpret

Off-peak survey timing (1/13/21)

* Some streams are sampled really early or late, as in the first week of July or in late September/early October (fall-run populations)
* I am getting rid of Chilkat River and Disappearance Creek fall-run samplings, Black River early season sampling, Herman Creek late sampling, and Saook Bay West Head early sampling in 2010 only (loss of 7 total data points). This leaves me with a final date range of 7/20-9/4 sampling dates to include, or approximately 6 weeks total
  + I chose to remove these because I thought that date would matter somewhat for some of my model covariates, particularly Cons\_Abundance and Pink\_Abundance. If salmon are arriving really early in July or during the fall (for which I don’t have as much data), then there may not be many other salmon around for abundance to be a metric of stream comparison by exploratory salmon
  + These are only being deleted in the R code, not in the Master\_dataset, just in case I wish to revisit these streams. I will also make a note of this change in my project log

2/21/21

Change in model type from logistic regression to a negative binomial mixed effects model 🡪 Curry pointed out in a recent meeting that the response (straying index) that I was using in my logit model made the response observations non-independent. That is to say, the proportion of strays in one stream would depend on the proportion in another. So, I have changed my response variable to counts, or the average number of hatchery strays observed in a stream (averaged to account for differences in the number of surveys) and am fitting this as a generalized linear mixed effects model with a negative binomial distribution instead of as a logistic regression. The negative binomial distribution is best for count data which is overdispersed (and mine is!)

So, the response is the average number of strays in a stream-year = total # of hatchery strays detected in that stream in that year / total # of surveys in that stream in that year

I have also rounded the response variable, Avg\_number\_strays, because the glmer.nb( ) function to fit the model doesn’t like to have a non-integer response. I don’t think this will make a large difference in my output because rounding means that a stream would have 0 strays instead of 0.49 strays for example, or 18 strays instead of 17.8 strays, so attractive streams will still easily emerge in the results.

8/23/21

I recently updated a few data points in *Pink\_Abundance.csv*, and therefore also in *Master\_dataset.csv* and reran my model. The reason I did this is because I noticed one small data error for Rowan Creek in 2010 and because I decided to assign average pink salmon abundances by stock group to make missing data estimates more accurate (for more information see *Pink\_Methods\_Detail.docx*). This resulted in minor changes to coefficient estimates, AICc, values, and p-values, but nothing major. The one change that did merit more attention was that in the second-best model, which I am including in my analysis, Pink\_Abundance and CV\_flow had a correlation of 0.532, surpassing my 0.5 cutoff established earlier on. In the top model, their correlation was 0.427 and therefore not of concern.

Instead of removing one of the two correlated covariates, I decided to leave both in and explore how this would maybe affect coefficient interpretation. As you can see in the document *FE\_correlation.docx*, Pink\_Abundance changes somewhat in model #2, but it still remains strongly positive overall. Considering this and in light of 0.532 not being that much greater than my 0.5 correlation cutoff, I will leave the two models as they are and mention the correlation in my discussion section

* December 2021 update: Pink\_Abundance was totally spurious, i.e., it clearly had a negative relationship with the response in exploratory data analysis, but was strongly positive in the final model for some reason, so I removed it from the model

#test for temporal autocorrelation in the residuals. None is apparent, assuming I correctly tested for residual autocorrelation here. I followed the example on pages 146-147 in Zuur et al. 2009, which said to use type = "normalized" in the *residuals* argument, but that returned an error which indicated that I could only use “deviance”, “pearson”, “working”, “response”, or "partial" residuals. From this website (<https://www.datascienceblog.net/post/machine-learning/interpreting_generalized_linear_models/>) I guessed that I should use “deviance” residuals and went with that

How to interpret final model coefficients for glmer.nb( )

Recall the log link for your glmer.nb!

* This means that the model you have above; the coefficient estimates + the data you input will give you the natural log of the response. Hence, you need to exponentiate the response variable to get the actual predicted [average] number of strays
* To interpret an individual coefficient estimate 🡪 Take mean\_flow. It has a negative relationship with the log response of -0.177. This means that if you increase mean\_flow by one standard deviation and hold the other covariates constant, you would multiply your predicted response by e-0.177 = 0.838, and the predicted response would therefore decrease a little (by about 16% to be precise; 1-0.84=0.16). Alternatively, (additively) the log response would decrease by 0.177. Logs are additive and the normal space is multiplicative
  + Holding other covariates constant: this threw me through a loop awhile back. Holding the other covariates constant may alternatively be thought of as interpreting a coefficient’s effect on the mean response conditional on the mean values of the other covariates. So, in the example model Y = 1.3 + 0.4X1 – 0.2X2, we would interpret X1’s coefficient estimate (0.4) as the change in the mean response (Y) when X1 increases by one unit and when X2 is held constant at its mean value 🡪 mean change in Y based on effect of β1 = 1.3 + 0.4\*(increase in one unit) – 0.2\*(mean value of X2)
  + For your model, you have scaled covariates (X1, X2, …), so the mean of each is 0. This means that you can simply interpret a coefficient estimate’s effect on the mean response as the value of β. E.g., the β estimate for WMA\_Releases\_by\_Yr is 0.412, which means (with all other covariates held constant at their mean value) that the mean log response will increase by 0.412 (add 0.412 to it). On the non-log, multiplicative scale, exp(0.412) = 1.51, which means your mean response would increase by 51%
* The intercept, 0.33342, is the unconditional expected mean of log *Avg\_number\_strays*, or e0.33342 = 1.396 strays expected on average
* **The quadratic term**: Recall that the CV\_flow coefficient is 0.531 and I(CV\_flow^2) is 0.633. This means that as for every increase in SD of Pink\_Abundance, the average number of strays increases by 0.531, and then this slope (0.531) increases by 0.633 for each 1 SD increase

Interpreting the results of a GLMM (not a GLM)

* I thought that you could interpret the coefficients from a GLMM output as they were shown, but I did a quick google search to make sure and now I’m not so sure. This stack exchange page (<https://stats.stackexchange.com/questions/378260/reporting-glmer-nb-results>) says that “you have to be aware that the estimated coefficients you obtain have an interpretation conditional on the random effects. Most often this is not the interpretation you are looking for, but instead, the interest most often lies in a marginal / population interpretation.”
* Looking into marginal/population interpretation led me to this stack exchange (<https://stats.stackexchange.com/questions/16390/when-to-use-generalized-estimating-equations-vs-mixed-effects-models?noredirect=1&lq=1>) explaining the difference between generalized estimating equations (GEEs) and GLMMs. GEEs are for uncovering the population average of a covariate effect and GLMMs are for uncovering the individual-specific effect
  + I failed to understand what was meant by “individual-specific” effect from the page, but I would imagine that I want the population average effect instead? So, I need to either use a GEE or somehow convert my GLMM coefficients to the marginal/population interpretation? How do I do that?
  + Zuur 2009 page 303 briefly compares GEEs and GLMMs (and has separate chapters for both) and says that GEEs do not compare the distributional properties of the subjects themselves. For a GLMM, we estimate the variance of the distribution that each subject’s (the year, farm, classroom, etc) effect is drawn from
  + Section 5.2 of this guy’s lecture notes show how to calculate the marginal/population-level coefficients. It looks complicated and I am going to forgo for now: <http://www.drizopoulos.com/courses/EMC/CE08.pdf>
  + GEEs not recommended: <https://www.researchgate.net/post/When_do_you_apply_GLMM_vs_GEE#:~:text=If%20it%20is%20a%20conditional,is%20the%20way%20to%20go>).
  + **Zuur’s notes make it sound like it is maybe okay to interpret the GLMM coefficients? I’m going to do that for now because I don’t understand how to manipulate the coefficients for correct interpretation. I posted on slack for help, and no one answered. Ask Peter about it at next meeting**
  + **Peter says to not worry about all this^^**

How would I sort out the environmental (quality of habitat) versus conspecific effects of the Cons\_Abundance predictor?

* Cons\_Abundance can only directly tell me about conspecific effects. It may also *indirectly* tell me about environmental effects for which Cons\_Abundance serves as a proxy. But I would only know if it was the latter if Cons\_Abundance was either collinear with an environmental covariate (thereby ruling it out) or if it interacted with an environmental covariate, thereby indicating that there is a conspecific effect and possibly also an environmental effect (i.e., Cons\_Abundance = proxy for environment). I don’t know that I can know for sure unless I included a great many more environmental covariates than I have in order to isolate the two phenomena, since I’ve only included flow and flow was neither collinear nor an interaction with Cons\_Abundance. **What I can interpret for certain is that a significant Cons\_Abundance effect indicates either that there is a conspecific abundance effect on the response and/or that Cons\_A is a proxy for an environmental covariate(s) not included in the model**

Modeling process summarized (6/3/21)

#Needs updating. The list below is approximate. See Model\_fitting3.R for full process, which I commented out # pretty thoroughly

1. **Read in and explore data. Exploring data steps:**
   1. Check distribution of response variable – is it normal?
      1. If not normal, say you are likely to be using a Poisson or negative binomial model which would have a log link, then you should plot the relationship of your explanatory variables against the log of the response
   2. Are there correlations between explanatory variables? If >0.5, you may need to remove a variable or accept that there will be multicollinearity
   3. Plot relationships of response with each covariate individually. Assess what relationships there might be. Are they linear?
   4. Relationships between response and explanatory variables should be linear, if not, you may need a transformation or to add a quadratic term
      1. Are there outliers? Note that they are only potentially an issue if they are outliers in the xy space, not in the x space 🡪 test for influential outliers using cook’s distance
      2. Also explore differences between data groups (‘year’ in this case)
   5. Plot histograms of each explanatory variable. Are they approximately normally distributed?
      1. Note that they are not required to be, but it may aid in producing normal errors in the end
      2. **Ultimately decided NOT to address non-normality of explanatory variable distributions.** In a regression context, we make assumptions of the distribution of the residuals, but not of the predictors. Curry thinks having a more coherent interpretation of coefficients is more important than transforming for normality
   6. Check for spatial or serial correlation in the response
2. Scale (z-score) data across years and streams. Earlier on I had tried to fit the model without z-scoring and it returned a lot of warnings
3. Explore potential transformations of the response variable, the average number of strays. I rounded *Avg\_number\_strays* to the nearest even integer using round( ) for ease of using glmer functions, but otherwise left it untransformed
4. Attempt (and largely fail) to deal with overdispersion. Move on and fit model, see if it makes satisfactory predictions anyway
5. Fit negative binomial mixed effects model using glmer.nb( ) with Year as a random effect (intercept)
6. Check for correlation of fixed effects, remove fixed effects that are highly collinear (>0.5 correlation). This resulted in the removal of Dist\_nearest\_H
7. Test for possible inclusion of complex regressors (quadratic terms and/or interactions) based on significant (>2) AICc change. CV\_flow quadratic term was added
8. Use MuMIn:dredge( ) to identify top model(s) based on delta AIC. Keep top handful of models. In my case, these were the top two models which included 59% of the total weight, while the 3rd best would only have added another 10% (and then sequentially less after that). Also compare to null model (glm.nb(Avg\_number\_strays ~ 1, data = scaled\_data))
9. ~~Average the predictions, NOT the coefficient estimates of the top two models. These are the results~~
10. Assess diagnostics for top model only for simplicity. Diagnostics included
    1. Qqplot and histogram of the residuals to check for normality
    2. Scatterplots of residuals against each covariate to check for homogeneity with specific variables
    3. Qqmath plot to check to normal distribution of random effects
    4. Check for temporal autocorrelation using acf plot
    5. Check residuals vs fitted values plot for homogeneity
11. Look at model results (predicted # of strays)! Identify which streams are attractive and which are mispredictions
12. Try to assess cause of model mispredictions
13. Cross-validate model (70-30) and evaluate using mean absolute error
14. Analyze results/make figures based on coefficient estimates

12/31/21

Franz’s comment on my final project report from FISH 604: Everything looked good, and he was particularly impressed by my analytical rigor. The only thing left for more to consider is described in Franz’s words below:

The only remaining issue is that you are now using the average number of strays across multiple surveys, which results in streams with more surveys in a given year having a more precise (lower variance) observation, which is not consistent with the variance assumption of the negative binomial. You could use the total number of hatchery fish sampled on a given year and stream as the count and the number of surveys as an effort offset. However, since the total number of hatchery fish was collected over multiple surveys, each reflecting a sampling event, the better approach would actually be to use the hatchery count for each sampling event as an independent estimate, i.e. as one data point. The number of data points would therefore be the total number of surveys conducted. The independent variables for each year/stream would be repeated for each survey conducted in that year/stream. That would best reflect the actual sampling. An additional benefit would be that you could also include a seasonal patterns (e.g. a quadratic function of Julian day) to effectively account for possibly lower counts in surveys conducted outside the peak season. Depending on variability among surveys within a stream, you may also get better estimates of the effects (tighter CIs, better predictions.

I cannot do any of these things he suggests here because 1) Including number of surveys as an offset would add another covariate and I already am at the limit of covariates I can include for successful model convergence (I already tried at one point). 2) The second suggestion would make the response variable non-independent (survey results earlier in year in a stream not independent from survey results later in that same year). I would need another random effect or repeated measures design to address this, which would further complicate interpretation of my results. Furthermore, including multiple responses within a stream and year (for multiple surveys at one site in a year) would require duplicating the covariate values. This would be inappropriate for conspecific and pink salmon abundance, because the number of chum and pink salmon would obviously be very different throughout the course of the season, and I cannot know what this abundance would be because that data is unavailable.

* I started to explore and think about this method proposed by Franz in R. See Model\_fitting4.R and the *Analysis* folder

Another thing Franz suggested trying was modeling estimated total strays in a stream and year as a response. This would be the estimated pHOS \* the escapement in the stream for a given year, that way I’m actually predicting the total # of strays, not the # detected in a survey. Then you would include the number of surveys as a log(offset) in the model to deal with variation in effort. I chose NOT to do this because the escapement data themselves are estimates, which would introduce a lot of additional noise to my data. Furthermore, I checked to see if Franz’s suggested response, estimated total strays, was correlated with the response I ultimately used (Avg\_number\_strays) and their correlation was 0.60, indicating that they are largely providing the same information anyway.

cor.test(f\_scaled$Avg\_number\_strays, f\_scaled$Est\_total\_strays)

[1] 0.5845 # 60% correlation!

3/16/22

Question asked of me at the 2022 Alaska Chapter AFS Conference:

**What is the proportion of variation explained by WMA\_Releases\_by\_Yr?** Or by the model, for that matter? I wasn’t entirely sure how to answer this question. As it turns out, calculating an R2 (the proportion of variation explained by covariates) is somewhat nebulous for mixed effects models, given that there are many ways to define it and some theoretical issues as well (larger models have smaller or negative R2 values if you use the typical R2 of an OLS model). See [Nakagawa and Schielzeth 2013](https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/j.2041-210x.2012.00261.x) for more info. Nakagawa and Schielzeth provide direction for how to approximate a marginal (fixed effects only) and conditional (fixed + random effects) R2 value, but it seems like in general it is not recommended to calculate an R2 at all for a glmer (see <https://stats.stackexchange.com/questions/111150/calculating-r2-in-mixed-models-using-nakagawa-schielzeths-2013-r2glmm-me> and <https://stats.stackexchange.com/questions/92221/is-it-worth-reporting-small-fixed-effect-r2-marginal-r2-large-model-r>)

* Instead, it is recommended to compare models using deviance, which can be calculated as -2\*log(L). The lower the deviance, the better the model fit (<https://stats.stackexchange.com/questions/423515/what-does-deviance-mean-in-lmer>)
* For the strength, or “proportion of variance” explained by an individual covariate then, it is best to compare how strong the effect size is (<https://stats.stackexchange.com/questions/92221/is-it-worth-reporting-small-fixed-effect-r2-marginal-r2-large-model-r>)
* Update March 2022: Megan and Curry (co-authors) also want to see an R2, but for the whole model, not individual covariates. I calculated a pseudo-R2 using MuMIn::rsquared.GLMM( ) function, which was developed based on the recommendations in the Nakagawa and Schielzeth 2013 paper. This is reported in my modeling script and paper. I don’t report any pseudo-R2 values for individual covariates because this still doesn’t make sense to me and I haven’t seen it done anywhere else. Also, Curry and Megan didn’t ask for them

3/21/22: Addressing Curry’s questions and suggestions on chum\_v5 chapter 1 manuscript

*“I’m guessing you did your covariate selection using ML and then generated final model estimates using REML, following standard practice, but I agree it would be good to include a sentence stating that directly. Correct?”*

**Why I did not use REML to generate the final model after the ML model selection process**: REML is not defined for glmer models and this is intentional because REML only makes sense for Gaussian linear mixed effects models, not for *generalized* mixed effects models. I don’t totally understand why this is the case but several authors (Douglas Bates, Ben Bolker) of the lme4 package indicate this in their responses to people in a mailing list looking to calculate REML estimates for glmers. See

* <https://stat.ethz.ch/pipermail/r-sig-mixed-models/2009q1/002104.html#:~:text=The%20bottom%20line%20is%20that%20REML%20only%20makes%20sense%20for%20linear%20mixed%2Deffects%20models>
* <https://stat.ethz.ch/pipermail/r-sig-mixed-models/2012q3/018797.html>
* <http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#:~:text=provide%20more%20information.-,REML%20for%20GLMMs,-While%20restricted%20maximum>
* <https://stats.stackexchange.com/questions/269224/how-to-obtain-reml-estimates-in-a-glmer-optimizing-random-effects-structure-in>

*“You seem to have a lot of zeros. Was the zero-inflated negative binomial considered, but was unnecessary? That is to say was the observed number of zeros consistent with you assumed error distribution? Here again a bit more justification in the methods and results would head off potential reviewer comments, even just a quick sentence about residuals.”*

**Why I did not incorporate zero-inflation into the model**: This is also really hard to do in a generalized linear mixed effects modeling framework. Most existing packages for incorporating zero inflation into a model are for fixed effects models only, unless you want to venture into glmmTMB or glmmADMB land… Before I went down that rabbit-hole, I decided to see if it would even be necessary to incorporate zero-inflation into the model structure. I did not ultimately think it was. The two plots below show the deviance and pearson residuals for the top model, where the spread of residuals +/- 0 is pretty even

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Another useful diagnostic might be this plot:

Chart, histogram

Description automatically generated

To create this histogram, I selected all rows of data (n = 49) where the observed attractiveness index was 0, then I plotted the model predicted indices that corresponded with each of those 49 zeros. The model predicts 0 and 0.5 for 20 of the true zeros (recall that I rounded the response variable for modeling, so many of the observed zeros are actually 0.5 as well), and indices < 2.5 for almost all the rest. Thus, I am not too worried about zero-inflation affecting my results. See “Manu\_figs.R” for code

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*“While I applaud the cross-validation approach to evaluating out-of-sample predictive performance, it is kind of hard to grasp what MAE=6.29 means in the broader context. I would suggest potentially using mean absolute % error, and comparing MAPE with the null model which only included year effects. MAPE would also have the benefit of placing errors on a relative scale.”*

**Why I am continuing to use MAE instead of MAPE:** There would be zeros in the denominator, so there would be a lot of undefined values of MAPE. Because MAPE is calculated as (observed-predicted)/observed, and I have many observed 0 values, those would be undefined. The symmetric MAPE is a possible alternative, calculated as (predicted-observed)/((observed-predicted)/2), but maybe [not recommended](https://en.wikipedia.org/wiki/Symmetric_mean_absolute_percentage_error#:~:text=A%20limitation%20to%20SMAPE%20is%20that%20if%20the%20actual%20value%20or%20forecast%20value%20is%200%2C%20the%20value%20of%20error%20will%20boom%20up%20to%20the%20upper%2Dlimit%20of%20error.%20(200%25%20for%20the%20first%20formula%20and%20100%25%20for%20the%20second%20formula).). I will instead continue to use MAE, but what I will do differently than before is report multiple MAE values in my paper for different points in the data, i.e., MAE for all values and MAE for observed vales < 20

From section 7.4 in my “Model\_fitting3.R” code, cross-validation results:

|  |  |  |
| --- | --- | --- |
|  | MAE for all values | MAE for obs vals < 20 |
| Top candidate model (bm1u) | 6.27 | 3.86 |
| Second best model (bm2u) | 6.44 | 3.90 |
| Null model (avg\_strays ~ (1|Year)) | 9.93 | 5.65 |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

*“The methods highlight that an influential outlier was removed. The text will need to justify why you believe this to be an outlier, and the specifics of the context. Do you believe the observation of the covariate was incorrect in some way? Did you conclude post-hoc that this outlier was driving the relationship?”*

**The outlier was removed because I concluded post-hoc that it was influencing the relationship**

mean(mean\_mae) #for f\_scaled (full dataset including large obs. values), the mean

#MAE is 7.23. For sub\_f\_scaled, it is only 4.31

^^I no longer remove an outlier! (May 2022)

4/27/22

After my defense, Megan asked about whether there might potentially be spurious relationships between any of my covariates and stream attractiveness, whereby attractiveness might be explained just by the location of a stream on the landscape. She noted that the attractive streams were located along a NW-SE axis, and so recommended that I run a PCA on latitude and longitude of modeled streams, and then plot + regress the model covariates against PC1. I did this and found that there was maybe some correlation between latitude/longitude (PC1) and the number of fish released within 40km. Note that in all plots, the dotted line is lm(variable ~ PC1) and PC1 correlates with increasing longitude and decreasing latitude (hence the NW-SE axis)

Chart, scatter chart

Description automatically generated

But not between PC1 and the other two covariates

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

We concluded that the possible spatial pattern in WMA\_releases\_by\_Yr ~ PC1 was probably accounted for within the WMA\_Releases\_by\_Yr covariate since larger values of that covariate acknowledge the proximity of release sites. But Megan and I weren’t 100% sure on this. Given that earlier on (in my Model\_fitting3.R script), I had tested for spatial autocorrelation in the response variable using Moran’s I (and found no significant evidence of autocorrelation), I will not worry about this for now. See PCA\_spatial\_analysis.R for the PCA code, including the code that created the above figures

5/6/22

Chris Habicht and Curry Cunningham raised the issue at my defense that there might be some bias in my model response variable (the average number of strays), whereby fewer hatchery strays are detected when proportionally less of the total dead are sampled. I.e., if there are more dead fish in a stream than technicians can sample in a day, not all hatchery fish will be detected.

There is an additional complication that in 2008–2011, there was a sampling cap of 192 fish, where if there were more than 192 fish in a stream, then technicians would stop sampling beyond the 192-fish threshold (see Piston and Heinl 2012 methods). Then in 2013–2015 in the non-fitness streams (i.e., all streams except for Fish Creek, Sawmill Creek, Admiralty Creek, and Prospect Creek), there was a threshold of 384 fish. See page 95 within appendix C of the 2015 Alaska Hatchery Research Project [report](https://www.adfg.alaska.gov/static/fishing/PDFs/hatcheries/research/pwssc_hw_2015_appendice_a-e.pdf) for documentation of this. Then, in 2017–2019, all streams were sampled exhaustively almost every day, hence the technicians likely detected all or the majority of hatchery strays.

In the **dead\_count\_analysis.R** script, I analyzed survey data downloaded from the “Stream Surveys” tab under “Reports” at hatcherywild.org for 2013–2019 streams. I found that there is evidence for bias in the response variable and I tailored stream survey data so that I could account for the bias. I accounted for the bias by dividing the number of strays detected in 2008–2015 streams by the proportion of fish sampled:

This inflates the number of strays in a survey where not all fish were sampled to reflect the true number of hatchery strays more accurately, proportional to the number of fish sampled versus the total number of dead.

**In case this isn’t complicated enough^^,** the DeadCount column in survey\_data.csv (downloaded directly from hatcherywild.org) gives the total dead, which is the number of previously sampled + previously unsampled (new dead) fish. We wanted to check for bias and adjust for the bias by assessing the number of fish sampled relative to the number of previously unsampled (new dead) fish only. To be able to move forward, I assumed (after discussing with Curry and Peter) that in 2008–2015 surveys, the DeadCount column gave the number of new dead fish only. We assume this because most (90%?) of the surveys in the same stream and year are at least a week apart, therefore previously sampled carcasses would have been washed out or dragged away by predators by the time technicians returned to survey again.

Then for 2017–2019 surveys, streams were sampled either daily or every other day, so the dead count likely included many previously sampled fish. For these surveys I subtracted the number of dead from the previous survey and assumed that the difference was the number of new dead. The exception to this rule was if the number of dead in the previous survey was greater than the number of dead in the survey immediately following. This situation tended to occur when there were rain/high waters (as mentioned in the “comments” section of survey\_data.csv), which likely would have washed out all the previously sampled carcasses, leaving only new dead. If this occurred, I did not subtract the previous survey’s dead count.

And for the final complication, 2008–2011 surveys had no dead count data at all. Given that there was a sampling threshold of 192 fish, I assumed that for streams where the season total was < 192 fish that technicians were able to sample 100% of carcasses. For these streams I set dead count = number of specimens, which meant I did not adjust for any potential bias later on. For streams where more than 192 fish were sampled, I fit a linear model of dead count ~ number of specimens using 2013–2015 data (because 2013–2015 surveys were spaced a week or more apart just like in 2008–2011) and used this model to predict missing dead counts for 2008­–2011 streams.

Note that for some surveys, the number of specimens would be greater than the dead count. This is because sometimes technicians would kill and sample post-spawn fish that were still technically alive, and these were counted as live fish instead of dead. For these surveys the proportion sampled would be greater than 1. Curry, Peter, and I decided to round these down to 1, or 100% of fish being sampled.

This analysis^^ and everything I’ve described here should be documented in the dead\_count\_analysis.R script itself and accompanying documentation. The output from this script is *survey\_data8.rds*, which is the cleaned-up version of the survey\_data.csv that I downloaded directly from hatcherywild.org “Stream Survey” reports. *survey\_data8.rds* also contains the complete and/or corrected dead counts for all stream surveys (including 2008–2011 surveys) based on all the methods I described on the previous page.

The other output from dead\_count\_analysis.R is *new\_response\_var.rds*, which, as the name suggests, contains the average effective number of strays to be used as my new model response variable to account for sampling bias. The average effective number of hatchery strays is the total effective number of hatchery strays in stream in a year divided by the number of surveys in that stream. The total effective number of strays is the sum of the effective number of strays from all surveys in the stream in that year, and the effective number of strays from each survey is the raw number of strays in each survey divided by the proportion sampled in that survey (equation on previous page).