STAT 3503/8109 Lecture 6 Notes

Edoardo Airoldi Scribe: Srikar Katta*

Fall 2020: October 5, 2020

Introduction

Modeling Consideration

So far, we have covered likelihood, mixture models, the transformation theorem, and the properties of the expectation and variance operators - all part of the first part of this course, labeled under "modeling." Essentially, this section of the course is there to help you communicate and write models with statisticians, researchers, data scientists, stakeholders, and anyone else involved. This starts with the 2x2 table we discussed earlier and then some formulas like the complete likelihood, which you use in the presence of latent variables, and the proper likelihood. How do we get from the complete to the proper likelihood? We marginalize out the latent variables from a joint probability density function by using integration techniques. While this is calculus, the only information necessary to apply this is understanding that the integral allows us to find the "area under the joint probability distribution" with respect to our latent variables, thereby changing the function - the expression - that we are working with to be only in terms of known variables.

What is important in this first section is not the calculus but rather the intuition. We have a model statement and a setting, which allows us to create our 2x2 table. And then we can write the likelihood and so on and so forth. As a researcher, this is important to understand to condense the assumptions we are making and being able to easily communicate this information to collaborators.

A part of communication is relating probability distributions to your variables. For example, someone might say that income has a log-normal distribution or that being accepted into a college follows a Binomial distribution or say some obscure probability distribution you are unfamiliar with. That all is fine as long as you remember that there is a relationship between probability distributions as we demonstrated with the transformation theorem that will allow us to attack these theory problems head on without having to worry about the names of distributions. All you need are the quantities involved, the possible values these quantities adopt, and whether they are theoretically distribution or not. Does the specific name of the distribution matter? Not necessarily.

*Please share any comments or suggestions with Srikar Katta at srikar@temple.edu

Example 0.1. *We are given quantities x, μ, θ and the following:*

$$\begin{aligned} x &\sim Normal(\mu, \sigma^2), x \in \mathbb{R} \\ OR\ x &\sim T_{(k)}, a\ \textit{T-distribution with k-degrees of freedom} \\ \mathbb{P}(X|\theta) &= function(X, \theta) \end{aligned}$$

In this problem, suppose that the probability distribution function has some obscure name. This should not waiver your resolve because all you need to do is find the mathematical expression for this probability distribution and the range of values that X adopts. Then you can write down the likelihood and compute an integral if needed.

Inference

Suppose we are given the following:

	Observed	**Unobserved**
Variable	Observed RV	Latent RV
Constant	Known Constants	Unknown Constants

Now, how do we estimate the unknown constants? There are a few techniques that we will discuss later but just mention here:

- If there is a likelihood, we can use maximum likelihood estimation
- If we have moments, we can use method of moments estimator
- If we have latent random variables and the likelihood is unavailable in closed form, and we can conduct expectation maximization

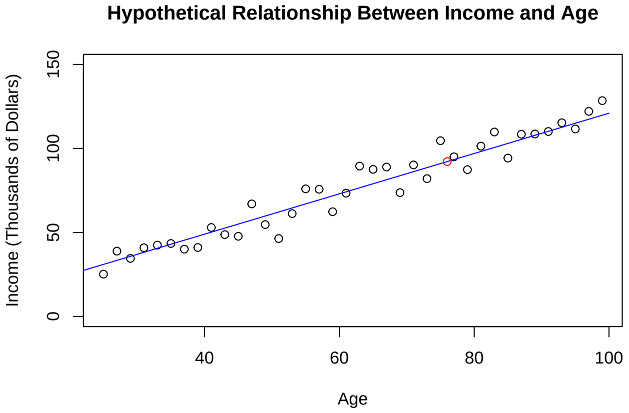
The second part of this class, known as inference, will be about estimating the latent random variables and unknown constants. These calculations are inherently tied to calculus but this course does not require or even emphasize the calculations; rather, we hope that you are able to offer the intuition behind these problems, *when* calculus should be applied, and what potential techniques you can use to estimate these unknowns. The actual calculation of integrals belongs in a calculus course, which this is not. This course serves to build your skills and intuition for data science. Many people can compute integrals but not everyone can understand statistics and data science.

Models

From here on, we will be discussing different statistical models that we can use in research or industry by providing a problem and model statement and finding the likelihood.

Linear Regression (Normal Distribution Version)

Consider you have a single variable regression, and we are interested the relationship between age and income. Suppose a scatter plot looks something like the following:



These circles are points that are observed and suppose we want to predict what the income for some aged 75 is. They may not be in the dataset, bu we can still interpolate it (red point). One way we can write the model down to predict income at age of 76 would be as follows:

$$\begin{aligned} i &= 1 \dots n \\ \epsilon_i &\sim Normal(0, \sigma^2) \\ y_i &= \alpha + \beta x_i + \epsilon_i \\ \implies y_i - (\alpha + \beta x_i) &\sim Normal(0, \sigma^2) \\ \implies y_i &\sim Normal(\alpha + \beta x_i, \sigma^2) \end{aligned}$$

so our observed data would look as follows:

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n).$$

First, let us write out the 2x2 table for this:

	Observed	**Unobserved**
Variable	$y_1 \dots y_n$	NA
Constant	$x_1 \dots x_n, n$	α, β, σ^2

One question that may arise here is, "why is Y_i a variable?" Using the transformation, we know that Y_i, a function of ϵ_i which is a random variable, must also have variation. So, let us find the likelihood:

$$\begin{aligned} Likelihood &= \mathbb{P}(y_1 \dots y_n | \alpha, \beta, \sigma^2, x_1 \dots x_n) \\ &= \prod_{i=1}^{n} \mathbb{P}(y_i | \alpha, \beta, \sigma^2, x_i) \\ &= \prod_{i=1}^{n} Normal(y_i | \alpha + \beta x_i, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2\sigma^2}(y_i - \alpha - \beta x_i)^2} \end{aligned}$$

We do not know α, β - our unknown constants. Using techniques we will discuss later, we can find estimates for these unknown constants which will allow us to predict this missing value using the likelihood.

Random Effects

The random effects model is a variant of the linear regression model such that instead of observing (y_i, x_i) and treating x_i as constant, we now draw x_i from some random distribution:

$$\begin{aligned} i &= 1 \dots n \\ \epsilon_i &\sim Normal(0, \sigma^2) \\ x_i &\sim Normal(\mu, \gamma^2) \\ y_i &= \alpha + \beta x_i = \epsilon_i. \end{aligned}$$

In one version, we observe $(x_1, y_1), \dots, (x_n, y_n)$. In another, we observe only $y_1 \dots y_n$.

Random Effects: Observed Independent Variables

First, let's create the 2x2 table for this situation. We have to classify the following variables: $x_1 \dots x_n, y_1 \dots y_n, \alpha, \beta, \sigma^2, \mu, \gamma^2$

	Observed	**Unobserved**
Variable	$x_1 \dots x_n, y_1 \dots y_n$	NA
Constant	NA	$\alpha, \beta, \sigma^2, \mu, \gamma^2$

So, the likelihood, which is by definition the probability of the observed random variables given the constants, would read as follows:

$$\begin{aligned}
Likelihood &= \mathbb{P}(x_1 \dots x_n, y_1 \dots y_n | \alpha, \beta, \sigma^2, \mu, \gamma^2) \\
&= \prod_{i=1}^{n} \left[\mathbb{P}(x_i|\mu,\gamma^2)\mathbb{P}(y_i|\alpha,\beta,x_i,\sigma^2)\right] \\
&= \prod_{i=1}^{n} \left[Normal(x_i|\mu,\gamma^2)Normal(y_i|\alpha,\beta,x_i,\sigma^2)\right] \\
&= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\gamma^2}}e^{\frac{-(x_i-\mu)^2}{2\gamma^2}} \frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(y_i-\alpha-\beta x_i)^2}{2\sigma^2}}\right].
\end{aligned}$$

Random Effects: Latent Independent Variables

First, let's create the 2x2 table for this situation. We have to classify the following variables: $x_1 \dots x_n, y_1 \dots y_n, \alpha, \beta, \sigma^2, \mu, \gamma^2$

	Observed	**Unobserved**
Variable	$y_1 \dots y_n$	$x_1 \dots x_n$
Constant	NA	$\alpha, \beta, \sigma^2, \mu, \gamma^2$

This is still a random effects model because our independent variables, these $x_1 \dots x_n$ have some theoretical, random variation. The idea of it being observed or unobserved should have no bearing on it being a random effects model or not. What is required to understand if it is a random effects model is the model statement. The problem statement, again, simply just allows us to determine what quantities are observed. Now let us find the likelihood. Notice, we have unobserved variables so to find the proper likelihood, we must integrate out the latent random variables from the joint probability distribution.

$$\begin{aligned}
Likelihood^{proper} &= \mathbb{P}(observedrv|constants) \\
&= \mathbb{P}(y_1 \dots y_n|\alpha,\beta,\sigma^2,\mu,\gamma^2) \\
&= \int_{latent} Likelihood^{complete} d(latent) \text{ - complete likelihood with latent rv integrated out} \\
&= \int_{x_n} \dots \int_{x_1} \mathbb{P}(x_1 \dots x_n, y_1 \dots y_n|\alpha,\beta,\sigma^2,\mu,\gamma^2) dx_1 \dots dx_n \\
&= \int_{x_n} \dots \int_{x_1} \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\gamma^2}}e^{\frac{-(x_i-\mu)^2}{2\gamma^2}} \frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(y_i-\alpha-\beta x_i)^2}{2\sigma^2}}\right] dx_1 \dots dx_n,
\end{aligned}$$

which is a closed form solution because it is the integral of a normal distribution. So, the likelihood is still a function of $\alpha, \beta, \sigma^2, \mu, \gamma^2$ just as before, but the only difference is that does not depend on $x_1 \dots x_n$. Here, we are simply building up intuition for the *approach* and not emphasizing the actual calculations themselves, so this actual calculation is of little relevance currently. What truly matters is the method of finding the proper likelihood by integrating out the latent random variables from the complete likelihood to find the proper likelihood. Taking the integral of an expression with respect to the $x_1 \dots x_n$ will return an expression that is not involved with any of $x_1 \dots x_n$.

Missing Data Example

Suppose we administered a survey to 100 people on age, which we will refer to as x_i, and income, which we will refer to as y_i. However, some people forgot to respond to certain questions and we are left with an incomplete dataset:

- Persons 1 to 97 reported all information
- Persons 98 and 99 did not report age but reported income
- Person 100 reported age but not income

So, our dataset reads as follow: $(x_1, y_1)...(x_{97}, y_{97}), (?, y_{98}), (?, y_{99}), (x_{99}, ?)$ Additionally, we are also given the following model statement:

$$\begin{aligned} i &= 1...100 \\ \epsilon_i &\sim Normal(0, \sigma^2) \\ x_i &\sim Normal(\mu, \gamma^2) \\ y_i = \alpha + \beta x - i + \epsilon_i \end{aligned}$$

Notice, since both our independent and dependent variables are variable and our dependent variable is a linear function of our independent variables, this is a random effects model. But now we have missing data. This section will detail how to approach such situations.

First, let us create the 2x2 table for this case:

	Observed	**Unobserved**
Variable	$x_1...x_{97}, x_{100},\ y_1...y_{99}$	x_{98}, x_{99}, y_{100}
Constant	NA	$\alpha, \beta, \mu, \alpha, \gamma^2$

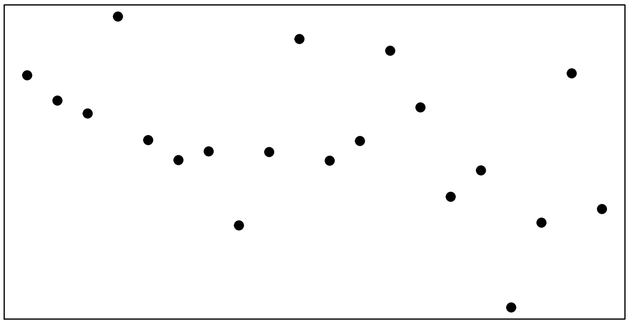
The direct consequence of this 2x2 table, as usual, is the likelihood:

$$\begin{aligned} likelihood &= \mathbb{P}(observedrv|constants) \\ &= \mathbb{P}(y_1...y_{99}, x_1...x_{97}, x_{100}|\alpha, \beta, \mu, \sigma^2, \gamma^2) \\ &= \int_{latent} Likelihood^{complete} d(latent) \text{ - complete likelihood with latent rv integrated out} \\ &= \int_{x_{98}} \int_{x_{99}} \int_{y_{100}} \mathbb{P}(y_1...y_{100}, x_1...x_{100}|\alpha, \beta, \mu, \sigma^2, \gamma^2) dy_{100} dx_{99} dx_{98} \\ &= \int_{x_{98}} \int_{x_{99}} \int_{y_{100}} \prod_{i=1}^{100} \left[\frac{1}{\sqrt{2\pi\gamma^2}} e^{\frac{-(x_i-\mu)^2}{2\gamma^2}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_i-\alpha-\beta x_i)^2}{2\sigma^2}} \right] dy_{100} dx_{99} dx_{98} \\ &= likelihood(\alpha, \beta, \mu, \sigma^2, \gamma^2), \end{aligned}$$

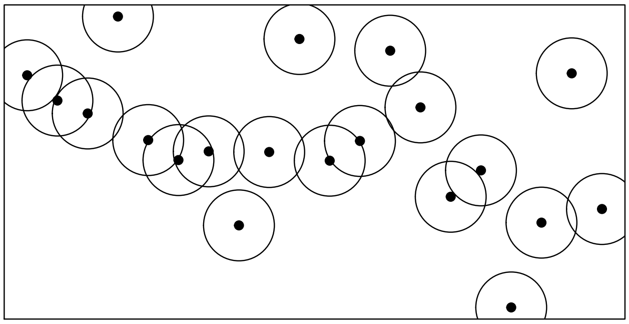
an expression for the likelihood that does not have x_{97}, x_{98}, y_{100}.

Latent[1] Space Models

Suppose a group of 20 people are sitting in a room as follows:



We want to draw a social network around these 20 people and identify friendships, but we are not sure if they are friends or not. We need some way to model the probability that these people know each other. Well, if they are sitting close together, there is a higher chance they know each other than compared to if they were sitting on opposite ends of the classroom. So, we can draw circles around each person, and if these circles overlap, we will draw a line between them, known as and edge, that says that these people are connected.



This latent space model is a technique for modeling the data generating process for these edges. How would we put this intuition into a model? Suppose we treat x_i as a constant. Then, the data generating process would look like this:

$$i = 1 ... n$$

$$x_i \in \mathbb{R}^2 \text{ - a 2-D space, social position of } i \text{ if } x_i \text{ is constant}$$

$$OR\ x_i \in Normal\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \Sigma_i\right) \text{ - a 2-D space, social position of } i \text{ if } x_i \text{ is variable}$$

$$(i,j)\ in\ \{(1,2),(1,3),...,(1,n),(2,3),(2,4),...,(n-1,n)\}$$

$$\mathbb{P}((i,j) = 1 | x_i, x_j) = \alpha - |x_i - x_j|$$

where (i,j) represents the existence of an edge between x_i and x_j, and $|x_i - x_j|$ is the distance between the two points. Now, α is just some constant that we can change that could be likened to the radius of the

[1] This does not refer to the latent/unknown variables or constants. This is called a latent space model because the space we are working with can be an unobservable.

circle around x_i and is just the baseline probability of α. So, the smaller the distance between x_i and x_j, the more likely they are to be connected. People who use this model only see the edges between two people, and they can estimate the latent position of people fairly well.

Changing the problem just slightly, assume the same model statement but now consider a new problem statement: suppose you work at a social media company and you notice that person i is posting very often. If you want to recommend posts to person j and you notice an edge between them, then you can recommend person i's post.

Or, the social media company can see what posts persons i and j make, and based on how similar the content is, it can recommend a post that person j liked to person i and a post that person i liked to person j. In this situation, the latent space is their social media content.

One Version of Latent Space Model

Suppose we are given the following model statement:

$$\begin{aligned} i &= 1...n, j = 1...n, j > i \\ x_i &\sim Normal(0, \sigma^2) \text{ - so } x_i \text{ belongs to a line instead of a 2-D space} \\ log\left(\frac{\eta_{ij}}{1-\eta_{ij}}\right) &= \alpha - |x_i - x_j|, \alpha \in [0, \infty) \\ y_{ij} &\sim Bernoulli(\eta_{ij}). \end{aligned}$$

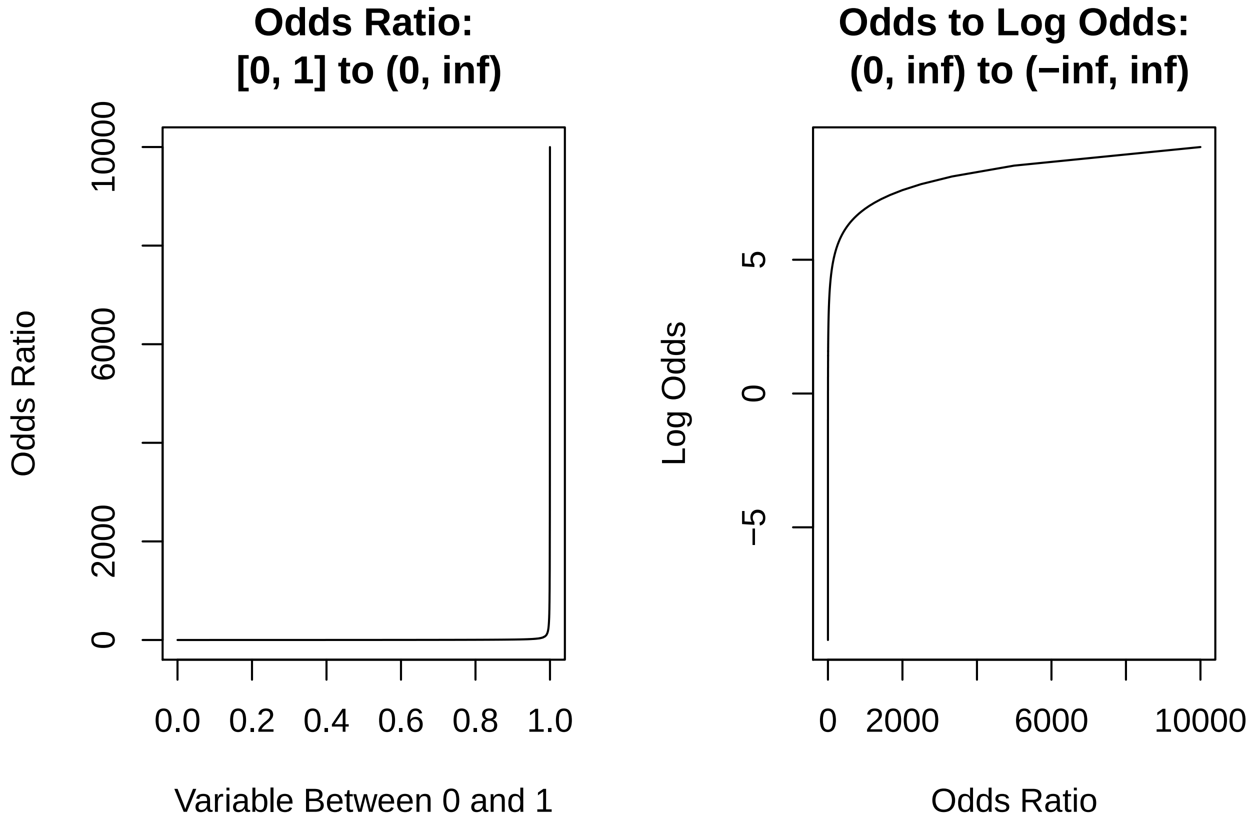
Let us unpack this. First, recognize that y_{ij} represents an edge (connection) between x_i and x_j, where x_i and x_j represent the social position of two people on a line instead of a 2-D plane this time. In other words, x_i and x_j are friends or they have similar content in their social media posts or are just connected in some way. However, this connection has some variability and is selected from a Bernoulli random variable with probability η_{ij}. Notice, η_{ij} must be bounded between 0 and 1 because it is a parameter for the Bernoulli distribution. There must exist a relationship between η_{ij} and x_{ij} because η_{ij} is the only determinant of y_{ij} and there is a relationship between y_{ij} and x_i and x_j, namely $\alpha - |x_i - x_j|$, where α is just some positive value that is proportional to some baseline probability and $|x_i - x_j|$ is the distance between x_i and x_j.

First, observe that x_i, x_j must be in $\mathbb{R}$ because they are selected from a normal distribution which goes from $(-\infty, \infty)$, so that means that $|x_i - x_j|$, the distance between x_i and x_j must belong to $[0, \infty)$. And because α is just some value that is proportional to some baseline probability, it goes from $[0, \infty)$. Then, if α is 0 but $|x_i - x_j|$ is ∞, then $\alpha - |x_i - x_j| = -\infty$. But if $\alpha = \infty$ and $|x_i - x_j| = 0$, then $\alpha - |x_i - x_j| = \infty$. So, $\alpha - |x_i - x_j|$ goes from $(-\infty, \infty)$.

So, we can connect η_{ij}, a number between 0 and 1 to a real-valued number by applying the log-odds transformation to η_{ij}. To find the log-odds of η_{ij}, we first divide the probability of success (η_{ij}) by the probability of failure ($1 - \eta_{ij}$), giving us the odds [2], so this would then go from 0 to ∞.

[2] Odds can be likened to a ratio: the ratio of your favored sports team winning may be 3-to-2, which just means they have a 60% chance of winning, or 9-to-1, a 90% chance of winning

Now, we can apply the log-transformation to the odds, a function that maps values from $(0, \infty)$ to $\mathbb{R}$. So, $log(\frac{\eta_{ij}}{1-\eta_{ij}})$ belongs to the real line and that means that we can represent values in $\alpha - |x_i - x_j|$ as log odds of η_{ij}, a parameter for the Bernoulli distribution that y_{ij} belongs to.



Some might be wondering why we need η_{ij} since it just seems like an extra step and that has log odds. We know there needs to be a relationship between x_i, x_j, α and y_{ij}, so why not just write the model statement directly and skip over the η_{ij}? For people with a modeling background, the log-odds become very familiar after some time, and we can illustrate why. Notice that η_{ij} has its own closed form solution as well:

$$\begin{aligned}
log\left(\frac{\eta_{ij}}{1-\eta_{ij}}\right) &= \alpha - |x_i - x_j| \\
\iff \frac{\eta_{ij}}{1-\eta_{ij}}(1-\eta_{ij}) &= e^{\alpha-|x_i-x_j|}(1-\eta_{ij}) \\
\iff \eta_{ij} &= e^{\alpha-|x_i-x_j|}(1-\eta_{ij}) \\
\iff \eta_{ij} &= e^{\alpha-|x_i-x_j|} - \eta_{ij}e^{\alpha-|x_i-x_j|} \\
\iff \eta_{ij} + \eta_{ij}e^{\alpha-|x_i-x_j|} &= e^{\alpha-|x_i-x_j|} \\
\iff \eta_{ij}(1+e^{\alpha-|x_i-x_j|}) &= e^{\alpha-|x_i-x_j|} \\
\iff \eta_{ij} &= \frac{e^{\alpha-|x_i-x_j|}}{1+e^{\alpha-|x_i-x_j|}}.
\end{aligned}$$

So, the first line, the log odds of x_i and x_j being connected can be expressed the difference of a value proportional to some baseline probability and the distance between x_i and x_j. But, η_{ij}, the parameter for the probability that x_i, x_j are connected can be represented as that final equation, which is a less intuitive to interpret. As before, $\alpha - |x_i - x_j|$ is the difference between some value proportional to a baseline probability and $|x_i - x_j|$ is just the distance between the two points. So, if the distance between points is small, then $\alpha - |x_i - x_j|$ is very large. Then, $e^{\alpha - |x_i - x_j|}$ would be very large, so $\frac{e^{\alpha-|x_i-x_j|}}{1+e^{\alpha-|x_i-x_j|}} \approx 1$. However, if the distance between the two points was very large, then $\alpha - |x_i - x_j| < 0$, so $e^{\alpha-|x_i-x_j|}$ would be very small, less than 1. So $\frac{e^{\alpha-|x_i-x_j|}}{1+e^{\alpha-|x_i-x_j|}} \approx 0$. We can technically rewrite this model using this closed form for η_{ij}:

$$\begin{aligned} i &= 1...n, j = 1...n, j > 1 \\ x_i &\sim Normal(0, \sigma^2) \\ y_{ij} &\sim Bernoulli\left(\frac{e^{\alpha-|x_i-x_j|}}{1+e^{\alpha-|x_i-x_j|}}\right) \end{aligned}$$

However, this is much less intuitive than the other model once you are familiar with log odds ratios a little bit.

Then, the likelihood for this model would be as follows:

$$\begin{aligned} likelihood &= \mathbb{P}(y_{12}...y_{n-1,n}|\alpha) \\ &= \int_{x_n} ... \int_{x_1} (y_{12}...y_{n-1,n}, x_1...x_n|\alpha) dx_1...dx_n, \end{aligned}$$

but this is also less intuitive. Specifying η_{ij} simply clarifies the relationship between $\alpha - |x_i - x_j|$ and y_{ij}.