MATH3070
Natural Resource Mathematics
Regression and Maximum Likelihood

1 Probability & Statistics

So far we have mostly been concerned about projecting future population trends and solving for the best policies to manage these populations, given some underlying mathematical model. In the first few weeks, the parameters of these models were given to you. However, in real life, you may be asked to figure out what these parameters are yourself before you ever start simulating population trajectories and determining optimal management strategies. How does one do this? There are many possible methods. We will explore three of the most common methods (1) regression (2) maximum likelihood, and if we have time (3) Bayesian approaches. We will also explore how to not only estimate parameters that govern dynamic processes but also (4) how to estimate population size in the wild, given count data. Lastly, we will return to population dynamics and consider stochastic models, whose solutions and properties will require many of the concepts we learn during this section on statistics.

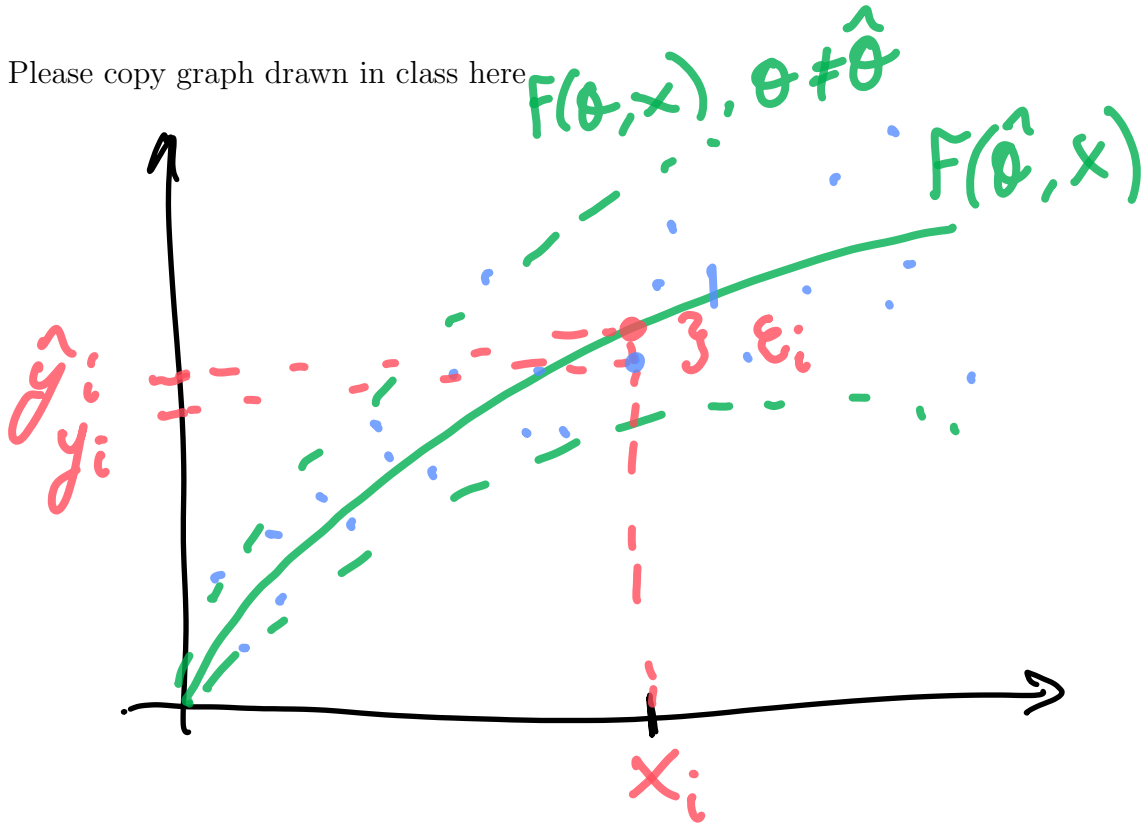
2 Least Squares

The simplest and most common method for fitting a curve to data is called "Least Squares". The principle behind the method is to choose parameters that minimize the sum of squared differences between the fitted values (the curve) from the actual observed values. To do this let us define some notation. In general, I will tend to use a "$-$" to signify the mean (or average), "$\hat{}$ " to mean an estimate (or predicted value, e.g., something you calculated), if there is no symbol above the variable, it means it is either real data or the true value of some parameter. I will let x be the independent variable and y be the dependent variable (sometimes known as the response variable). Consider a data set of n total observations. Each observation contains an observation of both the independent variable and dependent variable, forming a pair (x_i, y_i) where $i \in \{1, 2, ..., n\}$

- x_i = the independent variable in the i^{th} observation
- y_i = the dependent variable in the i^{th} observation
- $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ (note that this is simply the mean of your data, not an expectation)
- $F(\theta, x)$ = the function you are fitting to the data, where θ is the parameter, potentially a vector of m parameters $(\theta_0, \theta_1, ...\theta_m)$, and x is the independent variable
- $\hat{\theta}$ is your estimate of θ after the fitting (this is what we call an estimator, but we will get to that later)

- $\hat{y}_i$ = the predicted value $F(\hat{\theta}, x_i)$, i.e. the value you would have guessed for y, after you did your fitting, if someone told you $x = x_i$

Please copy graph drawn in class here



Note that f is a general function, in this case, it could be linear, logistic, Beverton-Holt or any of the other functions you have seen so far in this course. The goal of the general problem is to minimize the following quantity

$$SSE(\theta) = \sum_{i=1}^{n} [\, y_i - F(\theta, x_i) \,]^2 \,, \tag{1}$$

with respect to the parameter(s) θ. Let's consider a very simple case exponential growth of a population governed by the discrete time model $x_{t+1} = \theta x_t$. If you are given data of the values of the population size through time x_0, x_1, ... x_n. Then your response variable is just the population in the next time step, and your independent variable is the population size in the current time step. So we have the model. So $y = (x_1, ..., x_n)$. In this case,

$$SSE(\theta) = \sum_{i=1}^{n} [\, y_i - \theta x_{i-1} \,]^2 \,,$$

exercise: Find the value of θ that minimises SSE. Hint: use the main principle from calculus that we have been using all semester to maximize and minimize functions.

Now let's consider **linear least squares**. Here our model is

$$F(\theta, x) = \theta_0 + \theta_1 x. \tag{2}$$

Given you want to minimize SSE with respect to θ_1 and θ_2, can you figure out, what estimates for these parameters $\hat{\theta}_1$ and $\hat{\theta}_2$ achieve this goal?.

$$SSE = \sum_{i=1}^{n} (y_i - \theta_0 + \theta_1 x_i)^2$$

$$\Rightarrow \frac{\partial SSE}{\partial \theta_0} = \sum 2(y_i - \theta_0 - \theta_1 x_i)(-1) = 0 \Rightarrow \frac{\partial^2 SSE}{\partial \theta_0^2} = \sum_{i=1}^{n} 2 = 2n$$

$$\frac{\partial SSE}{\partial \theta_1} = \sum 2(y_i - \theta_0 - \theta_1 x_i)(-x_i) = 0 \Rightarrow \frac{\partial^2 SSE}{\partial \theta_1^2} = \sum_{i=1}^{n} 2x_i^2$$

For these to be a minimum, we need

$$\left(\frac{\partial^2 SSE}{\partial \theta_0^2}\right)\left(\frac{\partial^2 SSE}{\partial \theta_1^2}\right) - \left(\frac{\partial^2 SSE}{\partial \theta_0 \partial \theta_1}\right)^2 > 0$$

$$\frac{\partial^2 SSE}{\partial \theta_0 \partial \theta_1} = \sum_{i=1}^{n} 2x_i$$

$$\Rightarrow \left(\frac{\partial^2 SSE}{\partial \theta_0^2}\right)\left(\frac{\partial^2 SSE}{\partial \theta_1^2}\right) - \left(\frac{\partial^2 SSE}{\partial \theta_0 \partial \theta_1}\right)^2 = 4n \sum x_i^2 - \left(\sum 2x_i\right)^2$$

$$= n\left(\sum 4x_i^2\right) - \left(\sum 2x_i\right)^2$$

$$= n \sum (2x_i)^2 - \left(\sum 2x_i\right)^2$$

note Cauchy-Schwarz: $\left(\sum u_i v_i\right)^2 \leq \left(\sum u_i^2\right)\left(\sum v_i^2\right)$

let $u_i = 2x_i$, $v_i = 1 \Rightarrow \left(\sum 2x_i \cdot 1\right)^2 \leq \left(\sum (2x_i)^2\right)\left(\sum 1\right) = n \sum (2x_i)^2$

$\therefore n \sum (2x_i)^2 \geq \left(\sum 2x_i\right)^2 \Rightarrow n \sum (2x_i)^2 - \left(\sum 2x_i\right)^2 \geq 0$

$\therefore$ we have a minimiser by the double derivative test.

your solution (fill in during lecture):

answer:

$$\hat{\theta_0} = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2} \tag{3}$$

$$\hat{\theta_1} = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2} \tag{4}$$

We have discussed regression with only one independent variable. You can solve the multivariable case analytically as well. However, the calculations get messy without the use of linear algebra, so we will skip that here, but if you are familiar with matrices, I do suggest looking up "general linear least squares" for the elegant multivariable solution using linear algebra.

Note that if f is nonlinear, often we cannot obtain analytic solutions for the parameters, however, we can use computational optimization methods to solve for them numerically.

2.1 Simple linear regression

In the above section, we simply considered the deterministic optimization problem of minimizing SSE. We have not really introduced any of the concepts we learned last week about probability and statistics yet. This is why I was careful to call what we were doing "least squares" and not "regression." In this section, we will show that we can obtain a bit more information by using concepts from probability. We will focus on linear regression here

In this section, we will need a model

The standard linear model:

$$y_i = \theta_0 + \theta_1 x_i + \epsilon_i, \tag{5}$$

where $i \in \{1, 2, ..., n\}$ and ϵ_i is an iid. random variable with $E[\epsilon_i] = 0$ and $Var[\epsilon_i] = \sigma^2$.

Aside: Before we state the next theorem we must discuss "bias."

Definition: The *bias* of an estimator, $\hat{\theta}$, is

$$bias[\hat{\theta}] = E[\hat{\theta}] - \theta. \tag{6}$$

.

We say an estimator is *unbiased* if $bias[\hat{\theta}] = 0$ or equivalently $E[\hat{\theta}] = \theta$. If an estimator is biased, that means it is systematically higher or lower than the true value on average. Bias is different from an estimator being imprecise. The estimator can have a very large or small variance regardless of the bias.

Theorem 1: The estimators $\hat{\theta_0}$, and $\hat{\theta_1}$ given in (3), under the assumptions of the standard linear model, are unbiased estimators of θ_0, and θ_1.

In addition, under the assumptions of the standard linear model, we can also calculate the variance and covariance of these estimators. Below are the formulas for these values. I will leave the proof of the formulas for the variance estimators as an exercise. Although, you will not be required to know how to prove the covariance formula, as we will not have time to cover it. However, I wanted to at least point it out because it is important to note that the two estimators are not independent of one another. In other words, the estimate for $\hat{\theta}_0$, depends on the estimate of $\hat{\theta}_1$.

$$Var[\hat{\theta}_0] = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2} \tag{7}$$

$$Var[\hat{\theta}_1] = \frac{n\sigma^2}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2} \tag{8}$$

$$Cov[\hat{\theta}_0, \hat{\theta}_1] = \frac{-\sigma^2 \sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2}. \tag{9}$$

Now if in addition to the assumptions of the standard linear statistical model, we also assume that $\epsilon_i \sim N(0, \sigma)$, then it can be shown that the estimators are normally distributed. It is from these calculations that we can talk about p-values (for example, you can compute whether the slope is significantly different from zero). Note that the normality assumption is required to do these calculations and hence we often check for normally distributed **residuals** when performing a regression analysis. The residual is just $y_i - \hat{y}$, the distance between the observed dependent variable and the fitted value (the point on the line).

However, the quantities for the variance of the estimators require that you know the variance of the error σ, which you do not know in practice. To obtain an unbiased estimate for σ, it is useful to define the following quantity

$$RSS = \sum_{i=1}^{n} \Big[y_i - F(\hat{\theta}, x_i) \Big]^2, \tag{10}$$

which is called the **Residual Sum of Squares**. An unbiased estimate of σ can be written as

$$s^2 = \frac{RSS}{n-2}, \tag{11}$$

which can be plugged into the formulas in equations (7)-(9).

Because we have an analytic formula for all the quantities of interest for linear F, if F is nonlinear we often try to transform the data so that it follows a linear model where we can use all of the above expressions. See the Ricker Population Model example

2.2 Ricker Population Model Example

Consider the population dynamic model

$$N_{t+1} = N_t e^{r+bN_t}. \tag{12}$$

Given observed time series data for N, we would like to estimate r and b from that data. One option, rewrite the equation as

$$\log\left(\frac{N_{t+1}}{N_t}\right) = r + bN_t. \tag{13}$$

Then your dependent variable is $\log\left(\frac{N_{t+1}}{N_t}\right)$ and your independent variable is N_t, $\theta_0 = r$ and $\theta_1 = b$. **Question for you:** what have we assumed here in order to use the standard linear model and perform simple linear regression?

3 The Method of Maximum Likelihood

The method of maximum likelihood is a powerful method for parameter estimation in many contexts, not only curve fitting, but also parameter estimation for basically any data arising from probability distributions. Once, again the concept of setting derivatives equal to zero plays a central role. Suppose the set of continuous random variables $X_1, X_2, ...X_n$ have a joint density distribution

$$f(x_1, x_2, ..., x_n|\theta), \tag{14}$$

where θ is some parameter. Now, we define the *likelihood function* of that parameter, $\mathcal{L}(\theta)$ as

$$\mathcal{L}(\theta) = f(x_1, x_2, ..., x_n|\theta). \tag{15}$$

Note that everything we present in this section also works for discrete random variables in which case, just replace the probability density function, $f(x_1, , ..., x_n|\theta)$, with the probability mass function $p(x_1, , ..., x_n|\theta)$. If the random variable is discrete, $\mathcal{L}(\theta)$ can be interpreted as the probability of observing the given data if the true value of the parameter is θ.

The ***maximum likelihood estimate (MLE)*** of θ is the value of θ that maximizes $L(\theta)$. In other words, it is the value of θ that makes the observed data most likely.

If the X_i are iid then their joint density is just the product of the marginal densities

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(x_i|\theta). \tag{16}$$

For simple f one can often obtain an analytic expression for the MLE if instead, they optimize the **log likelihood** which we will define with a lower case l. **Please note that**

I am using log **here to denote** $\log_e$ **or** ln **which is typical in higher level math papers. You are free to use** ln **if you prefer that notation.** Formally,

$$l(\theta) = \log[L(\theta)] = \sum_{i=1}^{n} \log[f(x_i|\theta)] \tag{17}$$

Why is maximizing this function l equivalent to maximizing the likelihood, L, itself? **Your answer here:**

log is a monotonic function, so maximising the log of a function is the same as maximising the function.

3.1 Example: Searching for species

Say you are searching for an invasive pest species. A simple model for this is that you encounter invasive species at a random rate λ. You would like to know the distribution for the amount of time you will have to wait between spotting an individual pest. This is just an exponential distribution. Assume T is the amount of time between spottings and has the *pdf*, $f(t) = \lambda e^{-\lambda t}$. Given data on times between spottings, $t_1, ..., t_n$, as realisations of the random variable T, how would you estimate λ?

$$\mathcal{L}(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda t_i}, \tag{18}$$

meaning we have a log likelihood,

$$l(\lambda) = \sum_{i=1}^{n} [\log(\lambda) - \lambda t_i]. \tag{19}$$

Calculating the derivative and setting it equal to zero yields,

$$\begin{aligned} \frac{\partial l(\lambda)}{\partial \lambda} &= \sum_{i=1}^{n} [1/\lambda - t_i] \\ &= n/\lambda - \sum_{i=1}^{n} t_i \\ \implies \hat{\lambda} &= \frac{n}{\sum_{i=1}^{n} t_i}. \end{aligned}$$

Clearly $\hat{\lambda}$ is the MLE of λ as $l(\lambda)$ is a concave function.

Whenever you do an MLE estimator calculation, you should ask yourself if the estimator is biased. The estimators may, or may not, be unbiased. It turns out this estimator is

biased, but it is easy to correct for. We have not taught all of the tools you need to show this analytically [so I will stop here, as it requires more probability theory]. However, you can show that the estimator is biased numerically (which we might do in one of the practicals).

3.2 Example: Normal Distribution (helpful for assignment on regression)

Let $X_i, ..., X_n$ be i.i.d. $N(\mu, \sigma^2)$. Find the MLE estimators for μ and σ

$$\mathcal{L}(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left[\frac{x_i - \mu}{\sigma}\right]^2\right), \tag{20}$$

$$l(\mu, \sigma) = -n\log(\sigma) - \frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2. \tag{21}$$

The partials are

$$\frac{\partial l(\mu, \sigma)}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu).$$
$$\frac{\partial l(\mu, \sigma)}{\partial \sigma} = \frac{-n}{\sigma} + \sigma^{-3}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Setting the first partial equal to zero we have

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}. \tag{22}$$

and setting the second partial equal to zero we obtain

. $$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{23}$$

Are these estimators unbiased? Yes. And it is rather straightforward to show. Do it!

Your solution here:

$\mathbb{E}(\hat{\mu}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{n\mu}{n} = \mu \Rightarrow \mathbb{E}(\hat{\mu}) - \mu = 0$
$\Rightarrow \hat{\mu}$ has bias $= 0$

4 Estimating Population Size

In the above sections, we have mostly been concerned about fitting dynamic models to data. These all required accurate counts of population size through time. Surely it would be nearly impossible to count every single animal out in the field, so how does one actually estimate population abundance? In general, there are two approaches: (1) field sampling and (2) mark-recapture analysis. Count data just involves going out into the field and counting individual animals. If you know the amount of area, and time you surveyed you can obtain estimates of the population size. This assumes you know how good you are at detecting individuals, which in real life we rarely know. There are ways for accounting for this in count data, but an even better approach is to use *Mark-Recapture* analysis. The basic foundation of Mark recapture analysis is Maximum likelihood estimation, which we have already learned so far. There are whole courses taught in this subject in biology departments. We will only cover the basics. However, we will focus on the math behind "why" the methods work, which is sometimes glossed over in biology classes.

Consider a researcher who captures and tags t individual animals in the wild (t for tag) and then releases them. She goes out a second time and captures c individual animals (c for capture). In the second capture, there may be some animals with tags from the first capture. Let r be the number of animals with tags in the second capture (r for recaptured) and let N be the true total population size.

- N: population size
- t: number of tagged individuals from the first capture
- c: number of animals captured a second time (recaptured)
- r: number of animals recaptured, which have a tag from the first capture

The probability of recapturing r individuals in this model is given by the hypergeometric distribution. Meaning it has the *pdf*

$$f(r|N,t,c) = \frac{\binom{t}{r}\binom{N-t}{c-r}}{\binom{N}{c}}, \quad r \in \{0,1,...,\min(c,t)\}. \tag{24}$$

To understand this distribution we now must go over some notation. Reminder, $\binom{n}{k}$ is the number of possible ways to choose k distinct individuals from a population of size n, and is given by the formula,

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}. \tag{25}$$

Now let us go back to the formula for the hypergeometric distribution. The first bit is the number of ways to choose r individuals out of t individuals you tagged the first time. This is multiplied by the number of ways you can choose $c-r$ untagged individuals out of $N-t$ untagged individuals available to be caught. Together this multiplication gives you the total number of ways you can arrive at r tagged individuals. To turn this into a probability, now all you have to do is divide this number by the total number of ways you can choose any combination of c individuals from a population of size N. In other words, the denominator is the number of possible outcomes, and the numerator is the number of possible outcomes that match your observed number of tagged individuals after the recapture, r)

Now that we understand the model, we can use it to estimate N, when this parameter is unknown. In a real research situation, t and c are part of the sampling design. These quantities are known. You observe r after the second capture. So N is the only unknown quantity. From our model (24)

$$\mathcal{L}(N) = f(r|N) = \frac{\binom{t}{r}\binom{N-t}{c-r}}{\binom{N}{c}}. \tag{26}$$

But that equation has a bunch of factorials in it. How does one take the derivative of a factorial to find the optimal N that maximizes $\mathcal{L}$? Note that we want N to also be a natural number, as fractions of animals do not make sense. Taking the natural logarithm of equation (26) is not going to help either. So we will not be able to use calculus to find N. Instead, we must use other methods to maximize $\mathcal{L}(N)$ with respect to N.

One solution to finding the MLE is to look at successive ratios of the likelihood with respect to N,

$$R(N) = \frac{\mathcal{L}(N)}{\mathcal{L}(N-1)}. \tag{27}$$

Our goal is to find the value for N, $\hat{N}$, such that $R(\hat{N}) > 1$ for all $N < \hat{N}$ and $R(\hat{N}) < 1$ for all $N > \hat{N}$. Then $\hat{N}$ is the MLE.

$$
\begin{aligned}
R(N) &= \frac{\frac{\binom{t}{r}\binom{N-t}{c-r}}{\binom{N}{c}}}{\frac{\binom{t}{r}\binom{N-1-t}{c-r}}{\binom{N-1}{c}}} = \frac{\binom{N-t}{c-r}\binom{N-1}{c}}{\binom{N}{c}\binom{N-1-t}{c-r}} \\
&= \frac{(N-t)!(N-1)!(N-t-c+r-1)!(N-c)!}{(N-t-c+r)!(N-c-1)!(N-t-1)!N!} \\
&= \frac{(N-t)(N-c)}{(N-t-c+r)N}.
\end{aligned}
$$

This quantity exceeds one if

$$
\begin{aligned}
(N-t)(N-c) > (N-t-c+r)N &\iff \\
N^2 - tN - cN + ct > N^2 - tN - cN + rN &\iff \\
N < \frac{ct}{r}.
\end{aligned}
$$

It can also be shown that $R(N) < 1$ if $N > ct/r$. So the MLE is ct/r. This estimate is called the Lincoln-Peterson estimator in the ecology literature. The estimate makes sense because if we set $\hat{N} = ct/r$, then $t/\hat{N} = r/c$, in other words, it says the proportion of caught animals with a tag is equal to the proportion of animals with a tag in the wild population being sampled from.