The Role of Probability in Regression Models

FW8051 Statistics for Ecologists

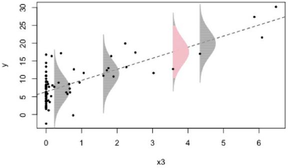
Department of Fisheries, Wildlife and Conservation Biology

Learning Objectives

Understand the role of random variables and common statistical distributions in formulating modern statistical regression models.

- Will need to know something about other statistical distributions
- Will need to have an understanding of basic probability theory
  - Probability rules and random variables
  - Expected Value
  - Variance
- How to work with probability distributions in R. . .

Linear Regression $y_i = \underbrace{\beta_0 + x_i\beta_1}_{\text{Signal}} + \underbrace{\epsilon_i}_{\text{noise}}$

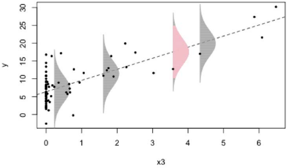


[1]

- Estimated errors, $\hat{\epsilon}_i$ given by vertical distance between points and the line
- Find the line that minimizes the errors

[1] Code and example from Jack Weiss's Ecol563: http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture4.htm. Normal distributions, above, extend to 3σ (pink = 2σ, with 1σ in gray)

Linear Regression $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$



Instead of errors, think about the normal distribution as a data-generating mechanism:

- The line gives the **expected** (average) value
- Normal curve describes the variability about this expected value.

Generalizing to other probability distributions

Replace the normal distribution as the data-generating mechanism with another probability distribution, but which one?

Leads us to...

- Discrete and continuous random variables
  - Probability mass functions (discrete random variables)
  - Probability density functions (continuous random variables)

See handout for probability rules and distributions!

Sample Space and [Frequentist] Probability

Sample space = the set of all possible outcomes that could occur.

Discrete variables:

- age class = (fawn, adult)
- dice = (1,2,3,4,5,6)

Continuous variables (range of possible values)

- age = $(0, \infty)$
- (x, y) such that x,y falls within the continental US

The probability of event A, P(A), is the long run frequency or proportion of times the event occurs.

Random variables

A random variable is a numeric quantity (or numerical event) that changes from trial to trial in a random process.

It is essentially a mapping that takes us from random events to numbers.

- Example: X = number of heads in two coin flips
- Possible events: {HH, TH, HT, TT} (all equally likely)
- Sample space of X = {0, 1, 2}

Discrete Random Variables

A random variable is discrete if it can take on a finite (or countably infinite[2]) set of possible values.

- X = Number of birds seen on a plot
- Y = (0 or 1), representing whether or not a moose calf survives its first year
- G = the species richness value obtained at a beach in the Netherlands {0, 1, 2, ...}

[2] can be put into a 1-1 correspondence with the positive integers

Continuous Random Variables

A random variable is continuous if it has values within some interval.

- T = the age at which a randomly selected adult white-tailed deer dies
- W = Mercury level (ppm) in a randomly chosen walleye from Lake Mille Lacs

Probability Mass Function: Discrete Random Variables

A probability mass function, $p(x)$ assigns a probability to each value of a discrete random variable, X.

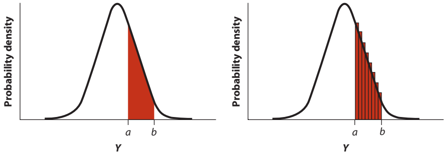
- Example: X = number of heads in two coin flips
- Possible events: {HH, TH, HT, TT} (all equally likely)
- Sample space of X = {0, 1, 2}

x	0	1	2
p(x)	1/4	1/2	1/4

Note: for any probability mass function $\sum p(x) = 1$

Continuous Distributions, Probability Density Function f(x)

For continuous variables, we define probabilities as areas under a curve, e.g., P(a $\leq X \leq$ b):

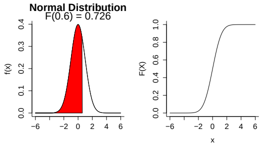


- $P(x < X < x + \triangle x) \approx f(x)\triangle x$
- Probability of any point, P(X = x) = 0
- P(a $\leq X \leq$ b) = P(a $< X <$ b)

Cumulative Density Function F(x)

Probability density function, $f(X)$

Cumulative distribution function, $F(X) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx$



- Unlike probabilities f(x) can be greater than 1
- $\int f(x)dx = 1$ (area under the curve is one)
- $F(x)$ goes from 0 to 1

Mean of a Discrete Random Variable

The mean for a discrete random variable with probability function, p(x), is given by:

$$E[x] = \sum xp(x)$$

Example: Calculate $E[x]$, where X = sum of two dice

```
x<-2:12
px<-c(1:6,5:1)/36
sum(x*px)
```

```
[1] 7
```

Total on dice	Pairs of dice	Proba
2	1+1	1/36
3	1+2, 2+1	2/36
4	1+3, 2+2, 3+1	3/36
5	1+4, 2+3, 3+2, 4+1	4/36 =
6	1+5, 2+4, 3+3, 4+2, 5+1	5/36 =
7	1+6, 2+5, 3+4, 4+3, 5+2, 6+1	6/36 =

Variance and Standard Deviation

The variance for a discrete random variable with probability function, p(x), and mean $E[x]$ is given by:

$$var(x) = E(X - E(X))^2 = \sum(x - E[x])^2 p(x) = E[x^2] - (E[x])^2$$

The standard deviation is $\sigma = \sqrt{var(x)}$

For continuous random variables

Mean: $E[x] = \mu = \int_{-\infty}^{\infty} xf(x)dx$

Variance: $\int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx$

Normal Distribution $X \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Parameters:

- $\mu = E[X]$
- $\sigma^2 = Var[x]$

Characteristics:

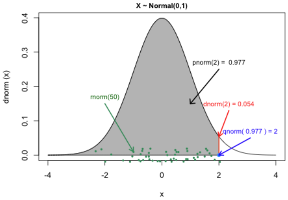
- Mean and variance are independent (knowing one tells us nothing about the other)...this is unique!
- X can take on any value (i.e., the range goes from $-\infty$ to ∞)
- R normal functions: dnorm, pnorm, qnorm, rnorm.
- JAGS: dnorm

Distributions in R

For each probability distribution in R, there are 4 basic probability functions, starting with either - d, p, q, or r:

- d is for "density" and returns the value of f(x) - probability density function (continuous distributions) - probability mass function (discrete distributions).
- p is for "probability"; returns a value of F(x), cumulative distribution function.
- q is for "quantile"; returns a value from the inverse of F(X); also know as the quantile function.
- r is for "random"; generates a random value from the given distribution.

Functions in R



Use this graph, and R help functions if necessary, to complete Exercise 9.1 in the companion book.

Why is the Normal Distribution so Popular

- Central limit theorem (as n gets large, $\bar{x}, \sum x$) become normally distributed
- Model for measurements that are influenced by a large number of factors that act in an additive way

Other notes:

- In JAGS, WinBugs, specified in terms of precision $\tau = 1/\sigma^2$
- In R, specified in terms of σ not σ^2.
- Often used for priors (Bayesian analysis) to express ignorance (e.g., N(0,100) for regression parameters).

log-normal Distribution: $X \sim \text{Lognormal}(\mu, \sigma)$

- X has a log-normal distribution of if $\log(X) \sim N(\mu, \sigma^2)$
- μ and σ are the mean and variance of $\log(X)$ not X
- Range: > 0
- R: dlnorm, plnorm, qlnorm, rlnorm with parameters `meanlog` and `sdlog`
- $E[X] = \exp(\mu + 1/2\sigma^2)$
- $Var(X) = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$
- $Var(X) = kE[X]^2$

Motivation for log-normal

CLT: if we sum a lot of independent things, then we get a normal distribution.

If we multiply a lot of independent things, we get a log-normal distribution, since:

$$log(X_1 X_2 \cdots X_n) = log(X_1) + log(X_2) + \ldots log(X_n)$$

Possible examples in biology? population dynamic models...

Lognormal Distirbution

Explore briefly in R:

```
curve(dlnorm(x, meanlog=0,sdlog=2), from=0, to=1000)
eps<-rlnorm(10000,meanlog=0, sdlog=2)
mean(eps)
var(eps)
```

Compare to the expressions for the mean and variance as a function of (μ, σ):

- $E[X] = \exp(\mu + 1/2\sigma^2)$
- $Var(X) = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1)$

Bernouli Distribution: $X \sim$ Bernouli(p)

$$f(x) = P(X = x) = p^X(1-p)^{1-x}$$

Discrete random variable with two possible outcomes

x	0	1
p(x)	1-p	p

- One parameter, p, the probability of 'success' = $P(X = 1)$
  - $0 \leq p \leq 1$
- $E[X] = \sum xp(x) = 0(1-p) + 1p = p$
- $Var[x] = \sum(x - E[x])^2 p(x)$
  $= (0-p)^2(1-p) + (1-p)^2 p = p(1-p)$
- JAGS and WinBugs: dbern
- R has only Binomial distribution (next)

Binomial random variable: $X \sim$ Binomial(n, p)

A binomial random variable counts the the number of "successes" (any outcome of interest) in a sequence of trials where

- The number of trials, n, is fixed in advance
- The probability of success, p, is the same on each trial
- Successive trials are independent of each other

Formally, a binomial random variable arises from a sum of *independent* Bernoulli random variables, each with parameter, p:

$$Y = X_1 + X_2 + \ldots X_n$$

Binomial: $X \sim \text{Binomial}(n, p)$

- $E[X] = np$
- $Var(x) = np(1-p)$
- In R: dbinom, pbinom,qbinom,rbinom
- `size` = n and `prob` = p when using these functions.

Examples:

- X = Number of heads in 2 coin flips (n = 2, p = 0.5)
- Y = number of males in a clutch, class, herd
- Z = number of animals detected among N present

Calculating Binomial Probabilities

YAHTZEE! Count the number of sixes in five dice rolls

On each roll:

- Success (S) = a "6"
- Fail (F) = any other number

X = number of S's in 5 trials:

- P(s) = p = 1/6
- n = 5

P(X = 5)? = P(SSSSS) = P(S)P(S)P(S)P(S)P(S) $= \frac{1}{6}^5$ = 0.00013

P(X = 0) = $P(F)^5 = \frac{5}{6}^5$ = 0.4019

Calculating Binomial Probabilities

X = number of S's in 5 trials:

- p = 1/6
- n = 5

P(X = 1)

= P(SFFFF) + P(FSFFF) + P(FFSFF) + P(FFFSF) + P(FFFFS)

$= 5\frac{1}{6}^1\frac{5}{6}^4$ = 0.419

- 5 = number of arrangements with one S and four F
- Probability of each arrangement = $\frac{1}{6}\{\frac{5}{6}\}^4$

Binomial Probability Function

For a binomial random variable with n trials and probability of success p on each trial, the probability of exactly k successes in the n trials is:

$P(x = k) = \binom{n}{k} p^k (1-p)^{n-k}$

$\binom{n}{k} = \frac{n!}{k!(n-k)!}$ with $n! = n(n-1)(n-2)\cdots(2)1$

Calculate P(X = 3) in the YAHTZEE example (n = 5, p = 1/6)

$= \binom{5}{3}\frac{1}{6}^3\frac{5}{6}^2 = \frac{5\cdot4\cdot3\cdot2\cdot1}{(3\cdot2\cdot1)(2\cdot1)}\frac{1}{216}\frac{25}{36}$ = 0.0322

Free Throws

Raymond Felton's free throw percentage during the 2004-2005 season at North Carolina was 70%. If we assume successive attempts are independent, what is the probability that he would hit **at least** 4 out of 6 free throws in 2005 Championship Game (he hit 5)?

$P(X \geq 4) = P(X = 4) + P(X = 5) + P(X = 6)$

$= \binom{6}{4} 0.7^4 0.3^2 + \binom{6}{5} 0.7^5 0.3^1 + 0.7^6$

```
choose(6,4)*(0.7)^4*(0.3)^2+choose(6,5)*(0.7)^5*(0.3)+
  0.7^6
```

```
[1] 0.74431
```

```
sum(dbinom(4:6, size=6, p=0.7))
```

```
[1] 0.74431
```

```
pbinom(3, size=6, p=0.7,lower.tail=FALSE)
```

```
[1] 0.74431
```

Multinomial Distribution

$$X \sim \text{Multinomial}(n, p_1, p_2, \dots, p_k)$$

- Records the number of events falling into each of k different categories out of n trials.
- Parameters: $p_1, p_2, \dots, p_k$ (associated with each category)
- $p_k = 1 - \sum_{i-1}^{k-1} p_i$
- Generalizes the binomial to more than 2 (unordered) categories
- R: dmultinom, pmultinom, qmultinom, rmultinom.
- JAGS: dmulti

Multinomial distribution

$X = (x_1, x_2, \dots, x_k)$ a multivariate random variable recording the number of events in each category

If $(n_1, n_2, \dots, n_k)$ is the observed number of events in each category, then:

$P((x_1, x_2, \dots, x_k) = (n_1, n_2, \dots, n_k)) = \frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$

Poisson Distribution: $N_t \sim Poisson(\lambda)$

Let N_t = number of events occurring in a time interval of length t. What is the probability of observing k events in this interval?

$$P(N_t = k) = \frac{\exp(-\lambda t)(\lambda t)^k}{k!}$$

Events in 2-D space, if events occur at a constant rate, the probability of observing k events in an area of size A:

$$P(N_A = k) = \frac{\exp(-\lambda A)(\lambda A)^k}{k!}$$

If A or t is constant:

$$P(N = k) = \frac{\exp(-\lambda)(\lambda)^k}{k!}$$

Poisson distribution

- Single parameter, λ = `lambda`.
- $E[X] = Var(x) = \lambda$
- R: dpois, ppois, qpois, and rpois.
- JAGS: dpois

Examples:

- Spatial statistics (null model of "complete spatial randomness""")
- Can be motivated by random event processes with constant rates of occurrence in space or time
- Binomial$(n, p) \to$ Poisson$(\lambda = np)$ as $n \to \infty$ if $p \to 0$ (such that $np \to$ a constant)

Poisson distribution

Suppose a certain region of California experiences about 5 earthquakes a year. Assume occurrences follow a Poisson distribution. What is the probability of 3 earthquakes in a given year?

```
dpois(3, lambda=5)
```

```
## [1] 0.1403739
```

```
5^3*exp(-5)/(3*2)
```

```
## [1] 0.1403739
```

Geometric Distribution

Number of failures until you get your first success.

$$f(x) = P(X = x) = (1 - p)^x p$$

- Parameter = p (probability of success)
- Range: $\{0, 1, 2, \ldots\}$
- $E[x] = \frac{1}{p} - 1$
- $Var[x] = \frac{(1-p)}{p^2}$
- *geom

Negative Binomial: Classic Parameterization

X_r = Number of failures, x, before you get r successes; $X_r \sim$ NegBinom(p)

- Total of $n = x + r$ trials
- Last trial is a success (p)
- The preceding $x + r - 1$ trials had x failures (equiv. to a binomial experiment)

$P(X = x) = \binom{x+r-1}{x} p^{r-1}(1 - p)^x p$

or

$P(X = x) = \binom{x+r-1}{x} p^r (1 - p)^x$

- $E[x] = \frac{r(1-p)}{p}$
- $Var[x] = \frac{r(1-p)}{p^2}$

Ecological Parameterization

Express p in terms of mean, μ and r:

$$\mu = \frac{r(1-p)}{p} \Rightarrow p = \frac{r}{\mu+r} \text{ and}$$
$$1 - p = \frac{\mu}{\mu+r}$$

Plugging these values in to $f(x)$ and changing r to θ, we get:

$$P(X = x) = \binom{x+\theta-1}{x} \left(\frac{\theta}{\mu+\theta}\right)^{\theta} \left(\frac{\mu}{\mu+\theta}\right)^{x}$$

Then, let θ = dispersion parameter take on any positive number (not just integers as in the original parameterization)

Negative Binomial: $X \sim \text{NegBin}(\mu, \theta)$

$$P(X = x) = \binom{x+\theta-1}{x} \left(\frac{\theta}{\mu+\theta}\right)^{\theta} \left(\frac{\mu}{\mu+\theta}\right)^{x}$$

- $E[x] = \mu$
- $Var(x) = \mu + \frac{\mu^2}{\theta}$
- In r: *nbinom, with parameters (`prob` = p, size = n) or (`mu` = μ, `size` = θ)
- JAGS: dnegbin with parameters (p, θ)

Overdispersed relative to Poisson (Var(x)/E[x] = 1 + $\frac{\mu}{\theta}$) versus 1 for Poisson

Poisson is a limiting case (when $\theta \to \infty$)

Negative Binomial

Its appeal for use as a probability generating mechanism in ecology includes the following.

- Allows for non-constant variance typical of count data.
- It often fits zero-inflated data well (and much better than a Poisson distribution).
- It respects the discreteness of the data (no need to transform).
- It can be motivated biologically - e.g.:

If: $X_i \sim \text{Poisson}(\lambda_i)$, with $\lambda_i \sim \text{Gamma}(\alpha, \beta)$, then X_i has a negative binomial distribution.

Continuous Uniform

If observations are equally likely within an interval (A,B):

$$f(x) = \frac{1}{b-a}$$

- Two parameters (a, b)
- Model of ignorance for prior distributions
- $E[x] = (a+b)/2$
- $Var(x) = \sqrt{(b-a)^2/12}$
- *unif
- JAGS: dunif(lower, upper)

Gamma Distribution: $X \sim$ Gamma(α, β)

$$f(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} \beta^{\alpha} \exp(-\beta x)$$

- Range 0 to ∞
- $\Gamma(\alpha)$ is a generalization of the factorial function (!) that we've seen earlier
- α and β are parameters > 0.
- $E[x] = \frac{\alpha}{\beta}$
- $Var[x] = \frac{\alpha}{\beta^2}$
- R: *gamma

Beta Distribution: $X \sim$ Beta(α, β)

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}$$

- ranges from 0 to 1.
- α and β are parameters > 0.
- $E[x] = \frac{\alpha}{\alpha+\beta}$
- $Var[x] = \frac{\alpha}{(\alpha+\beta)^2(\alpha+\beta+1)}$
- R: *beta

Exponential: $X \sim$ Exp(λ)

$$f(x) = \lambda \exp(-\lambda x)$$

- Range 0 to ∞
- $\lambda > 0$
- $E[x] = \frac{1}{\lambda}$
- $Var[x] = \frac{1}{\lambda^2}$
- R: *exp

Distributions

How do we choose an appropriate distribution for our data? (Zuur et al. ch 8.7.1):

- Presence-absence (0,1) at M sites $\rightarrow$ Bernoulli distribution
- Counts, fixed number of sites/trials/etc $\rightarrow$ Binomial distribution
- Counts, in a fixed unit of time, area
    - Normal distribution if the counts are large
    - Poisson distribution: if $E[Y|X] \approx Var(Y|X)$
    - If $Var(Y|X) > E(Y|X)$: Negative Binomial, Quasipoisson, Poisson-normal model
- Continuous response variable: normal distribution (usual default)
    - gamma (if Y must be > 0)
    - lognormal (if skewed data)
- Time to event: exponential, Weibull

Other useful information

For a diagram showing links between distributions, see:

Diagram of distribution relationships

- {http://www.johndcook.com/distribution/_chart.html}

See handout with distributions (note that some can be written in multiple ways):

For example, gamma: $f(x) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} \beta^{\alpha} \exp(-\beta x)$

$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} \exp(-x/\beta)$