Generalized linear model

General Linear Models

$$y_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} + \varepsilon_i$$

- T test / U test
- ANOVA
- Simple linear regression
- ANCOVA
- Multiple linear regression

Generalized Linear Model

$$g(\mu_i) = \beta_0 + \beta_1 x_{1i} + ... + \beta_k x_{ki} + \varepsilon_i$$

GLM is an extension of general linear model that deals with ordinal and categorical response variables.

There are three components that are common to all GLMs (McCullagh & Nelder 1989) :

- Random component
- Systematic Component
- Link Function

McCullagh, P., and J. A. Nelder 1989. Generalized linear models. Chapman and Hall.

Random Component:

The random component: refers to the probability distribution of the response Y.

Case 1. $(Y_1, Y_2, \ldots, Y_N)$ might be normal. In this case, we would say the random component is the normal distribution. This component leads to ordinary regression and analysis of variance models.

Case 2. If the observations are Bernoulli random variables (which have values 0 or 1), then we would say the link function is the binomial distribution. When the random component is the binomial distribution, we are commonly concerned with logistic regression models or probit models.

Case 3. Quite often the random variables $Y_1, Y_2, \ldots, Y_N$ have a Poisson distribution. Then we will be involved with Poisson regression models or loglinear models.

Systematic Component

The systematic component involves the explanatory variables $x_1, x_2, \cdots, x_k$ as linear predictors:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Link Function

The third component of a GLM is the link between the random and systematic components**.** It says how the mean **μ** = E(Y) relates to the explanatory variables in the linear predictor through specifying a function g(μ):

$$g(\mu_i) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

g(μ) is called the link function.

Generalized Linear Models

- The y_i's are allowed to have a distribution from the exponential family of distributions.

- The link function $g(\mu_i)$ is any monotonic function and defines the relationship between μ_i and $x\beta$.

$$g(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}$$

Logistic Regression

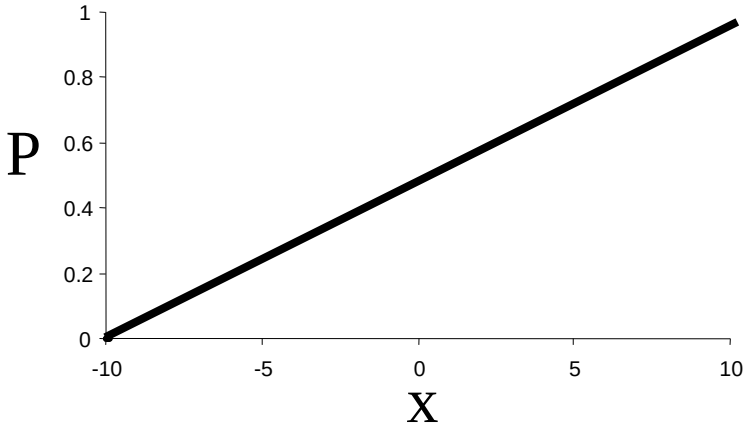
Dependent variable is binary

To assess the effects of multiple explanatory variables (which can be numeric and/or categorical) on the dependent variable

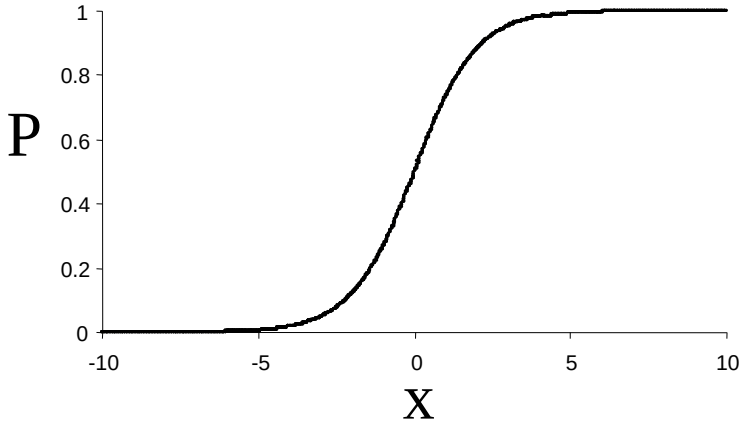
McCullagh, P., and J. A. Nelder 1989. Generalized linear models. Chapman and Hall.

Why use logistic regression

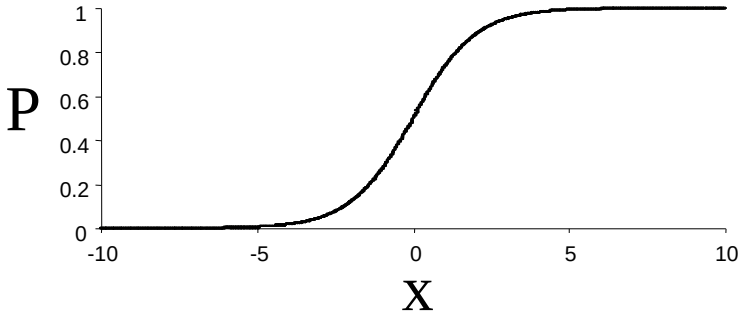
Dependent variable is binary



Linear function



Logistic function



Probit regression function

$$P(y_i = 1 \mid x_i) = p_i = \int_{-\infty}^{x_i} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2} t^2) dt$$

$$P(y_i = 1 \mid x_i) = p_i = \frac{1}{1 + e^{-(x_i)}}$$

$$P(y_i = 0 \mid x_i) = p_i = \frac{1}{1 + e^{(x_i)}}$$

Logit transformation

$$P(y_i = 1 \mid x_i) = p_i = \frac{1}{1+e^{-(x_i)}}$$

$$= \frac{e^{x_i}}{1+e^{x_i}}$$

$$1 - p_i = 1 - \frac{e^{x_i}}{1+e^{x_i}} \quad = \frac{1}{1+e^{x_i}}$$

$$Odds = \frac{p_i}{1-p_i} = e^{x_i}$$

$$\text{logit}\,(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = x_i$$

Model meanings – nest site use of birds

$$Odds = \frac{p_i}{1 - p_i} = e^{x_i}$$

The response variable was the odds of a site having a nest, where odds are calculated as p/(1-p) and p is the proportion of sites have a nest. The statistical model was:

$Odds = exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_n X_n) + error$

where *n* is the number of explanatory variables and *error* is consistent with a binomial distribution for *p*. The log of the odds is known as the logit transform of *p*.

Advantages of Logit

- Properties of a linear regression model
- Logit(P): between - ∞ and + ∞
- Probability (P) constrained between 0 and 1

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta x \qquad \frac{P}{1-P} = e^{\alpha+\beta x}$$

- Directly related to odds of event

Assumptions

To be satisfied	Does not matter
Dependent variable is binary or dichotomous	Linearity (the population means of the dependent variables at each level of the independent variable are not on a straight line)
The cases are independent	Homogeneity of variance (the variance of the errors are not constant)
The independent variables are not linear combinations of each other	Normality (the errors are not normally distributed)

Example

- Risk of developing coronary heart disease (CD) by age (< 60 and > 60 years old)

CD ~ age

CD	> 60 (1)	< 60 (0)
Present (1)	28	23
Absent (0)	11	72

Odds of disease among the old = 28/11
Odds of disease among the young = 23/72 **Odds ratio = 7.97**

R code

```
# Logistic regression
# Risk of developing coronary heart disease (CD) by age (<60 and >60 years old)
coronary1 <- data.frame(CD = rep(1, 28), age = 'old')
coronary2 <- data.frame(CD = rep(0, 11), age = 'old')
coronary3 <- data.frame(CD = rep(1, 23), age = 'young')
coronary4 <- data.frame(CD = rep(0, 72), age = 'young')
coronary <- rbind(coronary1, coronary2, coronary3, coronary4)
coronary <- rbind(coronary3, coronary4, coronary1, coronary2)
fit <- glm(CD ~ age, data = coronary, family = binomial())
summary(fit)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.9343	0.3558	2.626	0.00865 **
ageyoung	-2.0755	0.4289	-4.839	1.31e-06 ***

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.1412	0.2395	-4.765	1.89e-06 ***
ageold	2.0755	0.4289	4.839	1.31e-06 ***

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1 \times \text{Age} = -1.1412 + 2.0755 \times \text{Age}$$

Logistic regression model

	Coefficient	SE	Coeff/SE
Age	**2.0755**	**0.4289**	**4.839**
Constant	**-1.1412**	**0.2395**	**-4.765**

- β = increase in **logarithm of odds** for a one unit increase in x
- Test of the hypothesis that $\beta = 0$ (Wald test)

$$\chi 2 = \frac{\beta^2}{\text{Variance } (\beta)} \qquad (1\,\text{df})$$

$$\text{Odds ratio} = e^{2.0755} = 7.97$$

$$\text{Wald Test} = 4.839^2 \text{ with 1df } (p < 0.05)$$

$$95\%\ \text{CI} = e^{(2.0755 \pm 1.96 \times 0.4289)} = 3.4, 18.5$$

Interpretation of the coefficients in terms of the odds ratio – An Example

- Presence of a predator depending on prey density.
- 17 observations, 14 presence and 3 absence.

Prey density	Presence
10	0
10	1
10	1
11	0
11	1
11	1
11	1
11	1
12	0
12	1
12	1
12	1
12	1
12	1
12	1
12	1
12	1

```
pre1 <- data.frame(prey = c(10:12),  predator = rep(0, 3))
pre2 <- data.frame(prey = rep(c(10:12),  c(2, 4, 8)),  predator = rep(1, 14))
pre <- rbind(pre1,  pre2)
fit <- glm(predator ~ prey,  data = pre,  family = binomial())
summary(fit)
```

Variables in the Equation					
	B	S.E.	Wald	df	Exp(B)
Prey density	**0.6931**	**0.8072**	**0.7372**	**1**	**2.0**
Constant	-6.2383	8.9794	0.4826	1	0.00195

Interpretation of the coefficients in terms of the odds ratio – An Example

$$\ln\left(\frac{\mathrm{P}}{1-\mathrm{P}}\right) = \alpha + \beta \times density = \alpha + 0.69 \times density$$

- $e^{\beta} = 2$

$$\frac{\mathrm{P}}{1-\mathrm{P}} = e^{\alpha} \times e^{0.69 \times density} = e^{\alpha} \times 2^{density}$$

- So: increasing the prey density by one unit increases the odds of predator presence by a factor of 2 (increase in 100%), so that odds ratio is:

(odds after increasing prey density)/ (odds before increasing prey density) = 2

- If we look at the data we can see that this model predicts perfectly:

	Predator				
Prey density	**1**	**0**	**P(presence)**	**P(absence)**	**Odds of presence**
10	2	1	2/3=0.66	1/3=0.33	0.66/0.33=2
11	4	1	4/5=0.8	1/5=0.2	0.8/0.2=4
12	8	1	8/9=0.888	1/9=0.111	0.888/0.111=8

Marginal effect of a change in X

- $\ln[p/(1-p)] = \alpha + \beta X + e$

  The slope coefficient (β) is interpreted as the rate of change in the "log odds" as X changes … not very useful.

- We are also interested in seeing the effect of an explanatory variable on the probability of the event occurring

- $p = 1/[1 + \exp(-\alpha - \beta X)]$

  The marginal effect of a change in X on the probability is:

  $\partial p/\partial X = \beta p(1-p)$

$$= \beta \times \frac{1}{1+e^{-(\alpha+\beta X)}} \times \frac{1}{1+e^{(\alpha+\beta X)}}$$

Basically, the size of the 'marginal effect' will depend on two things:

- β coefficient
- The initial value of X

Marginal Effects: βxP(1-P)

- Passing or failing an exam as a function of the number of hours of study
- Previous study indicated the estimates of α and β were:

$$\alpha = -5 , \ \beta = 0.3$$

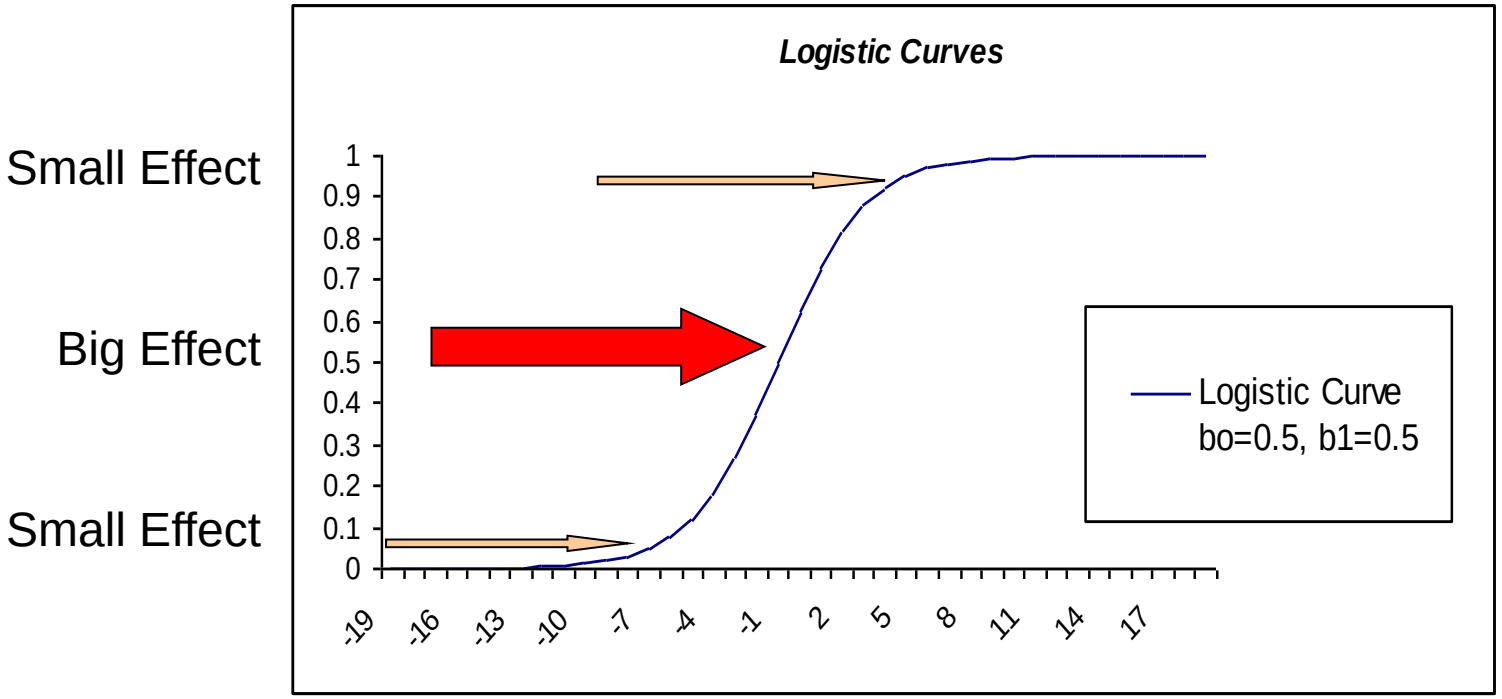
- So what's the effect of studying one more hour in the probability of the event occurring:

$$\beta \times \frac{1}{1+e^{-(\alpha+\beta X)}} \times \frac{1}{1+e^{(\alpha+\beta X)}}$$

Initial hours of study (x)	P	1-P	P(1-P)	Marginal effect
5	0.029	0.971	0.028	0.009
10	0.119	0.881	0.105	0.031
15	0.378	0.622	0.235	0.071
20	0.731	0.269	0.197	0.059
25	0.924	0.076	0.070	0.021
30	0.982	0.018	0.018	0.005

The importance of the initial value of X in the marginal effect

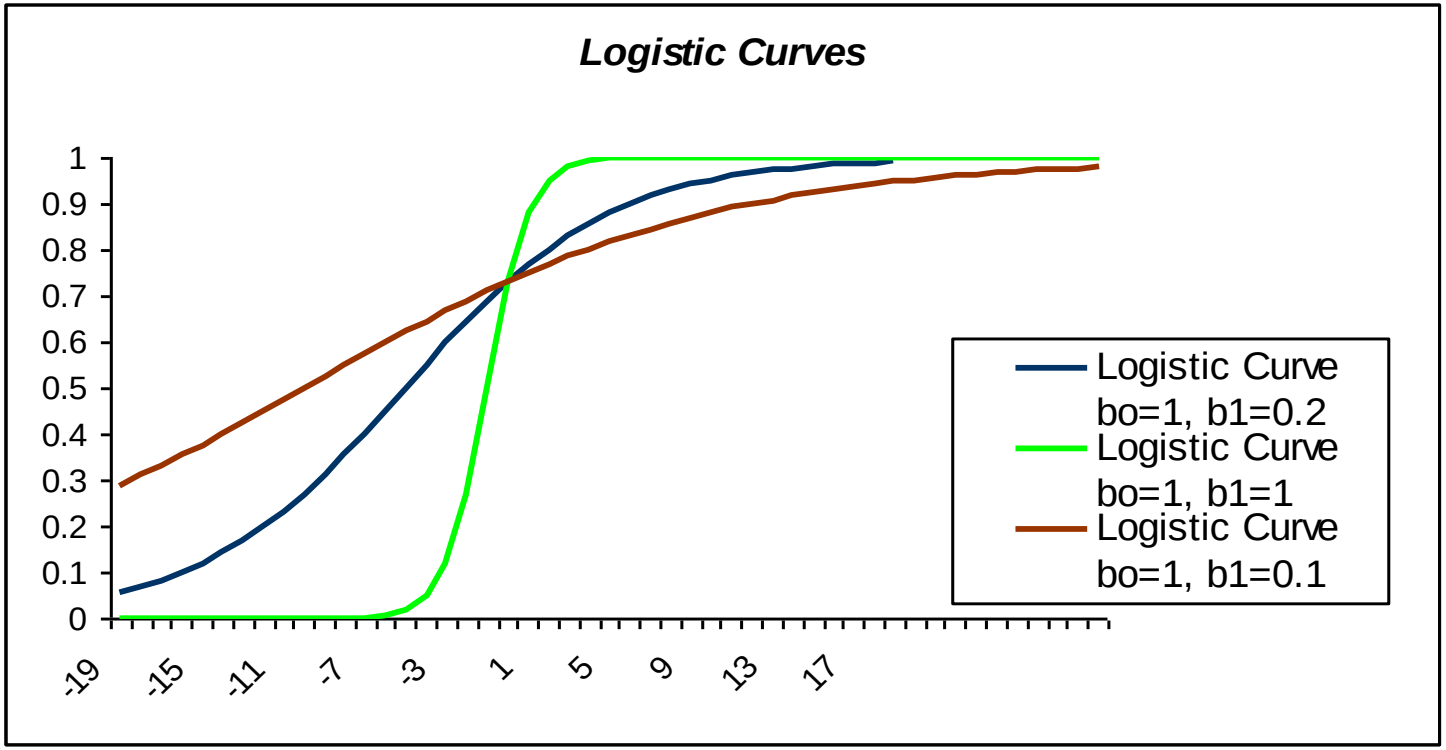
- Even with the same β, the marginal effect is different depending on where we are evaluating the change.



- Starting the change from the central values of X will have a higher impact on the probability of the event occurring than starting from very low or very high values of X

How important is β in determining the marginal effect?

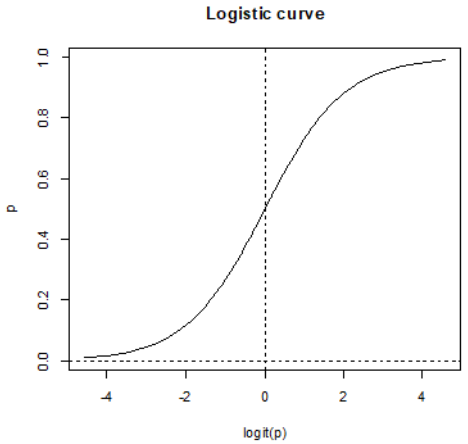
- As we have already seen, the larger is β, the steeper is the curve:



- So, the larger the β, the larger the impact of an increase in X on the probability of the event occurring

R code for logistic regression - glm

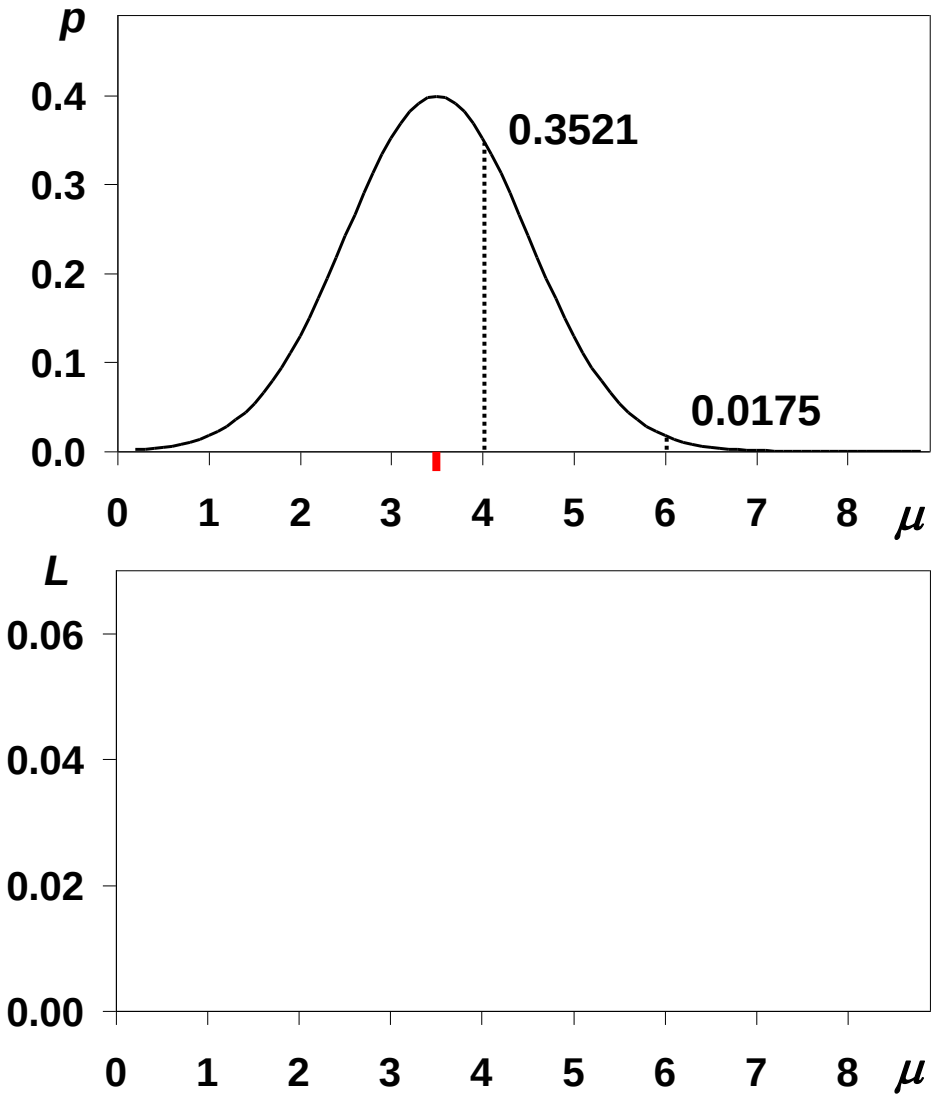
```
#Logistic regression curve
logit = function(p) log(p/(1 - p))
p = seq(0.01, 0.99, by = 0.01)
plot(logit(p),p, type = "l", main = "Logistic curve")
abline(h = 0, lty = 2); abline(v = 0, lty = 2)

#Logistic regression example
head(trees); use=sample(0:1,length(trees$Girth), rep=TRUE)
data1=cbind(trees, use)
names(data1); table(use); summary(data1$Girth); sd(data1$Girth)

fit = glm(use~Girth+Height+Volume,data=data1,family=binomial())
summary(fit) # display results
confint(fit) # 95% CI for the coefficients
exp(coef(fit)) # exponentiated coefficients
exp(confint(fit)) # 95% CI for exponentiated coefficients
pred = predict(fit, type="response") # predicted values
res = residuals(fit, type="deviance") # residuals
x11(); plot(pred, res)
plot(data1$Girth, pred)
op = par(mfrow = c(2, 2), pty = "s"); plot(fit)
```



How to estimate model coefficients

Maximum likelihood estimation (MLE)

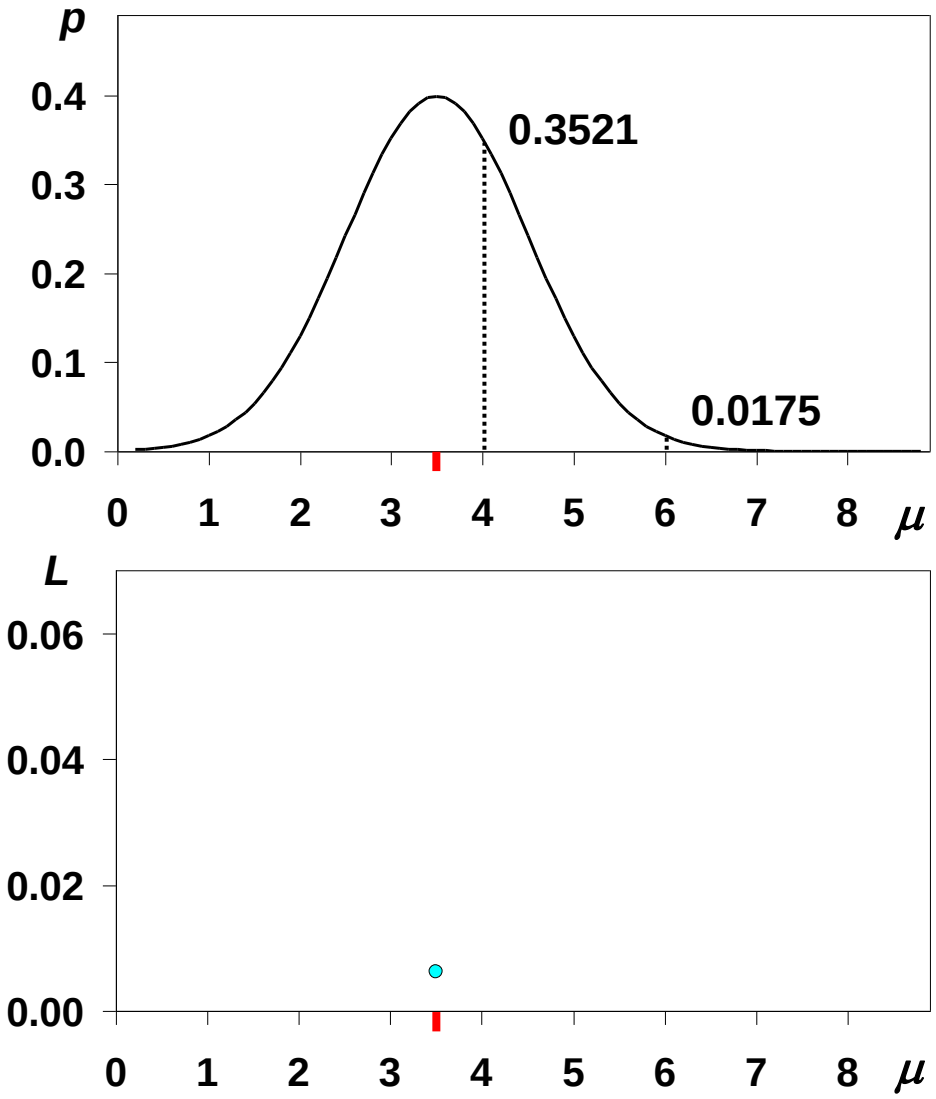
- Most statistical methods are designed to *minimize error*.
  - Choose the parameter values that minimizes predictive error: $|y - y'|$ or $(y - y')^2$
- Maximum likelihood estimation seeks the parameter values that are *most likely* to have produced the observed distribution.



Use sample (4, 6) to estimate population mean (SD=1)

μ	$p(4)$	$p(6)$
3.5	0.3521	0.0175

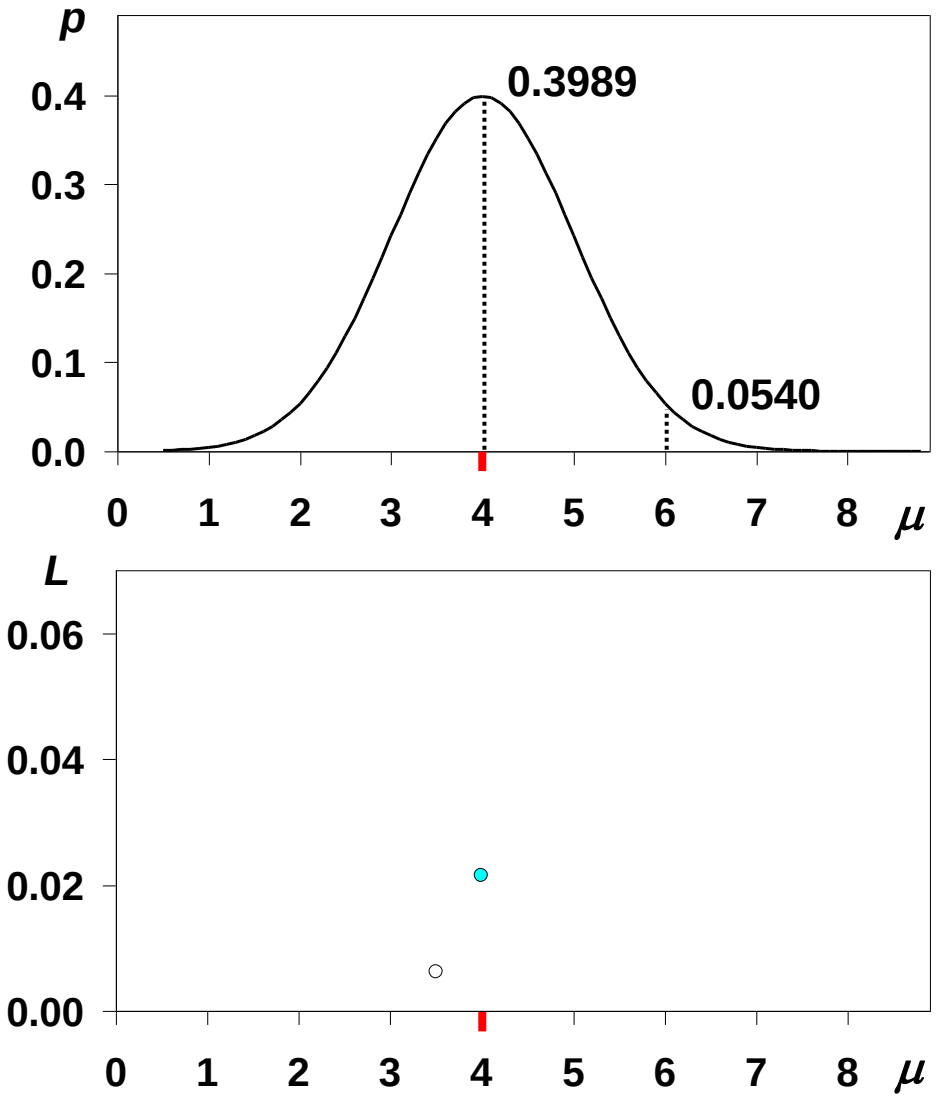
Suppose initially you consider the hypothesis $\mu = 3.5$. Under this hypothesis the probability density at 4 would be 0.3521 and that at 6 would be 0.0175.



Use sample (4, 6) to estimate population mean (SD=1)

μ	$p(4)$	$p(6)$	L
3.5	0.3521	0.0175	0.0062

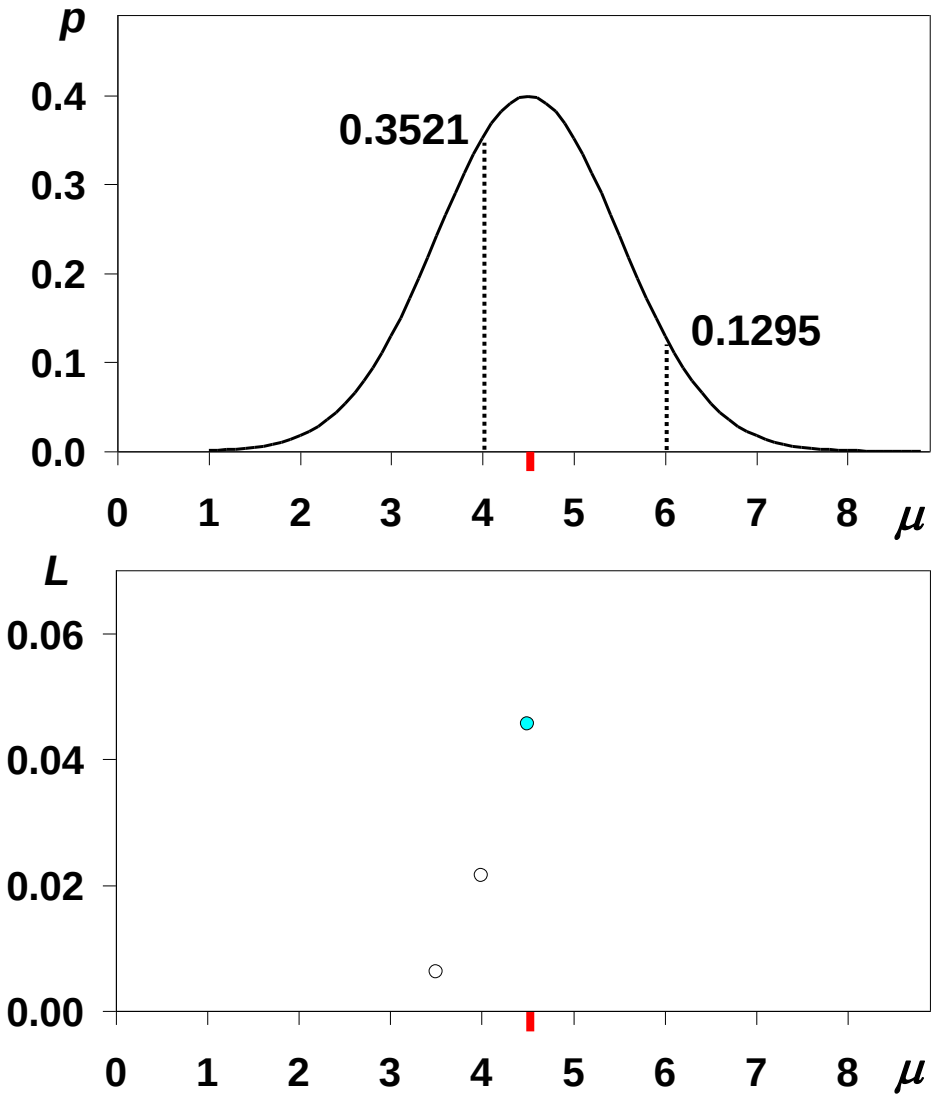
The joint probability density, shown in the bottom chart, is the product of these, 0.0062.



Use sample (4, 6) to estimate population mean (SD=1)

μ	$p(4)$	$p(6)$	L
3.5	0.3521	0.0175	0.0062
4.0	0.3989	0.0540	0.0215

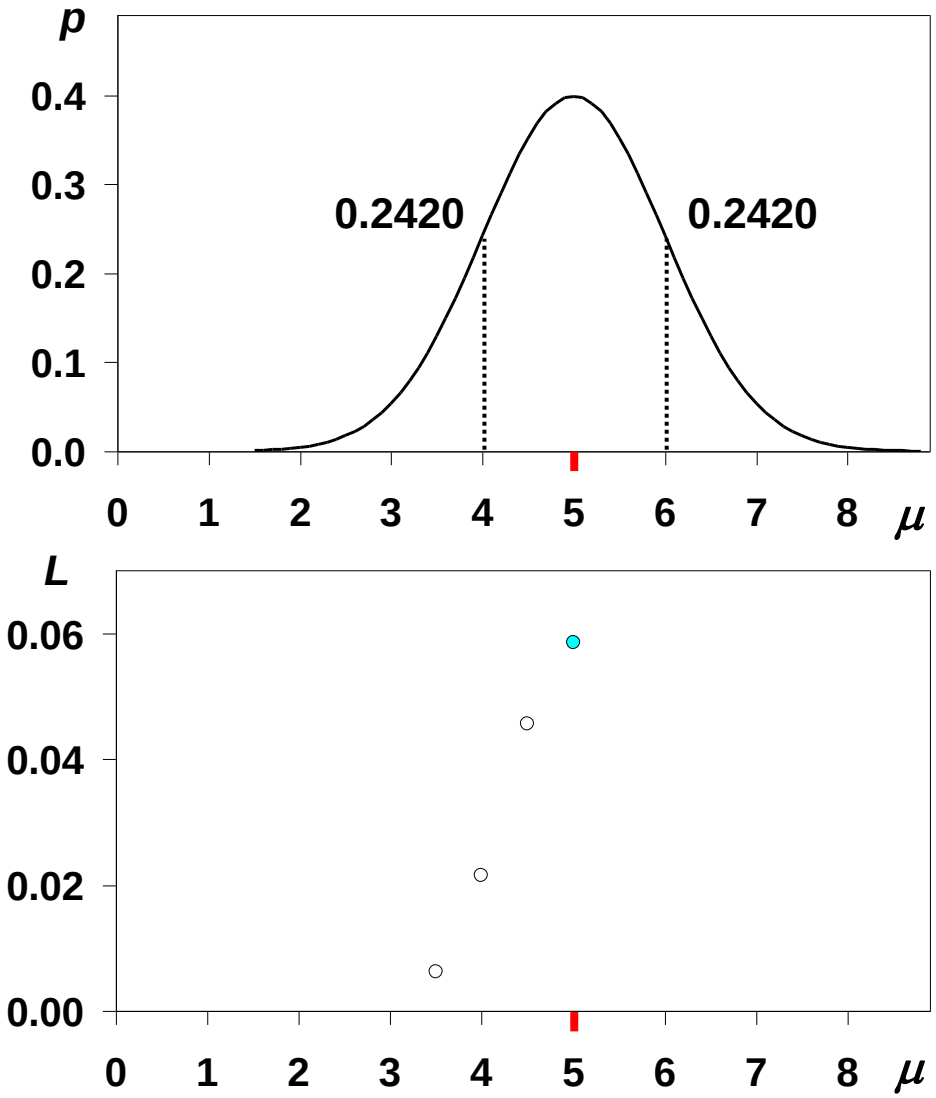
Next consider the hypothesis $\mu = 4.0$. Under this hypothesis the probability densities associated with the two observations are 0.3989 and 0.0540, and the joint probability density is 0.0215.



Use sample (4, 6) to estimate population mean (SD=1)

μ	$p(4)$	$p(6)$	L
3.5	0.3521	0.0175	0.0062
4.0	0.3989	0.0540	0.0215
4.5	0.3521	0.1295	0.0456

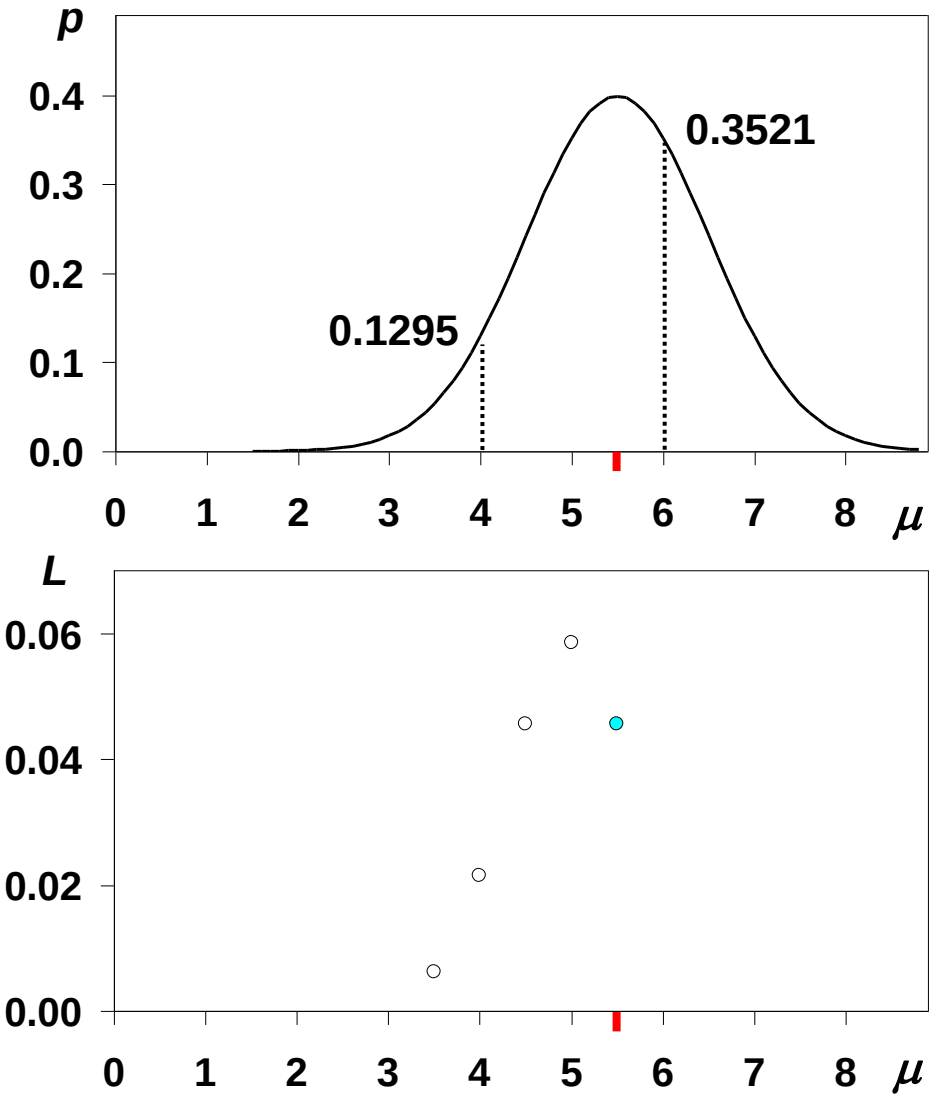
Under the hypothesis $\mu = 4.5$, the probability densities are 0.3521 and 0.1295, and the joint probability density is 0.0456.



Use sample (4, 6) to estimate population mean (SD=1)

μ	$p(4)$	$p(6)$	L
3.5	0.3521	0.0175	0.0062
4.0	0.3989	0.0540	0.0215
4.5	0.3521	0.1295	0.0456
5.0	0.2420	0.2420	0.0585

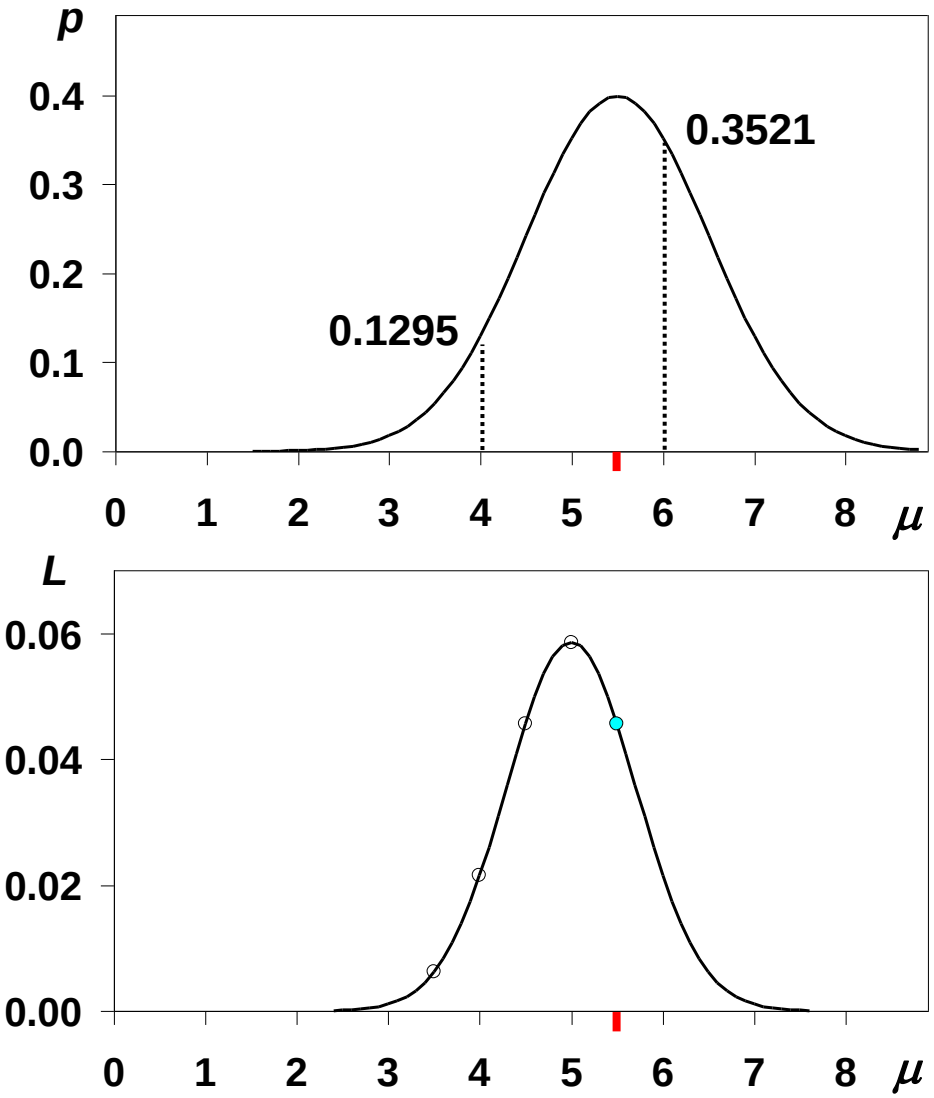
Under the hypothesis $\mu = 5.0$, the probability densities are both 0.2420 and the joint probability density is 0.0585.



Use sample (4, 6) to estimate population mean (SD=1)

μ	$p(4)$	$p(6)$	L
3.5	0.3521	0.0175	0.0062
4.0	0.3989	0.0540	0.0215
4.5	0.3521	0.1295	0.0456
5.0	0.2420	0.2420	0.0585
5.5	0.1295	0.3521	0.0456

Under the hypothesis $\mu = 5.5$, the probability densities are 0.1295 and 0.3521 and the joint probability density is 0.0456.



Use sample (4, 6) to estimate population mean (SD=1)

μ	$p(4)$	$p(6)$	L
3.5	0.3521	0.0175	0.0062
4.0	0.3989	0.0540	0.0215
4.5	0.3521	0.1295	0.0456
5.0	0.2420	0.2420	0.0585
5.5	0.1295	0.3521	0.0456

The complete joint density function for all values of μ has now been plotted in the lower diagram. We see that it peaks at $\mu = 5$.

Joint density

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

$$f(X) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(X-\mu)^2}$$

$$f(4) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(4-\mu)^2} \qquad f(6) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(6-\mu)^2}$$

$$\text{joint density} = \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(4-\mu)^2}\right)\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(6-\mu)^2}\right)$$

In maximum likelihood estimation we choose as our estimate of μ the value that gives us the greatest joint density for the observations in our sample. This value is associated with the greatest probability, or maximum likelihood, of obtaining the observations in the sample.

Log likelihood

$$L(\mu \mid 4,6) = \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(4-\mu)^2} \right)\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(6-\mu)^2} \right)$$

$$\log L = \log\left[\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(4-\mu)^2} \right)\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(6-\mu)^2} \right) \right]$$

$$= \log\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(4-\mu)^2} \right) + \log\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(6-\mu)^2} \right)$$

$$= \log\left(\frac{1}{\sqrt{2\pi}} \right) + \log\left(e^{-\frac{1}{2}(4-\mu)^2} \right) + \log\left(\frac{1}{\sqrt{2\pi}} \right) + \log\left(e^{-\frac{1}{2}(6-\mu)^2} \right)$$

$$= 2\log\left(\frac{1}{\sqrt{2\pi}} \right) - \frac{1}{2}(4-\mu)^2 - \frac{1}{2}(6-\mu)^2$$

We will now choose μ so as to maximize this expression.

Maximize log likelihood

$$\log L = 2\log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}(4-\mu)^2 - \frac{1}{2}(6-\mu)^2$$

$$-\frac{1}{2}(a-\mu)^2 = -\frac{1}{2}\left(a^2 - 2a\mu + \mu^2\right) = -\frac{1}{2}a^2 + a\mu - \frac{1}{2}\mu^2$$

$$\frac{d}{d\mu}\left\{-\frac{1}{2}(a-\mu)^2\right\} = a-\mu$$

$$\frac{d\log L}{d\mu} = (4-\mu)+(6-\mu)$$

$$\frac{d\log L}{d\mu} = 0 \quad \Rightarrow \quad \hat{\mu} = 5$$

Thus we confirm that 5 is the value of μ that maximizes the log-likelihood function, and hence the likelihood function.

More observations

$$f(X_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(X_i-\mu)^2}$$

$$L(\mu \mid X_1,...,X_n) = \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(X_1-\mu)^2} \right) \times ... \times \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(X_n-\mu)^2} \right)$$

$$\log L = \log \left[\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(X_1-\mu)^2} \right) \times ... \times \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(X_n-\mu)^2} \right) \right]$$

$$= \log \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(X_1-\mu)^2} \right) + ... + \log \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(X_n-\mu)^2} \right)$$

$$= n \log \left(\frac{1}{\sqrt{2\pi}} \right) - \frac{1}{2}(X_1-\mu)^2 - ... - \frac{1}{2}(X_n-\mu)^2$$

Maximizing log *L* with respect to μ.

Estimate the mean (mu)

$$\log L = n\log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}(X_1 - \mu)^2 - \ldots - \frac{1}{2}(X_n - \mu)^2$$

$$\frac{d\log L}{d\mu} = (X_1 - \mu) + \ldots + (X_n - \mu)$$

$$\frac{d\log L}{d\mu} = 0 \quad \Rightarrow \quad \sum X_i - n\hat{\mu} = 0$$

$$\therefore \hat{\mu} = \frac{1}{n}\sum X_i = \overline{X}$$

We have demonstrated that the maximum likelihood estimator of μ is the sample mean.

Estimate the variance

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

$$\log L = n\log\left(\frac{1}{\sigma}\right)+n\log\left(\frac{1}{\sqrt{2\pi}}\right)+\frac{1}{\sigma^2}\left(-\frac{1}{2}(X_1-\mu)^2-\ldots-\frac{1}{2}(X_n-\mu)^2\right)$$

$$= -n\log\sigma+n\log\left(\frac{1}{\sqrt{2\pi}}\right)-\frac{\sigma^{-2}}{2}\sum(X_i-\mu)^2$$

$$\frac{\partial \log L}{\partial \sigma}=-\frac{n}{\sigma}+\sigma^{-3}\sum(X_i-\mu)^2$$

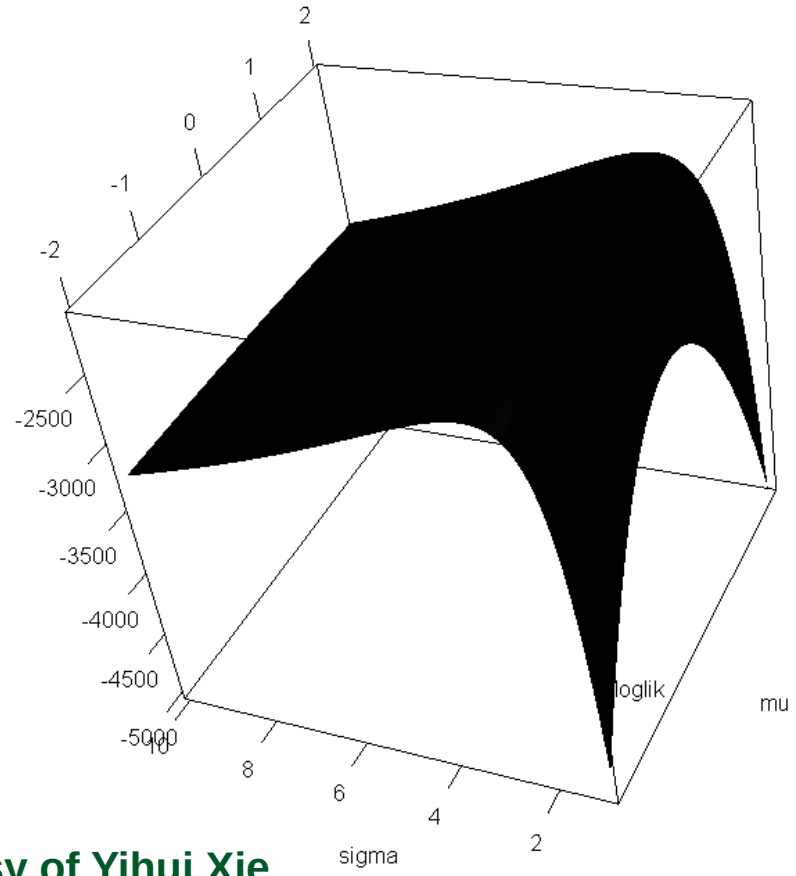
$$\frac{\partial \log L}{\partial \sigma}=0 \quad\Rightarrow\quad -\frac{n}{\hat{\sigma}}+\hat{\sigma}^{-3}\sum(X_i-\hat{\mu})^2=0$$

$$\therefore -n\hat{\sigma}^2+\sum(X_i-\bar{X})^2=0$$

$$\therefore \hat{\sigma}^2=\frac{1}{n}\sum(X_i-\bar{X})^2$$

It is biased estimation. The unbiased estimator is obtained by dividing by *n* – 1, not *n*.

Maximum likelihood function for a normal distribution n(0,2)




```
mu=seq(-2,2,len=100)
sigma=seq(1,10,len=100)
x=rnorm(1000,mean=0,sd=2)
loglik=matrix(apply(expand.grid(mu,sigma),1,function(z)sum(dnorm(x,mean=z[
1],sd=z[2],log=TRUE))),nrow=length(mu),ncol=length(sigma))
library(rgl)
#-------------------------------------------------------------------------#
open3d=function (..., params = get("r3dDefaults", envir = .GlobalEnv))
{
    rgl.open()
    args <- list(...)
    if (!is.null(args$material)) {
        params$material <- do.call(.fixMaterialArgs, c(args$material,
            Params = list(params$material)))
        args$material <- NULL
    }
    params[names(args)] <- args
    clear3d("material", defaults = params)
    params$material <- NULL
    if (!is.null(params$bg)) {
        do.call("bg3d", params$bg)
        params$bg <- NULL
    }
    do.call("par3d", params)
    return(rgl.cur())
}
#-------------------------------------------------------------------------#
open3d(windowRect=c(0,0,1024,768))
persp3d(mu,sigma,loglik)
M = par3d("userMatrix")
play3d(par3dinterp(userMatrix = list(M, rotate3d(M,
    pi/2, 1, 0, 0), rotate3d(M, pi/2, 0, 1, 0), rotate3d(M, pi,
    0, 0, 1))), duration = 10)
#movie3d(par3dinterp(userMatrix = list(M, rotate3d(M,
#    pi/2, 1, 0, 0), rotate3d(M, pi/2, 0, 1, 0), rotate3d(M, pi,
#    0, 0, 1))), duration = 5,fps=10,dir=getwd(),clean=F,convert=T)
#png(bg='black')
#par(mar=rep(0,4),fg='white',col.axis='white',col.lab='white',col.sub='white')
#persp(mu,sigma,loglik,col='lightblue',border=NA,theta=60,phi=30)
#dev.off()
```

 # Courtesy of Yihui Xie

```
mu=seq(-2, 2, len=100)
sigma=seq(1,10, len=100)
x=rnorm(1000, mean=0, sd=2)
loglik=matrix(apply(expand.grid(mu, sigma), 1, function(z) sum(dnorm(x, mean=z[1], sd=z[2],
        log=TRUE))), nrow=length(mu), ncol=length(sigma))
```

Likelihood Function

- Define a model by its pdf:
  - $f_X(x;\theta)$, where θ are the parameters of the pdf, constant across sampled data
- The likelihood function is:

$$L(x_1, x_2, ..., x_n; \theta) \equiv \prod_n f_X(x_i; \theta)$$

  - where there are n sets of sample data.
  - (Note, x and θ are often vectors)

Example: coin toss

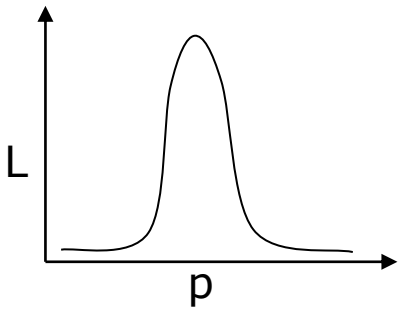
- n independent coin tosses, k heads
  - Binomial distribution, parameter p

$$L(x_1, x_2, \ldots, x_n; p) \equiv \prod_n f_X(x_i; p)$$

$$= \frac{n!}{(n-k)!k!} p^k (1-p)^{(n-k)}$$

  - Given data: 100 trials, 56 heads:

$$= \frac{100!}{44!56!} p^{56} (1-p)^{44}$$



  - Numerical solution yields max p=0.56

R code: maximum likelihood estimation (MLE)

```
# Negative log likelihood
negLL = function(psi) (-1)*dbinom(z, size=n, prob=psi, log=TRUE)

n = 5 # 5 independent coin tosses
z = 1 # 1 heads

# General-purpose optimization based on Nelder–Mead, quasi-Newton and
# conjugate-gradient algorithms. It includes an option for box-constrained
# optimization and simulated annealing
# Method "BFGS" is a quasi-Newton method (also known as a variable metric algorithm), specifically that published simultaneously in 1970 by
# Broyden, Fletcher, Goldfarb and Shanno. This uses function values and gradients to build up a picture of the surface to be optimized.
fit = optim(0.5, negLL, method='BFGS')

list(logLikelihood = fit$value, mle = fit$par)
```

```
$logLikelihood
0.8925742
$mle
0.2000002
```

Maximum likelihood estimates of parameters

- For MLE, the goal is to determine the most likely values of the population parameter value (e.g, μ, σ, β, ρ, …) given an observed sample value (e.g., x-bar, s, b, r, ….)

  - Any model's parameters (e.g., β in linear regression, a, b, c, etc. in nonlinear models) can be estimated using MLE.

Likelihood is based on the shape of the DV's distribution

- ANOVA, Pearson's *r*, *t*-test, regression… all assume that DV's residual is normally distributed.
  - Under those conditions, the LSE (least squares estimate) ***is*** the MLE.
- If the DV's residual is not normally distributed, the LSE is not the MLE.

MLE for logistic regression

For one observation

$$P(y_i) = p_i^{y_i}(1-p_i)^{1-y_i}$$

Likelihood function

$$L(\theta) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}$$

$$p = \frac{1}{1+e^{-(\theta X)}}$$

Estimating maximum likelihood

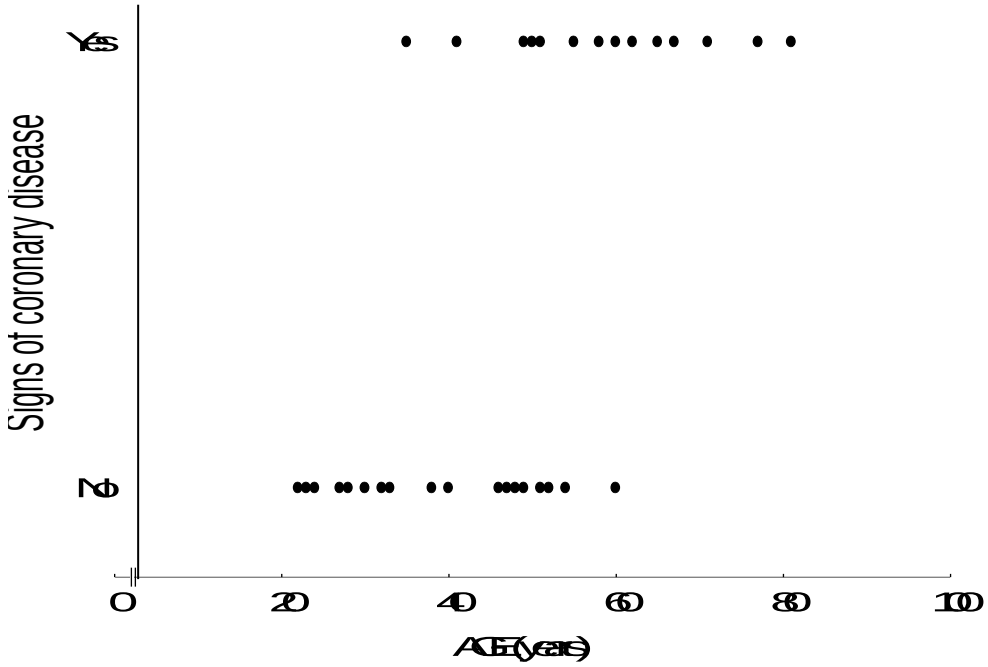
In logistic regression, MLE is an iterative algorithm which starts with an initial arbitrary "guesstimate" of what the logit coefficients should be, the MLE algorithm determines the direction and size change in the logit coefficients which will increase LL.

After this initial function is estimated, the residuals are tested and a re-estimate is made with an improved function, and the process is repeated (usually about a half-dozen times) until *convergence* is reached (that is, until LL does not change significantly).

$$L(\theta) = \prod_{i=1}^{n} p_i^{\,y_i}(1-p_i)^{1-y_i}$$

Why not ordinary least squares (OLS)?

- In logistic regression the relevant outcome is the probability of an event
- Since the dependent variable (*y*) is bounded between 0 and 1, OLS is inappropriate for binary outcome variables



Goodness of fit for the full model - likelihood ratio test (LR)

- We compare the value of the likelihood function in a model (with the variables) with the value of the likelihood function in a model (without the variables). The test:

$$LR = -2L\hat{L}_0 - (-2L\hat{L}_S) \Rightarrow \chi^2_k$$

where $L\hat{L}_0$ is the log likelihood value of the null model (only intercept included); $L\hat{L}_S$ is the log likelihood value of the full model (taking into account of all variable parameters).

  - The statistic is distributed as χ^2 with as many degrees of freedom as coefficients we are restricting

Log likelihood

```
ibis <-
"ID  Rice  Nest
1   2.5  1
2   1.3  1
3   3.6  1
4   4.2  1
5   0.8  0
6   2.1  0
7   0.9  0
8   1.5  0
9   1.1  0
10 0.3  0"
nests <- read.table(con <- textConnection(ibis), header=TRUE)
close(con)
```

```
fit.0 = glm(Nest ~ 1,      data=nests, family = binomial) # null model
fit.1 = glm(Nest ~ Rice, data=nests, family = binomial) # full model
summary(fit.1)

y.hat.0 = predict(fit.0, nests, type="response") # null model
```

```
1   2   3   4   5   6   7   8   9   10
0.4 0.4 0.4 0.4 0.4 0.4 0.4 0.4 0.4 0.4
```

```
y.hat.1 = predict(fit.1, nests, type="response") # full model
```

```
1       2       3       4       5       6       7       8       9       10
0.7533 0.1778 0.9718 0.9923 0.0669 0.5582 0.0821 0.2516 0.1221 0.0232
```

```
require(lmtest)
lrtest(fit.1); lrtest(fit.1, fit.0)
```

```
Model 1: Nest ~ Rice
Model 2: Nest ~ 1
  #Df  LogLik  Df  Chisq  Pr(>Chisq)
1   2  -3.4620
2   1  -6.7301  -1  6.5363   0.01057 *
```

```
logLik(fit.0) # -6.7301
log(0.4^4 * 0.6^6) # -6.7301
log(prod(y.hat.1[1:4]) * prod(1 - y.hat.1[6:10])) # -3.3926
```

Goodness of fit - Analogous R^2

$-2L\hat{L}_0$ Refer to total sum of square

$-2L\hat{L}_0 - (-2L\hat{L}_S)$ Refer to regression sum of square

Likelihood ratio index (LRI):

$$\mathrm{LRI} = \left(\frac{-2L\hat{L}_0 - (-2L\hat{L}_S)}{-2L\hat{L}_0}\right)$$

Goodness of fit - Analogous R^2

$$R^2 = LRI = \left(\frac{-2L\hat{L}_0 - (-2L\hat{L}_S)}{-2L\hat{L}_0} \right)$$

$$R^2_{adj} = \frac{R^2}{R^2_{\max}} = \frac{R^2}{1-(\hat{L}_0)^{2/n}}$$

```
# R code
library(rms) # required for lrm()
fit2 <- lrm(y ~ x1 + x2, data = data1)
fit2[[3]][10] # R square
```

Hosmer-Lemeshow goodness of fit

Hosmer-Lemeshow Statistic The Hosmer-Lemeshow Statistic is another measure of lack of fit. Hosmer and Lemeshow recommend partitioning the observations into 10 equal sized groups according to their predicted probabilities. Then

$$G^2_{HL} = \sum_{j=1}^{10} \frac{(O_j - E_j)^2}{E_j(1 - E_j/n_j)} \sim \chi^2_8$$

where

$$\begin{aligned} n_j &= \text{Number of observations in the } j^{\text{th}} \text{ group} \\ O_j &= \sum_i y_{ij} = \text{Observed number of cases in the } j^{\text{th}} \text{ group} \\ E_j &= \sum_i \hat{p}_{ij} = \text{Expected number of cases in the } j^{\text{th}} \text{ group} \end{aligned}$$

```
library(ResourceSelection)
model <- glm(y ~ x, family=binomial)
hoslem.test(model$y, fitted(model))
```

Goodness of fit - Wald Test

A Wald test is used to test the statistical significance of each coefficient (β) in the model. A Wald test calculates a *Z* statistic, which is:

$$z = \frac{\hat{\mathrm{B}}}{SE}$$

This z value is then squared, yielding a Wald statistic with a chi-square distribution.

However, several authors have identified problems with the use of the Wald statistic.

- Menard (1995) warned that for large coefficients, standard error was inflated, lowering the Wald statistic (chi-square) value.
- Agresti (1996) stated that the likelihood-ratio test was more reliable for small sample sizes than the Wald test.

```
library(aod)
wald.test(b = coef(model),
          Sigma = vcov(model),
          Terms = 2:3)
```

Stepwise Regression Analysis

- **Principle of Parsimony**: use the smallest number of parameters necessary to represent the data adequately
  - Goodness of fit increases with increasing K (number of parameters), trade-off
  - Lower K: underfit, miss important effects
  - Higher K: overfit, include spurious effects and “noise”
  - Parsimony – proper balance between these 2 effects so that you can repeat results across replications

Akaike's Information Criterion (AIC)

- is a number that you calculate for each model
- provides an estimate of the "distance" between the fitted model and the unknown mechanism that produced the data ("truth").
- the lower the AIC value, the better the model. AIC values are relative.
- can only be compared for exactly the same set of dependent variables.

AIC

AIC = -2 ln (likelihood) + 2K

K = number of parameters in the model, including 1 for the constant and 1 for the error term

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

K = 6

AIC

For least-squares regression, ANOVA, etc.

AIC = $n \ln(\sigma^2) + 2K$

σ^2= (sum of squared residuals)/n = MSE

For small samples ($n/K < 40$), use AIC_c, AIC corrected for small sample size

$$AIC_c = AIC + \frac{2K(K+1)}{n-K-1}$$

When to use AIC

- Mainly observational studies, especially with large numbers of variables
- Generally not in experimental studies because you are usually testing relatively few effects, and standard hypothesis testing works fairly well

Nest site selection of the crested ibis

Sample plots 35
Control plots 35
Habitat factors 11

Elevation (m)

Area of rice fields nearby (ha)

Human disturbance

Number of trees within 100 m^2

Mean tree height within 100 m^2 (m)

Nest position on the slope

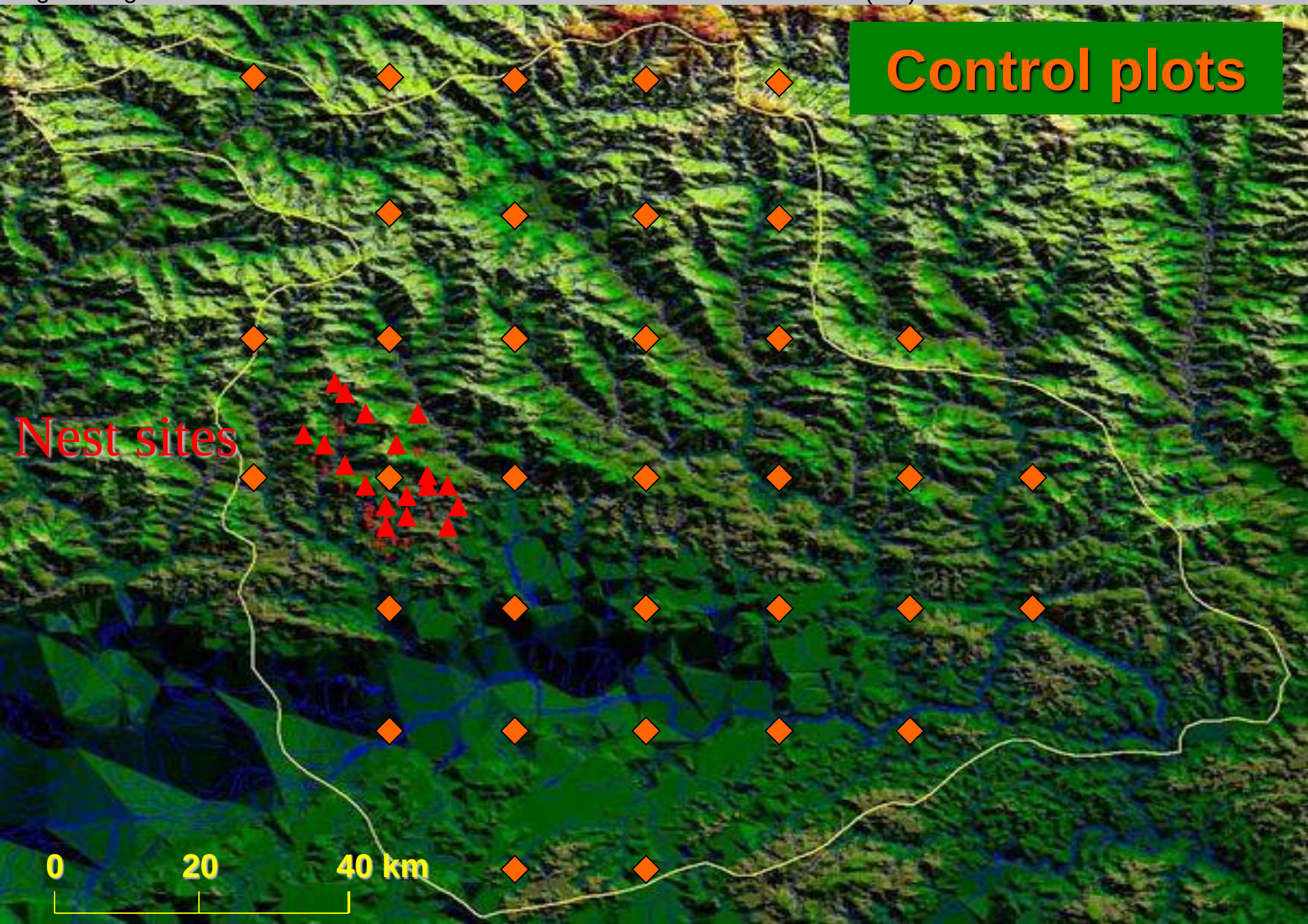
Slope aspect (°)

Slope gradient (°)

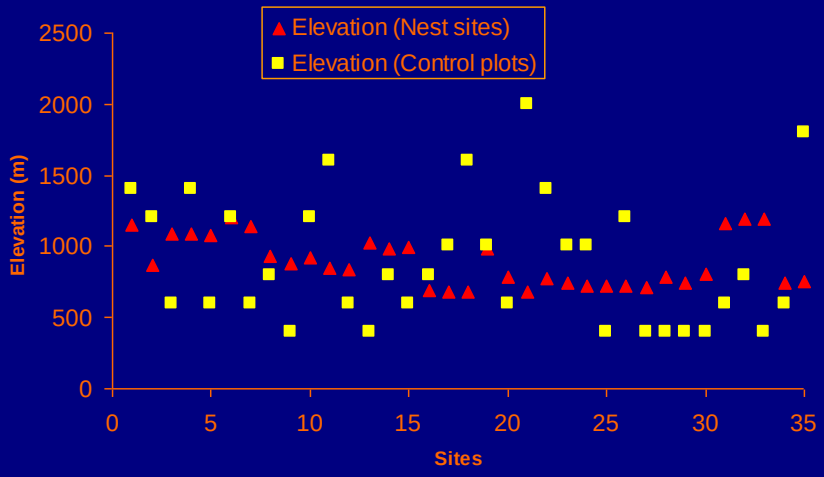
Nest tree height (m)

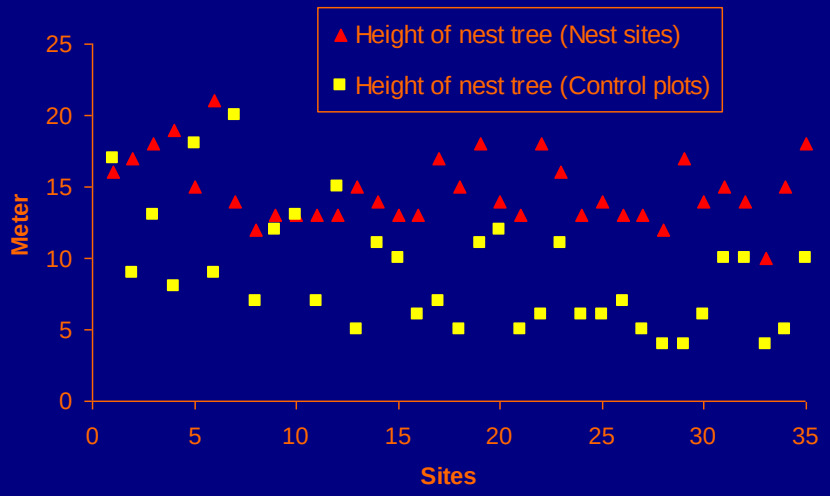
Nest aspect (°)

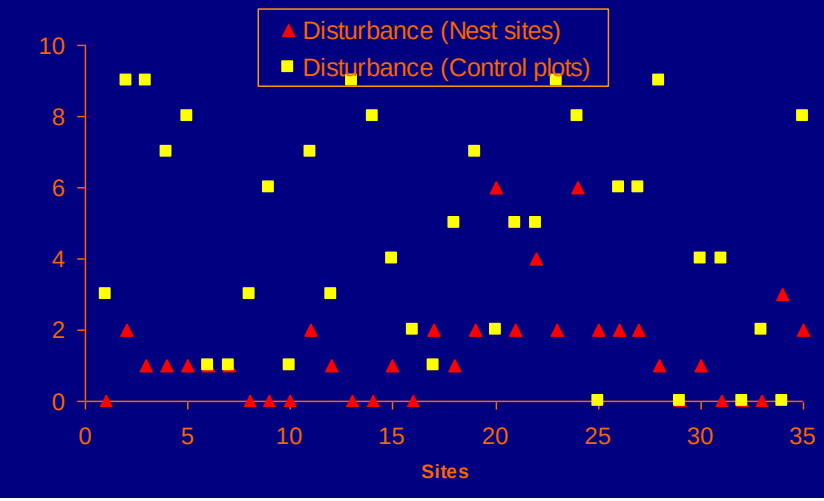
Coverage above the nest (%)

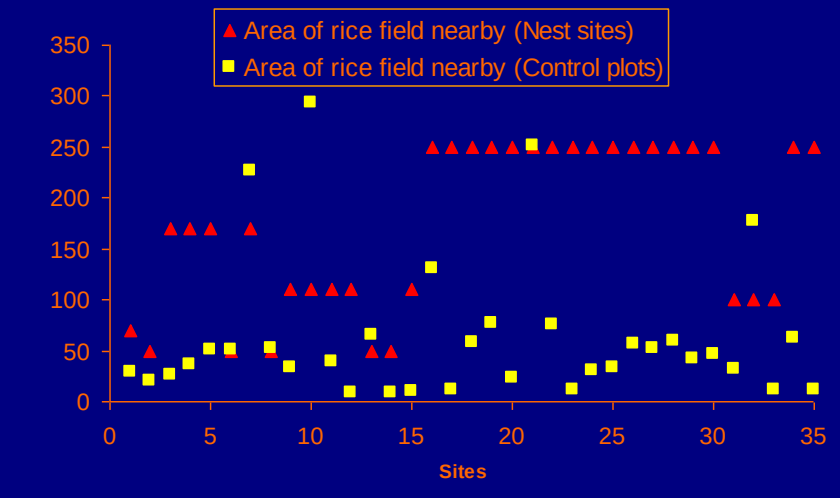

Control plots
Nest sites
0
20
40 km

Source data 1

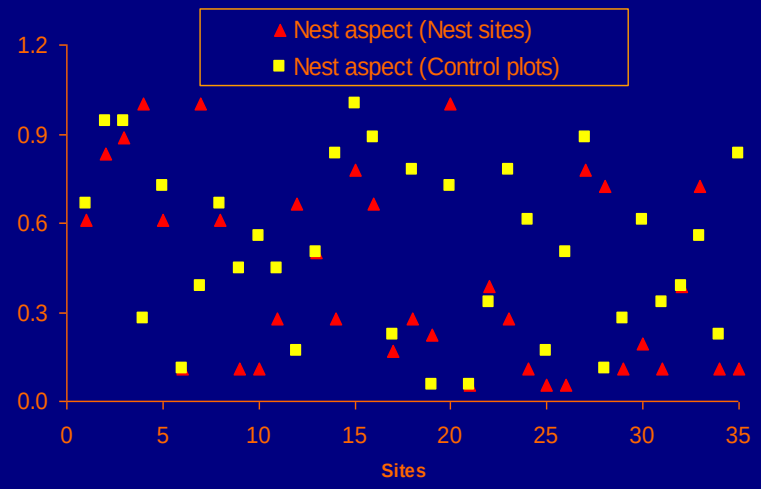


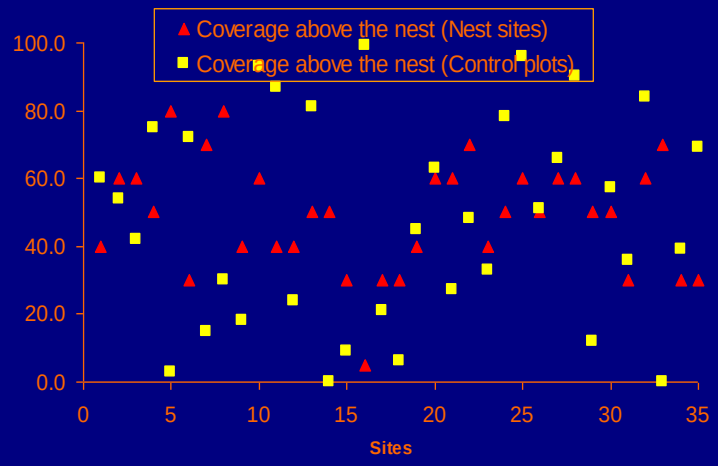


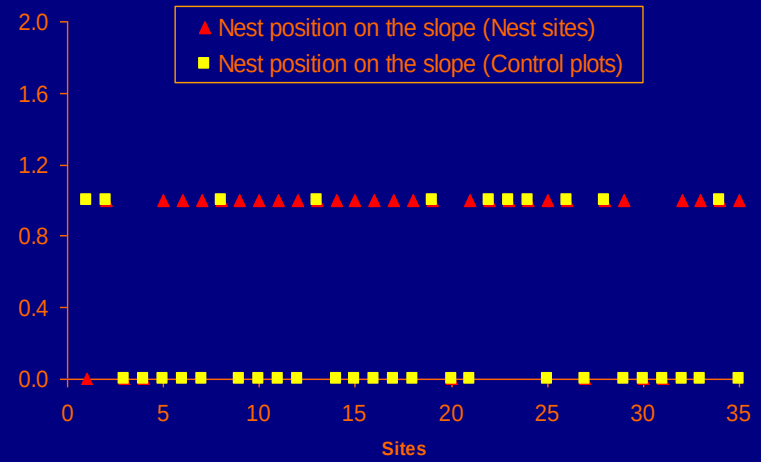


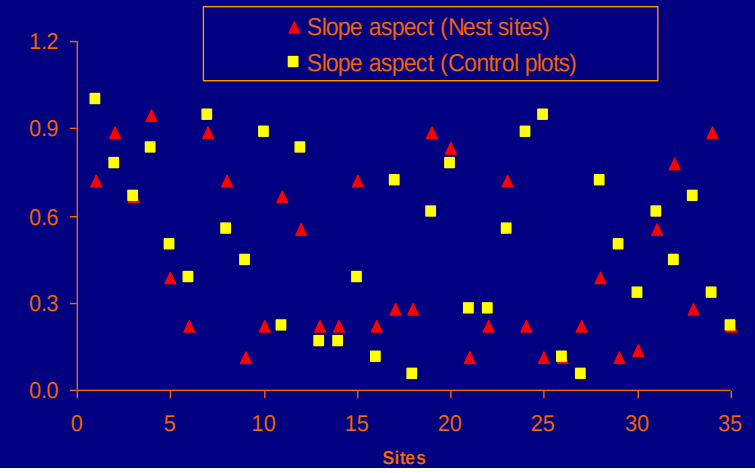


Source data 2

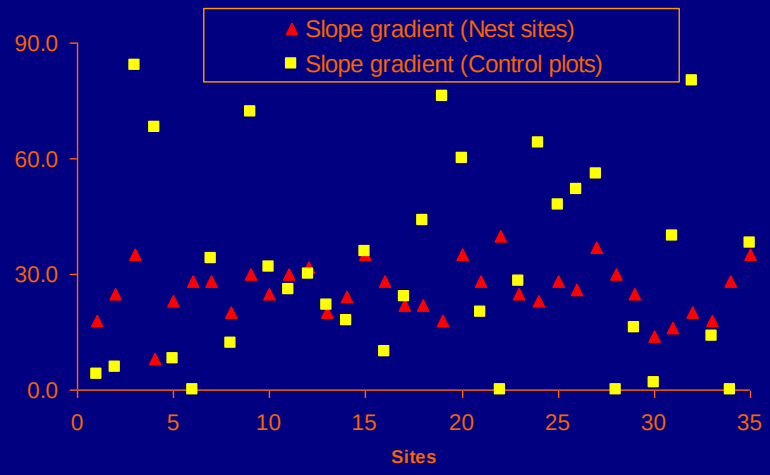


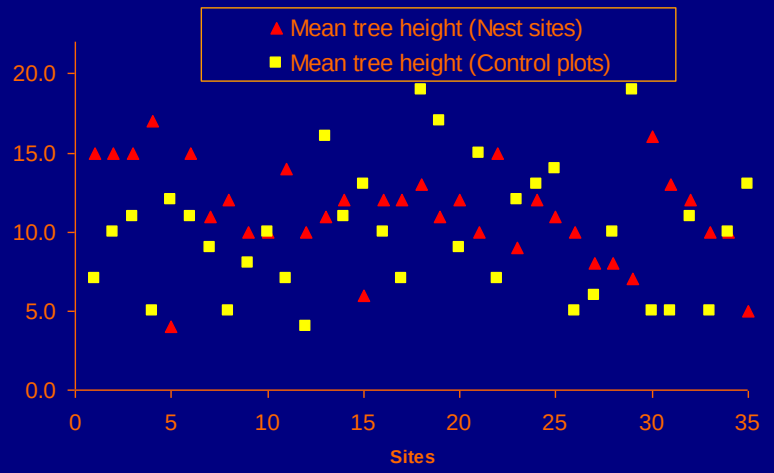


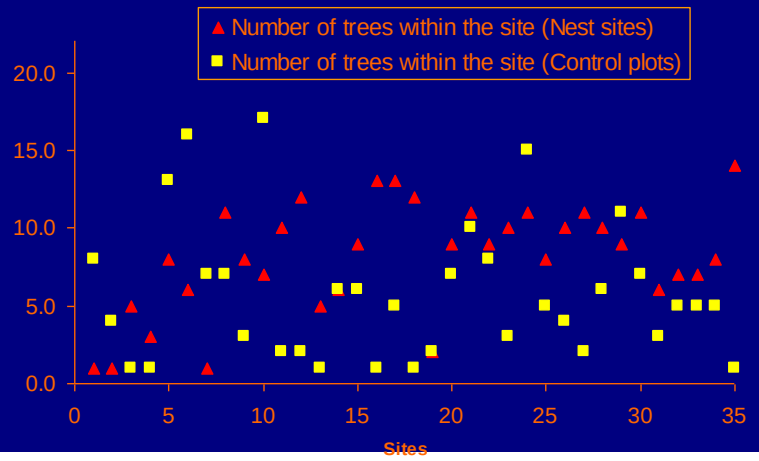




Source data 3







Correlation (VIF is better)

The Pearson correlations between the 11 habitat variables measured at 35 nest sites of crested ibis in Yang county, Shaanxi province, China. Mean values and standard deviations (S.D.) are also shown.

Habitat variables	Correlation coefficients											Mean	S.D.
	1	2	3	4	5	6	7	8	9	10	11		
1. Elevation (m)	1	-0.72*	-0.48*	-0.70*	0.21	-0.02	0.39*	-0.38*	0.162	0.34*	0.21	894.00	176.53
2. Area (ha) of rice fields within 1 km^2		1	0.53*	0.49*	-0.23	-0.08	-0.23	0.23	0.05	-0.21	-0.12	11.62	5.40
3. Human disturbance			1	0.22	0.06	-0.154	0.15	0.38*	0.10	-0.02	0.08	1.40	1.52
4. Number of trees within 100 m^2				1	-0.37*	-0.00	-0.52*	0.34*	-0.33	0.012	-0.25	8.11	3.53
5. Mean tree height within 100 m^2 (m)					1	-0.24	0.23	-0.34*	0.32*	0.11	-0.06	11.23	3.06
6. Nest position on the slope						1	0.03	0.22	-0.21	-0.00	-0.07	2.03	0.45
7. Slope aspect (South=1, North=0)							1	-0.15	0.18	0.55*	0.06	0.45	0.29
8. Slope gradient (°)								1	-0.05	0.10	0.01	25.69	7.01
9. Nest tree height (m)									1	-0.08	-0.23	14.80	2.36
10. Nest aspect (South=1, North=0)										1	0.32	0.43	0.32
11. Coverage above the nest (%)											1	49.00%	16.53%

Stepwise logistic regression for modeling nest site selection of crested ibis in Yang County, Shaanxi Province, China.

Step	Habitat features	Selection coefficients	Standard Error	P value for model selection	AIC
1	Nest tree height (m)	0.94	0.38	<.0001	63.356
2	Human disturbance	-0.99	0.40	0.0001	50.475
3	Slope aspect	-5.82	3.25	0.0013	41.727
4	Area of rice fields nearby (ha)	0.35	0.19	0.0109	36.252
5	Nest position on the slope	3.73	2.30	0.0478	34.336
6	Mean tree height within 100 m^2 (m)	0.28	0.27	0.0320	31.924
7	Nest aspect	54.9285	31.5378	0.0112	26.048
8	Slope gradient (°)	-0.4080	0.3602	0.2866	23.226
9	Coverage above the nest	0.5201	0.5586	0.0841	24.322
10	Number of trees within 100 m^2	-0.006830	0.00616	0.1160	25.764
11	Elevation (m)	0.07670	0.1328	0.1450	27.275

step(fit) ***# Stepwise logistic regression based on AIC***

Model equation (interaction and quadratic terms were ignored)

$$\text{logit}(p) = -20.99 + 0.94 \times \text{nest tree height} - 0.99 \times \text{human disturbance} + 3.63 \times \text{nest position} + 0.35 \times \text{rice paddy area} + \ldots$$

Probability of nest selection：

$$P = e^{\text{logit}(p)} / (1 + e^{\text{logit}(p)})$$

- R-Square 0.7380
- Max-rescaled R-Square 0.9840

李欣海, 马志军, 李典谟, 丁长青, 翟天庆, 路宝忠。2001。应用资源选择函数研究朱鹮的巢址选择。生物多样性, 9：352-358。

Assignment

General objectives: learn about logistic regression.

- Develop a dataset to perform:
  - logistic regression Y - X1, X2, X3, etc.
- Describe the dataset (e.g. sample and control plots), check the correlation between each two independent variables, describe the significance of each independent variables and overall model fit.