Advanced Linear Models

Prabhanjan N T
Lead Statistician
CustomerXPs Software Private Limited

- Bootstrapping Regression Models
- Nonlinear Regression Models
- Bayesian Regression Models

What is Bootstrap?

- Efron and Tibshirani (1998)
- The Aspirin Example of Chapter 1

	Heart Attacks (fatal plus non-fatal)	Subjects
Aspirin Group	104	11037
Placebo Group	189	11034

The odds ratio for heart attack rates is (104/11037)/(189/11034) = 0.55

	Heart Attacks Strokes	Subjects
Aspirin Group	119	11037
Placebo Group	98	11034

The odds ratio for strokes is (119/11037)/(98/11034) = 1.21

- The 95% confidence intervals give contradicting results

```
aspirin_overall=c(rep(1,104),rep(0,11037-104))
placebo_overall=c(rep(1,189),rep(0,11034-189))
aspirin_strokes=c(rep(1,119),rep(0,11037-119))
placebo_strokes=c(rep(1,98),rep(0,11034-98))
or_overall=or_strokes=c()
for(i in 1:1000)    {
bao=sample(aspirin_overall,11037,replace=T)
bpo=sample(placebo_overall,11034,replace=T)
bas=sample(aspirin_strokes,11037,replace=T)
bps=sample(placebo_strokes,11034,replace=T)
or_overall[i]=(sum(bao)/11037)/(sum(bpo)/11034)
or_strokes[i]=(sum(bas)/11037)/(sum(bps)/11034)
                        }
```

```
quantile(or_overall,c(0.025,0.975))
     2.5%     97.5%
0.4338026 0.6881324
quantile(or_strokes,c(0.025,0.975))
     2.5%     97.5%
0.9341606 1.5806812
```

Bootstrapping Regression Models

- Consider the linear regression model

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon$$

- Suppose that for the above model, the fitted model is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p$$

- Let the estimated residuals from the above model be $\hat{\epsilon}$
- **X**: design matrix, order $n * (p+1)$
  - Draw a sample (with replacement) from $\hat{\epsilon}$ and denote it ϵ^*
  - Calculate the output **Y** using $Y^* = X'\hat{\beta} + \epsilon^*$
  - Now, calculate β^* for the model $Y^* = X'\beta + \epsilon^*$
- Repeat the above steps a large number of times
- *This technique bootstraps the residuals*

Example. *Galton Data Set*. Bootstrapping the Residuals

```
library(UsingR)
cplm=lm(child~parent,galton)
resid=lm(child~parent,galton)$residuals
bcoef=matrix(nrow=100,ncol=2)
for(i in 1:100)      {
newy=cbind(rep(1,nrow(galton)),galton$parent)%*
     %cplm$coefficients+sample(resid,nrow(galton),replace=T)
bcoef[i,]=lm(newy~galton$parent)$coefficients
}
quantile(bcoef[,2],c(0.025,0.975))
    2.5%     97.5%
0.5595697 0.7282335
```

Since 0 does not lie in the above 95% Bootstrap Confidence Interval , we reject the null hypothesis and conclude that the height of father has significant effect on the height of son.

- Let us consider the data points of a regression model
- $(y_1,\boldsymbol{x}_1)$, $(y_2,\boldsymbol{x}_2)$, ..., $(y_n,\boldsymbol{x}_n)$
- Obtain a sample of size n with replacement from the above n points $(y_1^*,\boldsymbol{x}_1^*)$, $(y_2^*,\boldsymbol{x}_2^*)$, ..., $(y_n^*,\boldsymbol{x}_n^*)$
- Obtain the regression coefficients of the above bootstrap sample
- Repeat the above two steps a large number of times

- In this second method, we are *bootstrapping the observations*

```
bcoef=matrix(nrow=100,ncol=2)
for(i in 1:100) {
tempgalton=galton[sample(1:nrow(galton)
,nrow(galton),replace=T),]
bcoef[i,]=lm(child~parent,tempgalton)
$coefficients
}
quantile(bcoef[,2],c(0.025,0.975))
     2.5%     97.5%
0.5646895 0.7142975
```

Again, 0 does not lie in the 95% Bootstrap confidence interval.

Nonlinear Regression

- Linear Regression model gives poor fit
- Transformations failed
- Data is intrinsically nonlinear
- Ritz and Streibig (2009)
- Huet, Bouvier, Poursat, and Jolivet (2004)
- Seber and Wild (2003)
- Bates and Watts (1988)
- The R package is “nlrwr”

Nonlinear Regression Models

The predictor and the response may be related as

$$y = f(x, \beta)$$

The mean is expected to be centered around $f(x, \beta)$, and we thus refer to *f* as the *mean function*

If the form of the function *f* is known to the data analyst, the nonlinear models are then particularly useful

The nonlinear regression model is

$$y_i = E\{f(y_i | x_i)\} + \epsilon_i = f(x_i, \beta) + \epsilon_i$$

As with linear models, we assume that the errors are normally distributed.

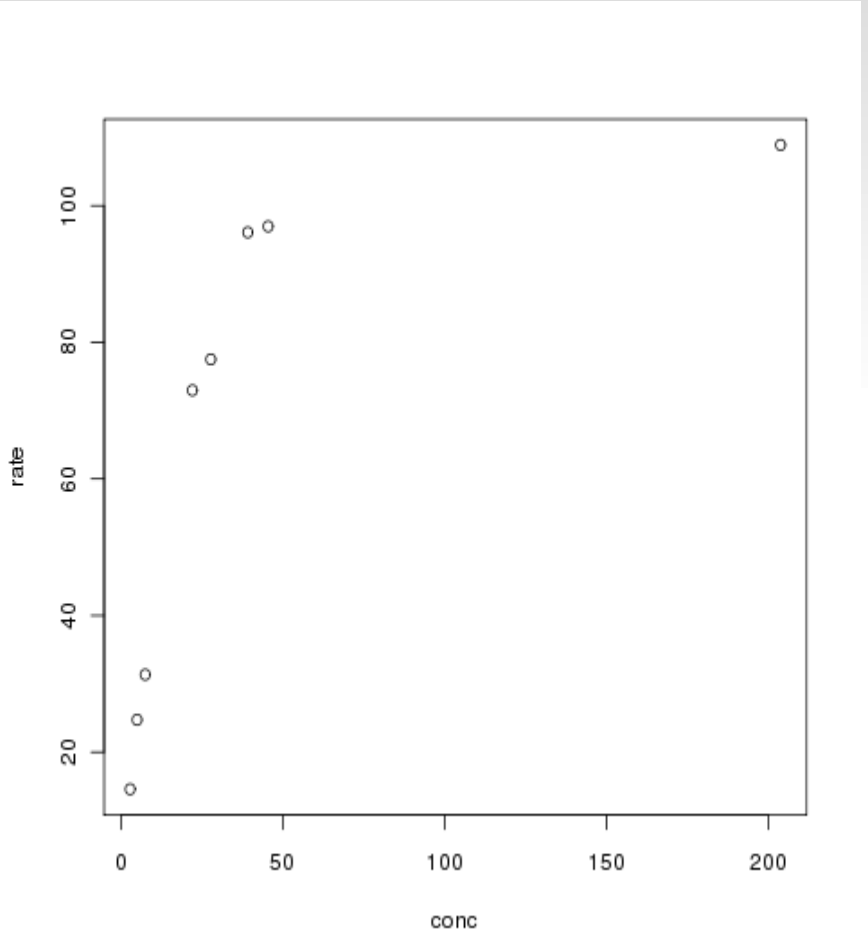
The basic idea remains the same as in Linear Regression Model, viz minimize the expression listed below:

$$RSS(\beta)=\sum_{i=1}^{n}(y_i-f(x_i,\beta))^2$$

We consider the simple example from Ritz and Streiberg (2009). The variables are the following:

- *uptake rate*: regressand
- *initial substrate concentrate*: predictor

```
library(nlrwr)
data(L.minor)
plot(L.minor)
```



The Michaelis-Menten Model

The Michaelis-Menten model for the uptake rate as a function of the substrate concentration is given by

$$rate = f(Concentrate, (K, V_m)) = \frac{V_m \times Concentration}{K + Concentration}$$

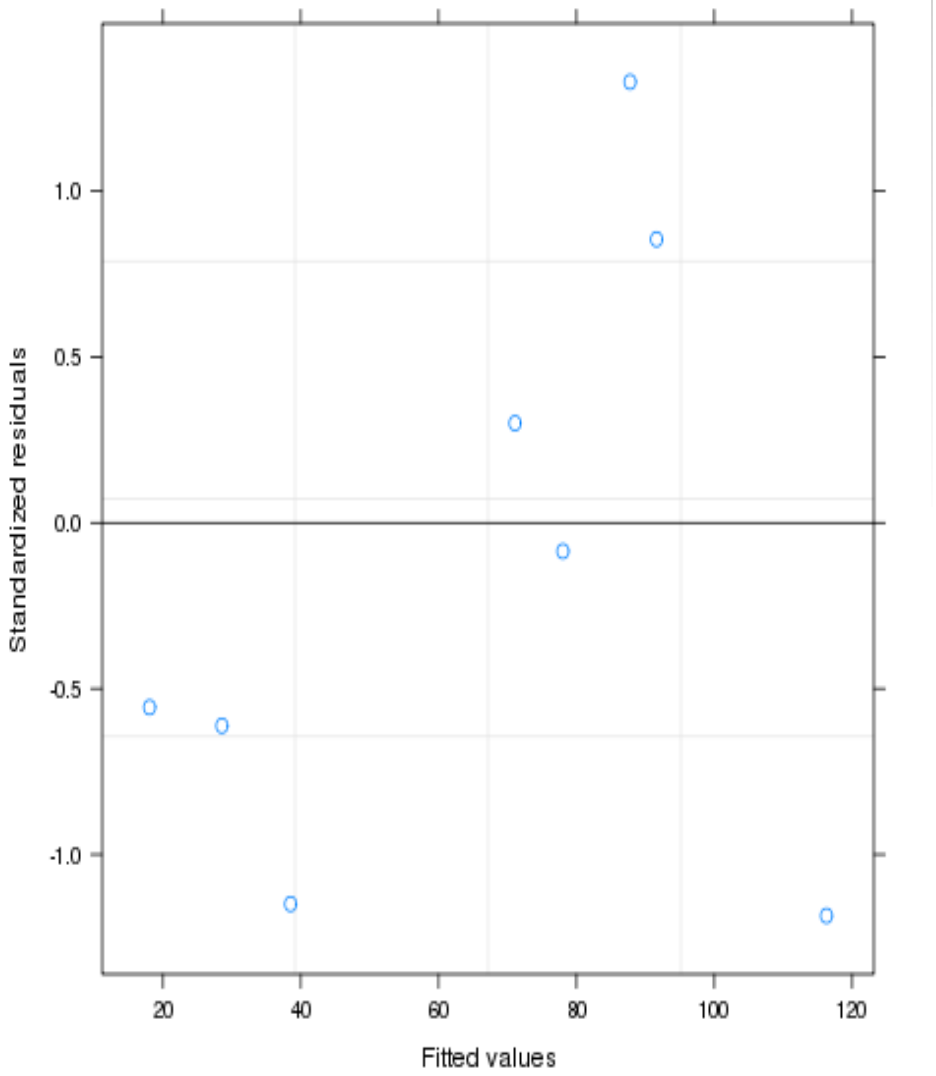
$$y = f(x, (K, V_m)) = \frac{V_m \times x}{K + x}$$

V_m and K are the unknown parameters.

When the concentration is 0, the uptake rate is also 0. As concentration increases, the rate reaches the maximum possible value from below.

```
nlm=nls(rate ~ Vm*conc/(K+conc),data=L.minor,start=list(K=20,Vm=120),trace=T)
624.3282 :   20 120
244.5460 :   15.92382 124.57148
234.5198 :   17.25299 126.43877
234.3595 :   17.04442 125.96181
234.3533 :   17.08574 126.04671
234.3531 :   17.07774 126.03016
234.3531 :   17.07930 126.03338
234.3531 :   17.07899 126.03276
summary(nlm)
Formula: rate ~ Vm * conc/(K + conc)
Parameters:
   Estimate Std. Error t value Pr(>|t|)
K    17.079      2.953   5.784  0.00117 **
Vm  126.033      7.173  17.570 2.18e-06 ***
Signif. codes:  0 â***â 0.001 â**â 0.01 â*â 0.05 â.â 0.1 â â 1
Residual standard error: 6.25 on 6 degrees of freedom
Number of iterations to convergence: 7
Achieved convergence tolerance: 8.144e-06
plot(nlm)
```

deviance(nlm)
[1] 234.3531



Bayesian Regression

- Albert (2009)
- Consider the linear regression model

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$$

where the errors are assumed to be iid normal with finite variance σ^2

- Assume that the parameters $(\boldsymbol{\beta}, \sigma^2)$ have the noninformative prior, that is

$$g(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}$$

- The joint density of $(\boldsymbol{\beta}, \sigma^2)$ is given by

$$g(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}) = g(\boldsymbol{\beta} | \boldsymbol{y}, \sigma^2) g(\sigma^2 | \boldsymbol{y})$$

- The posterior distribution of the regression parmeter given the data and the variance is multivariate normal with mean $\hat{\boldsymbol{\beta}}$ and variance-covariance matrix $V_\beta \sigma^2$, where

$$\hat{\beta} = (X'X)^{-1} X'y, V_\beta = (X'X)^{-1}$$

- The marginal posterior distribution of σ^2 is inverse gamma *((n-k)/2,S/2)*, where

$$S = (y - X'\hat{\beta})'(y - X'\hat{\beta})$$

- The computation is then done as follows:
  - Simulate a value of σ^2 from its marginal posterior distribution $g(\sigma^2 | y)$
  - Simulate value of $\boldsymbol{\beta}$ from the conditional posterior density $g(\boldsymbol{\beta} | y, \sigma^2)$

Example. ***Bird Extinct Case Study***

Following measurements are made on breeding pairs of landbirds from 16 islands around Britain for several decades:

Time: the average time for extinct of a species

Nesting: the average number of nesting pairs

Size: the size of the species group

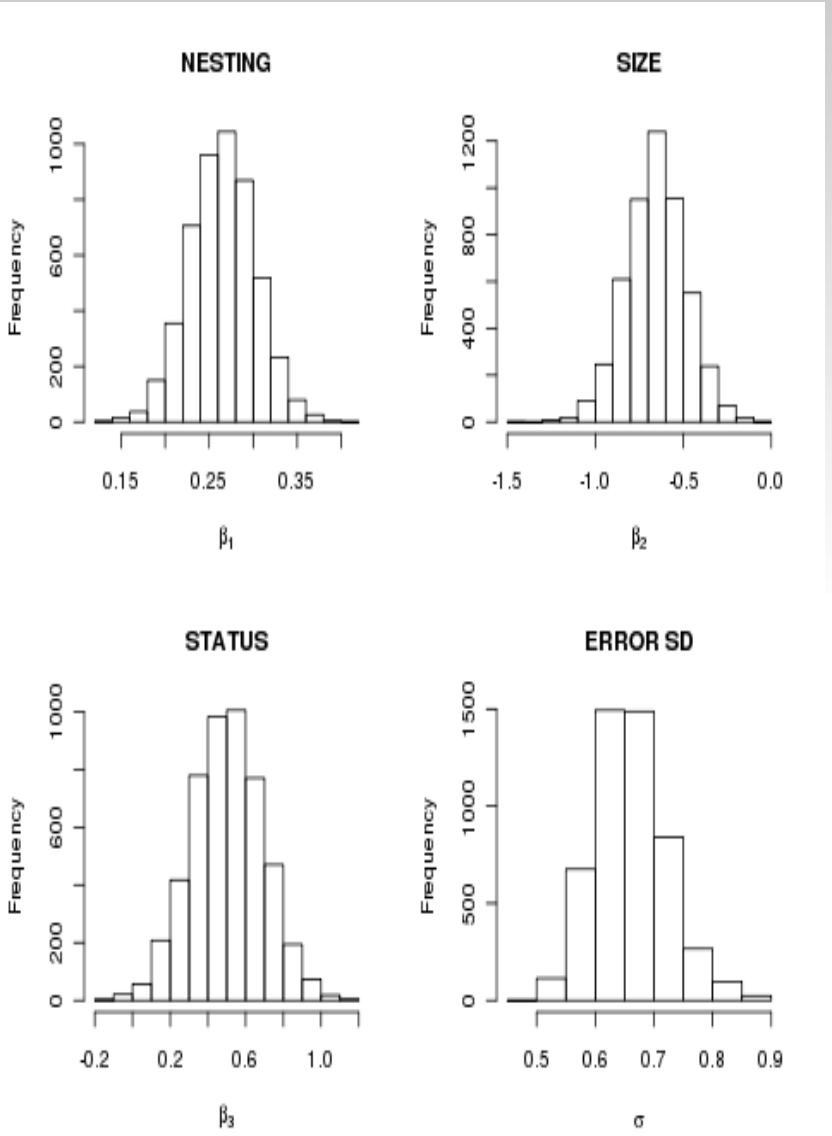
Status: migrant or resident indicator

Goal of the study: to model the average time to extinction of a species based on the covariates Nesting, Size, and Status

```
extinctlm=lm(logtime~nesting+size+status,data=birdextinct,x=T,y=T)
summary(extinctlm)
Call:
lm(formula = logtime ~ nesting + size + status, data = birdextinct,
    x = T, y = T)
Residuals:
    Min      1Q  Median      3Q     Max
-1.8410 -0.2932 -0.0709  0.2165  2.5167
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.43087    0.20706   2.081 0.041870 *
nesting      0.26501    0.03679   7.203 1.33e-09 ***
size        -0.65220    0.16667  -3.913 0.000242 ***
status       0.50417    0.18263   2.761 0.007712 **
---
Signif. codes:  0 â***â 0.001 â**â 0.01 â*â 0.05 â.â 0.1 â â 1
Residual standard error: 0.6524 on 58 degrees of freedom
Multiple R-squared: 0.5982,     Adjusted R-squared: 0.5775
F-statistic: 28.79 on 3 and 58 DF,  p-value: 1.577e-11
```

We now use the design matrix *X* and the output vector from the R object "extinctlm" to obtain samples from the posterior distribution, and the R function "blinreg" from the R package "LearnBayes".

```
theta.sample=blinreg(extinctlm$y,extinctlm$x,500
0)
par(mfrow=c(2,2))
hist(theta.sample$beta[,2],main="NESTING",
xlab=expression(beta[1]))
hist(theta.sample$beta[,3],main="SIZE",
xlab=expression(beta[2]))
hist(theta.sample$beta[,4],main="STATUS",
xlab=expression(beta[3]))
hist(theta.sample$sigma,main="ERROR SD",
xlab=expression(sigma))
```



Of course, we need to infer about the estimated regression coefficients:

```
apply(theta.sample$beta,2,quantile,c(.025,.975))
      X(Intercept)  Xnesting      Xsize   Xstatus
2.5%    0.02043236 0.1921894 -0.9961065 0.1345796
97.5%   0.84305283 0.3382079 -0.3226247 0.8794357
quantile(theta.sample$sigma,c(.025,.975))
     2.5%     97.5%
0.5527534 0.7991237
```