Regression Model Selection

Likelihood function

In the multiple linear regression model, we have

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

where:

$\beta_0, \beta_1, \ldots, \beta_{p-1}$ are parameters
$X_{i1}, \ldots, X_{i,p-1}$ are known constants
ε_i are independent $N(0, \sigma^2)$
$i = 1, \ldots, n$

As we can see, the error term is follow normal distribution with mean 0 and variance σ^2, then we can write

$$Likelihood = (2\pi\sigma^2)^{-n/2}\exp\left(-\sum_{i=1}^{n}\frac{\left(Y_i - \beta_0 - \beta_1 X1 - \ \ldots\ - \beta_{p-1} Xp-1\right)^2}{2\sigma^2}\right)$$

When replace σ^2 by its estimate (s^2_n= SSE/n) and the model coefficients by their estimates, we get

$$Likelihood = L = (2\pi SSE/n)^{-n/2}\exp(-\sum_{i=1}^{n}\frac{\left(Y_i - \hat{Y}_i\right)^2}{2\mathrm{SSE/n}})$$

then

-2Log L = n[log(2 π)+ log(SSE/n)+1]

Akaike's An Information Criterion

Generic function calculating Akaike's 'An Information Criterion' for one or several fitted model objects.

Akaike's Information Criterion is usually calculated with software. The basic formula is defined as:
AIC = -2(log-likelihood) + 2K
Where:

- K is the number of model parameters (the number of variables in the model plus the intercept).
- Log-likelihood is a measure of model fit. The higher the number, the better the fit. This is usually obtained from statistical output.

or small sample sizes (n/K < ≈ 40), use the second-order AIC:
AICc = -2(log-likelihood) + 2K + (2K(K+1)/(n-K-1))
Where:
n = sample size,
K= number of model parameters,
Log-likelihood is a measure of model fit.

formula *-2*log-likelihood + k*npa*,

where *npar* represents the number of parameters in the fitted model,

k = 2 for the usual AIC,

or *k = log(n)* (*n* being the number of observations). This can be used when n/p<40.

i.e

AIC=-2log L + 2(p+1),

Where

L is the likelihood function
p is the number of parameters in linear model and we add one because we have sigma^2 to be estimated.

How to know if the model is best fit for your data?

The most common metrics to look at while selecting the model are:

STATISTIC	CRITERION
R-Squared	Higher the better *(> 0.70)*
Adj R-Squared	Higher the better
F-Statistic	Higher the better
Std. Error	Closer to zero the better
t-statistic	Should be greater 1.96 for p-value to be less than 0.05
AIC	Lower the better
BIC	Lower the better
Mallows cp	Should be close to the number of predictors in model
MAPE (Mean absolute percentage error)	Lower the better
MSE (Mean squared error)	Lower the better
Min_Max Accuracy => mean(min(actual,	Higher the better

In this part we use AIC as a criterion to the model selection.

Example: Use the mtcars data to select the best model

https://gist.github.com/seankross/a412dfbd88b3db70b74b

Using
Stepwise: Backward selection
Stepwise: Forward selection
Stepwise: Combination of Forward and Backward selection

R code:

```
data(mtcars)
d=mtcars
head(d)
FitAll=lm(mpg~., data=d)
summary(FitAll)

p=length(FitAll$coef)
n=length(d$mpg)
SSE=sum((FitAll$res)^2)

AIC=n*(log(2*pi)+1+log(SSE/n))+2*(p+1)


logL=as.numeric(logLik(FitAll))
AIC=-2*logL+2*(p+1)

AIC(FitAll)
```

```
#========== Backward=============
step(FitAll, direction="backward")

#======== Forward ====================================
Fitstart=lm(mpg~1,data=d)
summary(Fitstart)
step(Fitstart, direction="forward", scope=formula(FitAll))
#============== Both ================================

Fitstart=lm(mpg~1,data=d)
summary(Fitstart)
step(Fitstart, direction="both", scope=formula(FitAll))
#=====================================================
```

R:Code

```
mydata=read.table("Dwaine Studios.txt",header=TRUE)
Y=mydata$Y
X1=mydata$X1
X2=mydata$X2
n=length(X1)
one=as.vector(rep(1, n))
X=cbind(one,X1,X2)
Model=lm(Y~X1+X2)
library(QuantPsyc)
lm.beta(Model)
```