Multiple Regression Selection and Stepwise Model Fitting

Joseph J. Luczkovich

February 18, 2014

1 Introduction

In this section, I will introduce some procedures to find the best fitting linear model with multiple regression. We will examine the R^2 and the Akaike Information Criterion (AIC), then implement automated procedures to find the best model (stepwise regression) and introduce the subject of residual analysis.

2 R^2 as a measure of fit

As we have learned, the R^2 is a standard, often-used measure of the goodness of fit of linear regression models. It can be interpreted as "percent of variance explained". This measure is similarly used in mutiple linear regression to find the best models when considering multiple candidate predictors for inclusion in the model. One criterion often used is to maximize the ajusted R^2.

2.1 Multiple and Adjusted R^2

The bottom of the output you will find the overall (Multiple R^2) and adjusted R^2. The multiple R^2 always increases when you add model predictors - more terms allow you to explain more variance, even if they explain just a little bit more. The adjusted R^2 is the appropriate measure of fit to examine for use in multiple regression model selection, because it is adjusted for the number of model predictors used, and it sometimes gets lower when you add more model predictors. This is because more is not always better, especially if they do not contribute much in explanatory power or are correlated with variables already in the model.

2.2 Example model selection with R^2 in R

In this example, the data from NCDMF Program 120 trawling data for pinfish, *Lagodon rhomboides* are used. The response variable is mean nummber of

pinfish per 1-min (75 m) trawl taken at each of 72 stations for the period 2000-2004. Also taken at each station are predictors water depth ("DEPTH", m), distance to nearest inlet ("DISINLE", km), bottom salinity ("BSALIN"), bottom temperature ("BTEMP" in °C), percent land use change in the watershed around each station ("PLUCHAN") and human population in the watershed in year 2000 US Census data ("POP2000"). The investigators want to know if any of these are good predictors of pinfish catches. First, we read in the data from SYSAT format *.syd file. We next fit the full model using the **lm()** function. We will examine the R^2 of successive models with deleted predictor terms to find a good subset of predictors.

```
> library("foreign", lib.loc="C:/Program Files/R/R-3.0.2/library")
> pf<-read.systat(file="C:/Users/luczkovichj/Dropbox/CRM7008/Lecture5/pinfish_final_edited2.
> fit1<-lm(AVG0004~DEPTH+DIS_INLE+BSALIN+BTEMP+PLU_CHAN+POP_2000,data=pf)
> summary(fit1)

Call:
lm(formula = AVG0004 ~ DEPTH + DIS_INLE + BSALIN + BTEMP + PLU_CHAN +
    POP_2000, data = pf)

Residuals:
    Min      1Q  Median      3Q     Max
-10.301  -3.215  -1.040   2.519  27.430

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 49.899916  24.885768   2.005  0.04918 *
DEPTH       -5.594722   1.624685  -3.444  0.00102 **
DIS_INLE    -0.060843   0.049370  -1.232  0.22231
BSALIN      -0.448649   0.161198  -2.783  0.00707 **
BTEMP       -1.122581   0.990986  -1.133  0.26153
PLU_CHAN    -0.015482   0.036873  -0.420  0.67599
POP_2000    -0.000042   0.000181  -0.232  0.81723
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.994 on 64 degrees of freedom
  (27 observations deleted due to missingness)
Multiple R-squared:  0.2365,        Adjusted R-squared:  0.1649
F-statistic: 3.305 on 6 and 64 DF,  p-value: 0.006768
```

What is the adjusted R^2 for this fit of the full model? Next, try a reduced model:

```
fit2<-lm(AVG0004~DEPTH+DIS_INLE+BSALIN+BTEMP+PLU_CHAN,data=pf)
summary(fit2)
```

What is the adjusted R^2 for this fit of the reduced model, after eliminating one predictor, POP2000? Did it go up or down? Keep trying to remove terms till you get the highest R^2. That is the likely best model. Record the βs for the terms included in that best model and interpret them in words. What does this equation mean (in English)?

3 Akaike Information Criterion

This is an alternative measure of model adequacy developed by Akaike. It represents the information content in the residuals, or entropy. As a consequence, we wish to keep this value low in model selection, i.e., minimize the AIC. The information should be in the predictors, not the residuals, the theory goes. To see the AIC for a model fit, it is simple in R:

```
AIC(fit1)
```

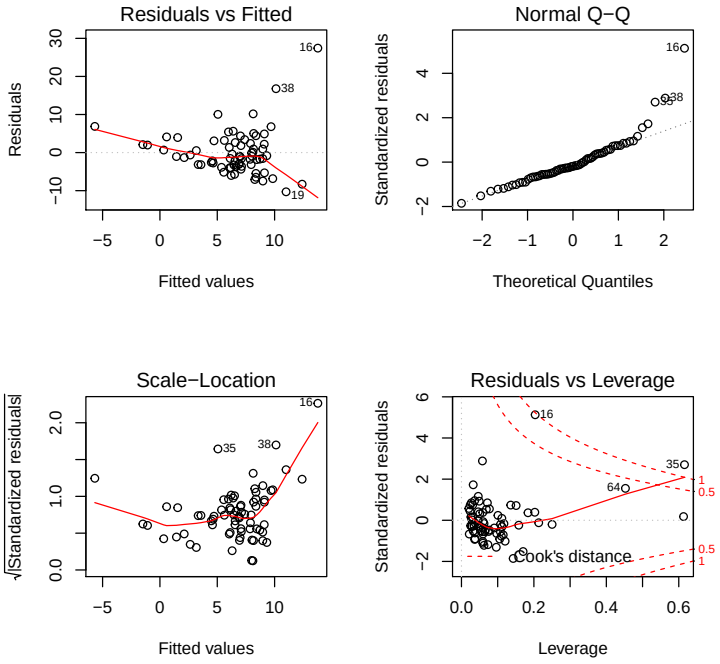
```
AIC(fit2)
```

Notice that the AIC is lower for fit2, just as the R2 was higher for fit2. What is the AIC for the best fit model you got above?

3.1 Residual plots

Look at some residual plots to see if there are some things you can do to get a better fit. First I change the par() function to get all four residual plots on one page, arranged in a 2 x 2 table.

```
par(mfrow=c(2,2))
plot(fit1)
```



Are there some unusual extreme points? Yes. You can reduce the influence of these outliers (and improve R^2 and lower the AIC of the fit) by applying some transformation to the pinfish catches. But which transformation?

3.2 BoxCox Transformation

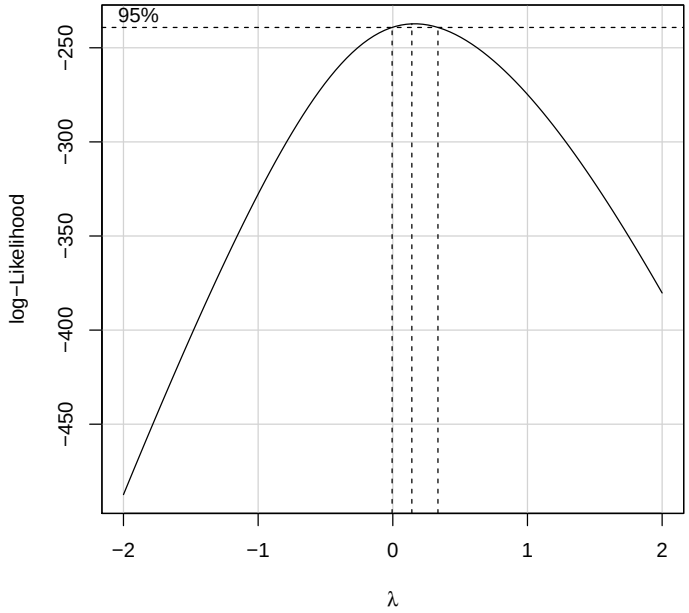
"...essentially, all models are wrong, but some are useful" - George E. P. Box. George Box and David Cox wrote an influential paper to describe a series of power transformations using a parameter λ. They wrote a paper together partly because they always wanted to have their names on the same paper, i.e., because it sounded cool. But it is also a very useful way to decide what is the best transformation to use. Using the correct transformation indicated by λ will transform the data to approximately a normal distribution.

Can you improve the fit of this pinfish catch model by doing a transformation of the response variable (pinfish catches in trawls)? Try using the **boxCox()** function in the **car** package to see what is the best transformation. This function will produce a Box-Cox transformation of your data and a graph showing a curve of λ values in the power function of transformations. Read the likelihood curve produced by the **boxCox()** function when it is maximized, i.e. read it at the peak to find λ. When $\lambda = 0$, the best transformation to use is a log transformation.

Figure 1: George E. P. Box

When $\lambda = -2$	use $1/Y^2$
When $\lambda = -1$	use $1/Y$
When $\lambda = -0.5$	use $1/sqrtY$
When $\lambda = 0$	use $log(y)$
When $\lambda = 0.5$	use $\sqrt{Y}$
When $\lambda = 1$	No transformation needed
When $\lambda = 2$	use Y^2

```
> library("car", lib.loc="C:/Program Files/R/R-3.0.2/library")
> boxCox(fit1)
```



The result suggests that we use a log transformation or possibly a $\sqrt{}$ transformation of pinfish catches. I will use log transform:

```
> fit1.log<-lm(log10(AVG0004+1)~DEPTH+DIS_INLE+BSALIN+BTEMP+PLU_CHAN+POP_2000,data=pf)
> summary(fit1.log)

Call:
lm(formula = log10(AVG0004 + 1) ~ DEPTH + DIS_INLE + BSALIN +
    BTEMP + PLU_CHAN + POP_2000, data = pf)

Residuals:
     Min       1Q   Median       3Q      Max
-0.60091 -0.18171 -0.01851  0.19010  0.76184
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.004e+00  1.205e+00   2.493   0.0153 *
DEPTH       -3.539e-01  7.866e-02  -4.500 2.94e-05 ***
DIS_INLE     3.152e-03  2.390e-03   1.319   0.1919
BSALIN      -1.073e-02  7.804e-03  -1.375   0.1739
BTEMP       -7.088e-02  4.798e-02  -1.477   0.1445
PLU_CHAN    -1.363e-03  1.785e-03  -0.763   0.4481
POP_2000    -7.355e-06  8.762e-06  -0.839   0.4044
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2902 on 64 degrees of freedom
  (27 observations deleted due to missingness)
Multiple R-squared:  0.368,        Adjusted R-squared:  0.3088
F-statistic: 6.212 on 6 and 64 DF,  p-value: 3.534e-05
```

```
> AIC(fit1.log)

[1] 34.43272
```

Note that the adjusted R^2 went up. Also notice that the AIC improved, which actually declined.

4 Stepwise regression

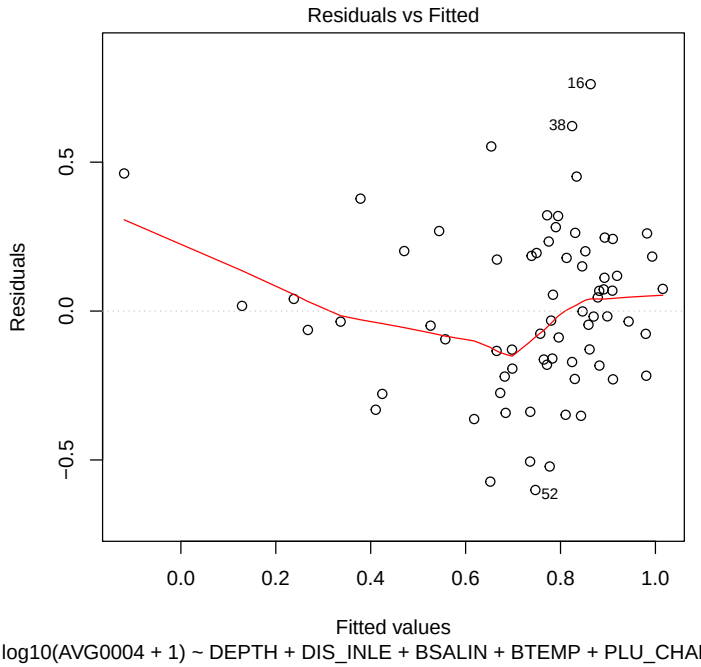
There are automated procedures for selecting the best subset of predictors in your multiple regression models. Stepwise regression is one such method. It starts with the full model and removes each predictor one at a time (backwards stepwise regression), or with the simplest model and adds the predictors one at at time (forwards stepwise regression). The default stepwise regression does backwards till it finds the best model. The criterion used in R for stepwise regression is AIC - if the AIC does not decrease, the search for the best model will stop. You can also specify a scope parameter, which determines the range of model predictors that can be considered, and is entered as a formula that determines which model terms should be considered for adding or dropping. When you specify a scope, the default is both forward and backward searches.

```
> step(fit1.log)

Start:  AIC=-169.06
log10(AVG0004 + 1) ~ DEPTH + DIS_INLE + BSALIN + BTEMP + PLU_CHAN +
    POP_2000

           Df Sum of Sq    RSS     AIC
```

```
- PLU_CHAN  1   0.04906 5.4384 -170.41
- POP_2000  1   0.05933 5.4486 -170.28
- DIS_INLE  1   0.14646 5.5358 -169.15
<none>                 5.3893 -169.06
- BSALIN    1   0.15919 5.5485 -168.99
- BTEMP     1   0.18381 5.5731 -168.68
- DEPTH     1   1.70499 7.0943 -151.54

Step:  AIC=-170.41
log10(AVG0004 + 1) ~ DEPTH + DIS_INLE + BSALIN + BTEMP + POP_2000

           Df Sum of Sq    RSS     AIC
- POP_2000  1   0.05300 5.4914 -171.72
- DIS_INLE  1   0.13397 5.5723 -170.69
<none>                 5.4384 -170.41
- BSALIN    1   0.16649 5.6049 -170.27
- BTEMP     1   0.23036 5.6687 -169.47
- DEPTH     1   1.75471 7.1931 -152.56

Step:  AIC=-171.72
log10(AVG0004 + 1) ~ DEPTH + DIS_INLE + BSALIN + BTEMP

           Df Sum of Sq    RSS     AIC
- DIS_INLE  1   0.11457 5.6060 -172.26
<none>                 5.4914 -171.72
- BTEMP     1   0.21141 5.7028 -171.04
- BSALIN    1   0.26827 5.7597 -170.34
- DEPTH     1   1.83276 7.3241 -153.28

Step:  AIC=-172.26
log10(AVG0004 + 1) ~ DEPTH + BSALIN + BTEMP

         Df Sum of Sq    RSS     AIC
<none>                5.6060 -172.26
- BTEMP   1   0.19449 5.8004 -171.84
- BSALIN  1   0.95935 6.5653 -163.04
- DEPTH   1   2.25214 7.8581 -150.28

Call:
lm(formula = log10(AVG0004 + 1) ~ DEPTH + BSALIN + BTEMP, data = pf)

Coefficients:
(Intercept)        DEPTH       BSALIN        BTEMP
    3.23603     -0.38884     -0.01868     -0.07111
```

Let's look at this output from the stepwise regresssion and interpret it. This is a backwards stepwise regression, so it starts with the full model. The AIC of the full model is -169.06. The output shows the what happened when each of the predictors was removed: AIC is -170.41 when "PLUCHAN" was removed; this AIC is smaller, so it is better. The next step was to remove "POP2000", and the AIC dropped again. The final model from the stepwise regression is with "DEPTH", "BSALIN", and "BTEMP" included; the AIC is lowest at -172. 26. At this point you can see teh final reduced model with the coefficients (βs). This is the best-fitting multiple regression model. Here is the residual plot:

```
> plot(fit1.log)
```



There are still some extreme values in this plot, the plot is funnel-shaped, and the red line showing the trend in the residuals is not straight across at $e_i = 0$. Sometimes, our data are messy. There are likely to be other variables that are not in the full model (like current patterns and predator abundances), so that the fit would improve greatly if these were known.