Multiple regression with Forward Selection

Srikar

2021-09-02

Introduction:

Multiple regression is a statistical technique that can be used to analyze the relationship between a single dependent variable and several independent variables. The objective of multiple regression analysis is to use the independent variables whose values are known to predict the value of the single dependent value. Each predictor value is weighed, the weights denoting their relative contribution to the overall prediction.

R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable.
The P value is defined as the probability under the assumption of no effect or no difference (null hypothesis), of obtaining a result equal to or more extreme than what was actually observed. The P stands for probability and measures how likely it is that #any observed difference between groups is due to chance.

Forward selection is a type of stepwise regression which begins with an empty model and adds in variables one by one. In each forward step, you add the one variable that gives the single best improvement to your model. It is one of two commonly used methods of stepwise regression; the other is backward elimination, and is almost opposite. In that, you start with a model that includes every possible variable and eliminate the extraneous variables one by one. Forward selection typically begins with only an intercept. One tests the various variables that may be relevant, and the 'best' variable—where "best" is determined by some pre-determined criteria—is added to the model

Objective:

To predict the weight of the fish using the given regressors

Data Description:

This dataset is a record of a certain type of fish known as Perch Fish. It contains 5 independent variables and 1 dependent variable i.e., the weight of the fish (in grams). The dependent variables are type of species, vertical length (in cms), diagonal length (in cms), cross length (cms), height and width (in cms), The dataset contains a sample of 50 fishes in a market.

```r
library(readxl)
dat<- read_excel("C:/Users/Srikar/Desktop/SS/R/Sem 5/Linear Regression/Practi
cal 5/Data.xlsx")
head(dat,10)
```

```
      Length1 Length2 Length3 Height Width Weight
##      <dbl>   <dbl>   <dbl>  <dbl> <dbl>  <dbl>
##   1    23.2    25.4    30      11.5  4.02    242
##   2    24      26.3    31.2    12.5  4.31    290
##   3    23.9    26.5    31.1    12.4  4.70    340
##   4    26.3    29      33.5    12.7  4.46    363
##   5    26.5    29      34      12.4  5.13    430
##   6    26.8    29.7    34.7    13.6  4.93    450
##   7    26.8    29.7    34.5    14.2  5.28    500
##   8    27.6    30      35      12.7  4.69    390
##   9    27.6    30      35.1    14.0  4.84    450
## 10    28.5    30.7    36.2    14.2  4.96    500
```

```r
names(dat)
```

```
## [1] "Length1" "Length2" "Length3" "Height"  "Width"   "Weight"
```

Here the column names represent the following:

Length1 - Vertical Length
Length2 - Diagonal Length
Length3 - cross Length
height - Height
Width - Width

Procedure:

1) Constructing the regression model

*mod=**lm**(dat$Weight~.,data=dat)*

The model obtained is:

$$Y=-515.24+10.53X1+ 102.73X2+-98.25X3+49.49X4+ 79.76X5$$

where X1,X2,X3,X4 and X5 are Length1,Length2,Length3,height and Width respectively.

```
summary(mod)


## Call:
## lm(formula = dat$Weight ~ ., data = dat)
```

Residuals:				
Min	1Q	Median	3Q	Max
-171.02	-34.82	-10.19	31.8	176.48

Coefficients:					
Estimate	***Std.***	***Error***	***t***	***value***	***P-value***
(Intercept)	-515.24	92.19	-5.589	1.35E-06	***
Length1	10.53	47.25	0.223	0.824699	
Length2	102.73	50.01	2.054	0.045928	*
Length3	-98.25	26.95	-3.645	0.000702	***
Height	49.49	15.44	3.205	0.002515	**
Width	79.76	39.23	2.033	0.048072	*

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 63.65 on 44 degrees of freedom
## Multiple R-squared:  0.9574, Adjusted R-squared:  0.9526
## F-statistic:   198 on 5 and 44 DF,  p-value: < 2.2e-16
```

We see that that range of residuals is large which means that values (in grams) will differ from the observed value.

We observe that the intecept p-value is below the significance value (0.05) and hence we can say that the intecept is significant in the prediction. This means that if the values of th e regressors were all zero, the intercept would tell us the mean estimate of the dependent variable (Weight).Since height, width and length of fish can't be 0, the intecept value has no real meaning.

The p-values of all the regressors except length1 show significance. That means these var iables describe the linear relationship with the independent variable. Since most of the var iables show significance, we can say that model is a good fit. Also the overall p-value is a lso lesser than significane level (0.05) and hence we can say that the model is a good fit.

The R-squared value 0.9574 which means that 95.74% of the variation is explained by th e regressors. The adjusted R-Squared shows the variation explained by the regressors that truly contribute to the known variation.

To find out which varibales really contribute to the the model, we can test it out by forwa rd selection method.

2) Selecting the best regressors using forward selection method:

i) Starting with no regressors~

```
f1=lm(dat$Weight~1,data=dat)
summary(f1)
```

```
##
## Call:
## lm(formula = dat$Weight ~ 1, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -466.66 -306.41   20.84  229.59  533.34
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   466.66      41.35   11.29 3.13e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 292.4 on 49 degrees of freedom
```

As we previously observed, the intercept is significant but has no true meaning as height, width and length of fish can't be 0

#ii) Testing the other regressors

```
step(f1, direction="forward", scope=formula(dat))
```

```
## Start:  AIC=568.8
## dat$Weight ~ 1
##
##           Df Sum of Sq     RSS      AIC
## + Weight   1   4189147       0 -3160.81
## + Length2  1   3917741  271406   433.97
## + Width    1   3881654  307493   440.21
## + Length3  1   3876858  312289   440.98
## + Height   1   3810056  379091   450.68
## <none>                 4189147   568.80
##
## Step:  AIC=-3160.81
## dat$Weight ~ Weight
```

```
## Warning: attempting model selection on an essentially perfect fit is nonse
nse
```

```
##           Df  Sum of Sq        RSS     AIC
## <none>                  1.6210e-26 -3160.8
```

```
## + Width    1 4.4926e-28 1.5760e-26 -3160.2
## + Length2  1 2.9300e-28 1.5917e-26 -3159.7
## + Height   1 7.8700e-29 1.6131e-26 -3159.1
## + Length3  1 6.2420e-29 1.6147e-26 -3159.0

##
## Call:
## lm(formula = dat$Weight ~ Weight, data = dat)
##
## Coefficients:
## (Intercept)       Weight
##  -1.286e-13    1.000e+00
```

#We observe that only Width, lenght2 and height and length3 truly give the estimate values. Using only these variables and excluding length1, we will construct a new model.

```
nmod=lm(dat$Weight~dat$Length2+dat$Length3+dat$Height+dat$Width)
```

The new model obtained is:

Y=-521.02+112.24X1+-96.94X2+47.98X3+ 76.55X4

where X1,X2,X3 and X4 are Length2,Length3,height and Width respectively.

```
summary(nmod)
```

```
##
## Call:
## lm(formula = dat$Weight ~ dat$Length2 + dat$Length3 + dat$Height +
##     dat$Width)
##
```

Residuals:				
Min	***1Q***	***Median***	***3Q***	***Max***
-170.655	-32.194	-8.474	33.177	176.274

Coefficients:					
Estimate	Std.	Error	t	value	P-value
(Intercept	-521.02	87.53	-5.953	3.67E-07	***
dat$Lengt	112.24	25.82	4.346	7.81E-05	***
dat$Lengt	-96.94	26.02	-3.726	0.000542	***
dat$Heigh	47.98	13.74	3.492	0.001086	**
dat$Widtl	76.55	36.1	2.121	0.039492	*

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.97 on 45 degrees of freedom
## Multiple R-squared:  0.9574, Adjusted R-squared:  0.9536
## F-statistic: 252.8 on 4 and 45 DF,  p-value: < 2.2e-16
```

#We see that all variables are now significant. The r-squared has also increased sinc e we removed an insignificant factor.

#3) Constructing the test data for prediction

```
set.seed(123)
x1<-rnorm(50,29.41)
x2=rnorm(50,33.846)
x3=rnorm(50,12.47)
x4<-rnorm(50,4.80)
df<-data.frame(x1,x2,x3,x4)

A=predict(nmod,df)
df1=data.frame(dat$Weight,A,dat$Weight-A)
head(df1,10)
```

	Observed	Estimated	Difference
1	242	282.1336	-40.13365
2	290	334.7471	-44.74712
3	340	391.8784	-51.87838
4	363	438.2986	-75.29863
5	430	428.045	1.954956
6	450	478.5225	-28.52254
7	500	552.4795	-52.47949
8	390	420.1983	-30.19833
9	450	486.3316	-36.33162
10	500	477.7511	22.248854

From this table, we can compare the observed and expected values are close to each other .As the R-squared value explains only 95.36 of the variation. The rest of the variation is explained by chance or unknow causes.

Conclusion:

The newly obtained model which is the best fit for predicting the weight of the fish is :

Y=-521.02+112.24X1+-96.94X2+47.98X3+ 76.55X4 with an R-squared value of 95.36 of the known variation.