Multiple Regression

Robin Donatello

Last Updated 2017-11-13 21:13:33

Contents

Introduction

Variable selection methods are used mainly in exploratory situations where many independent variables have been measured and a final model explaining the dependent variable has not been reached.

To do variable selection you need:

1. A general test,
2. Selection criteria, and
3. A selection process.

Consider a model with P variables and you want to test if Q additional variables are useful.
$H_0 : Q$ additional variables are useless, i.e., their β's all $= 0$
$H_A : Q$ additional variables are useful

Ex: Y = FEV1, X1 = ht, X2 = age, X3 = ethnicity, X4 = location.
Test H_0: location does not matter.

A General Test

Likelihood Ratio (Deviance) Test

- Deviance = -2 log likelihood
- Under H_0, the *full model*, the deviance $= D_0, df_0 = N - P - 1$
- Under H_a, the *reduced model*, the deviance $= D_a, df_a = N - P - Q - 1$
- LR (deviance) test statistic is:
- $D_0 - D_a$ is distributed approximately as χ^2 with Q degrees of freedom under H_0 for large N.

If we assume normally distributed residuals, the LR test becomes an exact F=test.

$$F = \frac{(SSR_{red} - SSR_{full})/(df_{full} - df_{red})}{SSR_{full}/df_{full}}$$

Likelihood

Let X be a random variable with pdf f and that depends on the parameter θ. The function $\mathcal{L}(\theta|x) = f_\theta(x)$ then is called the *Likelihood function*. It is the likelihood of θ given the outcome x. Many analyses rely on maximizing this function (Maximum likelihood estimate or MLE), but commonly do so by first taking the log of this function. Hence the *log likelihood*.

Example: Testing adding Q variables to a model

Consider a model to predict depression using age, employment status and whether or not the person was chronically ill in the past year as covariates.

```
full_model <- lm(log(cesd+1) ~ age + chronill + employ, data=depress)
pander(summary(full_model))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.196	0.17	12.92	2.284e-30
age	-0.01495	0.003978	-3.758	0.0002078
chronill	0.3261	0.1159	2.814	0.005235
employPT	0.3544	0.2034	1.742	0.08251
employUnemp	0.3484	0.6787	0.5133	0.6081
employRetired	0.4801	0.4843	0.9912	0.3224
employHouseperson	0.3788	0.1667	2.273	0.02379
employStudent	0.3721	0.2135	1.743	0.08241
employOther	0.7734	0.2648	2.921	0.00377

Table 2: Fitting linear model: log(cesd + 1) ~ age + chronill + employ

Observations	Residual Std. Error	R^2	Adjusted R^2
294	0.9491	0.09904	0.07375

The results of this model show that age and chronic illness are statistically associated with CESD (each p<.006). However employment status is a mixed bag.

Recall that employment is a categorical variable, and all the coefficient estimates shown are the effect of being in that income category has on depression *compared to* being employed full time. For example, the coefficient for PT employment is greater than zero, so they have a higher CESD score compared to someone who is fully employed.

```
exp(.379)
```

```
## [1] 1.460823
```

Specifically while holding all other variables constant, someone who is working part time has 46% higher CESD score as someone who is working full time.

Since only a small constant was added to the CESD score, we can interpret the exponentiated coefficient as the fold change as seen previously with log(Y).

But what about employment status overall? Not all employment categories are significantly different from FT status. To test that employment status affects CESD we need to do a global test that all β's are 0.

$H_0 : \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$
H_A: At least one β_j is not 0.

We fit the reduced model, the one without employment category.

```
red_model <- lm(log(cesd+1) ~ age + chronill, data=depress)
```

and conduct a global F test by running an `anova()`. *Not to be confused with* `aov()`

```
anova(full_model, red_model)
```

```
## Analysis of Variance Table
##
## Model 1: log(cesd + 1) ~ age + chronill + employ
## Model 2: log(cesd + 1) ~ age + chronill
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    285 256.74
## 2    291 270.25 -6   -13.505 2.4986 0.02261 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that as a whole, employment significantly predicts CESD score.

This only is valid for nested models Meaning all variables in the reduced model are present in the full model.

Selection Criteria

Coefficient of Determination

If the model explains a large amount of variation in the outcome that's good right? So we could consider using R^2 as a selection criteria and trying to find the model that maximizes this value.

The residual sum of squares (RSS in the book or SSE) can be written as $\sum(Y - \hat{Y})^2(1 - R^2)$. Therefore minimizing the RSS is equivalent to maximizing the multiple correlation coefficient.

Multiple R^2 Problem: The multiple R^2 *always* increases as predictors are added to the model.

Adjusted R^2 Ok, so let's add an adjustment, or a penalty, to keep this measure in check. $R^2_{adj} = R^2 - \frac{p(1-R^2)}{n-p-1}$

Information Criteria

Mallows Cp

- Compares MSE of a reduced model to the full model.
- Penalized function, as P increases Cp decreases.
- Many investigators recommend selecting those independent variables that minimize the values of Cp.

Akaike Information Criterion (AIC)

- A penalty is applied to the deviance that increases as the number of parameters p increase.
- AIC $= -2LL + 2p$
- Smaller is better

Bayesian Information Criterion (BIC)

- A different penalty function
- BIC $= -2LL + p * ln(n)$
- Compare nested and non-nested models
- BIC identifies the model that is more likely to have generated the observed data.
- Smaller is better

Selection Process

We want to choose a set of independent variables that both will yield a good prediction using as few variables as possible. In many situations where regression is used, the investigator has strong justification for including certain variables in the model.

- previous studies
- accepted theory

The investigator may have prior justification for using certain variables but may be open to suggestions for the remaining variables.

The set of independent variables can be broken down into logical subsets

- The usual demographics are entered first (age, gender, ethnicity)
- A set of variables that other studies have shown to affect the dependent variable
- A third set of variables that *could* be associated but the relationship has not yet been examined.

Partially model-driven regression analysis and partially an exploratory analysis.

What to watch out for

- Use previous research as a guide
- Variables not included can bias the results
- Significance levels are only a guide
- Perform diagnostics after selection
- ***Use common sense***:
  - A sub-optimal subset may make more sense than optimal one
- Blind use of automated variable selection (forward/backward) is discouraged. See [here] and [here] In addition to the almost dozen entries in the textbook, see the following resources regarding areas of concern.