Variable Selection

Robin Donatello

Last Updated 2017-11-13 22:50:30

Contents

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

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.

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. $R^2_{adj} = R^2 - \frac{p(1-R^2)}{n-p-1}$

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

Wald test

The Wald test is used for simultaneous tests of Q variables in a model

- 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

This can be done in R by using the `regTermTest()` function in the `survey` package.

```
library(survey)
## regTermTest(model.name, "variable name to test") # not run
```

Example 1: Employment status on depression score

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(cesd ~ age + chronill + employ, data=depress)
pander(summary(full_model))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.48	1.502	7.646	3.191e-13
age	-0.133	0.03514	-3.785	0.0001873
chronill	2.688	1.024	2.625	0.009121
employPT	6.75	1.797	3.757	0.0002083
employUnemp	1.967	5.995	0.328	0.7431
employRetired	4.897	4.278	1.145	0.2533
employHouseperson	3.259	1.472	2.214	0.02765
employStudent	3.233	1.886	1.714	0.08756
employOther	7.632	2.339	3.263	0.001238

Table 2: Fitting linear model: cesd ~ age + chronill + employ

Observations	Residual Std. Error	R^2	Adjusted R^2
294	8.385	0.1217	0.09704

The results of this model show that age and chronic illness are statistically associated with CESD (each

p<.006). However employment status shows mixed results. Some employment statuses are significantly different from the reference group, some are not. So overall, is employment status associated with depression?

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.

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.

```
regTermTest(full_model, "employ")
```

```
## Wald test for employ
##  in lm(formula = cesd ~ age + chronill + employ, data = depress)
## F =  4.153971  on  6  and  285  df: p= 0.0005092
```

- Confirm that the degrees of freedom are correct. It should equal the # of categories in the variable you are testing, minus 1.
  - Employment has 7 levels, so $df = 6$.
  - Or equivelantly, the degrees of freedom are the number of *beta*'s you are testing to be 0.

The p-value of this Wald test is significant, thus employment significantly predicts CESD score.

Example 2: Blood Pressure

Consider a logistic model on smoking status (0= never smoked, 1=has smoked) using gender, income, and blood pressure class (`bp_class`) as predictors.

$$logit(Y) = \beta_0 + \beta_1(\text{female}) + \beta_2(\text{income}) + \beta_3(\text{Pre-HTN}) + \beta_4(\text{HTN-I}) + \beta_5(\text{HTN-II})$$

```
bp.mod <- glm(smoke ~ female_c + income + bp_class, data=addhealth, family='binomial')
pander(summary(bp.mod))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.046	0.1064	9.836	7.881e-23
female_cFemale	-0.6182	0.07617	-8.117	4.798e-16
income	-3.929e-06	1.411e-06	-2.785	0.005346
bp_classPre-HTN	0.07289	0.08206	0.8882	0.3745
bp_classHTN-I	-0.02072	0.1093	-0.1895	0.8497
bp_classHTN-II	0.02736	0.1888	0.1449	0.8848

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	4853 on 3728 degrees of freedom
Residual deviance:	4769 on 3723 degrees of freedom

It is unlikely that blood pressure is associated with smoking status, all groups are not statistically significantly different from the reference group (all p-values are large). Let's test that hypothesis formally using a Wald Test.

```
regTermTest(bp.mod, "bp_class")
```

```
## Wald test for bp_class
##  in glm(formula = smoke ~ female_c + income + bp_class, family = "binomial",
##     data = addhealth)
## F =  0.428004  on  3  and  3723  df: p= 0.73294
```

The Wald Test has a large p-value of 0.73, thus blood pressure classification is not associated with smoking status.

- This means blood pressure classification should not be included in a model to explain smoking status.

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 model diagnostics after selection to check model fit.
- ***Use common sense***: A sub-optimal subset may make more sense than optimal one