Model selection for linear models with unbalanced design

February 5, 2013

1 Unbalanced designs

For unbalanced designs, it is no longer true that the total variation in the response variable,

$$\mathrm{SSD}_T = \sum_{i=1}^{n} (Y_i - \bar{Y})^2, \tag{1}$$

can be decomposed into components corresponding to variation explained by different factors and numerical explanatory variables. Thus, certain percentages of the total variation can no longer be attributed to variation explained by each factor or numerical explanatory variable of a given model. Which variables that should be included in the model can also no longer be determined from the single analysis of variance table produced by `anova( )`.

It is still true, however, that the sum of squares representing the total variation can be decomposed into two parts; variation explained by the model and residual variation,

$$\mathrm{SSD}_T = \mathrm{SSD}_{\mathrm{model}} + \mathrm{SSD}_{\mathrm{res}}, \tag{2}$$

that is,

$$\sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{y}_i)^2, \tag{3}$$

where $\hat{y}_i$ is the expected response of the i'th observation based on the estimated model. For a model with p parameters including the intercept, it follows that

$$\frac{\mathrm{SSD}_T}{\sigma^2}, \quad \frac{\mathrm{SSD}_{\mathrm{model}}}{\sigma^2}, \quad \text{and } \frac{\mathrm{SSD}_{\mathrm{res}}}{\sigma^2}, \tag{4}$$

are chi-square with $n-1$, $p-1$ and $n-p$ degrees of freedom, respectively, under the null hypothesis

$$H_0 : Y_i = \mu + e_i, \tag{5}$$

that is, a model in which none of explanatory variables are included.

This null hypothesis can be tested against a given fitted model by using the ratio of between the two last independent chi-square distributed quantities of equation (4) divided by their respective degrees of freedom

$$F = \frac{\frac{\mathrm{SSD}_{\mathrm{model}}}{\sigma^2} \Big/ (p-1)}{\frac{\mathrm{SSD}_{\mathrm{res}}}{\sigma^2} \Big/ (n-p)} = \frac{\mathrm{MS}_{\mathrm{model}}}{\mathrm{MS}_{\mathrm{res}}} \tag{6}$$

which is F-distribuded with $p-1$ and $n-p$ degrees of freedom under H_0. H_0 is then rejected for a sufficiently large observed value of F indicating that a large proportion of the total variation is explained by the model.

Suppose that we are working with the data set

```
> trainingset
           y         x1        x2 f1
1  10.507078 0.25521840 0.7055308  a
2   9.782901 2.86878143 3.2946667  d
3  11.475879 1.28364151 1.6474991  b
4  10.678075 2.25265237 2.6209034  b
5  11.421004 3.15204571 2.9759809  c
6  11.909168 2.69234357 3.0451013  c
7   9.622157 0.04698267 0.6971617  a
8  11.968766 1.35175719 1.3708832  b
9  12.747906 4.22431362 3.7346717  c
10 11.858765 3.19981161 3.5966922  c
```

The F-value of 4.32 in the last line in the summary for the model with all three variables `x1`, `x2` and `f1` included,

```
> fullmodel <- lm(y~x1+x2+f1,data=trainingset)
> summary(fullmodel)

Call:
lm(formula = y ~ x1 + x2 + f1, data = trainingset)

Residuals:
         1          2          3          4          5          6          7 
 3.457e-01  3.849e-18  1.598e-01 -4.124e-01 -8.383e-01  1.828e-01 -3.457e-01 
         8          9         10 
 2.526e-01  3.545e-01  3.010e-01 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  10.7623     0.6872  15.662  9.7e-05 ***
x1            0.9780     0.7473   1.309    0.261    
x2           -1.2054     0.8930  -1.350    0.248    
f1b           1.2843     0.8683   1.479    0.213    
f1c           2.0015     1.5569   1.286    0.268    
f1d           0.1864     1.5626   0.119    0.911    
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 0.6022 on 4 degrees of freedom
Multiple R-squared: 0.844,Adjusted R-squared: 0.649 
F-statistic: 4.328 on 5 and 4 DF,  p-value: 0.09037 
```

indicates that we can reject the above null hypothesis (5) in favor of the fitted model (H_1) so we can conclude that some of the explanatory variables have a significant effect on the response. It remains to determine which of the variables we should include in the final model.

2 Tests between different nested alternatives: `drop1()`, `add1()`

Suppose that we want to test a given model H_0 (e.g. the model, in symbolic notation, `y ~ x1 + x2`) against an extended model H_1 (e.g. the model obtained by adding the factor `f1`, `y ~ x1`

+ x2 + f1), that is, does the factor f1 have an effect on y in a model that already contains x1 and x2? Let p_0 and p_1 be the number of estimated parameters under H_0 and H_1, respectively.

In general, when we add a term to a model, the residual sum of squares will always decrease. In general, a test of H_0 versus H_1 can in such cases be based on the test statistic

$$F = \frac{\frac{\mathrm{SSD}_{\mathrm{res},H_0} - \mathrm{SSD}_{\mathrm{res},H_1}}{\sigma^2} \Big/ (p_1 - p_0)}{\frac{\mathrm{SSD}_{\mathrm{res},H_1}}{\sigma^2} \Big/ (n - p_1)} \tag{7}$$

which is F-distributed with $p_1 - p_0$ and $n - p_1$ degrees of freedom under the null hypothesis that the additional term has no effect on the response.[1] Again, we reject this null hypothesis if this statistics takes a large value. This will occur if we observe a large change in the residual sum of squres when adding the extra term making the numerator large.

Using drop1( ) we carry out tests of this kind of different reduced models tested against a given fitted model corresponding to removal of individual terms one at the time.

```
> drop1(fullmodel,test="F")
Single term deletions

Model:
y ~ x1 + x2 + f1
       Df Sum of Sq    RSS     AIC F value  Pr(F)
<none>               1.4507 -7.3051
x1      1   0.62127 2.0720 -5.7406  1.7130 0.2607
x2      1   0.66091 2.1117 -5.5511  1.8223 0.2484
f1      3   2.73843 4.1892 -2.7008  2.5168 0.1969
```

The residuals sums of squares (the column named RSS) lists the residual sums of squares under the full model, and under different reduced models obtained by removing x1, x2 and f1, respectively.

Focusing on f1, the observed difference between the sum of squares between the reduced and the full model (appearing in the numerator of equation (7)) becomes $4.1892 - 1.4507 = 2.7384$ (appearing in the column Sum of Sq. The change in number of parameter $p_1 - p_0$ is one less than the number of levels of the factor f1, that is, 4-1, and is listed in the column named Df. The residual degrees of freedom $n - p_1$ is $10 - 6$ since 6 parameters (including the intercept) are being estimated under the full model. For f1 this gives the observed F-value of

$$F^* = \frac{2.7384/3}{1.4507/4} = 2.5168 \tag{8}$$

listed in the column F value and a P value of

```
> pf(2.5168, df1=3, df2=4, lower.tail=F)
[1] 0.1969
```

listed in the column Pr(F).

According to the above table, x1 is the least significant term of the full model. Thus a reasonable first step is to remove this term from the model and use drop1( ) on the reduced model to see if the significance of the other variables have changed.

```
> reduced <- lm(y~x2+f1,data=trainingset)
> drop1(reduced,test="F")
```

[1] This result holds more generally also if H_0 is nested within H_1. By this we mean that H_0 can be seen as a special case of H_1. In this sense, (6) is a special case of (7).

```
Single term deletions

Model:
y ~ x2 + f1
       Df Sum of Sq    RSS     AIC F value  Pr(F)
<none>              2.0720 -5.7406
x2      1    0.0898 2.1618 -7.3165  0.2166 0.6612
f1      3    5.2447 7.3167  0.8757  4.2187 0.0776 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

For this particular data set, we see that x2 should also be removed. Without x2 in the model, the factor f1 becomes significant and should not be removed.

```
> reduced2 <- lm(y~f1,data=trainingset)
> drop1(reduced2,test="F")
Single term deletions

Model:
y ~ f1
       Df Sum of Sq    RSS     AIC F value   Pr(F)
<none>              2.1618 -7.3165
f1      3    7.1373 9.2991  1.2733  6.6032 0.02496 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Having removed x2, the effect of other variables, in this case x1, needs to be reconsidered because x1 may have a significant effects in a model without x2. In general, we may specify a list of variables to be considered for addition to the current model using add1( ) as follows.

```
> add1(reduced2, .~.+x1+x2+f1,test="F")
Single term additions

Model:
y ~ f1
       Df Sum of Sq    RSS     AIC F value  Pr(F)
<none>              2.1618 -7.3165
x1      1  0.050125 2.1117 -5.5511  0.1187 0.7445
x2      1  0.089765 2.0720 -5.7406  0.2166 0.6612
```

Since f1 is already included in the model it is not considered for addition. We already know that x2 should not included. The final row in the table tells us that x1 don't have a significant effect in the model only including f1 since the P-value is 0.7445.

Note that if all variables are simple numerical variables or factors with only two levels, the F-tests produced by drop1( ) are equivalent to the t-tests produced by summary( ). You may verify that this is the case in the output above.

3 Simpler models give better predictions

The above data is a subset of a dataset containing $n = 20$ observations simulated from the model

$$Y = \mu + \alpha_1 x_1 + \alpha_2 x_2 + \beta_j + e, \quad e \sim N(0, \sigma^2) \tag{9}$$

with parameter values

$$\begin{aligned}\mu &= 10,\\ \alpha_1 &= 0.01,\\ \alpha_2 &= 0.2,\\ \beta_1 &= 0, \quad \beta_2 = 1, \quad \beta_3 = 1.5, \quad \beta_2 = -0.5,\\ \sigma &= 0.5.\end{aligned} \tag{10}$$

using the R-code

```
set.seed(3)
f1 <- factor(sample(c("a","b","c","d"),20,repl=T)) # a factor
x1 <- rnorm(20,(1:4)[f1])
x2 <- rnorm(20,x1,sd=.5)
y <- rnorm(20,
          mean=10 + 0.01*x1 + 0.2*x2 + c(0,1,1.5,-.5)[f1],
          sd=.5)
completedata <- data.frame(y,x1,x2,f1)
rm(x1,x2,f1)
trainingset <- completedata[1:10,]
validationset <- completedata[11:20,]
```

So in reality we know that x_1 and x_2 do have an effect on the response y. In many real world applications, we may have reason to believe that almost any variable have at least a small effect on a given response variable. Why should we then not include these in our estimated model?

While it is true that the full model will give the "best" predictions for the observed response for the subset of the data used for estimating the model, a good model should clearly be able to predict the response also of future observations. We may validate the different alternative models estimated from the first part of the simulated data set used above, the training set, by making predictions for the values of the explanatory variables given in the second part of the data set and comparing these predictions with the actual observed values in the validation set.

Predicted values for the validation set based on the full model estimated from the training set can be computed as follows

```
> predict(fullmodel,newdata=validationset)
       11        12        13        14        15        16        17        18
11.739122 12.483014 11.104000 12.939481  9.939196 11.572952 10.368710 11.169888
       19        20
10.329998 12.392251
```

The corresponding observed responses in the validation set (what we are trying to predict) can be referred to using

```
> validationset$y
 [1] 12.373928 12.360169 12.385249 11.876564 10.860277  9.717411 10.228094
 [8] 12.933177  9.634806 11.184458
```

The sum of squared differences between the observed and predicted values is a reasonable measure of the the overall prediction error

```
> sum((validationset$y - predict(fullmodel,newdata=validationset))^2)
[1] 12.5519
```

The same measure of overall prediction errors based on the above simpler reduced model containing only `f1` becomes

```
> sum((validationset$y - predict(reduced2,newdata=validationset))^2)
[1] 2.615882
```

that is, considerably smaller.

The better prediction of our reduced model may of course be an effect of chance. Nevertheless, if we repeated the above simulation many times, we would see that simpler models selected based on criterias similar to the ones used above, in the long run, would produce better predictions.

4 Several models may appear reasonable

Consider a simple dataset generated using the following code

```
set.seed(1)
x1 <- rnorm(40)
rho <- .98
x2 <- rnorm(40,rho*x1,sqrt(1-rho^2))
y <- rnorm(40,mean=10 + 1*x1,sd=1)
```

Here, x_1 and x_2 have been simulated from a binormal distribution with correlation equal to 0.98. We have then simulated the response y by assuming that y depends on x_1 only. If we now fit a model with both x_1 and x_2 as explanatory variables neither x_1 or x_2 appear to have a significant effect on y. This arise because of the colinearity between x_1 and x_2.

```
> summary(fullmodel)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.3048 -0.4472 -0.1376  0.6518  1.5887 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)   0.1147     0.1304   0.879   0.3850  
x1           -0.5264     0.7627  -0.690   0.4944  
x2            1.2499     0.7282   1.716   0.0944 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 0.8155 on 37 degrees of freedom
Multiple R-squared: 0.4431,Adjusted R-squared: 0.413 
F-statistic: 14.72 on 2 and 37 DF,  p-value: 1.978e-05
```

The output from `drop1()` gives the same P-values.

```
> drop1(fullmodel,test="F")
Single term deletions

Model:
y ~ x1 + x2
       Df Sum of Sq    RSS     AIC F value   Pr(F)  
```

```
<none>              24.606 -13.435
x1      1   0.31679 24.923 -14.923  0.4763 0.49439
x2      1   1.95933 26.566 -12.370  2.9462 0.09444 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

We may consider two different reduced models with only x_1 or x_2 included.

```
> summary(lm(y˜x1))

Call:
lm(formula = y ˜ x1)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4196 -0.5722 -0.1809  0.5359  1.7198

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1391     0.1329   1.046    0.302
x1            0.7581     0.1510   5.021 1.24e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.8361 on 38 degrees of freedom
Multiple R-squared: 0.3988,Adjusted R-squared: 0.383
F-statistic: 25.21 on 1 and 38 DF,  p-value: 1.244e-05

> summary(lm(y˜x2))

Call:
lm(formula = y ˜ x2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3480 -0.4986 -0.1195  0.6486  1.6018

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1225     0.1290   0.949    0.348
x2            0.7568     0.1396   5.420 3.56e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.8099 on 38 degrees of freedom
Multiple R-squared: 0.436,Adjusted R-squared: 0.4211
F-statistic: 29.37 on 1 and 38 DF,  p-value: 3.556e-06
```

In both models all variables included are significant and extensions of either model (the full model) are in both cases non-significant. So both models are reasonable choices of a “best” model. This occurs because x_1 and x_2 are highly correlated and thus contains almost the same information. In this case, we know that there is only a true causal path from x_1 to y. We have no way of knowing this from the data, however.

In real world applications, we may try to measure the same explanatory variable in several ways. Including different alternative measures is then seldom a good idea. Instead you should make a choice a priori between different alternative measures and use the one that you think most accurately represent what you want to include in the model.

5 Parsimony, Akaikes information criteria, `step()`

Model selection can be seen as a trade off between minimising the bias and variance of the predicted values. For a simple linear model, it is known that the estimators of the regression coefficients are unbiased, that is, $E(\hat{\beta}_i) = \beta_i$ for all the parameters. This implies that the expected value of our predictions

$$E(\hat{y}) = E(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k = y \tag{11}$$

is unbiased aswell if all relevant explanatory variables (here, $x_1, x_2, \ldots, x_k$) have been included in the model. If excluding some explanatory variable for which the true regression coefficient is nonzero, the predictions become biased, that is, $\text{Bias}(\hat{y}) = E(\hat{y}) - y \neq 0$. Thus, it would seem that a sensible model selection strategy, in terms of reducing the bias of our predictions, would be to include as many covariates as possible.

This argument is not in itself flawed but ignores the fact the variance of our predictions $\text{Var}(\hat{y})$ increases as we include more covariates and unknown parameters that we can only estimate in the model. Recall the formula for the variance of linear combinations of random variables, in our case, the $\hat{\beta}_i$'s in equation (11).

To see the trade off more explicitly, consider the mean squared prediction error, $E[(\hat{y} - y)^2]$. This quantitity can be decomposed into two components associated with the bias and variance of the predictions as follows

$$\begin{aligned} E[(\hat{y} - y)^2] &= E[((\hat{y} - E\hat{y}) + (E\hat{y} - y))^2] \\ &= E[(\hat{y} - E\hat{y})^2] + 2E(\hat{y} - E\hat{y})(E\hat{y} - y) + (E(\hat{y}) - y)^2 \\ &= \text{Var}(\hat{y}) + [\text{Bias}(\hat{y})]^2 \end{aligned} \tag{12}$$

Thus, a complex model with many parameters reduces the bias $\text{Bias}(\hat{y}) = E(\hat{y}) - y$ but comes with the cost of an increase in $\text{Var}(\hat{y})$. Conversely, a simple model for which the estimated parametes have small variance $\text{Var}(\hat{y})$ comes at the cost of large bias $\text{Bias}(\hat{y})$. The optimal model complexity, in terms of the above quantity, is thus a model with moderate number of parameters such that the sum of the bias and variance components is minimised (Fig. 1).

Akaike's information criterion (AIC) is a statistic used in model selection and is defined as

$$AIC = -2\ln L + 2p, \tag{13}$$

where $\ln L$ is the maximised log likelihood of a given fitted model and p is the number of estimated parameters. We seek to find the model with the smallest AIC-score. Note how L and $\ln L$ will increase (and hence $-\ln L$ will decrase) as we include more explanatory variables and parameters in a model L where as the second term $2p$ increases with the number of parameters. Minimising the AIC-score can thus be seen as a trade off between improving model fit (measured by the first term) and model complexity (the second term). The theoretical rationale of this model selection procedure is given in Akaike (1974), also see the book by Burnham & Anderson (2002).

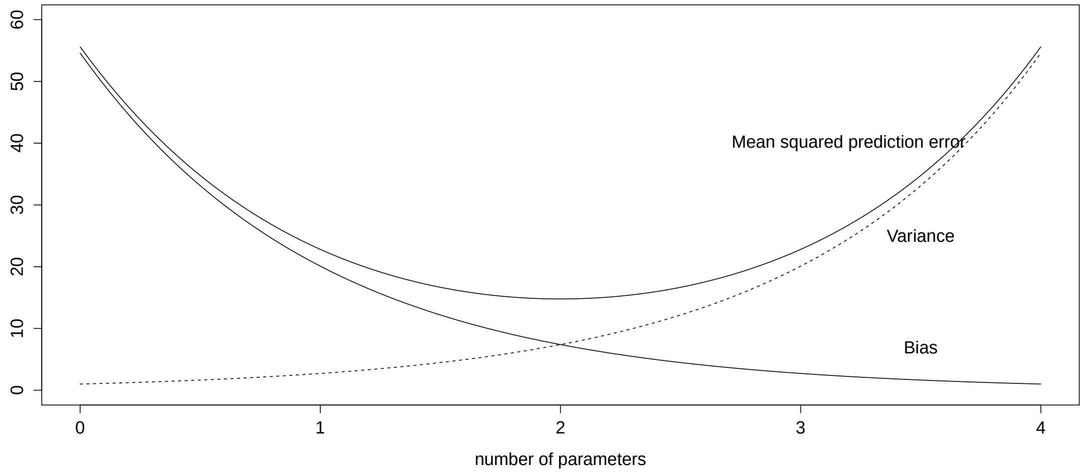

Figure 1: Choosing a parsimoneous model can be seen as a trade of between minimising the bias and variance of the predicted values.