18.5: Selecting the best model

Introduction

This is a long entry in our textbook, with many topics to cover. We discuss aspects of **model fitting**, from why model fitting is done to how to do it and what statistics are available to help us decide on the **best model**. Model selection very much depends on what the intent of the study is. For example, if the purpose of **model building** is to provide the best description of the data, then in general one should prefer the **full** (also called the **saturated**) **model**. On the other hand, if the purpose of model building is to make a predictive statistical model, then a **reduced model** may prove to be a better choice. The text here deals mostly with the later context of model selection, finding a justified reduced model.

From Full model to best Subset model

Model building is an essential part of being a scientist. As scientists, we seek models that explain as much of the variability about a phenomenon as possible, but yet remain simple enough to be of practical use.

Having just completed the introduction to multiple regression, we now move to the idea of how to pick best models.

We distinguish between a full model, which includes as many variables (predictors, factors) as the regression function can work with, returning interpretable, if not always statistically significant output, and a saturated model.

The saturated model is the one that includes all possible predictors, factors, and interactions in your experiment. In well-behaved data sets, the full model and the saturated model will be the same model. However, they need not be the same model. For example, if two predictor variables are highly **collinear**, then you may return an error in regression fitting.

For those of you working with **meta-analysis** problems, you are unlikely to be able to run a saturated model because some level of a key factor are not available in all or at least most of the papers. Thus, in order to get the model to run, you start dropping factors, or you start nesting factors. If you were unable to get more things in the model, then this is your “full” model. Technically we wouldn’t call it saturated because there were other factors, they just didn’t have enough data to work with or they were essentially the same as something else in the model.

Identify the model that does run to completion as your full model and proceed to assess model fit criteria for that model, and all reduced models thereafter.

In R (Rcmdr) you know you have found the full model when the output lacks “NA” strings (**missing values**) in the output. Use the full model to report the values for each coefficient, i.e., conducting the inferential statistics.

Get the estimates directly from the output from running the regression function. You can tell if the effect is positive (look at the estimate for sample — it is positive) so you can say — more samples, greater likelihood to see more cases of cancer.

Remember, the experimental units are the papers themselves, so studies with larger numbers of subjects are going to find more cases of diabetes. We would worry big time with your project if we did not see statistically significant and positive effects for sample size.

For illustration, here’s an example output following a run with the linear model function on an experimental data set.

The variables were

BMI = Dependent variable, continuous]

Age = Independent variable, continuous

CalsPDay = Independent variable, continuous

CholPDay = Independent variable, continuous

Sex = Independent variable, categorical

Smoke = Independent variable, categorical

```
lm(formula = BMI ~ Age + CalsPDay + CholPDay + Sex + Smoke +
Sex:Smoke, data = BMI) Residuals:
Min 1Q Median 3Q Max
```

```
-9.9685 -3.3766 -0.6609 2.5090 22.3482

Coefficients:
                          Estimate   Std. Error   t value    Pr(>F)
(Intercept)             25.9351297    3.7205047     6.971   1.708e-09 ***
Age                      0.890
CalsPDay                -0.0005757    0.0009882    -0.583    0.562
 CholPDay                0.0103521    0.0060722     1.705    0.093 .
 Sex[T.M]               -0.8529925    2.2209045    -0.384    0.702
 Smoke[T.Yes]           -1.1670159    1.9134734    -0.610    0.544
Sex[T.M]:Smoke[T.Yes]    0.9261469    2.8510680     0.325    0.746
```

The Y-variable was BMI, and the predictor variables included gender (male, female), smokers (yes, no), and the interaction, plus two measures of diet quality (calories per day and amount of cholesterol).

Question. Write out the equation in symbol form.

We see that none of the factors or covariates were statistically significant, so I wouldn't go on and on about positive or negative.

But, for didactic purposes here, imagine the P-value for `CholPDay` was less than 0.05 (and therefore statistically significant). We report the value (0.0103521 → I would round to 0.01), and note that those who had more cholesterol in their diet per day, those individuals tended to have higher BMI (e.g., the sign of the coefficient — and I related the coefficient back to the most important thing about your study — the biological interpretation).

Now's a good time to be clear about HOW you report statistical results. DO NOT SIMPLY COPY AND PASTE EVERYTHING into your report. Now, for the estimates above, you would report everything, but not all of the figures. Here's how the output should be reported in your Project paper.

Table 18.5.1. Coefficients from full model.

```
                     Estimate      SE       t   P-value
Intercept              25.935   3.721    6.97  < 0.0001
Age                    -0.007   0.051   -0.14    0.8892
Calories/Day           -0.001   0.001   -0.58    0.5622
Cholesterol/Day         0.010   0.006    1.71    0.0929
Sex                    -0.853   2.221   -0.38    0.7021
Smoke                  -1.167   1.915   -0.61    0.5440
Interaction Smoke:Sex   0.926   2.851    0.33    0.7463
```

Looks better, doesn't it?

Once you have the full model, use this model for the inferential statistics. Use the significance tests of each parameter in the model from the corresponding ANOVA table. Now, where is the ANOVA table? Remember, right after running the linear regression,

Rcmdr: Models → Hypothesis testing → ANOVA tables

Accept the default (**partial marginality**), and, Boom! Out pops the ANOVA table you should be familiar with.

From the ANOVA table you will tell me whether a Factor is significant or not. You report the ANOVA table in your paper. You describe it.

Now, the next step is to decide what is the best model. It then guides you to the next step which is to decide whether a better model (fewer parameters, Occam's razor) can be found. Identify the parameter from the ANOVA table with the highest P-value and remove it from the model when you run the regression again. Repeat the steps above, return the **ANOVA table**, checking the estimates and P-values, until you have a model with only statistically significant parameters.

Find the best model

Output from R follows:

```
Anova(LinearModel.1, type="II")
Anova Table (Type II tests)

Response: BMI
             Sum Sq   Df     F value        P
Age            0.62    1      0.0196    0.890
CalsPDay      10.79    1      0.3394    0.562
CholPDay      92.37    1      2.9065    0.093
Sex            1.52    1      0.0478    0.828
Smoke          8.84    1      0.2782    0.600
Sex:Smoke      3.35    1      0.1055    0.746
Residuals   2129.34   67
```

This is my full model and I would start anticipating the need to reduce my model because none of the factors are statistically significant. By the criterion that simple models are better, I would proceed first to drop the interaction. See below for more on selecting the best models.

But first, I want to take up an important point about your models that you may not have had a chance to think about. The order of entry of parameters in your model can effect the significance and value of the estimates themselves. The order of parameter model entry above can be read top to bottom. Age was first, followed in sequence by CalsPDay, CholPDay, and so on. By convention, enter the covariates first (the ratio-scale predictors), that's what I did above.

Here's the output from a model in which I used a different order of parameters.

```
Anova(LinearModel.2, type="II")
Anova Table (Type II tests)

Response: BMI
              Sum Sq    Df    F value     Pr(>F)
Sex             1.52     1     0.0478    0.82764
Smoke           8.84     1     0.2782    0.59964
Age             0.62     1     0.0196    0.88918
CalsPDay       10.79     1     0.3394    0.56215
CholPDay       92.37     1     2.9065    0.09286 .
Sex:Smoke       3.35     1     0.1055    0.74631
Residuals    2129.34    67
```

The output is the same!!! So why did I give you a warning about parameter order? Run the ANOVA table summary command again, but this time select **Type III type of test**, i.e., ignore marginality.

```
> Anova(LinearModel.2, type="III")
Anova Table (Type III tests)

Response: BMI
               Sum Sq    Df    F value       Pr(>F)
(Intercept)   1544.34     1    48.5929   1.708e-09 ***
Sex              4.69     1     0.1475     0.70214
```

```
Smoke           11.82     1     0.3720     0.54400
Age              0.62     1     0.0196     0.88918
CalsPDay        10.79     1     0.3394     0.56215
CholPDay        92.37     1     2.9065     0.09286 .
Sex:Smoke        3.35     1     0.1055     0.74631
Residuals     2129.34    67
```

The output has changed — and in fact it now reports the significance test of the intercept. This output is the same as the output from the linear model. Try again, this time selecting Type I, sequential:

```
> anova(LinearModel.2)
Analysis of Variance Table

Response: BMI
           Df   Sum Sq    Mean Sq  F value    Pr(>F)
Sex         1     0.68      0.681   0.0214   0.88402
Smoke       1     2.82      2.816   0.0886   0.76690
Age         1     3.44      3.436   0.1081   0.74333
CalsPDay    1     2.27      2.272   0.0715   0.78998
CholPDay    1    96.30     96.299   3.0301   0.08633
Sex:Smoke   1     3.35      3.354   0.1055   0.74631
Residuals  67  2129.34     31.781
```

Here, we see the effect of order. So, as we are working to learn all of the issues of statistics and in particular mode fitting, I have purposefully restricted you to **Type II analyses** — obeying **marginality** correctly handles most issues about **order of entry**.

Status check

Where are we??? Recall that the purpose of all of this effort is to find the best supported model. The question we are working on is whether the full (saturated) model is the best model or if a reduced model can be supported.

We go back to my first full model output from ANOVA.

Model 1:

```
Anova(LinearModel.1, type="II")
Anova Table (Type II tests)

Response: BMI
            Sum Sq   Df   F value    Pr(>F)
Age           0.62    1    0.0196   0.88918
CalsPDay     10.79    1    0.3394   0.56215
CholPDay     92.37    1    2.9065   0.09286 .
Sex           1.52    1    0.0478   0.82764
Smoke         8.84    1    0.2782   0.59964
Sex:Smoke     3.35    1    0.1055   0.74631
Residuals  2129.34   67
```

We have two factors (Sex, Smoke), three covariates (Age, CalsPDay, CholPDay), and one two-way interaction (Sex:Smoke). We would write our full model then as

$$BMI \sim Age + CalsPDay + CholPDay + Sex + Smoke + Sex : Smoke$$

Get and save in your output the ANVOA table for this Full model. Proceed to test a series of nested reduced models. Start by dropping the interaction terms, consistent with our Occam's razor approach.

Model 2:

```
Anova Table (Type II tests)

Response: BMI
            Sum Sq   Df   F value    Pr(>F)
Age           0.63    1    0.0201   0.88760
CalsPDay     10.45    1    0.3331   0.56572
CholPDay     96.30    1    3.0704   0.08424 .
Sex           1.52    1    0.0484   0.82650
Smoke         8.84    1    0.2819   0.59720
Residuals  2132.69   68
```

Next, reduce by identifying Factors or Predictors with the highest P-values. Looks like "Age" is next.

Model 3:

```
Anova Table (Type II tests)

Response: BMI
            Sum Sq   Df   F value   Pr(>F)
CalsPDay      9.87    1    0.3194  0.57381
CholPDay     97.78    1    3.1627  0.07974 .
Sex           2.55    1    0.0824  0.77488
Smoke         8.61    1    0.2786  0.59934
Residuals  2133.32   69
```

Next up, drop the "Sex" parameter.

Model 4:

```
Anova Table (Type II tests)

Response: BMI
            Sum Sq   Df  F value    Pr(>F)
CalsPDay     10.45    1   0.3424   0.56032
CholPDay     96.12    1   3.1502   0.08027 .
Smoke        12.42    1   0.4070   0.52557
Residuals  2135.87   70
```

Next? You would select CalsPDay, right?

Model 5:

```
Anova Table (Type II tests)

Response: BMI
            Sum Sq   Df   F value    Pr(>F)
```

```
CholPDay     90.24    1    2.9852   0.08837 .
Smoke        13.95    1    0.4613   0.49923
Residuals  2146.32   71
```

And finally, we remove the “Smoke” factor.

Model 6:

```
Anova Table (Type II tests)

Response: BMI
            Sum Sq   Df   F value   Pr(>F)
CholPDay     77.93    1    2.5974   0.1114
Residuals  2160.26   72
```

Oops!

What happened to Model 6? Nothing remains significant? Panic? What is the point??? Arggh, Dr Dohm…!!!

Easy there…. Take a deep breath, and guess what? Your best model needs to have significant parameters in it, right? Your best fit model then is Model 5. And that model will be your candidate for best fit as we proceed to complete our model building.

$$BMI \sim CholPDay + Smoke$$

Now we proceed to gain some support evidence for our candidate best model. We are going to use an information criterion approach.

Use a fit criterion for determining model fit

To help us evaluate evidence in favor of one model over another there are a number of statistics one may calculate to provide a single number for each model for comparison purposes. The criteria model evaluators available to us include **Mallow’s C_{p}\)**, **adjusted R^2**, Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC) to select best model.

We already introduced the **coefficient of determination** R^2 as a measure of fit – in general we favor models with larger values of R^2. However, values of R^2 will always be larger for models with more parameters. Thus, the other evaluators attempt to adjust for the parameters in the model and how they contribute to increased model fit. For illustrative purposes we will use Mallow’s C_p. The equation for Mallow’s C_p in linear regression is

$$C_p = \frac{Reduced\ SS_{residual}}{Full\ MS_{residual}} - [n - 2(p+1)]$$

where p is the number of parameters in the model. Mallow’s C_p is thus equal to the number of parameters in the model plus an additional amount due to lack of fit of the model (i.e., large residuals). All else being equal we favor the model in which the Cp is close to the number of parameters in the model.

In Rcmdr, select **Models** → **Subset model selection …** (Fig. 18.5.1)

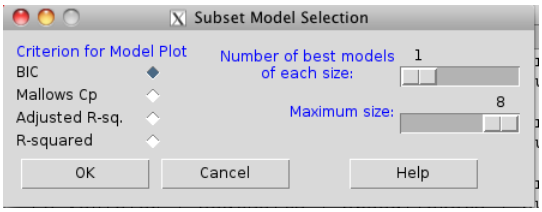

Figure 18.5.1: Rcmdr popup menu, Subset model selection…

From the menu, select the criterion and how many models to return. The function returns a graph that can be used to interpret which model is best given the selection criterion used. Below is an example (although for a different data set!) for Mallow’s C_p (Fig. 18.5.2).

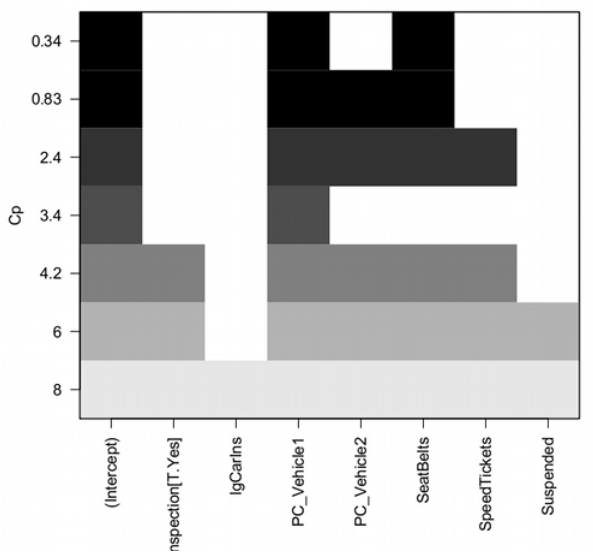

Figure 18.5.2: Mallow's C_p plot

Let's break down the plot. First define the axes. The vertical axis is the range of values for the C_p calculated for each model. The horizontal axis is categorical and reads from left to right: Intercept, Inspection[T.Yes], etc., up to Suspended. Looking into the graph itself we see horizontal bars — the extent of shading indicates which model corresponds to the C_p value. For example, the lowest bar, which is associated with the C_p value of 8, extends all the way to the right of the graph. This says that the model evaluated included all of the variables and therefore was the saturated or full model. The next bar from the bottom of the graph is missing only one block (lgCarIns), which tells us the C_p value 6 corresponds to a reduced model, and so forth.

Cross-validation

Once you have identified your Best Fit model, then, your proceed to run the diagnostics plots. For the rest of the discussion we return to our first example.

Rcmdr: **Models → Graphs → Basic diagnostic plots**.

We'll just concern ourselves with the first row of plots (Fig. 18.5.3).

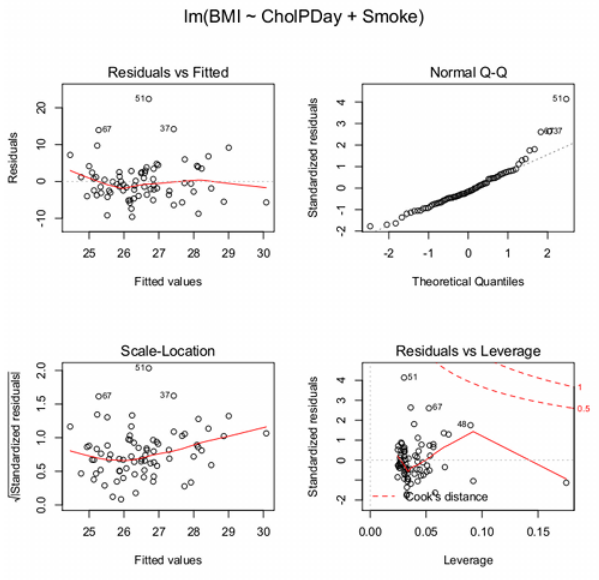

Figure 18.5.3: Diagnostic plots.

The left one shows the residuals versus the predicted values — if you see a trend here, the assumption of linearity has been violated. The second plot is a test of the assumption of normality of the residuals. Interpret them (residuals OK, Residuals normally distributed? Yes/No), and you're done. Here, I would say I see no real trend in the residuals vs. fitted plot, so assumption of linear

fit is OK. For normality, there is a tailing off at the larger values of residuals, which might be of some concern (and I would start thinking about possible leverage problems), but nothing dramatic. I would conclude that our Model 5 is a good fitting model and one that could be used to make predictions.

Now, if you think a moment, you should identify a logical problem. We used the same data to "check" the model fit as we did to make the model in the first place. In particular if the model is intended to make predictions it would be advisable to check the performance of the model (e.g., does it make reasonable predictions?) by supplying new data, data not used to construct the model, into the model. If new data are not available, one acceptable practice is to divide the full data set into at least two subsets, one used to develop the model (sometimes called the calibration or training dataset) and the other used to test the model. The benefits of cross-validation include testing for influence points, over fitting of model parameters, and a reality check on the predictions generated from the model.

Questions

[pending]