Biostatistics in Research Practice - Regression II

Simon Crouch

29th January 2008

Categorical Predictors and Dummy Variables

You may wish to build a statistical model in which one or more of your explanatory variables is *categorical*. That is, the variable indicates which class the data record corresponds to rather than representing a numerical value. For example, instead of having a continuous explanatory variable `weight` measured in kilograms, one might have the categorical variable `weight_category` with values `heavy`, `medium` and `light`. One might *code* these values as 0, 1 and 2 in a dataset, but it's important to remember that these codes are purely artificial constructs and shouldn't, in general, be considered as true numerical values. The question arises as to how to represent this type of variable in a regression model. One shouldn't use the code (0, 1 and 2) directly as that doesn't represent any sort of true numerical value. One has to use what are called *dummy variables*. In our example of `weight_category`, one defines three new explanatory variables `weight_category_1`, `weight_category_2`, `weight_category_3` where each takes the value zero or one for each subject according to which weight category the subject is in. So for a `medium` weight subject, `weight_category_1=0`, `weight_category_2=1`, `weight_category_3=0`. It turns out that these three variables contain too mcuh information between them and you only need two of them to determine which category the subject is in. Therefore one chooses one of the categories to be a baseline category (corresponding to a sort of "zero") and includes the dummy variables corresponding to the two other categories in the regression.

When one has categorical predictors in a regression, especially when they have three or more levels, it becomes harder to interpret the statistical significance of such predictors as they are each represented by more than one dummy variable. A statistical test of such a categorical predictor has to be performed by bundling together the statistical significance of all the corresponding dummy variables. We'll see how to do this using the *likelihood ratio test* in the practical.

Interpreting categorical predictors and their interactions with other categorical or continuous explanatory variables can get complicated. For now, just note that it is possible for a categorical predictor to be statistically significant overall, but for there to be no statistically significant difference between particular levels of the predictor. If this is the case, it might make sense to recategorize the predictor.

Model Building

Building a model that includes the appropriate explanatory variables in the appropriate functional form is, in general, tricky and to a certain extent is a craft rather than a science. It's important to emphasize that different reasonable models may be developed for the same dataset. One must assume that for real data we are unlikely to find the *true model* that has generated that data; we will be more than happy with a *correct model.* Or as the statistician George Box is reported as saying: "all models are wrong but some are useful".

If one already has proven scientific knowledge about the problem at hand, then one should use that knowledge, if possible, to determine which explanatory variables should be included in a model. However, it's often the case that such knowledge provides only a partial help or it may be that you're the first person trying to establish a plausible scientific theory! In either case, the design of your study should determine what is going to be measured and then you might use one of the following techniques.

- Backwards elimination: Starts with all the possible explanatory variables and their interactions in the model. Successively eliminates terms from the model one by one, at each stage eliminating the term that is "least significant" according to some criterion (such as size of p-value).
- Forward selection: Starts with no terms in the model. Successively adds terms to the model by choosing from the possible remaining terms the one that is "most significant" when added to the model so far.
- Stepwise selection: A combination of backwards elimination and forwards selection, of which there are a number of flavours. A common feature is that at each step a variable may be added but it might be removed at a later step, and vice versa.

Here's an example of backwards elimination. The response variable here is the murder rate in each of the 51 states in the US. You'll be using this dataset in the exercises.

```
             Estimate Std. Error t value P-value
(Intercept) -33.75475   18.96622  -1.780   0.0819 .
pctmetro      0.05105    0.03674   1.390   0.1715
pctwhite     -0.19596    0.07878  -2.487   0.0166 *
pcths         0.20000    0.22201   0.901   0.3725
poverty       0.58158    0.31437   1.850   0.0709 .
single        2.82440    0.57737   4.892 1.32e-05 ***

             Estimate Std. Error t value P-value
(Intercept) -19.78721   10.90079  -1.815   0.0760 .
pctmetro      0.04632    0.03629   1.277   0.2081
pctwhite     -0.17338    0.07453  -2.326   0.0245 *
poverty       0.36018    0.19562   1.841   0.0720 .
single        3.07633    0.50409   6.103 2.03e-07 ***

             Estimate Std. Error t value P-value
(Intercept) -14.93045   10.28357  -1.452  0.15318
```

```
pctwhite     -0.19704    0.07267  -2.711  0.00933 **
poverty       0.29287    0.18964   1.544  0.12921
single        3.18362    0.50035   6.363 7.59e-08 ***

             Estimate Std. Error t value P-value
(Intercept) -14.00841   10.41329  -1.345  0.18487
pctwhite     -0.20223    0.07363  -2.746  0.00846 **
single        3.50949    0.46018   7.626 8.12e-10 ***
```

Each of these techniques may be performed manually. Many software packages allow you to perform these techniques automatically, however, I would strongly advise you to do things manually until the point that you are very familiar with the data you are analysing. In real data it is often the case that there is a choice over which term is eliminated next (see below, where we discuss collinearity), and a rational guide to the choice is almost invariable better than a mechanical one.

If one is attempting to build an *explanatory* model, convention seems to set a p-value of 0.05 as being the threshold for eliminating (or adding) variables. In the case of a *predictive* model, convention seems to set the threshold higher (often at about 0.1). One should remember that these are only conventions! There may be good reasons for you to set different thresholds (based on the consequences of particular types of error, for example).

One should also be cautious in the use of such techniques. Chucking a hundred possible explanatory variables into a backwards selection algorithm that ends up with a model with three explanatory terms is quite likely to end up with a model that has serious problems. This is a complex area, but roughly speaking, you should start with no more than $n/10$ to $n/20$ possible explanatory terms where n is your sample size. (For a detailed and sophisticated exposition of the problems with such techniques, see chapter four of Harrell's "Regression Modelling Strategies").

Another problem that you will run into if you use too many explanatory variables in a linear regression model is that of *overfitting.* Roughly speaking, the more variables you throw into a model, the more likely you are to improve the fit of the model. However, what you may end up doing is fitting your model to the sample and you fail to obtain a model that generalizes to the population from which the sample was drawn. The effect will be that predictions from your model are hopeless and you will also find that if you draw a new sample from your population and refit the model, you will get completely different coefficients.

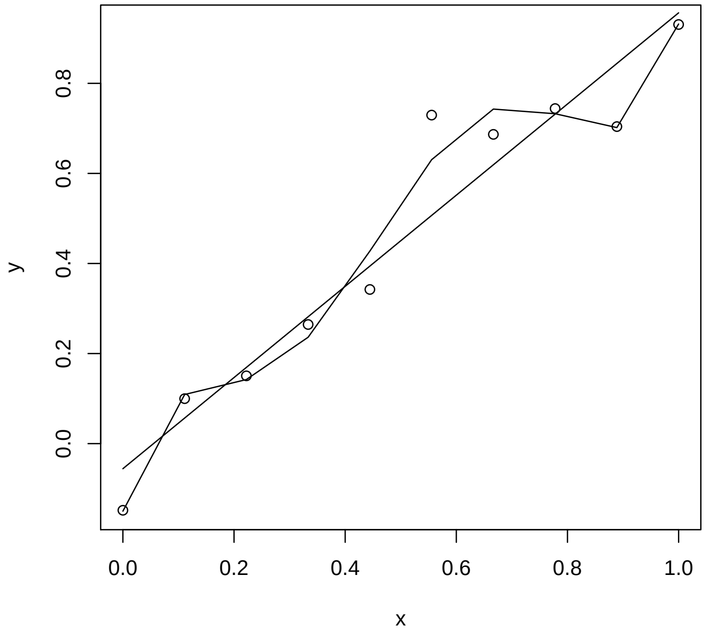

Figure 1: Which is the Better Model?

It is always wise to observe *Ockham's Razor* (also known as the *law of parsimony*): "entia non sunt multiplicanda praeter necessitatem" (or "entities should not be multiplied beyond necessity"). In other words: favour the simplest model that you can get away with. (William of Ockham was 14th century English logician most famous for his work in nominalism. It appears that he never actually stated the razor in the form we have it (although he did say "numquam ponenda est pluralitas sine necessitate") and that the principle was a commonplace at the time).

Problems in Interpretation

In the planning, execution, analysis and interpretation of a study it is important to bear in mind whether it is an *observational* study or an *experimental* study.

The important feature of an experimental study, from our point of view, is that it is possible for the experimenter to assign the values of the explanatory variables (for example treatments or exposures) to the experimental units. A typical experimental setup would involve randomizing patients to one of three groups. The first group might get treatment A, the second treatment B and the third treatment C. If the experiment has been correctly carried out and analysis has shown that there is a significant difference in outcomes between treatments A, B and C, then one may rationally infer that the different treatments have *caused* the difference in outcomes.

In contrast, in an observational study, the experimenter has no such control. One simply observes the values of the explanatory variables in the population under study. For example, in a study of people's weights one may be interested in the effect of height, gender and stress level on weight. One might be able to assign stress levels (fun for the experimenter, but there may be ethical problems!) but one simply cannot assign people's heights and genders.

There are a number of problems associated with this feature of observational studies that affect their interpretation:

- You are no longer able to rationally infer that statistically significant explanatory variables are *causes* of the observed response. All you can say is that there is an *association* between the explanatory variables and the outcome. Discovering whether the association is causal is, in general, hard and requires ingenious study design.
- The values of the explanatory variables you have chosen may themselves be causally determined by some variable(s) that you did not (or could not) measure.
- There may be associations between the explanatory variables themselves. This may cause problems in the analysis (collinearity).

It is also important in interpretation to be careful about the *Ecological Fallacy.* This is the mistake of attributing group level or averaged effects to individuals. Suppose a country, divided into provinces, has a population of either green people or orange people. Suppose that it is observed that the suicide rate is higher in provinces where the green:orange ratio is higher. Can we infer that green people are more likely to commit suicide? No. It may be that all the suicides are actually among orange people - and that they are more likely to commit suicide the more they are in a minority.

Further, one must be careful about *extrapolation* from a model. If one uses a model to make a prediction for the values of explanatory variables well within their range of observed values, one can be much more confident than if one makes a prediction well outside that range. The reason for this is that, a priori, one has no reason to suppose the same form of the model outside the range of observed explanatory variables as in.

Outliers and Influence

- An *Outlier* is a datapoint that does not fit the model. One way of spotting outliers is to inspect the residuals from a model and look for exceptionally large values.
- An *Influential* point is one whose removal from the dataset would cause a large change in the fit. In Stata you can get a measure of influence by calculating the *Cook's Distance* and looking for large values. You can also inspect the *dfbetas.* These show how much the estimated coefficients change if you omit each observation in turn.
- A point with high *leverage* is one that is unusual in explanatory variable space. In Stata you can get an idea of points with high leverage by calculating the "leverage values'.'

An influential point may or may not be an outlier and it may or may not have high leverage but it will tend to have at least one of these two properties.

One should always check to see whether there are outliers and/or influential points in our model.

As an outlier is basically a point that not consistent with our model for all the other points, it is a good idea to explain that fact or possibly even eliminate the point from our analysis. An outlier may be caused (for example) by

- A mistake in data measurement.
- A mistake in data entry.
- Something is wrong with your model.

Removal of an outlier from your dataset is a possible option, but you should be prepared to offer good scientific justification for doing so. In removing it, you may alter which coefficients are statistically significant in your model; you may even miss a new scientific discovery!

It's also a good idea to check for influential points. If they are having a big influence on your results, it's a good idea to make sure that you're confident in their reliability.

Collinearity

Collinearity occurs when there are (approximate) linear relationships between explanatory variables. (Sometimes you'll see the term *Collinearity* used when there's linear relationship between two explanatory variables and the term *Multicollinearity* when there is a linear relationship between three or more explanatory variables.) Collinearity can lead to a number of problems. For example:

- It makes parameters hard (or even impossible) to estimate and parameter estimates can be very sensitive to small changes in data values (so that rounding your data to a number of significant figures can change your parameter estimates!) A model with serious collinearity cannot be trusted.
- It makes interpretation hard. If one explanatory variable is a multiple of another, then both variables are attempting to do the work of one! How do you attribute effects between them in this case?
- It can mess up model building strategies. If collinearity is present then it can be a matter of chance which variable is eliminated from consideration at any stage.

For example, I've let the explanatory variable x take the values one through twenty and I've defined the response variable z to be x plus a random variable with mean zero and variance one. Then I've let y be equal to x. If I try to model z with both x and y, my software throws a fit. It cannot estimate the coefficient of y.

```
            Estimate Std. Error t value P-value
(Intercept) -0.05344    0.38697  -0.138    0.892
x            0.99264    0.03230  30.728   <2e-16 ***
y                 NA         NA      NA       NA
```

If I now add a little random noise to y, so that x and y are now approximately collinear, my software can estimate the coefficients of x and y but the estimates are clearly rubbish.

```
            Estimate Std. Error t value P-value
(Intercept)  -0.0491     0.3980  -0.123    0.903
x           -42.9197   186.7709  -0.230    0.821
y            43.9118   186.7684   0.235    0.817
```

In the exercises, you'll see an example of how collinearity can affect backwards elimination.

How do we detect collinearity? One of the first clues is that either estimates make no sense in the context of the problem or standard errors of estimates are highly inflated (look at the second example). Also, you may find that you see a high value of R^2 for your model, but none of the coefficient estimates for the explanatory variables are statistically significant. Fortunately, most statistical software provides collinearity diagnostics. In Stata, you simply use the `vif` command after you've performed the regression to see how variance estimates have been inflated by collinearity. Collinearity is reported by statistical software packages in a number of forms:

- Reporting the correlation between explanatory variables. A high correlation between explanatory variables should alert you of the possibility of collinearity.
- In terms of the *Condition Numbers* (these are the eigenvalues of $X^T X$, where X is the design matrix). If you spot a condition number greater than thirty, there's likely to be a problem with collinearity. In Stata, the command `collin` is available as an external command from various places (such as UCLA). Use `findit collin` if your version of Stata is web-enabled.
- In terms of *Variance Inflation Factors* (these show how the variance of the parameter estimates are being affected by collinearity). Again, large values indicate a problem.

Curing collinearity can be done in a number of ways. The simplest, if you detect it, is for you to decide, on the basis of the problem at hand, which of the collinear explanatory variables to eliminate.

Missing Data

One of the unfortunate problems that plagues real world data analysis is *Missing Data.* For any number of possible reasons you will often find that you have data records that are missing values for either some of the explanatory variables or for the response variable or for both. What should one do in these circumstances? The best answer is to not get into this situation in the first place, or if you have, do your best to recover complete observations! This may be the best answer, but unfortunately it's often an unhelpful answer; in practice there are a number of strategies.

- Remove data records from your data set if they have any missing data and analyze what's left. This is the traditional way of proceeding and you'll see this done a lot in the literature. Unfortunately it's a very poor way of proceeding because you are throwing away good data and therefore losing statistical power, but even worse, you will probably bias your results if you do so. The only circumstance under which you do not bias your results is if the missing data is *Missing Completely at Random (MCAR).* That is, the fact that any data value is missing is independent both of the actual missing value and of any observed value in the dataset. It is becoming less acceptable to publish an analysis that proceeds on this basis.

- Impute missing values with some reasonable guess of that value. For example, replace a missing covariate value with the mean of observed values for that covariate. Although this is an improvement, results will still tend to be biased.
- Impute the missing values using some model of the missingness based on the data you so have. We can do this if we assume that the missing value, although not MCAR, is *Missing at Random (MAR)*, that is, the fact that the value is missing is independent of its actual value, but may be dependent upon some or all of the observed values. If this is repeated, so that we obtain a number of plausible completions of our dataset, these can each be analyzed and the results combined in the technique of *Multiple Imputation.*
- If the data is neither MCAR or MAR, then the situation is much harder. In the first place, there is no statistical test for it (therefore you'll have to use subject knowledge to infer that it's occurring) and in the second place, to do anything about it you'll need some model of what's going on that is independent of the data you've gathered (so you may need a secondary study in order to proceed).

For data suitable for analysis by linear regression, the technique of multiple imputation is now considered standard and if you have missing data in such an analysis, you should seek expert help with how to proceed.

For the Future

In these two sessions we've been looking at multiple regression as a method of explaining the variation in data in terms of a fixed determinate effect plus a random effect. We've noted that in order to apply this model, we have had to make certain assumptions about our data. Indeed, inspecting residuals from a linear regression model makes up most of the hard work in developing a model. Linear regression models have the great advantage of simplicity (remember the principle of parsimony), but they cannot be used in all circumstances. It may be that after attempting to fit your data with a linear regression model, perhaps after trying a number of transforms of the explanatory variables and of the response variable, you simply cannot get satisfactory fit or diagnostics. It may be that you perceive that your data isn't really suitable for a linear regression model in the first place.

Fortunately, statisticians have developed a wide range of modelling techniques beyond linear regression that can be seen as building on the principles of linear regression but which involve successive relaxations of the conditions applying to linear regression. You will look at one or two of these later in this course, but here I'll give a quick overview of the kind of tools that are available.

- If it really makes little sense for your response variable to be modelled with a normal random error, then the class of *Generalized Linear Models* is available. In this class you find models suitable for count data (poisson regression) and for proportions (logistic regression) as well as a number of other response distributions.

- If your residuals are not independent of each other (in other words, your responses are not independent of one another, for example in repeated measurements on the same subject), then it's possible to model the covariance between observations using *Random Effects Models*. These are often built in to *Linear Mixed Effects Models*, that are analogous to linear regression.
- If your residuals are both not normal and not independent, then you can use the class of *Generalized Linear Mixed Models*.
- If it's not clear what the functional form of your covariates should be, then you can use *Generalized Additive Models*, for independent observations or *Generalized Additive Mixed Models* for dependent data. These techniques model appropriate functional forms for the covariates using general curve fitting techniques.