Biostatistics in Research Practice - Regression II Practical

Simon Crouch

29th January 2008

1 Categorical Explanatory Variables

Load the dataset `data02` which we used last week to investigate interactions.

- Repeat the regression with interaction that we did last week, with the commands
  - `generate xgp = x*gp`
  - `regress y x gp xgp`
- Now compare the result with the result you get if you use the incantation
  - `xi: regress y i.gp*x`

Now load the datset `data07`. This contains artificial data that represents blood pressure measurements `bp` on subjects in three weight categories `weight` under different lengths of exercise time `extime`.

- Do a scatter plot of blood pressure versus exercise time. What do you notice?
- Perform a linear regression with `regress bp weight extime`. Do you think that this is the right thing to do?
- Now do `xi: regress bp i.weight extime`. Look at the results: can you interpret them? Is this a more appropriate model?
- Harder: is there any evidence that there should be an interaction between exercise time and weight category in the model?

For the last question, you'll need to do a regression with `xi: regress bp i.weight*extime`. What do you notice about the p-values for the interaction terms? Do you think that you can use the p-values to decide whether the interaction term should be in the model?

In fact, we can bundle up the multiple terms that are being tested using a likelihood ratio test that compares the model with interaction with the model without interaction. To do this, we need to `store` the model results.

- `xi: regress bp i.weight extime` : first regression
- `est store A` : store the results in A.

- `xi: regress bp i.weight*extime` : second regression
- `est store B` : store the results in B.
- `lrtest A B` : likelihood ratio test.

2 Backwards Elimination

Load the dataset `crime` into Stata. This is real data about crime rates in the fifty one states of the USA from a particular year. List the data and look at the variables that you have available. `sid` is a numerical state identifier and `state` identifies the state using a standard letter code. `pct` mean "percentage" and `hs` stands for "high school". All the other names should be self explanatory.

- Now inspect the data itself (look at the raw numbers and plot histograms, for example). Does anything stand out at the moment about any of the variables?
- Now build a linear regression model using manual backwards elimination considering the murder rate `murder` as the response variable. First of all, consider which explanatory variables it makes sense to include in the starting model. Then perform the backwards elimination. Don't worry about residuals or other diagnostics at this stage. Use $p = 0.05$ as a criterion for elimination.
- Interpret your model. Do you think that the ecological fallacy might be relevant to your interpretation?
- Use the `vif` command to check for evidence of collinearity in your model.
- Did you include `crime` as an explanatory variable in your starting model? If you did, take it out and repeat the backwards elimination. If you didn't, repeat the backwards elimination with it in. In either case, do you notice a difference in the final model that you arrive at?

3 Collinearity

Load the dataset `data06` into SPSS. This is artificially generated data that is designed to show how unpleasant collinearity can be. In this dataset z is the outcome variable and x, y and t are explanatory variables.

- Perform a regression with x, y and t as explanatory variables. What do you notice about the coefficient p-values?
- Now remove t as an explanatory variable, so that x and y are left in. What do you notice about the coefficient p-values?
- Now put t back as an explanatory variable and remove x (so now you have y and t as the explanatory variables). What do you notice about the coefficient p-values?

What's going on? (Hint: use the corr command to look at correlations between the explanatory variables and use vif to calculate the variance inflation factors with x, y and t as explanatory variables and then with only x and y as explanatory variables).

4 More Collinearity

Now load the bodyfat dataset. This is real data relating body fat measurements to various body measurements. Perform a linear regression using bodyfat as the outcome and all the other variables as explanatory variables. (Don't do any variable elimination at this stage). What do you conclude? Now look at correlations between the explanatory variables and use vif to calculate variance inflation factors. Are you happy with the model you've got? Try removing the explanatory variable triceps and repeating the analysis. What do you conclude?

5 Outliers and Influence

Now return to the crime dataset. Redo the regression with the explanatory variables from the final model from the first question (after starting without crime amongst the explanatory variables), but this time calculate the predicted values, the standardized residuals, the Cook's distance and the leverage values. (Use the predict function, with the xb, rstandard, cooksd and leverage options, respectively. For example, predict lev, leverage.)

- Plot the residuals against fitted values. What do you think?
- Plot the Cook's distance against sid. What do you notice?
- Plot the leverage values against sid. What do you notice?

What do you conclude from these plots? What do you think is the best way to proceed? Proceed as you think fit. Arrive at a final model, justify your choice and interpret the model. Use the cprplot command to check the functional form of the explanatory variables in your model.

Hint: based on your plots you may wish to eliminate one of the data points and return to the start of the analysis to determine a new model. Use all the diagnostic techniques on the final model from this stage to determine whether you are satisfied with what you've arrived at. To remove a datapoint, use the drop command. For example, for Washington DC, use drop in 51.