Biostatistics in Research Practice

Regression II

Simon Crouch

University of York

29th January 2008

Reprise I

- We have observations $Y_1, Y_2, \ldots, Y_n$ of the values of some outcome variable of interest. For example, Y_i might be the weight of the ith person in a group.
- We also have the values of explanatory variables $X_1, X_2, \ldots, X_p$ for each observation Y_i. For example, p might be 2, with $X_{1,i}$ being the ith person's height and $X_{2,i}$ being the ith person's age.
- We want to model, or explain, the variation in the Y_i by using the values of the explanatory variables.

Reprise II

- We want the explanation to consist of a fixed, determinate bit that depends on the values of the explanatory variables plus a residual random bit ϵ_i for each observation.
- We want the determinate bit to be nice and simple, a linear combination for each observation Y_i

$$X_{1,i}\beta_1 + \ldots + X_{p,i}\beta_p$$

- We want the random bit ϵ_i to be normal, with zero mean and the same variance for each i, with each ϵ_i independent of the others.

Reprise III

Fitting a multilinear regression model simply means

- Finding the values of the $\beta_1, \ldots, \beta_p$ that minimizes the value of $\epsilon_1^2 + \ldots + \epsilon_n^2$.
- Checking that the residuals ϵ_i behave themselves.

Reprise IV

How did we check the residuals?

- Q-Q plots to check normality.
- Residual versus Fitted to check homoscedasticity.
- Residual versus Fitted (or Partial Residual Plots) to check functional form of explanatory variables.

Now we need to work out how to build a *good* model!

Categorical Predictors I

- Instead of *weight* measured in kilograms, *weight_category* with values *heavy*, *medium* and *light*.
- Can code these values as 0, 1 and 2. These are NOT true numerical values.
- Define *dummy variables*: *weightcategory_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.
- For a *medium* weight subject, *weight_category_1=0*, *weight_category_2=1*, *weight_category_3=0*.

Categorical Predictors II

- Choose one category as baseline.
- Include the other dummy variables in the regression.
- Coefficients represent the contribution each category makes relative to the baseline category.
- Testing for statistical significance becomes more complicated.

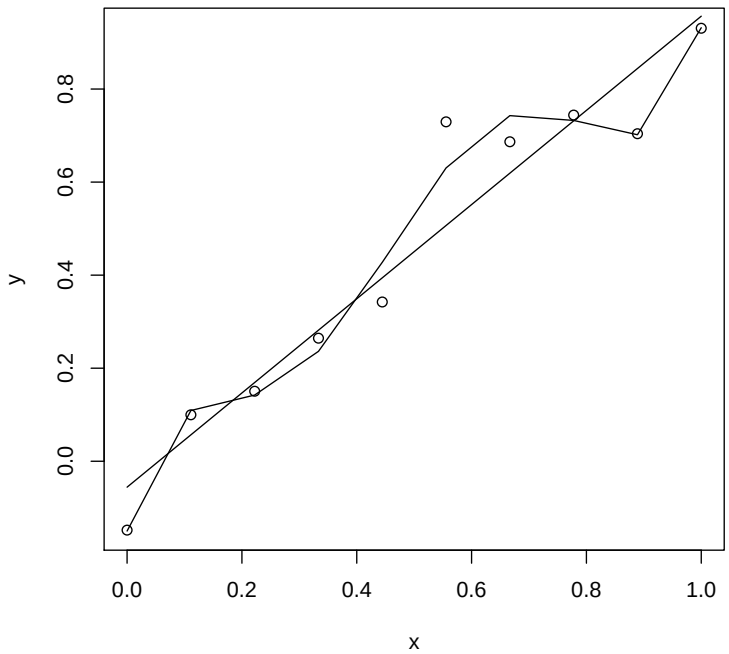
Variable Selection

- 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 a variable may be added but removed later and vice versa.

Variable Selection

- This can be done automatically, but best to do by hand.
- Suggest $p = 0.05$ elimination threshold for explanatory models.
- Suggest $p = 0.1$ elimination threshold for predictive models.
- Be careful not to overuse this technique.

Overfitting



Ockham's Razor (also known as the "law of parsimony"):
"entia non sunt multiplicanda praeter necessitatem"

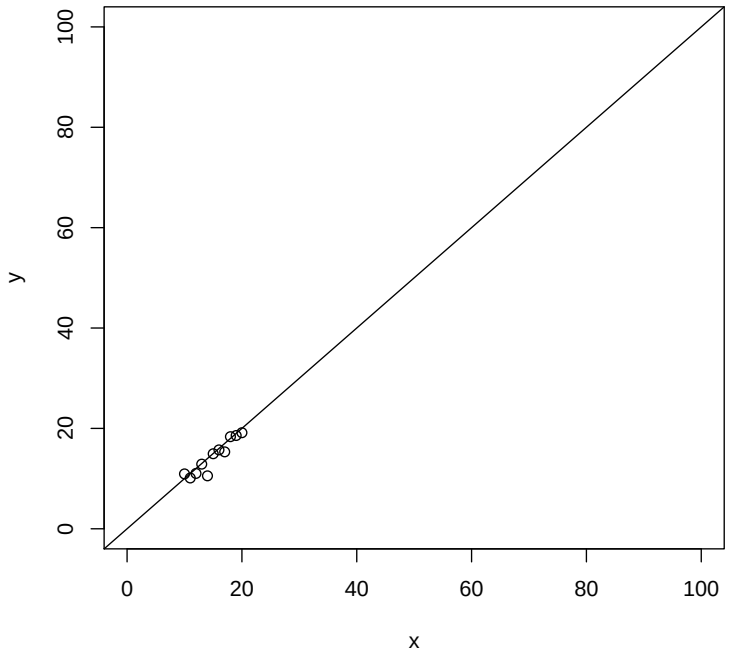
Experimental versus Observational Studies

- Experimental study
  - Control over the explanatory variables.
  - Causal Inference.
- Observational Study
  - No control over the explanatory variables.
  - Inference about Association.

The Ecological Fallacy

Beware the *Ecological Fallacy*. This is the mistake of attributing group level or averaged effects to individuals.

Extrapolation



Would it be reasonable to make a prediction for $x = 80$ based on this model?

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 SPSS, use the "Cook's distance" or the "dfbetas".
- A point with high *leverage* is one that is unusual in explanatory variable space. In SPSS you can get an idea of points with high leverage by saving the "leverage values"

Collinearity

Collinearity occurs when there are (approximate) linear relationships between 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. A model with serious collinearity cannot be trusted.
- It makes interpretation hard.
- It can mess up model building strategies.

The Effects of Collinearity

Explanatory	Estimate	Std. Error	t value	p-value
Intercept	−0.0534	0.387	−0.138	0.892
x	0.993	0.0323	30.7	tiny
y	NA	NA	NA	NA

Explanatory	Estimate	Std. Error	t value	p-value
Intercept	−0.0491	0.398	−0.123	0.903
x	−42.9	186.8	−.230	0.821
y	−43.9	186.8	0.235	0.817

Collinearity

How do we spot collinearity? Informally,

- Estimates that make no sense.
- High standard errors for estimates.
- Large R^2 but no significant explanatory variables.

Collinearity Diagnostics

How do we spot collinearity? In Stata we get some collinearity diagnostics to help.

- Correlation between explanatory variables.
- Condition Numbers (if collin available).
- Variance inflation factors.

Missing Data

- Remove data records from your data set if they have any missing data and analyze what's left. Only valid if the missing data is *Missing Completely at Random (MCAR)*.
- Impute missing values with some reasonable guess of that value.
- Impute the missing values using some model of the missingness. Valid if the missing data is *Missing at Random (MAR)*. We can then use *Multiple Imputation*.
- If the data is neither MCAR or MAR, then the situation is much harder.

For the Future

- If the response has non-normal residuals, use *Generalized Linear Models*.
- If residuals are not independent, 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.

Contact Details

- Simon Crouch, Epidemiology and Genetics Unit, Department of Health Sciences.
- simon.crouch@egu.york.ac.uk
- xt. 1938
- Room A/TB/113, Seehbohm Rowntree Building, Area 3.