Statistical Methods
Principles

Model Checking

Dr Eleni Matechou

matechou@stats.ox.ac.uk

References:

- F.L. Ramsey and D.W. Schafer "The Statistical Sleuth"
- A.C. Davison "Statistical Models"
- J.J. Faraway "Linear Models with R"
- J.J. Faraway "Extending the Linear Model with R"
- A Gelman, J.B. Carlin, H.S. Stern and D.B. Rubin "Bayesian Data Analysis"

Assumptions

Model inference, prediction, selection etc. usually rely on certain **assumptions**.

When the assumptions are violated the results can be seriously flawed.

Understanding, and **checking**, the model assumptions is vital for any valid analysis.

Clearly state any assumptions you make

For example, you have learned that in the normal linear model we assume that the errors are i.i.d. N(0, σ^2) and that the model is *correct* i.e. all the necessary variables have been included.

Will the assumptions hold **exactly**?

Probably not. Many distributional results rely on **asymptotics** which means they hold for large sample sizes.

Even if the asymptotics hold, real data will not be exactly how the assumptions state.

How *much* violation is acceptable?

> "We do not like to ask:
> "Is the model true or false?" since probability models in most data analyses will not be perfectly true....
>
> The more relevant question is:
> "Do the model's deficiencies have a noticeable effect on the substantive inferences?". Gelman et al. chapter 6

Diagnostics

How do we check if the assumptions of the model hold?

We perform **diagnostic checks**.

> "We may divide diagnostic methods into two types.
>
> - Some methods are designed to detect single case or small groups of cases that do not fit the pattern of the rest of the data. Outlier detection is an example of this.
> - Other methods are designed to check the assumptions of the model, such as the choice and transformation of the predictors, and those that check the stochastic part of the model, such as the nature of the variance about the mean response".
>
> Faraway (2) section 6.4.

Goodness-of-fit

Does the model fit well? This can be difficult to assess.

Comparing the observed to the **fitted** values should give an indication.

> **"If the model fits, then replicated data generated under the model should look similar to observed data."**
> Gelman et al. chapter 6

For certain data types, eg. contingency tables or binomial data, there exist goodness-of-fit tests, such as the residual deviance.

However, using these tests can only tell you if the model fits well or not and cannot suggest ways to improve the fit, something which is possible using, less formal but often more revealing, diagnostic figures.

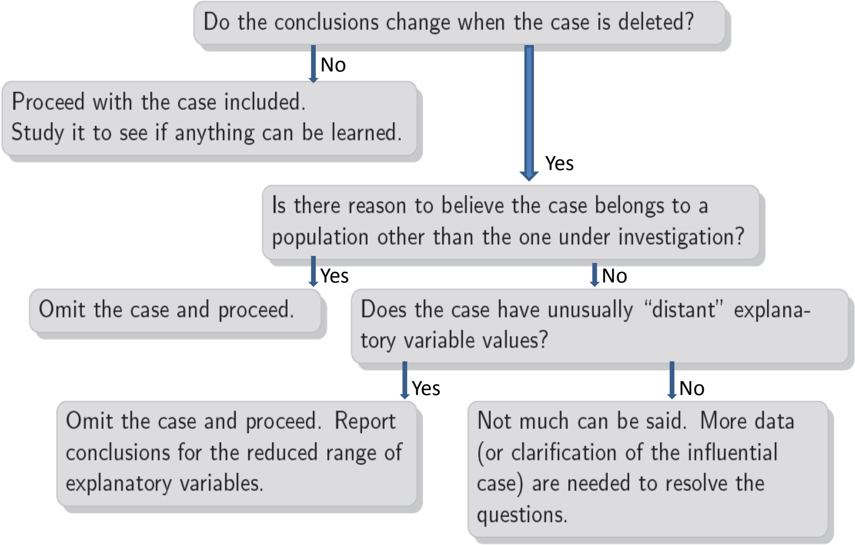
Influential observations

Are there observations which control/**influence** the fit more than we would like to?

This could lead to erroneous results which are driven by one or a small group of observations.

How much do our conclusions change if these observations are removed?

Influential points/outliers can mask other influential points/outliers, which is why leave-one-out methods do not always spot the problems.



Do the conclusions change when the case is deleted?

Proceed with the case included.
Study it to see if anything can be learned.

Is there reason to believe the case belongs to a population other than the one under investigation?

Omit the case and proceed.

Does the case have unusually “distant” explanatory variable values?

Omit the case and proceed. Report conclusions for the reduced range of explanatory variables.

Not much can be said. More data (or clarification of the influential case) are needed to resolve the questions.

Added variable plots

Added variable plots reduce the higher-dimensional regression problem to a series of two-dimensional plots and show leverage and influence of the observations on each coefficient of the model.

They can also indicate whether a variable should be added to the model, **after** the other variables have been added.

However, they can prove misleading when diagnosing other sorts of problems, such as nonlinearity.

Outliers

Are there observations that are not fitted by the model well?

An observation can be outlying for one model but not for another.

"There are two ways to deal with excessively influential observations:

- one is to use procedures that are robust/resistant to these observations
- the other is to examine them closely to see whether they are indeed influential, why they are influential and whether they provide some interesting extra information about the process under study."

Ramsey and Schafer chapter 11.

Cycles of model-fitting

Detailed model-fitting should be performed after the model assumptions and influential/outlying observations have been considered.

> “Often unexpected discrepancies between a fitted model and data will lead to further thought, and then to more cycles of model-fitting, checking and interpretation, iterated until a broadly satisfactory model has been found”.
> Davison section 8.7.

Transformations

Would variations of the model improve the fit?

There are cases where transforming the variables leads to a better fitting model which complies with the assumptions.

Examples include the log, square root, square transformations etc.

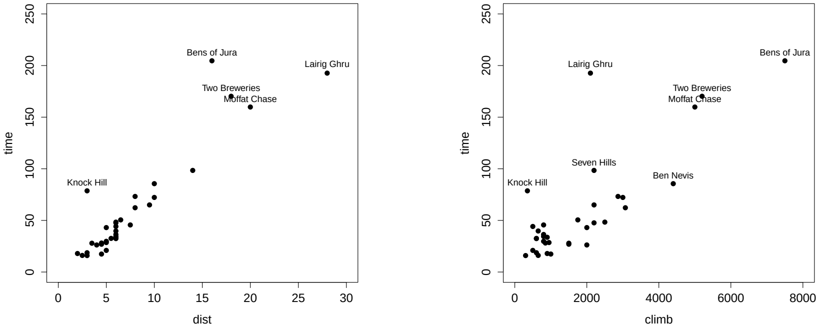
If several transformations result in a similar fit, then the transformation which makes interpretation of the results more straightforward should be preferred.

Example

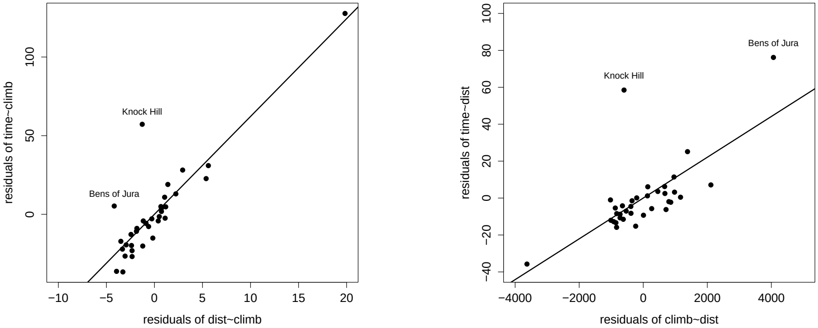
The *hills* data set in *library(MASS)* in **R**.

The response variable is the time it took to complete the race, in minutes, and the two potential explanatory variables are the total height gained during the route, in feet, and the distance on the map, in miles.

Does a linear model make sense?



Added variable plots



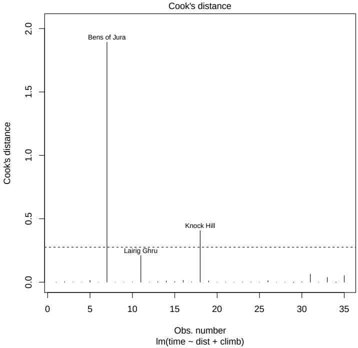
The slopes of these two simple linear regressions are equal to the coefficients in the multiple linear regression for the corresponding predictor variables.

The observations with $h_i > 2 \cdot (3/35)$ are:

```
 Bens of Jura   Lairig Ghru Two Breweries  Moffat Chase
        0.42          0.69           0.17          0.19
```

The observations with studentised residuals> 3 are:

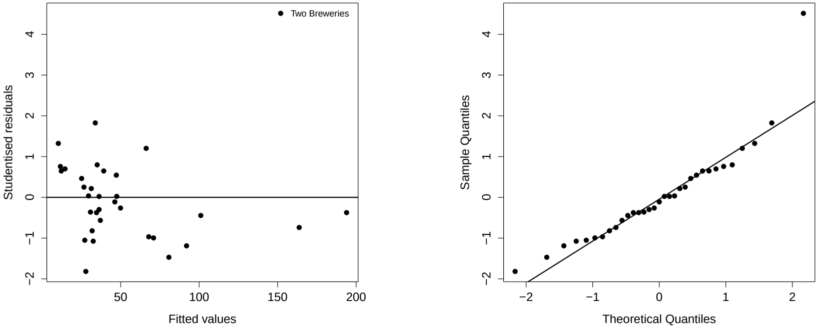
```
Bens of Jura   Knock Hill
        3.17         7.61
```



Bens of Jura is highly influential, much more than Knock Hill although the latter was further from the fitted line. Therefore, although Knock Hill is an outlier, it does not have the ability of Bens of Jura to pull the line towards itself.

Both these observations are removed and the model is refitted. **However**, this is done for demonstration purposes only and great care should be taken when data points are removed from the analysis.

Testing model assumptions



Two Breweries has appeared now as a possible outlier!
Is this observation influential? Check it on your own.