Model selection and diagnostics

todos

1. Recap

2. Apocrita:

https://docs.hpc.qmul.ac.uk/intro/login/

3. Crick Institute seminar, next Thurs, 1600-1700:

Mike Levine, Princeton

Model criticism and choice

1. Model selection
2. Assumptions of GLMs
3. Independence
4. Testing the other assumptions
5. Solutions to model problems

Principles for Model Choice

If we have multiple explanatory variables, we often need to make choices about which models to investigate, and which models to present in papers, and to rely on in reaching conclusions, in estimation etc.

‘There are three main principles of model choice:

economy of variables

multiplicity of p-values

preserve marginality

Economy of variables

"the simpler the better":
You should pick the simplest model that adequately explains the response

Models should have as few parameters as possible

Simple relationships between variables are preferred (linear, no interaction)

Models should be **simplified** to only include terms that are necessary. This is called a

Minimum adequate model

Multiplicity of *p*-values

We have already seen that "p = 0.05" is not a magic wand.

It simply means $Pr(x \geq x_{crit} \mid H_0)$

*B*ut H_o may or may not be true. Worse than this, multiple tests compound this as:

$P_{type\text{-}1} = 1\text{-}(1\text{-}a)^k$ for k tests at level a

Marginality

Marginality is about making sure that your tests are invariant to linear transformations of the data: e.g. If you halve all of the observations, the F-ratio will be identical. Don't worry too much about this, but preserving marginality means that:

main effects must precede interactions that involve the term in the model formula

If an interaction is significant, the main effects involved are important and should be retained, irrespective of significance

Do not test main effects for a term with a SS adjusted for an interaction involving the term

Options for selecting a model

Automated model choice: fit a model with all terms and let a software algorithm do the model selection for you

Start with a model with all terms and remove non-significant terms yourself

“All subsets”: analyse all possible variants and choose one

“Enlightened” model choice – use your knowledge of the system to select a small number of candidate models and select the best one.

Multi-model inference - fit a selection of possible models and calculate coefficients based on all models, weighted by how likely they are

Automated model Choice

A popular choice for choosing models, particularly in multiple regression (i.e. Multiple continuous explanatory variables, no categorical variables) is automated model choice, particularly **stepwise regression**

In FORWARDS stepwise regression, we start with single-factor models, and at each step add the variable with the highest F-ratio or lowest p-value. This process continues until no factors are sufficiently informative to pass a F-ratio/p-value cut-off. This is then the preferred model

In BACKWARDS stepwise regression, we start with all factors in the model, and remove factors with the highest p-values/lowest F-ratios in turn.

Problems with automated model choice

Focuses on purely STATISTICAL issues in model choice. Often scientific/practical issues are just as, if not more, important

Different approaches give different ‘best models’ – subjective choice between them, may not return best fitting model in any statistical sense

Lots of p-values: could include factors by chance

(I think) most statisticians would advise against these automated methods

Non-automated model choice

Has the same problems as automated procedures if you just follow a rule

Allows subjective decisions about whether to keep important terms in the model if you think while you do it

Allows model simplification by collapsing multiple levels of factors when they don’t have an effect

All subsets analysis

Can be useful when there are a few terms in the model

When there are lots of terms can become very unwieldy

“Enlightened” choice

Avoids lots of the problems associated with stepwise model selection

Can lead to unimportant terms being included

Favoured by lots of statisticians

Only really useful when you have a good idea of what’s going on in the system.

Which to use?

No simple answer.

When you know a reasonable amount about what is likely to be important and what isn't then use the approach where you select a few candidate models and compare them

When you are doing more exploratory analysis then use a stepwise approach

NB the above is only my opinion

p-values vs AIC

Two common approaches to comparing the goodness of fit between models

Compare two models using an F-test, a likelihood-ratio test or similar. If one model gives a significant reduction in explanatory power discard it.

Compare models using information-theory criteria such as AIC (Akaike Information Criterion).

p-values

To decide whether to retain a term in a GLM, fit a model containing the term and then carry out a “deletion test”

remove the term in question from the model and compare the goodness of fit of models with and without the term using a partial F-test

In R you can do this using the drop1() function. This will only give results for terms whose removal doesn’t violate marginality (i.e. it won’t give you a result for a main effect when the term is also present in an interaction term).

For an “ordinary” GLM specify an F-test to compare models e.g. drop1(mymodel,test="F")

AIC

AIC is defined as:

$$2k - 2\log(L)$$

Where k is the number of explanatory variables and L is the maximised likelihood of the model

The preferred model is the one with the lowest AIC value

As a rule of thumb, if two models have AIC values that differ by 2 or less then we cannot distinguish between them: they are both equally good descriptors of the data

Assumptions of the GLM

$$height_i = \alpha + \beta.rainfall_i + \gamma.elevation_i + \varepsilon$$

In each case, ε is independently sampled from a normal distribution with mean 0 and standard deviation *s* (estimated from the data) for every *y*

This formula tells us everything we need to know about the assumptions of GLM:

the error for each data point is independent

the standard deviation of the error is the same in every case

The errors are normally distributed

The relationship between covariates and response is linear and additive

Assumptions of GLM

We will discuss each of these four assumptions in turn:

Independence

Normality of error

Homogeneity of variance

Linearity/additivity

Independence

In many ways, independence is the most fundamental of the GLM assumptions

Formal definition: Data points are *independent* if knowing the error of one datapoint tells us nothing about the error of any other datapoint

Independence is a matter of experimental design

Effects of many forms of non-independence on tests are (a) unpredictable and (b) unfixable!

Normality of Error

There are two main approaches to testing for normality:

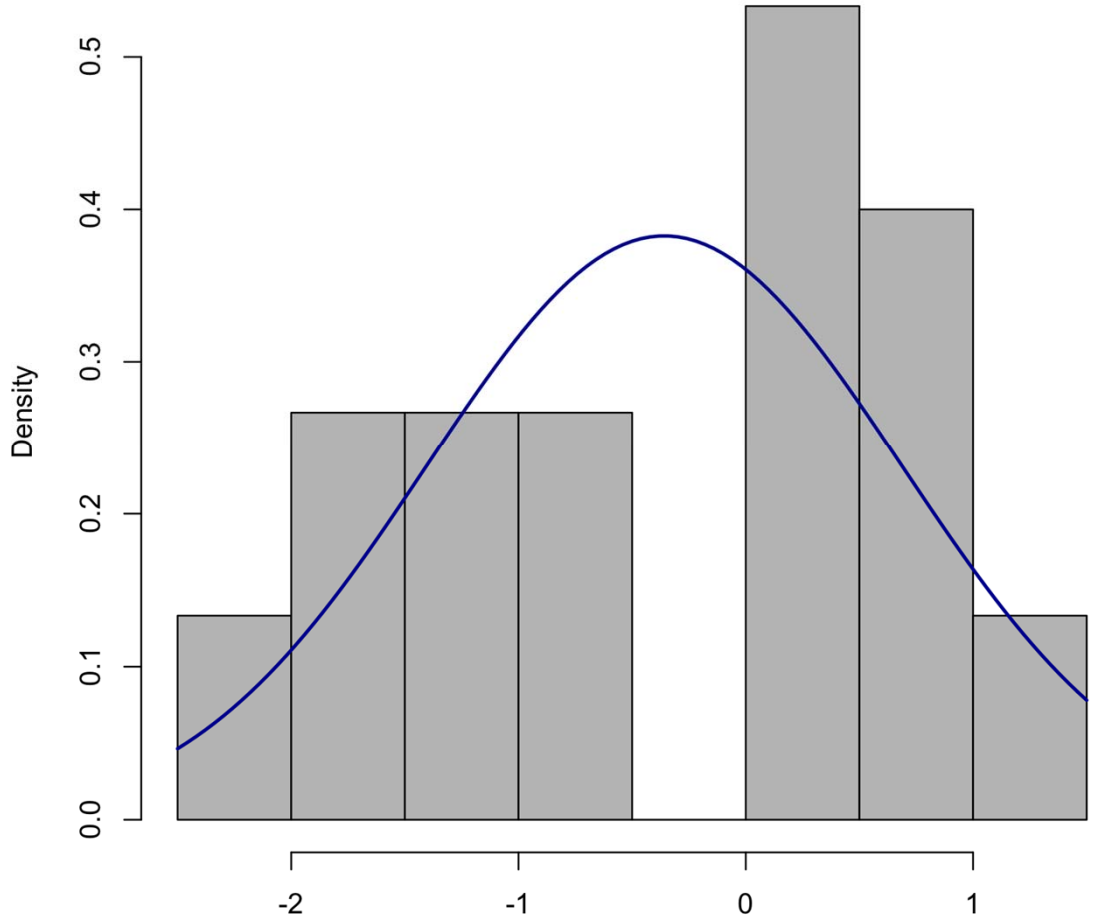
Graphical approaches:

Histogram of residuals
Normal quantile-quantile plot

Formal tests for normality.

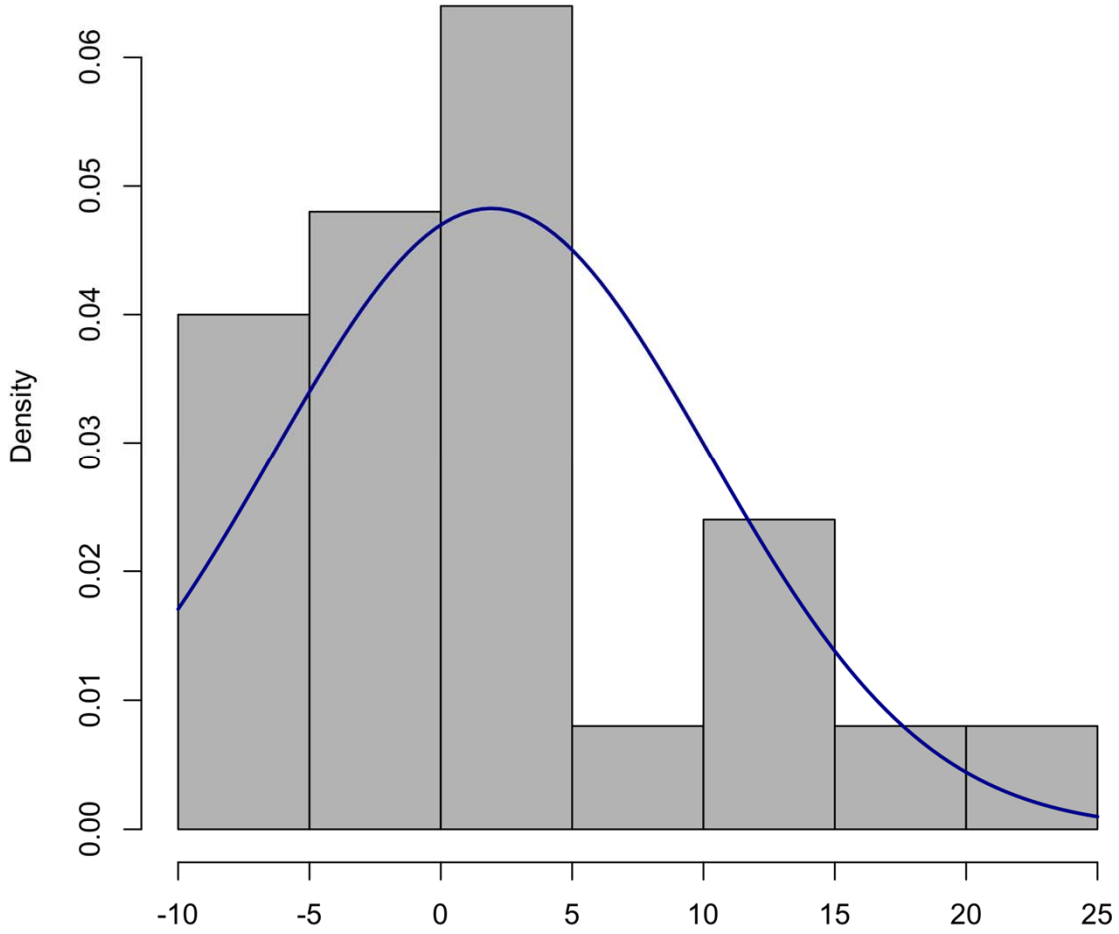
Normality of Error

Is this something to be worried about?



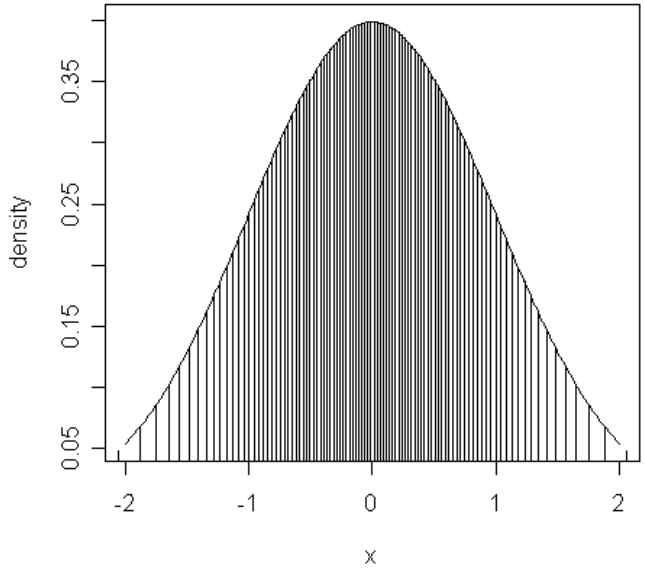
Normality of Error

Is this something to be worried about?



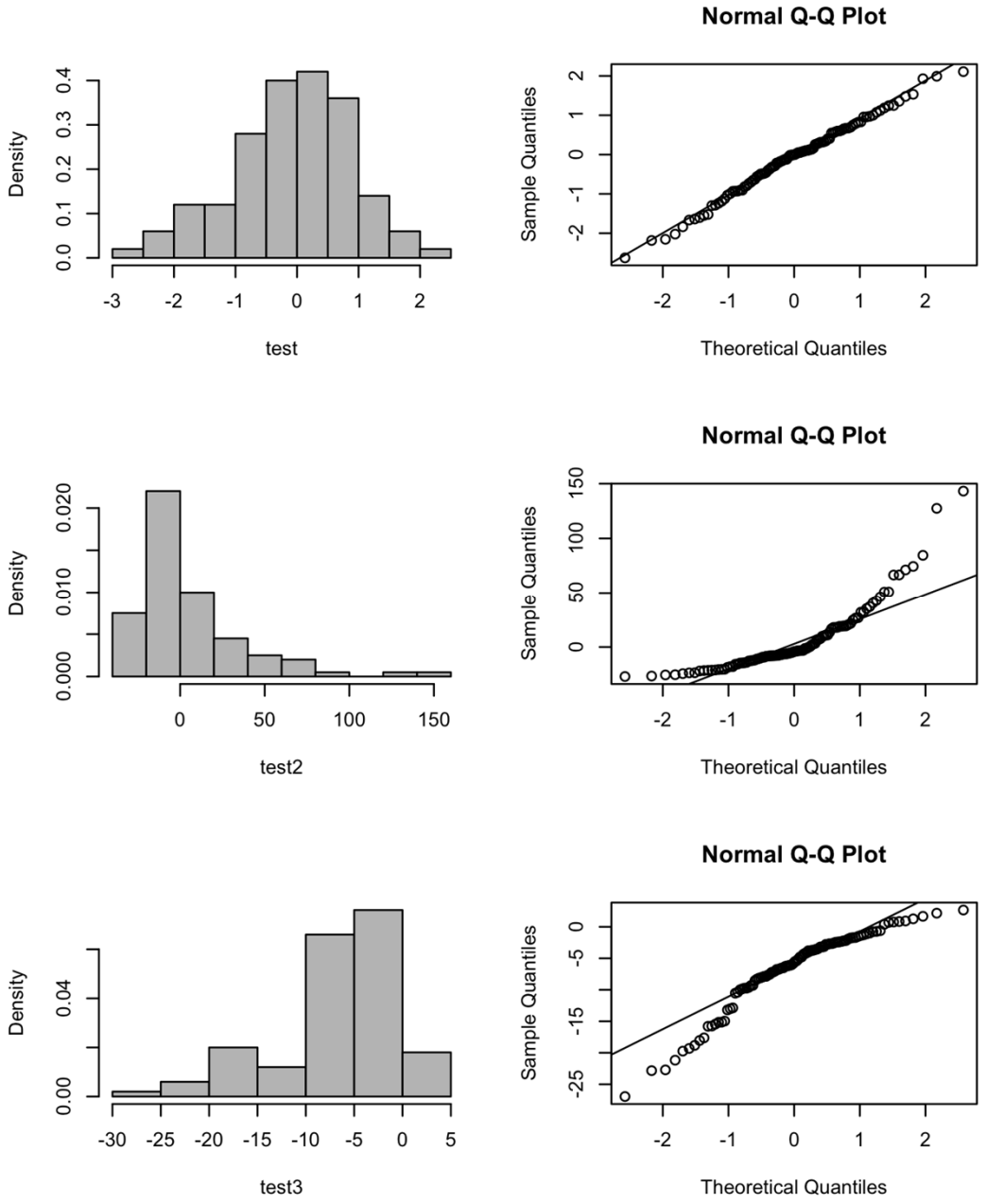
Q-Q plot

For a more sensitive look at the distribution of residuals, we plot our standardised residuals in ascending order along one axis, with the appropriate quantiles of a normal distribution on the other axis



If the data follows a normal distribution, the resulting plot should be a straight line

Normality of Error



Formal tests for normality

A number of tests exist for whether a set of residuals significantly departs from a normal distribution. The most common is the Shapiro-Wilks test.

There are a variety of these:

Anderson-Darling test
Ryan-Joiner test
Kolmogorov-Smirnov test

Formal tests for normality

```
> test4<-rnorm(100,10,3)
> ks.test(test4,"pnorm",mean(test4),sd(test4))

    One-sample Kolmogorov-Smirnov test

data:  test4
D = 0.0465, p-value = 0.982
alternative hypothesis: two-sided
```

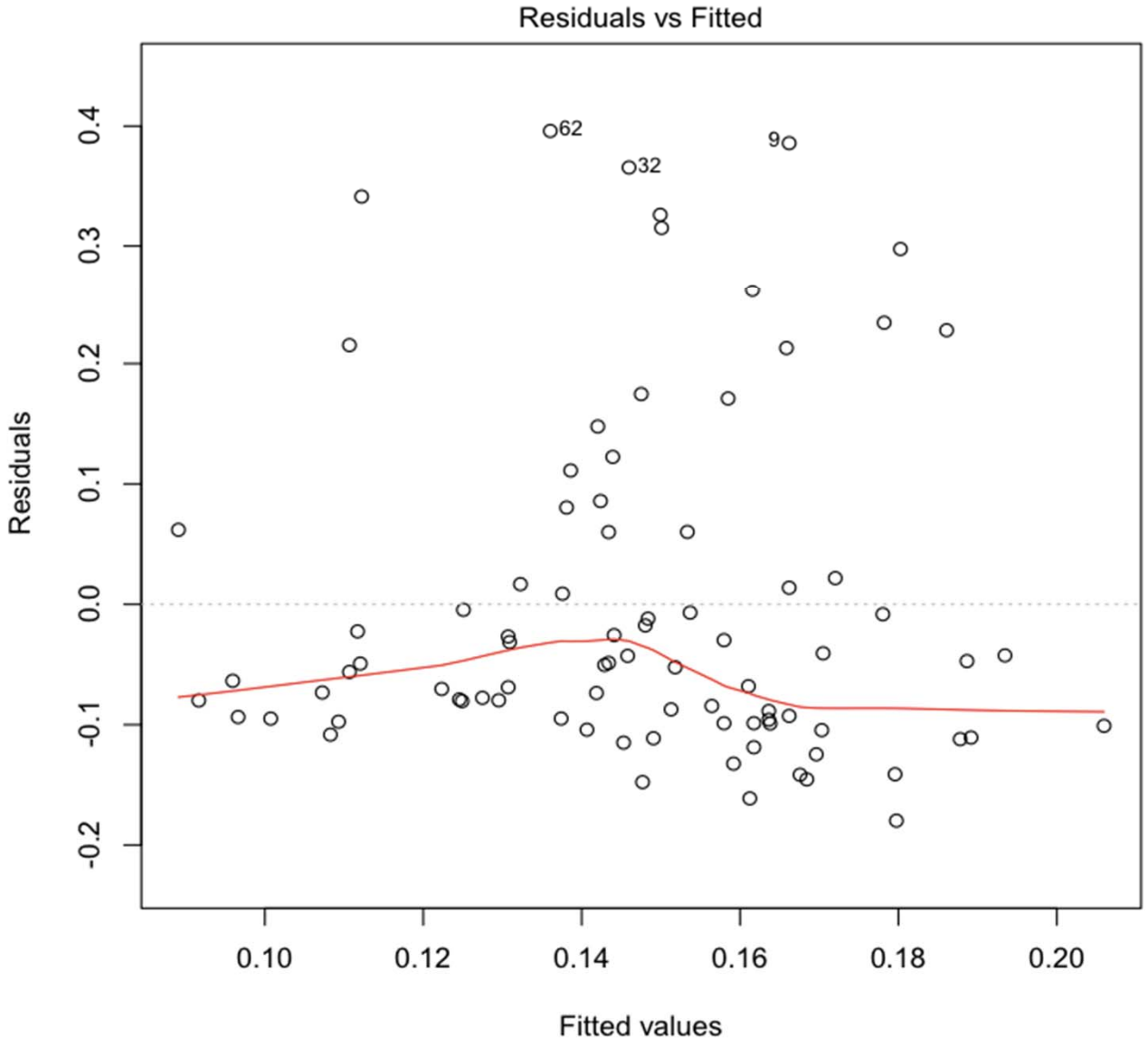
Homogeneity of Variance

This also has a fancy statistical name – homoscedasticity

A different graphical approach can help diagnose heteroscedasticity. Plotting residuals (or standardised residuals) against fitted values.

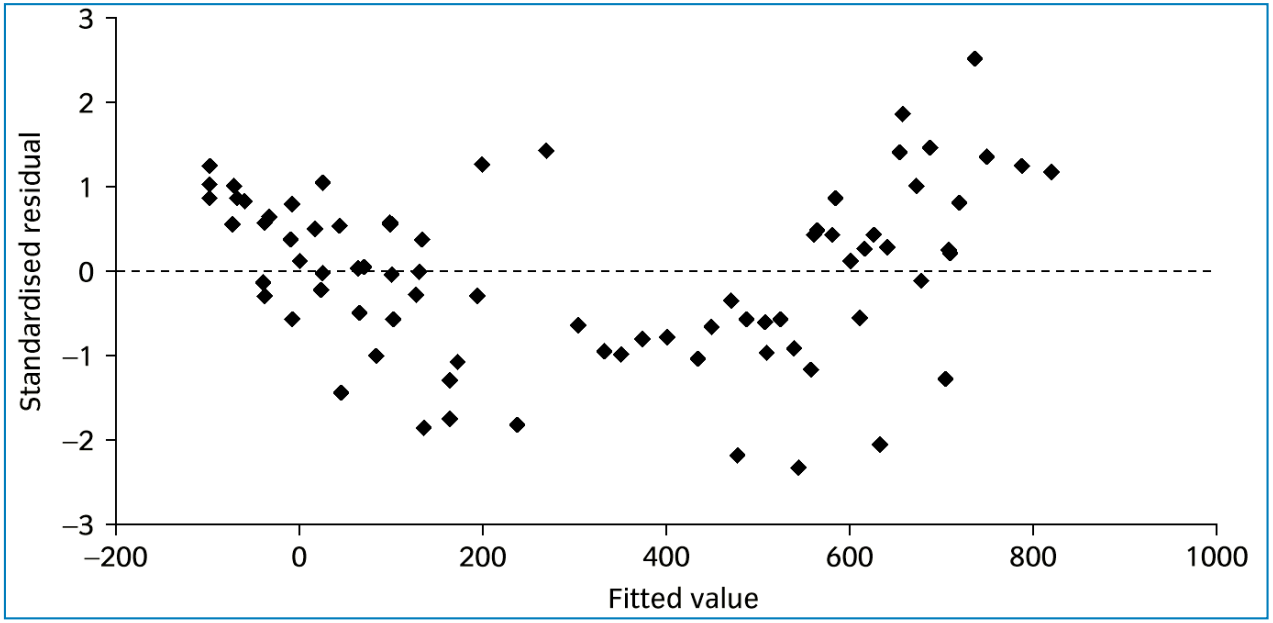
Also formal tests for homogeneity of variance for many circumstances

Homogeneity of Variance



Linearity

Plotting residuals (or standardised residuals) against fitted values can also help with diagnosing departures from linearity. If the points on this graph seem curved, then there may be a non-linear relationship between one or more explanatory variable and the response:



Formal tests for heteroscedasticity

You have already used a test for homoscedasticity that applies in this simplest case, perhaps without realising it: the F-test.

Recall that the F-test in an ANOVA table is testing whether the variance explained by part of the model (MS for a particular factor) is greater than that due to error (error MS). This is testing whether two variances are significantly different, and we can use the same test to compare variances between groups in our data.

Formal tests for heteroscedasticity

Other tests generalise this to multiple levels of a factor, and are thought to be more sensitive, more robust (e.g. to lack of normality), or both:

Levene's test

Bartlett's test

Brown-Forsythe test

Formal tests vs. plots of residuals

Statistics texts (and, indeed, statisticians) are somewhat divided over whether graphical approaches or formal statistical tests are to be preferred, but I think most do **not** recommend formal hypothesis tests in this context.

This seems counter-intuitive – surely an objective and powerful test is preferable to assessing things ‘by eye’?

Problems with formal tests

What is the meaning of a non-significant p-value? p-value is: p(data |null is true).

A high (insignificant) p-value doesn't mean the null hypothesis is, or even is likely to be true.. e.g. depends on power, and so on sample size.

A significant p-value just tells you that the residuals **are** non-normal/heteroscedastic. Not how large the departure from assumptions is.

Certainly not how important the departure is, in terms of how much it will effect the statistical test you are using

Problems with formal tests

Formal tests are very powerful at detecting departures from normality/ homoscedasticity

but GLM methods are quite robust to departures from these assumptions – particularly to non-normality.

Depends on the distribution – very asymmetric distributions can be problematic

Depends on how balanced the design is

Depends on size of effect

Depends on sample size: for sample size >30, GLM is pretty robust. Of course a big sample size means a test ismore likely to reject normality!

Always look at the plots of residuals

Even if you decide to rely on formal tests of assumptions, you should get in the habit of **always** looking diagnostic plots of residuals.

R will produce them for you nicely

To reiterate:

For every model you fit, look at:

histogram or qqplot of residuals
Residuals vs fitted values

Solutions to problems

- Transform variables
- Use tests with different, appropriate assumptions (next lecture)
- Use robust (often non-parametric) methods
- (for non-linearity) use models with polynomial terms

Solving non-linearity

Possible options:

Log transform the response variable

Introduce a polynomial term into the model

Introduce an appropriate interaction term into the model

Generalised additive models

Measure more explanatory variables

GLM is a (fairly) robust method

One of the most important things to bear in mind is:

While its good statistical practice to perform model criticism, it's too easy to push the 'panic button' and beat your badly behaved data with transformations, or reach the most exotic weapon in your statistical armoury.

GLM is a (fairly) robust method

GLM (and so ANOVA, regression) has been demonstrated to be fairly robust to departures from its assumptions

Significant departures are not necessarily important!

there are problems with statistical tests of assumptions

More complex methods are often **more** sensitive to violations of assumptions..

You need to develop statistical wisdom, as well as knowledge!

Summary

Model choice is mostly about finding a **minimal adequate model** for your data

But we must preserve marginality, and for designed experiments you might need to retain some parameters

should always test if data matches assumptions of GLM – graphical approaches are (probably) better

often, simple transformations can help fix problems

If this fails, try e.g. Box-Cox, look for robust methods, or methods with other assumptions

DON'T BE TOO PICKY – GLM IS QUITE ROBUST