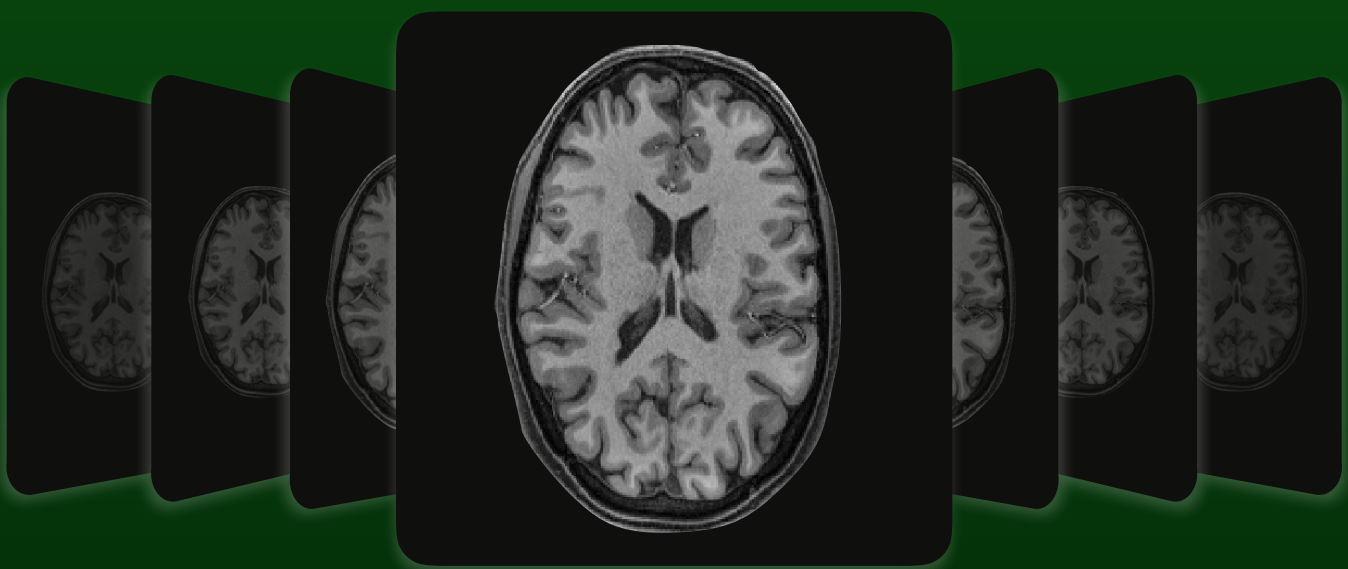
OXFORD NEUROIMAGING PRIMERS

Short introduction to the

General Linear Model for Neuroimaging



Mark Jenkinson
Janine Bijsterbosch
Michael Chappell
Anderson Winkler

Series editors:
Mark Jenkinson and Michael Chappell

PRIMER APPENDIX

List of Primers

Series Editors: Mark Jenkinson and Michael Chappell

Introduction to Neuroimaging Analysis
Mark Jenkinson
Michael Chappell

Introduction to Perfusion Quantification using Arterial Spin Labelling
Michael Chappell
Bradley MacIntosh
Thomas Okell

Introduction to Resting State fMRI Functional Connectivity
Janine Bijsterbosch
Stephen Smith
Christian Beckmann

List of Primer Appendices

Short Introduction to Brain Anatomy for Neuroimaging

Short Introduction to MRI Physics for Neuroimaging

Short Introduction to MRI Safety for Neuroimaging

Short Introduction to the General Linear Model for Neuroimaging

Copyright

Preface

This text is one of a number of appendices to the Oxford Neuroimaging Primers, designed to provide extra details and information that someone reading one of the primers might find helpful, but where it is not crucial to the understanding of the main material. This appendix specifically addresses the General Linear Model (GLM), as it is used in neuroimaging. In it we seek to go into more detail than we might in one of the primers, such as the Introduction to Neuroimaging Analysis, for those who want to understand more about how the GLM works and how to use it.

We hope that this appendix, in keeping with the series as a whole, will be an accessible introduction to the topic of the General Linear Model (GLM) for those without a background in the physical sciences. Hence, we have concentrated on key concepts rather than delving into any detailed mathematics. However, we also hope it is a good introduction to physical scientists meeting the GLM for the first time, perhaps before going on to more technical texts.

This appendix contains several different types of boxes in the text that are designed to help you navigate the material or find out more information for yourself. To get the most out of this appendix, you might find the description of each type of box below helpful.

Example Boxes

Example box: **GLM with**

Consider a simple example

These boxes provide specific examples to illustrate the general principles explained in the main text. It is expected that all readers will read the material in these boxes as the examples often contain further elaborations that are very helpful for gaining a good understanding of the topic.

Boxes

Box 2.1: **ICA-based deno**

Denoising based on indepe

These boxes contain more technical or advanced descriptions of some topics covered in this appendix. None of the material in the rest of the appendix assumes that you have read these boxes, and they are not essential for understanding any of the other material. If you are new to the field and are reading this appendix for the first time, you may prefer to skip the material in these boxes and come back to them later.

Summary and Further Reading

FURTHER READING

- Poldrack, R. A., Mumford

At the end, we include a list of summary points and suggestions for further reading. A brief summary of the contents of each suggestion is included, so that you can choose the most relevant references for you. None of the material in this appendix assumes that you have read anything from the further reading. Rather, this list suggests a starting point for diving deeper, but is by no means an authoritative survey of all the relevant material you might want to consult.

Mark Jenkinson, Janine Bijsterbosch, Michael Chappell and Anderson Winkler

Contents

1 Introduction

The General Linear Model, or GLM, is used for modelling and statistical hypothesis testing in nearly all areas of neuroimaging. This is due to its great flexibility - it can be used to analyse within-subject timeseries or between-subject data, and to remove components from the data identified as noise; all of which we will discuss briefly in this short introduction. It is therefore very important to obtain a practical understanding of how to use the GLM well if you are going to do neuroimaging analyses.

1.1 Linear modelling

At its core, the GLM is a way of modelling an observed signal in terms of one or more ***explanatory variables***, also known as ***regressors***. Signal here could mean the timeseries arising from a single imaging experiment, e.g., a BOLD timeseries in a given brain voxel, or equally it could be a series of measurements associated with individuals in a group, e.g., the cortical thickness in different patients at a given anatomical location. The GLM tries to explain this series of measurements in terms of one or more regressors (also called explanatory variables), which consist of series of values that represent patterns that we expect to be found in the measured signal.

The GLM is fundamentally a linear model, which means that it can scale the regressors and add them together in order to best explain the data. This is not the same as saying that it can only model straight lines, as many GLMs involve more complex relationships with time or subject ID. What remains linear is how the regressors can be combined together to explain the data.

The simplest GLM is one with a single regressor, and in this case the model only contains one parameter that is fit, which is the scaling value for this regressor; we will call this scaling β. This is closely related to Pearson's correlation, as correlation provides one way to measure the similarity of two signals (the regressor and the data in this case) while the GLM models how well one signal (the regressor) can fit another (the data). To determine what the best value of the scaling parameter is, the GLM examines the difference between the data and the scaled regressor (the fitted model). This difference is known as the ***residual error***, or more concisely just as the ***residuals***. In equation form the GLM can be expressed as:

$$Y = X\beta + \varepsilon,$$

where Y represents the data, X represents the regressor, β represents the scaling parameter and ε represents the residual errors. See Example Box "GLM with a single regressor" for an illustration of this.

Example box: GLM with a single regressor

Consider a simple example of modelling the average height of a group of individuals. This is done in the GLM by using a single regressor, together with its scaling parameter, β, to represent the average height. You can picture this by imagining the subjects lined up and an adjustable bar (like a high-jump bar) set to their average height. It is the value of β that can be adjusted, and so it represents the height of the bar, with the bar itself representing the regressor (in this case the regressor values are constant - i.e., the same for each individual - representing a horizontal bar). This is shown in Figure 1.1 along with an illustration of the differences between the model fit (the bar) and the individual data points (heights of the individual subjects). These differences represent the residual error and the best fit is defined as the one that has the least residual error.

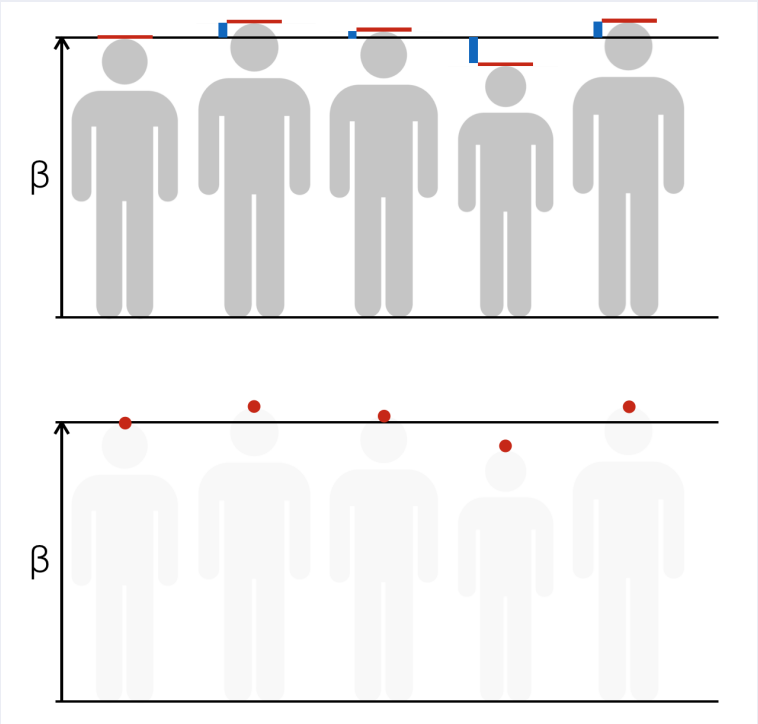


Figure 1.1: Illustration of GLM modelling for a set of group data - one datapoint per subject. Here we take heights for our data (top panel) and show how the GLM represents the group average by scaling a constant value (the horizontal bar - like a high-jump bar) by the value β. Changing β has the same effect as changing the height of the bar. The "best fit" is achieved when the differences between the individual heights (red lines) and the bar are minimised. These differences are called residuals and are shown in blue. The bottom panel shows the same data but with a more conventional view, with a single point representing each subject's data value.

It can be seen from Figure 1.1 in the Example Box "GLM with a single regressor" how the GLM can be used with one data value from each subject in a study. In that case the data were heights, but that can easily be replaced with any other data source, such as values derived from MRI (e.g., fMRI activation strength, cortical thickness, fractional anisotropy) at a specific location. When working with imaging data we would have a separate GLM for each location in the image - that is, for a particular voxel location we extract one value from each subject and analyse these values in one GLM. We then repeat this for every voxel location, running a separate GLM, but (typically) using the same regressors for all these GLMs, as it is the data that changes.

In a similar manner, the GLM can be used for timeseries analysis, such as for first-level fMRI data. In this case the data represents a timeseries, taken from one voxel for a single GLM. This is repeated for every voxel, building up an image of results, one voxel at a time (as illustrated in Figure 1.4). The regressors are usually predicted responses (the expected MRI signal resulting from neuronal activity) that relates to the timing of the stimuli (e.g., an event-related design) - see Figure 1.2.

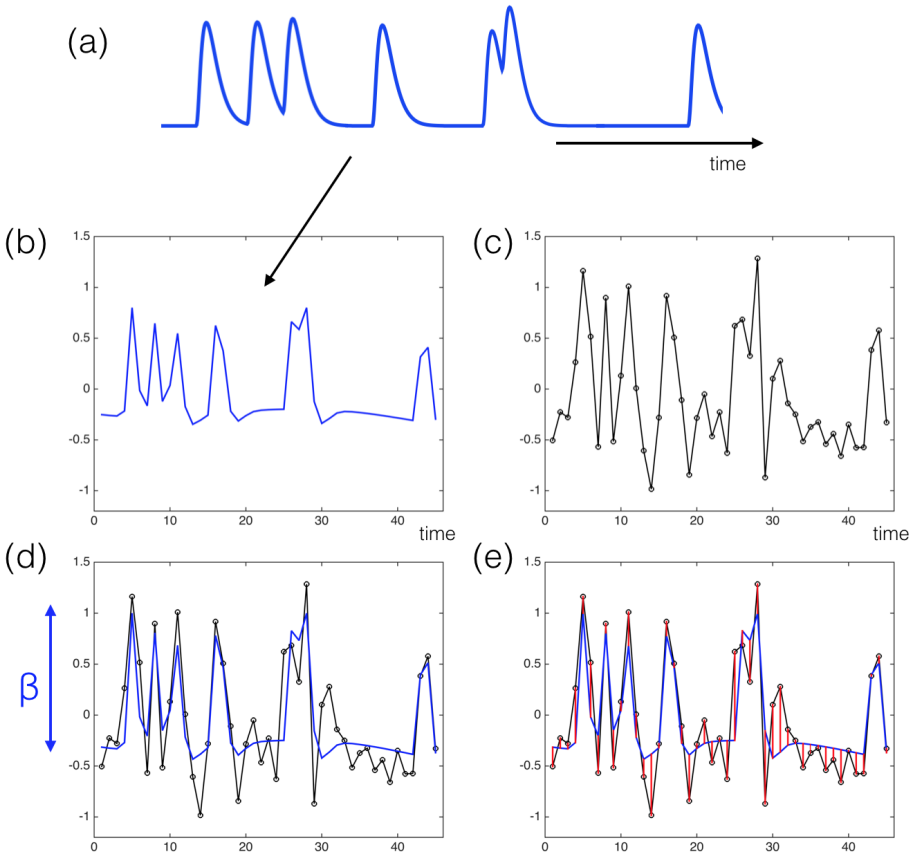


Figure 1.2: Illustration of GLM modelling for timeseries data. Each panel shows a signal changing with time (time here is the horizontal axis). In (a) we show an example of a predicted response (taking into account the hemodynamic response function) for a portion of an event-related design (seven discrete events) at high temporal resolution. In (b) we show the same predicted response, but sampled at a much lower temporal resolution, to match that of the fMRI data. In (c) we show some example data, as would be observed in one voxel. The process of GLM fitting is shown in (d) and (e), where the predicted response (i.e., the regressor - shown in blue) is scaled by a parameter, β, to match the data. The "best fit" occurs when the residuals (shown in red; i.e., the difference between the data and fitted model) are minimised.

Changing the scaling parameter will change the model fit and hence the residuals, and the best fit is the one that corresponds to the smallest residuals (quantified by the sum of squared values). This is known as minimising the residuals (or finding the least squared error) and can be done quickly using the GLM. The fitted or estimated parameter value is often denoted as $\hat{\beta}$ (note the hat) and represents a value that is calculated from the noisy data. This is in contrast to β (without the hat), which usually represents an ideal or theoretical value (e.g., the ground truth value that is not normally known in practice). For simplicity we will often use β to refer to either estimated values or ideal ones, as it is usually clear from the context, but when it is not we will use the symbol $\hat{\beta}$ to be

more explicit. When using the GLM it is not only the β value that we are often interested in, but also the uncertainty surrounding its estimation: we need to know both in order to do any statistical testing. The uncertainty in the value of any given $\hat{\beta}$ is affected by the noise (i.e., the size of the residuals) but also, in the case of multiple regressors, by whether individual regressors are correlated or not, something we will discuss further in section 3.

1.2 Multiple regression

Most GLMs contain more than one regressor. In this case, there is one scaling parameter for each regressor and the model fit consists of the addition of all the scaled regressors: this is a multiple regression model. The minimization of the residual error in this case requires finding the best values for all the separate scaling parameters. For example, with three regressors the GLM can be written as:

$$Y = X_1 \beta_1 + X_2 \beta_2 + X_3 \beta_3 + \varepsilon,$$

and the model fit is determined by finding the best values for β_1, β_2, and β_3 (also see Example Box "GLM with multiple regressors" for an illustration with two regressors). With the GLM these values can be calculated quickly and easily using matrix mathematics, but you do not need to know the details of how this is done in order to use the GLM effectively.

Example box: GLM with multiple regressors

Let us revisit the example of modelling the heights of individuals, but now extend it to two groups. In this case we will have a separate regressor for each group, modelling the average height of that group, i.e., we have a different bar for each group. Each regressor is associated with an individual β value and these are the values that are fit and will equal the average heights for the respective groups. See Figure 1.3.

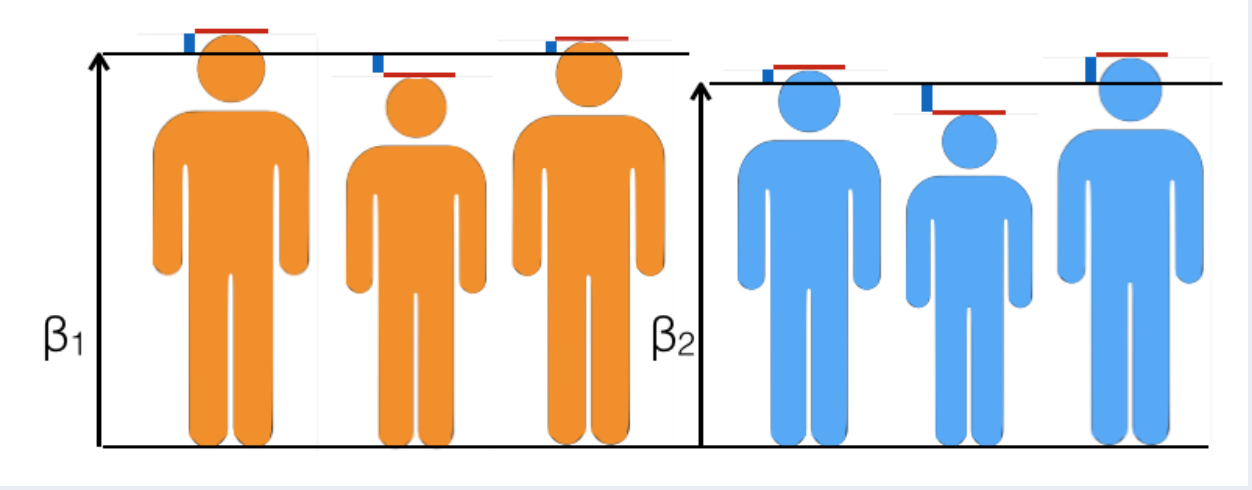


Figure 1.3: An illustration of GLM fitting for two group averages, each one controlled by separate scaling parameter (β1 and β2), with individual heights shown in red and the residuals shown as blue bars.

The analogy between what is shown in Figure 1.3 and the case of between-subject neuroimaging analyses is very straightforward - simply replacing the height values with appropriate values derived

from MRI data. In the case of within-subject timeseries analysis (such as a first-level fMRI analysis) the use of multiple regressors is more typically related to different conditions (i.e., types of stimuli); for example, one set of stimuli might correspond to showing picture of faces, while the other stimuli might correspond to showing pictures of houses. In such an experiment the two types of stimuli would be separated and each one modelled with one regressor. The β values in this case represent the average effect size of the corresponding stimulus type - see figure 1.4.

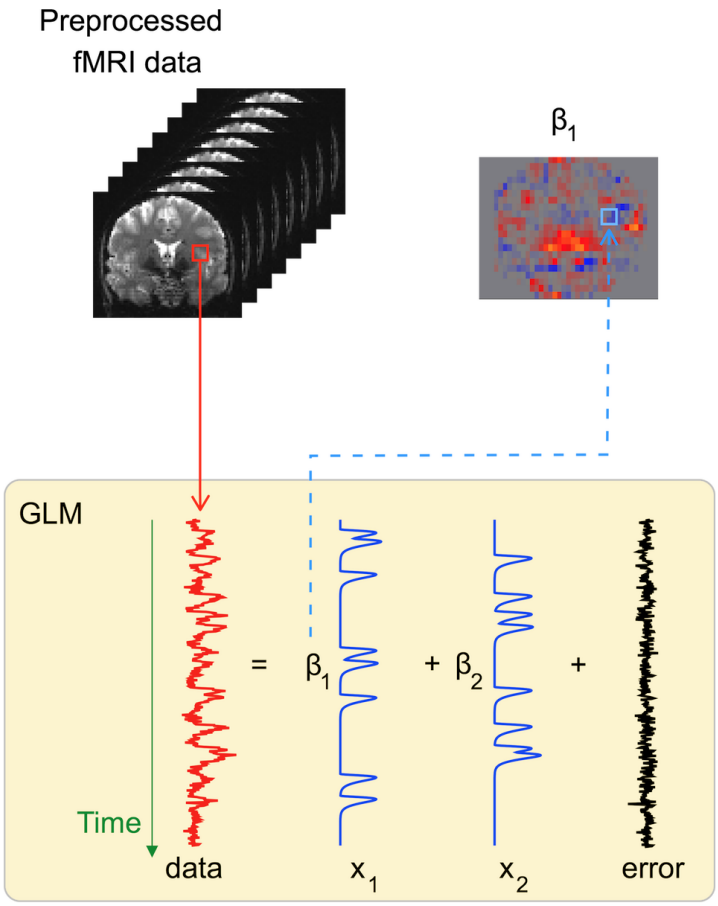


Figure 1.4: An example of the GLM applied to within-subject timeseries data. Here the data (from one voxel) is described as the linear combination of a model (X) containing a set of regressors (X_1 and X_2). One β value (representing an amplitude) is calculated for each of the regressors included in the model (i.e., for X_1 and for X_2), and what is left over are the residuals (i.e., the errors). This is typically done for all voxels separately, known as a voxelwise analysis, and the results (β values, or probabilities based on them) are stored, displayed, and possibly analyzed further in the form of voxelwise maps (images, as shown for β_1 in the top right).

Although you do not need to understand the mathematics behind the GLM, it is useful to be familiar with the main GLM equation and what the terms represent as well as the various names used for them. The equation used in the general case, with multiple regressors, is written in a matrix format as this is a concise way of grouping multiple elements together, as well as being useful technically. This equation is written as: $Y = X \beta + \varepsilon$; where Y represents the data from one voxel; X is a matrix that collects all the regressors together, with each regressor being one column; β is a vector (i.e., a set of numbers) that consists of all the individual scaling parameters; and ε is the residual error. Also note that each term goes by several alternative names: X is known as the ***design matrix*** or model,

and its individual columns are known variously as ***regressors***, ***covariates***, ***independent variables*** or ***explanatory variables (EVs)***; Y is the data or ***dependent variable***; β values are also called ***parameters*** or ***effect sizes***; $\hat{\beta}$ values are also called ***parameter estimates (PEs)***; and ε is called the ***residual noise, residual error*** or ***residuals***.

In a typical neuroimaging analysis, the GLM works with data from every voxel (or vertex); usually there are at least tens of thousands of such voxels (or vertices). For each one of these voxels, some measurement is available for different points in time or across different subjects, meaning that we have a series of data values at each voxel. For instance, the data could be a timeseries from a functional MRI scan of one subject and represents how the measured signal changed over time, or could come from a group of subjects with one value per subject (e.g., cortical thickness). Either way, the GLM models the data from one location in the brain and explains this in terms of the regressors. When we apply the GLM analysis to such a dataset, the regression is therefore performed separately for the data at each voxel - a ***voxelwise*** analysis[1] (i.e., for each separate analysis the dependent variable is the series of data values from a different voxel). This is also called a mass univariate analysis, which simply means that the same analysis is performed many times ("mass"), and is performed separately for every voxel in the brain ("univariate," as opposed to a multivariate analysis, which would take more than one voxel into account in the same analysis). Therefore, the "best fit" depends on the series of data values at each voxel, and the estimated β values are different for every voxel. The result of a whole-brain multiple regression analysis is a set of whole brain maps of β values (one map for each regressor). Each map contains one β value per voxel, as estimated from the multiple linear regression analysis performed on the series of data values at that voxel. In Figure 1.4, this would result in two maps, one for β_1 (as shown) and one for β_2 (not shown).

2 Inference, Probabilities and Statistics

In neuroimaging the GLM is typically used for hypothesis testing and probabilistic inference. These build on the basic modelling introduced in the previous section.

2.1 Hypothesis testing, false positives and false negatives

We will review hypothesis testing now as using the GLM in practice requires some understanding of this topic (if you have never encountered the null hypothesis or statistical testing before then you may want to also look at some more basic texts on this topic - see further reading). Hypothesis testing starts with the assumption that there are no signals (effects) of interest in the data and so what we measure is random noise. In addition to this, a specific question of interest and ***test statistic*** are chosen. The test statistic is a quantity derived from the fitted model (e.g., the ratio of a β value to its standard error). There are many possible statistics but some are most commonly used as they have been shown to have various interesting properties (e.g., ***t***-statistics, F-statistics). The hypothesis that there is no signal related to a given regressor is known as ***null hypothesis***. If large effects of interest are present in the data, the statistic values reflect this by becoming large (very different from zero) which would lead to the null hypothesis being rejected, based on some margin.

[1] For surface-based analyses this would be a vertexwise analysis with data coming from one vertex rather than one voxel.

In various fields, this margin is the ***confidence interval***, which can be calculated from the data and the model; in imaging, however, we tend to prefer ***p-values*** (defined below) as these are easier to display in the form of maps; confidence intervals and p-values provide roughly the same information regarding whether the null hypotheses should be rejected or retained.

If we know the ***probability distribution*** for each of the acquired datapoints (e.g., that it is a zero mean Gaussian), based on the assumption that no signal is present then we can calculate the probabilities (or p-values) of the observed data. It is possible to either take a ***parametric*** approach and assume that we know what the probability distribution of the noise is (e.g., assuming it is white, Gaussian noise) or we can use a ***non-parametric*** approach that makes little or no assumptions about the noise (e.g., permutation-based methods that estimate the distribution from the data). Either way, we end up with a number that tells us how likely it is to find a test statistic (e.g., a t value) that is at least as large[2] as the one calculated from the observed data, based on the assumption that there are no effects present (e.g., no neuronal activation). This number is the p-value. Once we calculate this probability we can decide whether the null hypothesis is too unlikely; i.e., a very low

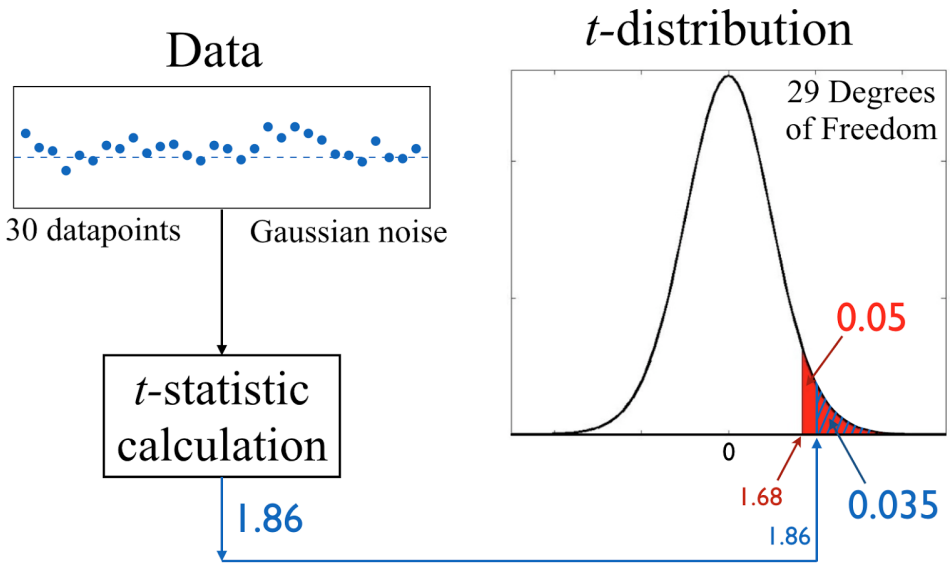


Figure 2.1: An illustration of hypothesis testing using a t-test. In the top left some data is shown and the statistical test applied here is designed to test if the mean value of the data is greater than zero (represented by the dashed line). It is assumed that the data can be accurately modelled by a mean value plus independent, white Gaussian noise. From the data values a t-statistic is calculated (formula not shown) and the result in this example is a value of 1.86. Given that the noise distribution is known (Gaussian) then the probability distribution of the t-statistic can be calculated theoretically under the assumption of the null hypothesis (i.e., that the true mean value is zero). This t-distribution is shown in the top right, and has 29 degrees of freedom (number of datapoints minus one degree of freedom due to estimating the mean value). The total area of the probability distribution is equal to 1, and the area of the distribution to the right of the calculated value (the t-statistic of 1.86) is 0.035 (shown as the hatched area). This represents the probability that a value equal to or greater than this can be generated purely by noise (using the assumption of the null hypothesis). By convention, this probability is compared to 0.05, which is taken as the standard false positive rate (represented by the red area, and corresponding to a threshold of 1.68 for this particular distribution). As the probability in this case (for this particular set of data) is less than 0.05, the null hypothesis is rejected and the test is considered to be statistically significant (i.e., the mean value is greater than zero).

[2] This describes a one-sided test, which is all that we will need for the application of the GLM in neuroimaging.

probability would indicate that the null hypothesis was unlikely to generate such a large test statistic, and hence we can reject the null hypothesis based on a pre-defined level, or margin - see Figure 2.1. Such a result, where the null hypothesis is rejected, is also called a ***statistically significant result***.

Rejecting the null hypothesis does not guarantee that there is a true effect present. For instance, there is always a chance that the null hypothesis was rejected when in fact it was actually true (i.e., there was no real effect present) and the rejection is a consequence of the noise containing some highly unlikely values. The probability distributions for the noise typically extend out to extreme values, and so no value is impossible due to noise, just increasingly unlikely as the values get far from zero. Therefore, any observed statistic value could have been generated by noise, and it is the p-value (probability) that gives us a handle on the likelihood of this. Similarly, if the null hypothesis is not rejected then there is a chance that there was a true effect present but it was not detected with this test and this data. These represent the two main errors that arise from null hypothesis tests: ***false positives*** (rejecting the null hypothesis when there was no true effect) and ***false negatives*** (not rejecting the null hypothesis when there was a true effect).

The probability associated with a false positive error is something that is set in advance as it is the level (threshold) for deciding when the probability of an observed statistic value (or greater) is small enough to be considered significant. This probability threshold is known as the false positive rate. Due to historical reasons it has become conventional to use a value of 0.05 (i.e., a 5 percent chance of such an error). The probability of a false negative is not usually known (and cannot be calculated without knowledge or assumptions about the size of the effect) but is related to the false positive rate such that decreasing the chances of one type of error, by varying this level (false positive rate), will increase the chances of the other type of error. Therefore, some tradeoff between the two types of errors is needed and a false positive rate of 0.05 is the established convention in neuroimaging and many other fields.

Statistical power is another important concept in statistical hypothesis testing. It refers to the chance of finding a true effect when it exists, which is inversely related to the chances of obtaining a false negative (i.e., missing a true effect). This depends on the relative strength of the effect of interest and the noise in the data. It can be estimated based on knowledge or assumptions about this relative strength.

The interpretation of the result from a hypothesis test depends on understanding these errors. For instance, when the null hypothesis is rejected it could either be because a real effect was detected or because it was a false positive. Similarly, when the null hypothesis is not rejected this could either be because there was no real effect or because it is was a false negative (where a real effect was present but not detected). As we know the chances of a false positive are 0.05, it is standard practice to interpret the rejection of the null hypothesis as evidence that a true effect was present, implicitly acknowledging that there is a chance that this is an error, which is why replication and validation are important in areas of science that rely on such statistical tests. However, as we do not know the chance of a false negative we do not interpret the opposite situation (not rejecting, or accepting, the null hypothesis) in a similar way. The standard practice is to not interpret a lack of rejection of the null hypothesis either one way or the other. It is tempting to interpret it as meaning that there was no effect present, but if the effect size of interest was not strong relative to the noise then there might be many instances where a true effect was present but not detected. This can also be expressed by saying that it is not possible to prove the null hypothesis. So be aware that the lack

of a statistically significant result (when the null hypothesis is not rejected) should not be used as an argument that no effect exists.

2.2 Contrasts

We will now consider some more details of how these hypothesis tests relate to the GLM. The first things to consider are how to specify the question of interest and the associated statistic. In the GLM, the model is specified by the regressors and is separate from the definition of the question, or questions, of interest. These questions (or alternative hypotheses) are expressed in the GLM by the use of contrasts, which combine together different parameters (effect sizes) to form mathematical inequalities. For example, the question of whether a positive effect exists (that is associated with one regressor, as compared to a zero or negative effect) can be expressed as $\beta 1 > 0$. Or to test if the effect size associated with one regressor is larger than another can be expressed as $\beta 1 > \beta 2$. Defining contrasts based on the parameters from the GLM provides the flexibility to ask many different questions from a single model (and typically several different contrasts are included in any analysis). It also conceptually separates the modelling from the formulation of the hypotheses, which makes it easier to correctly set up a GLM analysis.

A ***contrast*** in the GLM is defined by a set of weights, one for each β, which are used to specify an inequality. For example, if there are three parameters then the inequality will take the form:

$$c_1 * \beta_1 + c_2 * \beta_2 + c_3 * \beta_3 > 0.$$

The advantage of this form is that the weights, c_1, c_2 and c_3, become a simple list we can write down that entirely describes a given contrast. Although it is not necessary to understand the details of the mathematics behind the GLM fitting and inference, it is necessary to have a good working knowledge of contrasts and these inequalities. For example, if $c_1=1$, $c_2=0$, $c_3=0$ then this results in $\beta_1>0$, which simply tests if the effect size associated with the first regressor is positive. Alternatively, if $c_1=0$, $c_2=+1$, $c_3=-1$ then this results in $\beta_2-\beta_3>0$, or equivalently $\beta_2>\beta_3$. This tests whether the effect size associated with the second regressor is greater than the effect size associated with the third regressor. The ability to understand how the contrast values translate into inequalities and how these relate to questions or hypotheses is crucial for being able to use the GLM and interpret the results appropriately.

When writing down the contrasts we will often use a shorthand notation where we just list the values of the contrast parameters in order: e.g., [c_1 c_2 c_3] if there are three. For example, $c_1=-1$ and $c_2=+1$ would be written as [-1 +1] if there were two regressors, or [-1 +1 0] if there were three and $c_3=0$. You will find this notation used commonly both in papers and software implementations.

Example box: Simple contrasts

In this box we will give two examples of designs with various contrasts to help illustrate the different statistical questions (hypotheses) that can be expressed. The first example is a between-subject analysis involving two groups (patients and controls) where there is a regressor to model the mean of each group (analogous to Figure 1.3), the first regressor indicating membership of the patient group and the second the control group. In this case there are six common contrasts:

- [1 0] - tests if the mean of the patient group is positive (greater than zero)
- [0 1] - tests if the mean of the control group is positive
- [-1 1] - tests if the mean of the control group is greater than the mean of the patient group
- [1 -1] - tests if the mean of the patient group is greater than the mean of the control group
- [-1 0] - tests if the mean of the patient group is negative (less than zero)
- [0 -1] - tests if the mean of the control group is negative
- [1 1] - tests if the mean across both groups is positive

In the last example, [1 1], the quantity formed is the numerical average of the two mean values (the one for patients and the one for controls) and it represents a test of whether this average value is greater than zero. If this is significant it does not imply that both the individual group means are also significantly greater than zero, as it could be that one group mean is extremely large and so the average is still significantly greater than zero even if the other group mean was near zero or even mildly negative. As a consequence, this particular contrast is less often used, and it is more common to test the individual means (with [1 0] and [0 1]) along with other ways to combine these (see section 2.4).

This example also shows that the way the contrasts are constructed depends on the meaning and order of the regressors in the design matrix. It is good practice to specify the design matrix and the respective contrasts for the hypotheses being tested during the planning stage of an experiment.

The second example is a first-level fMRI analysis involving three stimulus types (pictures of people, pictures of animals, and pictures of plants) in addition to a baseline task of looking at a fixation cross. In this case a separate regressor is used to model the mean effect size of each of the three main stimulus types, with the baseline implicit. Common contrasts in this case include:

- [1 0 0] - tests if the mean effect size for pictures of people is positive (greater than zero)
- [0 1 0] - tests if the mean effect size for pictures of animals is positive
- [0 0 1] - tests if the mean effect size for pictures of plants is positive
- [1 -1 0] - tests if the mean effect size for people is greater than for animals
- [-1 1 0] - tests if the mean effect size for animals is greater than for people
- [1 0 -1] - tests if the mean effect size for people is greater than for plants
- [-1 0 1] - tests if the mean effect size for plants is greater than for people
- [0 1 -1] - tests if the mean effect size for animals is greater than for plants
- [0 -1 1] - tests if the mean effect size for plants is greater than for animals
- [-1 0 0] - tests if the mean effect size for people is negative (less than zero)
- [0 -1 0] - tests if the mean effect size for animals is negative
- [0 0 -1] - tests if the mean effect size for plants is negative
- [1 1 1] - tests is the mean effect size across all three stimulus types is positive

2.3 t-statistics

The contrasts that we have defined here can be used to form a particular statistic: the ***t-statistic***, also known as the Student's ***t***-statistic, and is essentially a ratio between the amplitude of the quantity of interest and the amplitude of the uncertainty (or standard error) of this quantity. It is the contrasts that specify the hypotheses of interest in the GLM, which might be β_1 itself or might be β_2-β_3, using the previous examples. The contrasts are used to compute a quantity that is a weighted addition of effect sizes; it is this quantity that is tested using the t-statistic. The calculation of the uncertainty in the estimation of this quantity involves the root mean square amplitude (i.e., sigma or standard deviation) of the residuals of the model, as well as the number of independent values in the residuals (known as the ***degrees of freedom***) and any covariance between regressors. These values can all be calculated as part of the model fitting stage. We will cover how correlations and covariances affect things in the next section and for now just discuss the simpler case where there are no correlations between regressors.

It is not necessary to understand the details of how the t-statistic is calculated, but it is useful to appreciate that it depends on the amplitude of the contrast of parameters (e.g., β_2-β_3) as well as the uncertainty (i.e., variance or standard deviation) of this. The latter is affected by the amplitude of the noise (the greater the noise the smaller the ***t***-statistic) as well as the number of degrees of freedom (the greater the degrees of freedom the smaller, or more significant, the associated statistical probabilities). The degrees of freedom are determined by how many independent data points are available to characterise the noise, which is normally equal to the total number of data points minus the number of regressors, although pre-processing steps such as demeaning and filtering also reduce the degrees of freedom - for further illustration see the Example Box "Degrees of Freedom". What is important to understand is that for reasonable statistical power, in order to give you a good chance of detecting true effects, you need a sizeable effect compared to the noise level, as well as a good number of degrees of freedom. Exact figures for what you need depend a lot on the nature of the data and the effect of interest, but you can find power calculation tools and design efficiency estimates that can give you some quantitative information on how much power you will have in a particular experiment, and we encourage you to use these tools when designing and analysing experiments.

Example box: Degrees of Freedom

To get some intuition for degrees of freedom imagine that, in order to maintain a healthy weight, you give yourself a lunchtime "constraint": you want to make sure that the average number of calories you consume at lunchtime across any one week (7 days) is 600 Kcal. During the holiday season, you go out for lunch with friends most days and indulge in high calorie options. However, when the last day of the week comes around, you are stuck with a very small plain salad in order to make sure you don't exceed your 600 Kcal average. In this example, you had six degrees of lunch-freedom, because you were free to choose whatever you wanted for 6 of the 7 days. If you had decided that you should average 600 Kcal across the entire month (instead of just across 7 days), then you would have had more degrees of lunch-freedom (i.e., you would have been free to choose whatever you wanted for more days). However, if you put on additional lunch constraints (such as avoiding meat at least 3 days a week), then you would end up reducing your degrees of lunch-freedom further in order to meet the additional constraints. The same happens in a multiple regression analysis; for every parameter that we want to estimate (every constraint we put on our lunch), we lose one degree of freedom. The number of degrees of freedom is important because it affects the accuracy with which we can estimate the βs for our model.

The ***t***-statistic values are converted into probability values using either known parametric distributions or using non-parametric, numerical methods (e.g., permutation-based methods). This conversion takes into account the statistic value as well as the degrees of freedom. Whether this conversion of t-statistic values to probabilities (where the probabilities are calculated under the null hypothesis assumption) is done with parametric or non-parametric methods is purely a choice made by the user or the developers of the analysis tool. The GLM itself can be used in either case and the formulation of the model (specifying regressors) and definition of the hypotheses or questions of interest (the contrasts) works in exactly the same way regardless of what inference technique (for conversion to p-values) is used.

One last thing to note about the t-statistic is that it is conventional to perform a one-sided statistical test. That is, the alternative hypothesis takes the form of $c \cdot \beta > 0$ rather than testing if $c \cdot \beta$ is negative, or testing if it is either negative or positive (i.e., non-zero). This is not something enforced by the GLM at all, but just a popular approach in neuroimaging (and in other areas). It is helpful in that it distinguishes between effects of different signs rather than mixing them together - as two-sided, or non-zero, tests do. However, it is not restrictive, as it is still possible to test for negative effects as well as positive effects. This is done by changing the sign of the contrast values to invert the effects when you are looking for negative changes. For example, if you want to test for a negative effect such as $\beta_1 < 0$ then this is simply formed using the equivalent test, $-\beta_1 > 0$ as the negative sign changes the direction of the inequality. This corresponds to using $c_1 = -1$ instead of $c_1 = +1$. In general, any positive test can be turned into a negative test (looking at the other side of the distribution) simply by multiplying all the contrast values by -1. For example, $\beta_1 > \beta_2$ corresponds to $\beta_1 - \beta_2 > 0$, and is specified by $c_1 = +1$ and $c_2 = -1$, while the opposite is $\beta_1 < \beta_2$ and corresponds to $\beta_2 - \beta_1 > 0$ or $-\beta_1 + \beta_2 > 0$, and is specified by $c_1 = -1$ and $c_2 = +1$.

2.4 F-statistics

The contrasts used with the GLM can also be combined together to form ***F-statistics*** rather than ***t***-statistics. These F-statistics are used to formulate hypotheses that concern whether any one of several quantities, or any combination of them, is significantly non-zero. This test is therefore a two-sided test, as it tests for whether the fundamental effects are non-zero (i.e., positive or negative), not just whether they are positive. Furthermore, the test will respond to any combination of non-zero effects and can be driven by a single strong effect or several weaker ones in combination. It is therefore quite different from a conjunction test where all effects must be strong in order to produce a result[3]. The nature of the F-test is to ask a question (or formulate a hypothesis) along the lines of: is effect A or effect B or effect C, or any combination of them, significantly non-zero?

An F-test can be based on multiple effects (such as three effects in the previous example) or even a single effect, where the latter is just performing a two-sided test on the single effect (i.e., responding to positive or negative effects). In terms of the GLM, an F-contrast is specified using a set of ***t***-contrasts (we will call the contrasts from the previous section ***t***-contrasts here in order to distinguish them) and it combines the quantities formed by the ***t***-contrasts. For example, one contrast might form β_1-β_2 and another might form β_2-β_3 and an F-contrast can include both of these contrasts, represented by a matrix of contrast values where each row is a single t-contrast - in this case the matrix would have two rows, with [1 -1 0] as the top row and [0 1 -1] as the bottom row. Setting these up in neuroimaging analysis packages is simple as it just requires defining a set of ***t***-contrasts.

Since the F-test is two-sided it does not matter whether a ***t***-contrast or its negative version is included, as both give equivalent results; e.g., β_1-β_2 or β_2-β_1. Including both of these contrasts is redundant and should be avoided, although most software implementations will gracefully handle this and effectively exclude one of them in the internal calculations. Determining exactly what kind of effects are included in an F-test can be tricky as it is not always obvious how they combine[4]. See Example Box “Formulating F-contrasts” for more information.

[3] Further discussion of conjunction tests is beyond this scope of this Appendix, but can be easily achieved by simply taking the intersection of the masks of significant voxels from a set of individual ***t***-statistic tests.

[4] For the mathematically inclined readers - it uses each t-contrast as a basis vector and it is the span of these vectors that defines the space used by the F-test.

Example box: **Formulating F-contrasts**

Here we will consider a range of examples for F-contrasts to illustrate the way that they work and to highlight the fact that they are not always very intuitive.

Let us start with a simple example: a design with two regressors where we are interested in a response to either of them. Including contrasts [1 0] and [0 1] will test for any non-zero effect related to β_1 or β_2 or any combination of them, which is fairly straightforward to understand. However, including contrasts [1 -1] and [1 1] would do the same, as combining the difference between the parameters and their mean value covers the same set of possibilities (any two values for β_1 and β_2 are uniquely determined by their difference and their mean).

Now let us consider an example with three regressors, which is more difficult. In this case including contrasts [1 -1 0] and [1 1 1] is not the same as including [1 0 0] and [0 1 0] since the latter does not include the third parameter (it always gets zero weight in the contrasts). Using [1 1 0] instead of [1 1 1] would give us the equivalent of the example above (although the third parameter would be entirely excluded). Alternatively, including [1 -1 0], [0 1 -1] and [1 1 1] is the same as including [1 0 0] , [0 1 0] and [0 0 1] since two differences and one mean can uniquely determine three parameter values, while [1 -1 0], [1 0 -1] and [0 1 -1] would not be the same since the third contrast here is actually redundant (it is the difference of the first two: i.e., $(\beta_1 - \beta_3) - (\beta_1 - \beta_2) = \beta_2 - \beta_3$) and so the mean contrast of [1 1 1] is crucial in this case. In fact the mean could be used to replace any of the three difference contrasts, as any one of them is redundant with respect to the other two.

Although the basic mathematical manipulations here are relatively straightforward it is still not easy or intuitive to check if the F-test you are formulating is asking the question you are really interested in. Hence it is most useful to stick to certain standard design patterns when using F-tests, as then it is easier to use them or adapt them for your particular purposes, whilst being confident that they are doing what you expect. If in doubt seek help from someone with more experience and/or mathematical background.

2.5 Multiple testing correction

When performing voxelwise analyses of brain images (using either t or F statistics) the number of voxels, and hence the number of statistical tests, is in the range of tens of thousands to millions. This causes a major problem for the false positive rate, since sticking with a 0.05 false positive rate for every test in isolation would result in many thousands of false positives across the brain. Therefore, a further correction to the false positive rate is needed to account for the large number of tests - that is, ***multiple testing correction*** (also often called, somewhat inaccurately, ***multiple comparison correction***).

There are many methods of multiple testing correction, and they are often combined with statistics that aim to take into account the spatial distribution of the signal. Using information about the spatial distribution of large statistic values is a powerful approach in neuroimaging as it takes advantage of the fact that real biological effects appear as connected clusters of voxels, whereas false positives due to noise tend to be scattered and unconnected. Examples of spatial statistics that

include multiple testing correction include parametric methods such as ones based on Gaussian Random Field theory (using cluster size) as well as non-parametric methods based on cluster-size, cluster-mass or spatial support (e.g., Threshold-Free Cluster Enhancement; TFCE).

3 Confounds, Corrections and Correlations

It is common to have ***covariates of no interest*** (e.g., a regressor representing age) where such covariates may affect the measurements, but you are not directly interested in this effect itself, only in another covariate or covariates (e.g., disease duration). There are three main ways that this can be partially compensated for: (1) only include subjects in the study that are exactly the same with respect to the covariates of no interest (e.g., the same age); (2) include subjects where the variation in the covariates of no interest is random but balanced between the groups (e.g., each group has the same average age and spread of ages); (3) include covariates of no interest in the analysis to "correct" or "adjust" for them. In practice the first option is often not feasible, and although the second option should be aimed for as much as possible, it is usually far from ideal (either because the ranges cannot be matched or because they might not be random enough in relation to the covariates of interest). Therefore the third option, in conjunction with the second, should normally be used. This section will cover the third option and what difficulties can arise, since there are several tricky issues that relate to this, especially when covariates are correlated with each other, which is normally the case.

3.1 Modelling

The first thing to discuss is whether to include covariates in the model (the GLM) or not. A covariate, whether it is of interest or not, can be included in the GLM but it is not necessarily a good idea to include all the covariates that you can possibly think of. For instance, it is easy to make a list of potential factors that might affect the measured quantity, such as age, educational status, caffeine levels, sex, anxiety, handedness, and so on. Including all possible factors is not only infeasible in most cases (e.g., measuring caffeine levels) but would be disastrous for the analysis, since we often have a limited number of subjects and having a large number of regressors would lead to very low statistical power (remember that each additional regressor "costs" one degree of freedom) or an ill-conditioned analysis (if the number of regressors is greater than the number of datapoints). Hence it is necessary to decide beforehand what are the most crucial factors that need to be accounted for, and only include these. This decision is not dependent on statistical theory or understanding the GLM (except for the fact that it is bad to have too many covariates) and so will rely on other issues related to understanding the experiment, relevant psychological issues, the biology and physiology, and the measurement process. Using previously published studies is a good guide for what covariates are usually included or not.

Once a decision has been made regarding what covariates to include, it is necessary to put all of these in the GLM. Regressors that are included in the model but are then not included in a contrast are called ***confounds***[5] and are important in the estimation, as they match parts of the measured signal that are related to them and effectively remove this variation that otherwise would be left in

[5] The definition of a confound is with respect to a particular contrast. When there are multiple contrasts a given regressor may be a confound for some contrasts but not for others.

the residuals. Reducing the residuals in this way is important, as the confounds remove structured signals of no interest and prevent them from affecting the estimation of the amount of random noise that is present. Getting an accurate measurement of the noise is important because the amount of noise affects the uncertainty in the estimation of each of the β values, on which the statistics are based. If there are non-random, or structured, signals left in the residuals (e.g., by not including key confounds) then the statistics will no longer be accurate or even valid. Typically, if confounds that influence the data are not included in the model then the estimate of the noise, that is taken from the residuals, will be incorrect and usually result in lower statistical power.

When the regressors in the GLM are not correlated with each other, the confounds only remove signal that otherwise would have been treated as residual noise. In general, including an additional regressor in a GLM will alter all the parameter estimates and their uncertainties, since the whole model is fit jointly. The addition of uncorrelated confounds does not affect the estimated β values, but will change the final statistics, due to having a better, and often lower, estimated variance for the noise and consequently a different uncertainty for β. Hence it is worthwhile including confounds even when they are uncorrelated although, as discussed above, only include important covariates that explain a substantial amount of variance in the signal.

The more common situation with confounds (and regressors in general) is that they are correlated with other regressors. When this happens, including confounds will change the estimated β values and the residuals and the uncertainty associated with the estimated β values. The change in β values is due to signals that are linearly related to both the covariates of interest and the confounds. For example, the covariates of disease duration and age are often correlated and the measurements being analyzed are likely to either correlate strongly with both of these covariates or with neither of them. This arises in group-level analyses as well as in timeseries analyses (as done for task fMRI); e.g., visual stimulation and head movement. In this situation there is also a change in the uncertainty due to the fact that more than one of the correlated regressors can explain the data. As the correlation between the regressors increases they become more similar to each other and the uncertainty related to how exactly to split the signals amongst them also increases. We will discuss some specific examples below, but these general principles apply broadly.

When there are correlated regressors the GLM will fit the '***shared signal***' (also referred to as '***shared variance***' although that term can be confusing since it is not necessarily related to noise) by splitting the signal across several regressors. The exact details of the split will depend on the '***unique***' parts of the regressors (i.e., where they differ) and how well the measured signal fits these parts. As the regressors become more similar, and more highly correlated, the splitting increasingly depends on small differences between the regressors and hence can be easily influenced by noise. At the same time, the uncertainty becomes larger since the fit is only being determined by the very small difference between the regressors, which will be more easily influenced by the noise. In all cases the overall fit will include both the 'unique' and 'shared' parts of the signal; i.e., the total signal, at all points, is represented by the combination of regressors, weighted by the respective β values. See Example Box "Fitting correlated regressors" for an illustration of this fitting.

Example box: Fitting correlated regressors

We will consider a simple example here, in order to illustrate what happens with correlated regressors in the GLM. The principles demonstrated here are general and apply to all situations involving correlated regressors.

Consider an analysis of timeseries in a task fMRI experiment where there is a visual stimulation that moves across a screen with variable speed. The relationship of interest for the experimenter is between the speed (which can be positive or negative, depending on whether it moves to the right or the left) and the brain activity (measured with BOLD fMRI). However, despite each subject being told to keep their head very still, it is likely that some head motion will exist and even after motion correction there may be changes in the signal related to head motion. It is also likely that the head motion is correlated with the stimulus timing. In this example the speed is sinusoidal and will form the first regressor (the one of interest). The second regressor is based on the head motion (which can be measured by motion correction), and both of these regressors are shown in Figure 3.1.

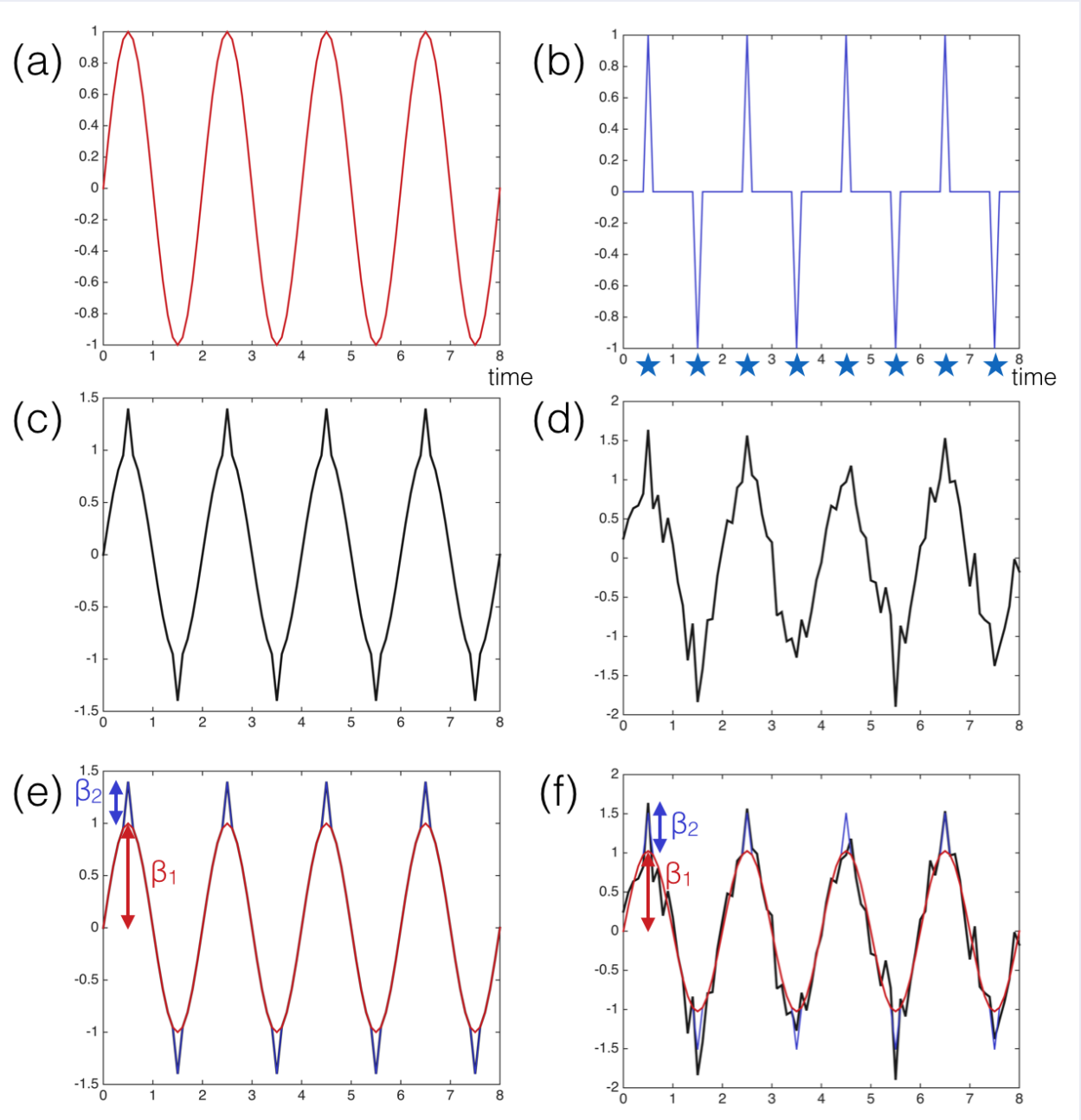


Figure 3.1

Example box: **Fitting correlated regressors**

Figure 3.1: Illustration of GLM fitting for correlated regressors. The example here is based on a GLM for a task-fMRI analysis with varying visual stimulus and correlated head motion; corresponding regressors shown in (a) and (b) respectively, and correlation is r=0.447. Examples of timeseries data (measured MRI signal) are shown in (c) without noise and in (d) with a small amount of added noise. In both cases the underlying ground truth is that the timeseries consists of 1 unit of the regressor in (a) and 0.4 units of the regressor in (b). Fits, using the GLM, are shown in (e) and (f) for the timeseries in (c) and (d) respectively - red is for the regressor shown in (a) and blue is for the regressor shown in (b). The amplitude of the fits are shown in the figures as β1 and β2, and it can be seen that the signal is fit across all timepoints. However, the amplitude of the fit for β2 really only depends on how much the signal departs from the fit of regressor (a) at the timepoints corresponding to the spikes (where there is shared signal) - these timepoints are shown as stars in (b). The head motion is somewhat idealised here (being equal magnitudes for each spike) as this is for illustrative purposes, but it is not far from what can happen in practice and the principles demonstrated apply generally.

From Figure 3.1 you can see how the combination of the two regressors fits the signal well at all points in the timecourse. The size of the fit for the second regressor, β_2, is based on the values of the signal at a small number of isolated points in time (where this regressor is non-zero - at the spikes), while the size of the fit for the first regressor, β_1, is based on the rest of the timepoints. In this case, it is these timepoints (non-spikes) that represent the 'unique' part of the first regressor, while both regressors contribute to the signal at the timing of the spikes and that represents the 'shared' part of the signal. An important point here is that although the β values might be determined by the 'unique' parts of the signal, the aim of the fit is still to be a good match for ***all*** parts of the signal so that it gives the closest fit possible across both the unique and shared parts.

The example shown in Figure 3.1 demonstrates how the two regressors can give an overall fit that matches well at all points, while the split between them (the individual values of β_1 and β_2) is determined by the 'unique' part of the signal. In this case the signals were very 'clean' and idealised and the split is obviously doing something highly sensible. However, in practice the regressors themselves are often less clean and may be formed from imperfect data (e.g., noisy estimates of head motion; test scores affected by how the person slept last night; etc.). In such cases the split between regressors can become influenced by these random or uninteresting aspects, making the individual β values less useful at answering key questions. Hence while correlated regressors are taken account of 'correctly' by the mathematics in the fitting and inference, they can still cause problems in the interpretation of the results. All of this is equally true for between-subject analyses, and correlated regressors (such as disease duration and age) are common in such analyses.

3.2 Inference

What happens with probabilities (i.e., inference) in the presence of correlated regressors is less straightforward compared to the fitting. If there are confounds without any correlation then the situation is simple and, as described above, the confounds only affect the residuals during the fitting

and hence typically increase the statistical power (due to the variance of the residuals being reduced). If the confounds are correlated with the covariates of interest then the splitting of the signal fit across the different regressors has a much bigger effect on the statistics, and in this case the analysis with the GLM takes a conservative approach, where the ***t*** and F statistics for any particular contrast are based on what is ***purely*** due to the covariates of interest and ***cannot*** be explained by the confounds. If there is any ambiguity (such as shared signal) then the result of the GLM analysis will be conservative in the sense that any ambiguities will be considered to have been driven by the confounds rather than the covariates of interest. For example, if a pattern of measurements matched equally well to either disease duration or age, then it would be considered (in the GLM analysis) to have been caused by the confound (e.g., age) rather than the covariate of interest (e.g., disease duration). Only patterns of measurements that cannot be explained by any of the other regressors (i.e., the unique parts of the signal associated with the covariates of interest) can drive the statistics.

As we saw in the previous section, when doing the GLM fitting, the shared signals are apportioned between correlated regressors and even in cases of strong correlation the β values can be high for one or more regressors. However, when it comes to calculating the statistics, if the correlation with the confounds is very high then it is unlikely that any signal can be confidently assigned to the covariates of interest, thus the uncertainty on the β values (or a contrast formed from them) will be large. In such cases it is unlikely that any statistically significant result can be found - that is, the statistical values (e.g., t-values) are small, due to the high uncertainty, and there is low statistical power. This is intuitively sensible, since if the confound is very similar (e.g., the pattern of age is very similar to disease duration; or the head motion is very similar to the stimulus) then you cannot be sure that an observed pattern in the measurements is caused by the covariates of interest rather than the confounds. In such cases it is often better, in a research setting, to avoid coming to a false conclusion (a false positive) and therefore preferable to treat such signals as if they had been caused by the confounds. Using the GLM will naturally do this as part of the inference, without you having to do anything special when creating the model or the contrasts.

One useful way to think about what the analysis using the GLM does[6] is to know that it gives you the ***same result that you would have got if you initially removed the confounds from both the data and the model***. For example, if you removed all signals matching the confound (e.g., head motion) from the data and also removed them from all the covariates of interest (e.g., visual stimulus) and then set up the GLM without the confound (as it has now been removed from everything) then the results for that contrast would be identical to what you would get with the original GLM (where the original data, covariates of interest and the confounds were all used). It is reasonably easy to see in this case why it is only the 'unique' parts and not the shared signals (between the covariates of interest and the confounds) that affect the statistics. If the confounds are very similar (i.e., highly correlated) to the covariates of interest then this can leave very little useful signal behind in the covariates of interest, and this is what makes it hard to find a good match to the observed signal and hence is what reduces the statistical power. See Example Box "Inference with correlated regressors" for an illustration of this.

[6] This is not normally what is happening in an analysis using the GLM internally but it is exactly equivalent mathematically.

In practice when using the GLM we do not normally remove the confounds during any of the calculations[7], but the effects are exactly the same as if we did. The precise details of the GLM calculations are beyond the scope of this appendix, as it is quite technical[8]. One thing to be aware of is that even though a β value might be large (as part of the fit for a shared signal) it might still lead to insignificant statistical values due to the fact that the uncertainty associated with this β value would be extremely high. This uncertainty reflects how sensitive the β value is to noise, and when the correlation is high it only takes a small amount of noise to make a large change in the β values of the fit. It is useful to have some intuition for the fact that correlation reduces the statistics because the uncertainty of the β values becomes high.

***Example box:* Inference with correlated regressors**

We will extend the example in the previous example box (visual stimulus and head motion). In Figure 3.2 you can see correlated regressors, as in Figure 3.1, and how they can be fit after removing the confound (head motion) from the covariate of interest (visual stimulus) as well as the data. It can be seen that removing the confound from the covariate of interest has got rid of the portions of shared signal (when the spikes in the head motion occur) and just leaves the unique portion of the covariate. In Figure 3.3 the same set of results are shown but for a slightly different head motion regressor (one that is even more highly correlated with the visual stimulus). It is even clearer in this second example how, after removing the confound from the covariate of interest, there is very little useful structured signal remaining, making it highly sensitive to the effects of noise and therefore difficult to estimate reliably, which leads to low statistical power.

[7] One exception to this is in permutation testing, where removing confounds initially is a common way of performing the calculations.

[8] For our mathematically inclined readers - it is related to the properties of the matrix inverse for the covariance matrix, since it is the variance calculation (and not the β fits) where correlation has a big effect.

Example box: **Inference with correlated regressors**

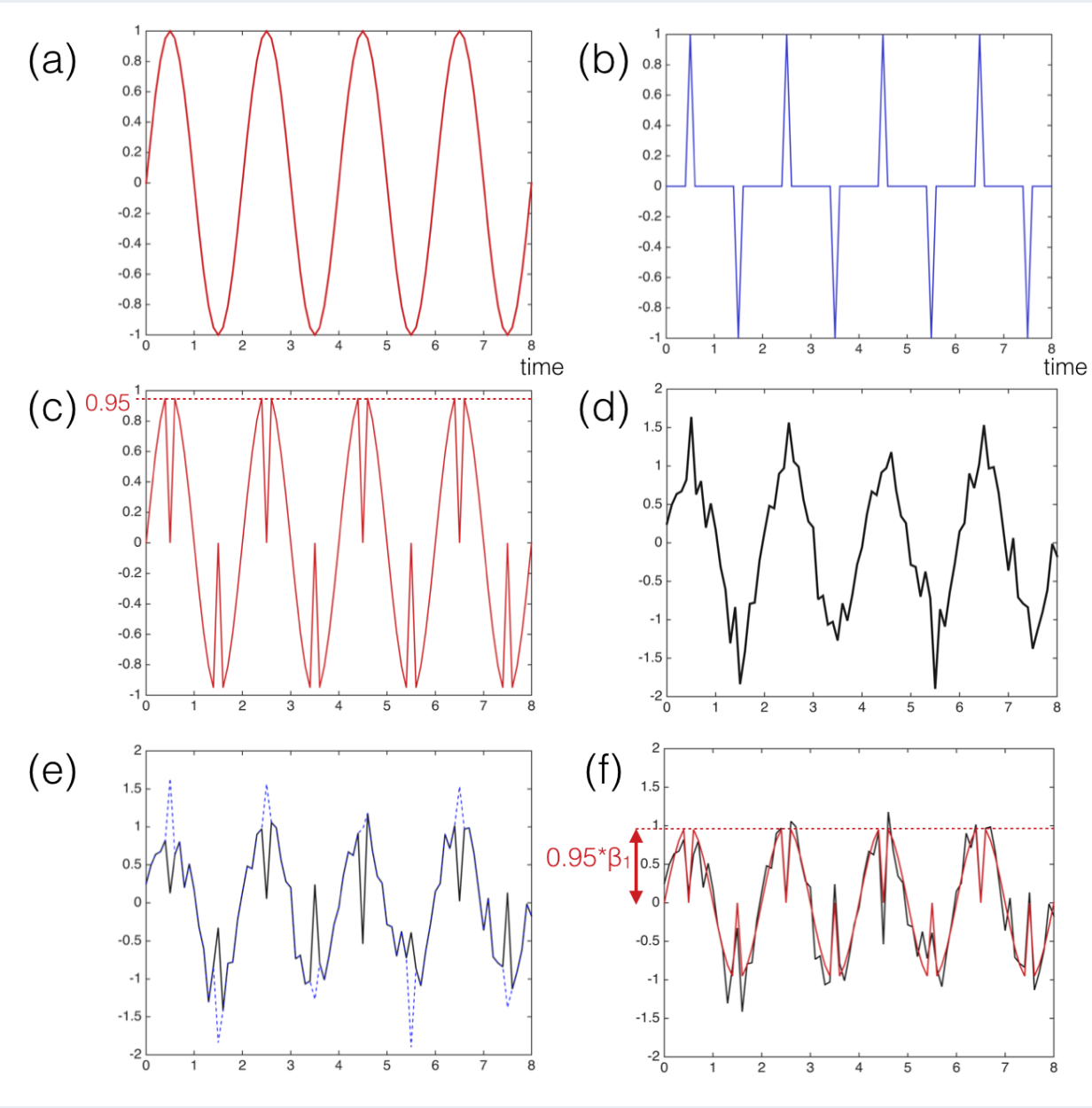


Figure 3.2: Illustration of GLM inference for correlated regressors. The example here is based on a GLM for a task-fMRI analysis with regressors associated with varying visual stimulus and (correlated) head motion shown in (a) and (b); where the correlation is r=0.447. In (c) the visual stimulus regressor is shown after removal of the head motion regressor - note how this only affects points where the head motion is non-zero and that the maximum value of the result is now 0.95 instead of 1.0. An examples of timeseries data (measured MRI signal) is shown in (d), that contains 1 unit of the regressor in (a) plus 0.4 units of the regressor in (b) and a small amount of noise. The result of fitting and removing the head motion regressor from the data is shown in (e) - dotted blue line shows the original data, and again this only differs at points where the head motion is non-zero. In (f) the final fit of the (reduced) GLM is shown, demonstrating how β_1 is used to scale the regressor in (c) to match the data in (e) - note that the maximum value is $0.95*\beta_1$ as the maximum of the regressor in (c) is 0.95. This results in a fit that is purely based on the part of the signal that is unique to the visual stimulus response and not related to head motion. Note that the value of β_1 here is exactly the same as we would have got for regressor (a) when fitting the full model; i.e., regressors (a) and (b) to data (d).

Example box: Inference with correlated regressors

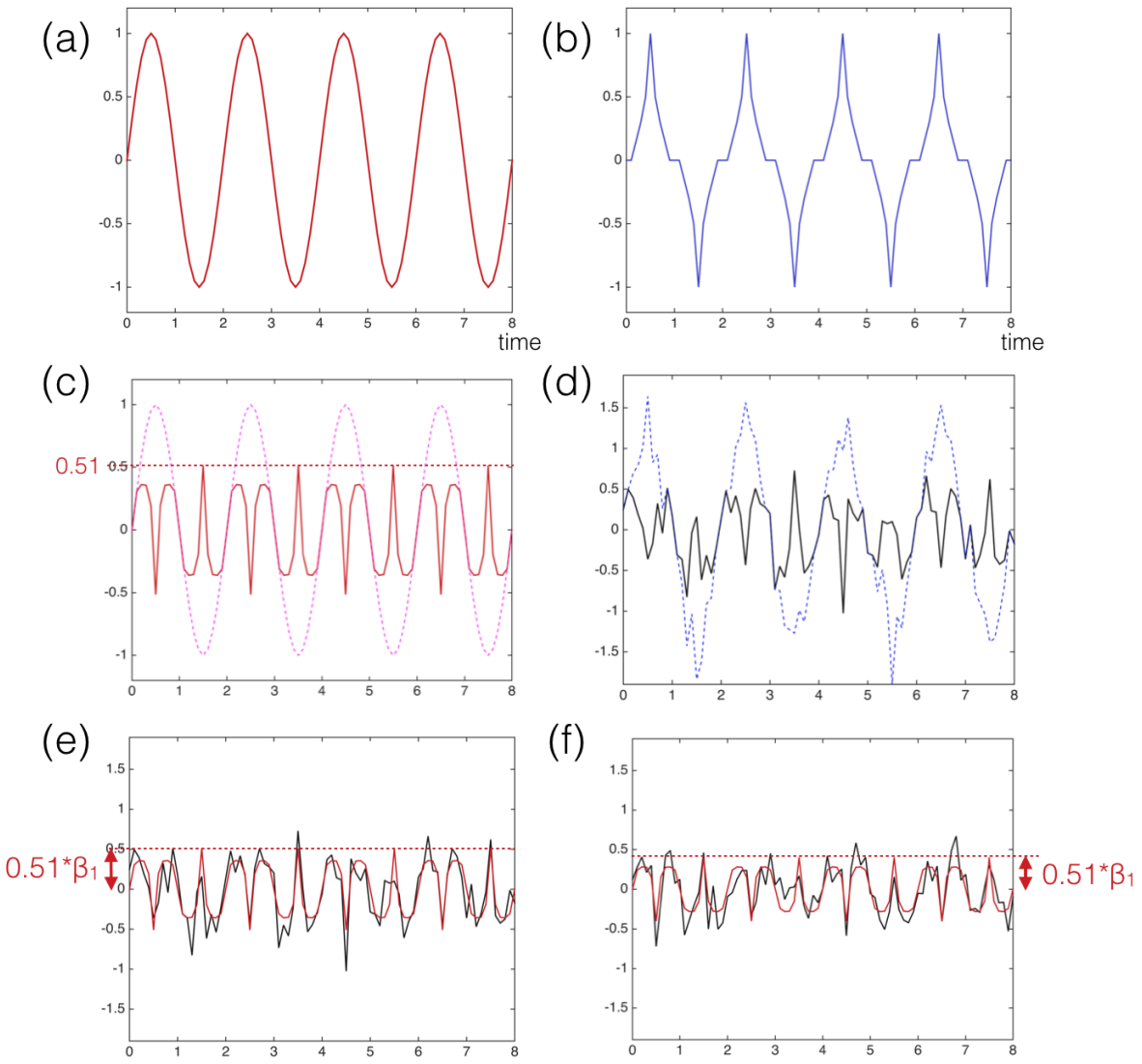


Figure 3.3: Illustration of GLM inference for correlated regressors. This follows the same format as in figure 3.2 where the original two regressors are shown in (a) and (b), where the head motion in this example is now more highly correlated: r=0.89. The result of removing the head motion regressor from (a) is shown in (c) and the result of removing it from the data is shown in (d). The fit of the GLM for (c) and (d) is shown in (e) - note that the maximum value is $0.51*\beta_1$ as the maximum of regressor (c) is 0.51. Another fit, for data with the same signal but different noise (original data not shown), is illustrated in (f); i.e., it is the equivalent of what is shown in (e) but based on a different version of (d) where there was a different example of the noise. By comparing (e) and (f) you can see how much the noise can affect the size of the fit. Because there is much less unique signal left in (c), the final fit is therefore based on less distinctive data and hence is more easily influenced by noise, since the noise represents a larger proportion of (c) compared with (a). This is reflected in the different values for β_1 in (e) and (f), which differ by around 20%, since small changes in the data can have a larger effect on the fit of regressor (c).

3.3 Denoising

A common use of confound regression (which is often implemented using the GLM) is to remove signals of no interest from the data - often called ***denoising*** the data. We introduced the idea of removing signals in the previous section to explain how the GLM statistics work. The difference here is that we are no longer interested in calculating a statistic (the uncertainty no longer matters to us), we are only interested in the signal after the contribution from noise (confounds) has been extracted. The confounds might be defined by separate information (e.g., age, motion measurements, physiological recordings, etc.) or derived from the MRI data itself (e.g., mean CSF signal, ICA components, etc.), but either way the principles of the method are the same.

If the unwanted signals are not correlated with those of interest, denoising is easy, we simply subtract off the fitted regressors that describe the 'noise'. However, when there are multiple covariates that are correlated, i.e. correlation between the 'noise' regressors and the others, then there are different ways that denoising can be done. The two main methods are: ***aggressive*** (or ***hard***) and ***non-aggressive*** (or ***soft***) denoising. If there is no correlation between the covariates then the two methods are the same.

The main difference between the two methods, for correlated covariates, is what happens to the 'shared' signal between the regressors of interest and the confound regressors. In the aggressive approach all the shared signal is treated as if it belonged to the 'noise' (i.e., confounds) and is completely removed - this is the same idea you met in the Example Box "Inference with correlated regressors". In the non-aggressive approach the shared signal is split between the covariates according to the estimates of the contribution in the GLM, which is the same concept that we met in the Example Box "Fitting correlated regressors". In this case, only the part of this signal associated with the 'noise' (confounds) is removed. As a consequence, some of the shared signal is left behind, according to the estimate from the GLM of what proportion relates to the covariates of interest. See Figure 3.4 for an illustration of these two methods.

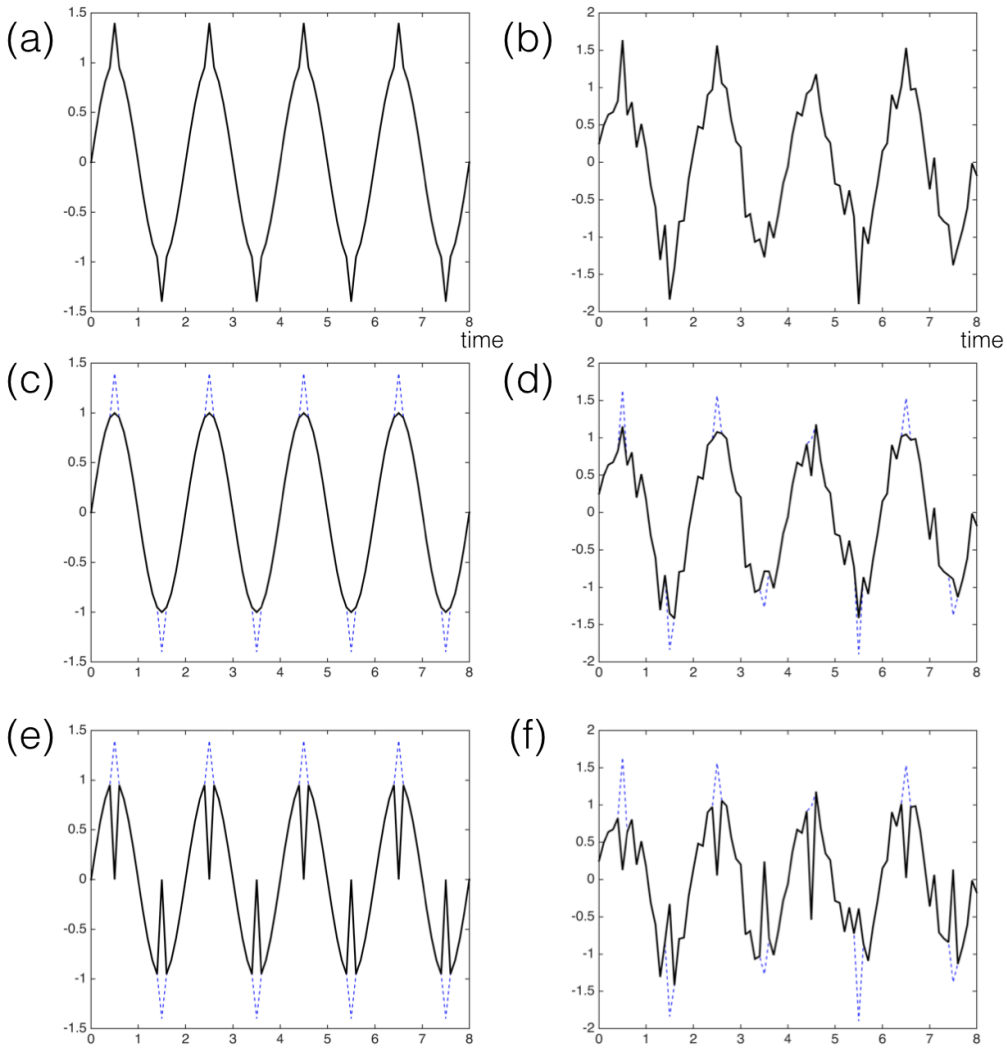


Figure 3.4: Illustration of aggressive and non-aggressive denoising. The first row shows example data without noise (a) and with noise (b), where the data consists of the same combination of visual stimulus response and head motion as shown in figure 3.2. Results of using non-aggressive (soft) denoising are shown in the second row, for the data in the row above: i.e., (c) is a denoised version of (a) and (d) is a denoised version of (b). Aggressive (hard) denoising results are shown in the third row, again based on the data in the first row. It can be seen that the aggressive denoising leads to zero or near-zero values at the points where the head motion is non-zero, whereas the non-aggressive denoising preserves the continuity across these points in time better, though its ability to reliably estimate this gets worse as the correlation between the original regressors increases.

An important point to know about denoising is that performing denoising of separate covariates in sequential steps leads to a different, and usually undesirable, outcome compared to doing a single denoising step involving all covariates together. This is because when a set of regressors are correlated then shared signal removed by one regressor may be reintroduced by another later on, often unintentionally. It is possible to correctly perform denoising with sequential steps, but the regressors in the later steps have to be adjusted to take into account the previous denoising steps - more precisely, the regressors used in any step need to be aggressively denoised by all the previous regressors (removing all the shared signals). Consequently it is a lot simpler and less error-prone to perform a single denoising step involving all the covariates together.

Box 2.1: ICA-based denoising

Denoising based on independent component analysis (ICA) is a special case of denoising since the independent components consist of both timecourses ***and*** spatial maps. For MRI data it is most common to apply ***spatial ICA*** since we have more voxels than timepoints, and this results in spatial maps that are uncorrelated (actually they are independent which is an even stronger condition, but this guarantees that they are also uncorrelated). This means that if regression is done in space, rather than time, then there is no correlation between components (regressors) and hence no difference in the denoising strategies, so that simple subtraction of the components considered as 'noise' is all that needs to be done. However, in practice some methods work directly with the timecourses. The results from using component timecourses in a GLM are the same as working with the full (space by time) components as long as ***all*** the timecourses are used together. Spatial maps resulting from the fits of this GLM will be the same as those generated by ICA and, consequently, in this case the non-aggressive denoising is the same as subtracting components. However, if not all the timecourses are used then this is equivalent to estimating the spatial maps in a different way and then the denoising is no longer constrained to the components estimated by ICA.

3.4 Demeaning

There is one particular instance of shared signal between covariates that is worth discussing as a special case. That is the mean signal (related to the intercept), which creates strong dependences and covariances unless it is removed where appropriate. The process of removing the mean is called demeaning and is commonly used to avoid the mean signal being shared amongst many covariates. Performing demeaning is very simple; the mean value of all entries in a regressor is calculated and this value is subtracted from each individual entry to create a new regressor that now has zero mean. For example, the values 8, 5, 9, 6, 7 have a mean of 7 and the demeaned values would be +1, -2, +2, -1, 0. Note that correlation analyses explicitly remove the mean, so that only fluctuations about the mean drive the correlation, but a GLM analysis does not do this automatically and so demeaning, when needed, must be done explicitly.

In all cases the mean signal needs to be either: (1) removed from the data and regressors; or (2) modelled by one or more regressors. For some datasets the mean signal is not useful or informative and in these cases the mean signal can either be removed from the data and not included in any regressors, or the mean can be left in the data and modelled as a covariate of no interest. A first-level fMRI analysis is one example of this case, as the mean value of the fMRI measurements is influenced by many physiological and scanner-related effects of no interest. If the mean signal is left in the data but not modelled then the unmodelled signal appears in the residuals and will bias the resulting statistics. Hence it is crucial to either remove the mean or model the mean signal when it is left in the data.

When the mean signal is of interest, or is present in the data, it needs to be modelled separately so that it does not affect the effect size estimate of other covariates. For example, in a between-subject analysis (e.g., functional connectivity strength) the mean is usually of great interest and is explicitly modelled by one or more regressors. If there is a single group of subjects then one mean regressor on its own (like the bar in Figure 1.1) is a common model. If there are two or more groups then the

overall mean across groups can be modelled by the same type of regressor plus another regressor for the group difference or it could be modelled by a set of regressors, one for each group. Both cases model the mean equally well and are alternative, but equivalent, options.

If other regressors are also being included in the model then it is important that these additional regressors are zero mean or otherwise they will stop the main regressors from correctly estimating the mean. For example, if age is used as a covariate of no interest then the age values should be demeaned. In this way the effects of changes in age (with respect to the average age) are corrected for, but without affecting the mean of the measurements.

Demeaning becomes more complicated when multiple groups are involved and covariates are modelled with different relationships (e.g., strength of correlations) in different groups. In this case a separate regressor is needed for each group, and there are two possible demeaning options: (1) demean the regressors separately; or (2) demean using all values (e.g., ages) as a single set.

For example, if there is a group of patients and a group of controls then one possible GLM model would include two regressors (one for each group mean) and one extra regressor to model age effects. This age regressor should be demeaned and would allow differences in the group means that were induced by age to be corrected for, whilst preserving the overall estimate of the mean across groups. Such a model assumes that the effect of age on the measured data (the slope of the linear relationship) is the same for patients and controls. Models that estimate influences of age in each group separately can also be formulated, but are beyond the scope of this introductory appendix.

Example box: Demeaning

In this box we will consider an example to provide specific illustrations of the general principles covered in this section. Take the simple case of a single group of subjects in a VBM analysis where we are interested to see if there is a relationship between a test score (e.g., a neuropsychological test, such as the mini-mental state examination; MMSE) and the gray matter density (in each location separately, as it will be a voxelwise test). The mean gray matter density across the group, which will be positive, and the change of this with test score, which we expect will be positive or zero, must be dealt with separately. As explained in the main text, the two ways of dealing with this are: (1) remove the mean from both the data and the model; and (2) model the mean and the change with test score separately. In option 1 we must demean the test scores before creating a regressor from them, as it is essential that both the data and the model have zero mean (i.e., do the same thing to the left and right hand side of the equation) - see Figure 3.5c. In option 2 it is necessary to include a mean regressor as well as a regressor formed from the test scores. This second regressor could be constructed from demeaned scores or the raw scores (prior to demeaning), however without demeaning the model does not cleanly separate the mean from the changes with test score. For example, if the mean values are left in the test scores then the "mean value" (modelled by the first regressor) represents the amount of gray matter density at a test score of zero - see Figure 3.5a. For MMSE this represents the extrapolation of the results to a subject with zero MMSE, which will not represent the group of subjects well (as studies are normally conducted with patients having MMSE of 20 or more) and so this is an unhelpful value to represent. If demeaning of the test scores is done then the first regressor represents the gray matter density corresponding what would be expected from a subject with an average MMSE (with respect to the group being studied) - see Figure 3.5b. This is therefore much more natural and useful as a model. You may see some similar analyses performed without demeaning, and for certain, specific contrasts (those only involving the test score) they lead to the same result and are completely valid. However, this is not the case for all contrasts and so using demeaned regressors creates a better model that cleanly separates effects and can be used to create interpretable hypotheses for either mean effects or for relationships with test scores. Consequently we recommend demeaning as the default approach.

Example box: **Demeaning**

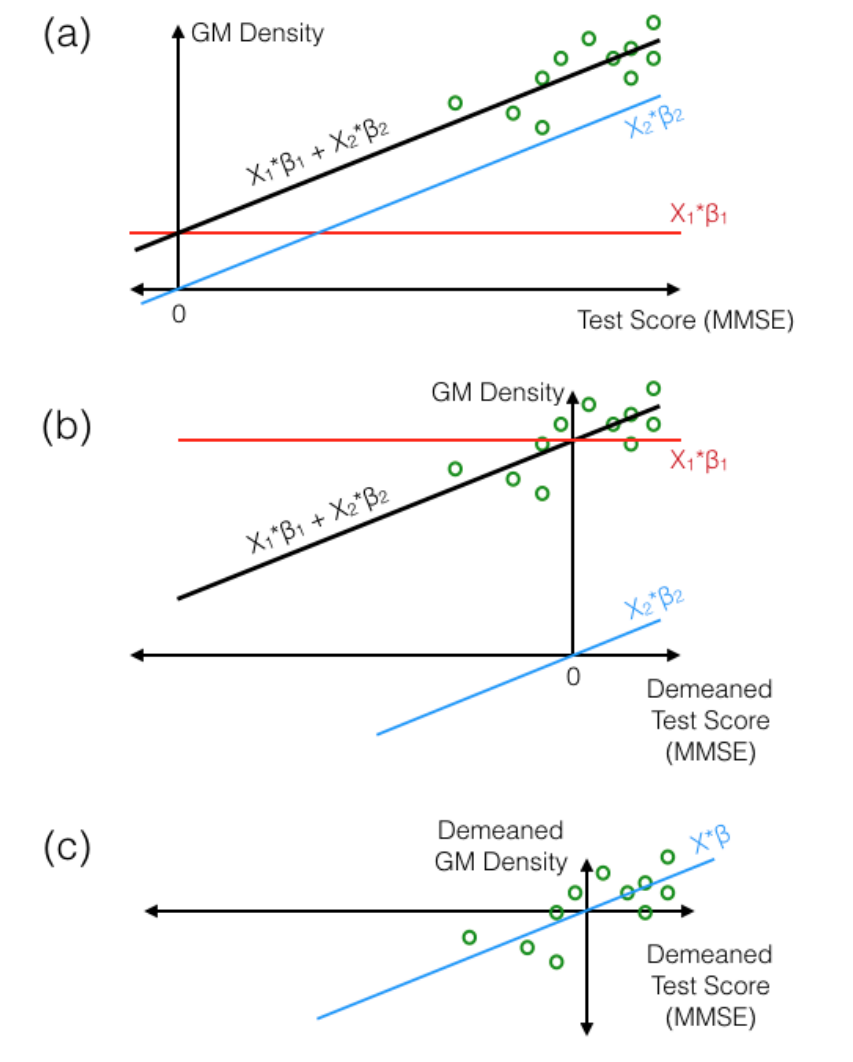


Figure 3.5: Illustration of three approaches to modelling VBM data (Gray Matter density) in relation to a test score (MMSE). In (a) two regressors are used - a mean regressor (red) and the test scores without demeaning (blue), in (b) the test scores are demeaned, and in (c) both the data and the test scores are demeaned, so that only one regressor is needed in this case. It can be seen that the slope of the linear relationship between test score and gray matter density is the same in all cases (represented by the β parameter associated with the test score regressor - shown in blue). In (a) and (b) the total fit is shown as a black line and is equal to the sum of the two individual fitted (scaled) regressors (in red and blue). In each case the blue regressor must pass through zero when the score in the regressor is zero, which leaves the mean regressor (red) with very different values: equal to the gray matter density modelled by the fitted line (black) at a test score of zero for (a) or the average test score (corresponding to a demeaned test score of zero) for (b). If the mean regressor is of interest then option (b) is the best model to use - especially if the other regressor is "correcting for" some covariate of no interest, such as age. Therefore we generally recommend demeaning in all cases as it will allow either the linear slope or the mean value (intercept) to be tested straightforwardly.

3.5 Orthogonalization

There is also another option within linear models to deal with correlations between variables, and that is ***orthogonalization***. This, however, is only rarely a good option since it makes very strong assumptions about the relationships between the correlated variables and these are typically unjustifiable. More precisely, it takes any shared signals and removes them from one of the covariates (the one being orthogonalized) in order to allow the other covariate(s) to explain these signals without sharing. This makes the covariates uncorrelated, or in other words, orthogonal. What this effectively assumes is that the shared signal could not possibly be caused by the covariate being orthogonalized, which is normally something that is not known.

Consider the two examples from the previous section: (1) activation invoked by a visual stimulus and correlated motion artefacts; and (2) effects related to disease duration and age. In the first case a signal that correlates with the visual stimulus timings or with the measured motion could be caused by either. Without more information it is not known which one caused the signal. The GLM will split any signal amongst the regressors, but also estimates its uncertainty in this split, becoming less and less certain of the split when the correlation increases. If orthogonalization is used then this removes the uncertainty in the split as all shared signals are only associated with the one regressor that is specified, i.e., chosen when doing the orthogonalization (e.g., visual stimulus). If done without proper justification, this will falsely inflate, or bias, the statistics. The same is true in the second example, as a measured signal that correlates with either disease duration or age could be due to either, without knowing any other information. Only in extremely rare circumstances can orthogonalization be justified and so it should generally be avoided.

There are two specific cases of orthogonalization are worth noting. The first case is demeaning, as orthogonalizing with respect to a mean regressor is the same as demeaning. That is, this type of orthogonalization removes the mean signal from a regressor. The second case is parametric designs when constant, linear, quadratic or other terms are included and should be made independent of each other. Orthogonalization can remove dependencies between such polynomial terms, which is desirable for easier interpretation. Although both demeaning and specifying parametric designs can be achieved using orthogonalization, it is usually simpler and less error prone to set up a parametric design with appropriate centering or, for demeaning, just to remove the mean directly from the regressors.

SUMMARY

- The GLM is a multiple regression model that consists of a set of regressors (collected together in a design matrix) that are scaled (by separate β parameters) and added together to model the signal. It is the β parameters that are fit by the GLM.
- Difference between the fitted model and the data is known as the residual noise and from this the uncertainties in the fitted β parameters can be calculated.
- GLM is used for between-subject analysis and within-subject, or timeseries, analysis.
- Fitting is done separately for each voxel in a typical voxelwise analysis (or vertexwise if the data is on a surface rather than a 3D/4D volume).
- Hypothesis testing is the basis of the stiatistical inference, where the false positive rate is set (conventionally at 0.05) while false negative rate is unknown (without some knowledge about the size of the true effects and noise).
- Some form of multiple testing correction is always required for voxelwise (or vertexwise) analyses.
- Contrasts are used to specify questions of interest = alternative hypotheses.
- A contrast (technically a ***t***-contrast) is specified by a set of numbers used to weight the parameters and add them together to form a scalar quantity, forming an inequality (for a one-sided test): e.g., $c_1\beta_1 + c_2\beta_2 > 0$.
- A statistic is formed using the contrast value and its uncertainty. The ***t***-statistic is the ratio of the contrast value and its standard error (related to the uncertainty).
- An F-test combines together a number of t-contrasts to formulate a question along the lines of whether A or B or C or any combination of them are are significantly non-zero.
- Correlation between covariates (regressors) makes the fitting and inference trickier to understand. The total fit always tries to represent all of the data as accurately as possible (minimal residual error) with shared signals split between correlated regressors.
- The individual β parameters are determined by the 'unique' parts of the signal, subject to fitting the overall signal well at all points. As the correlation between the corresponding regressors increases, the uncertainty in the split and hence the uncertainty in the estimated values of the individual β parameters also increases.
- Statistics for contrasts where the individual split is important, and the regressors are correlated, often have low power due to the increased uncertainty. Individual β values can still be high, but it is the uncertainty that reduces the statistical power.
- Denoising can use either an aggressive or a non-aggressive approach, although they are the same if the regressors are all uncorrelated. For correlated regressors the aggressive approach removes all shared signals, while the non-aggressive approach leaves in parts of the shared signals, as determined by the split of the model fit.
- Demeaning is necessary to formulate the correct hypothesis and be able to interpret the results easily. In general covariates should be demeaned and a separate regressor used to model the mean, or the mean be removed from both the data and the model regressors.
- Orthogonalization is a method that associates all shared signals with pre-specified regressors and requires proper justification (based on some external knowledge or theory) or otherwise it will bias the statistics.

FURTHER READING

- Field, A. (2017). Discovering Statistics Using IBM SPSS Statistics (5th ed.). SAGE Publications Ltd.
  - This is a general textbook on statistics, pitched at an introductory level for people with a non-technical background and based on practical illustration through SPSS.
- Poldrack, R. A., Mumford, J. A., and Nichols, T. E. (2011). Handbook of Functional MRI Data Analysis. Cambridge University Press.
  - This is a textbook primary about fMRI analysis, but includes information about the GLM as applied to neuroimaging, pitched at a slightly more technical level than this appendix.