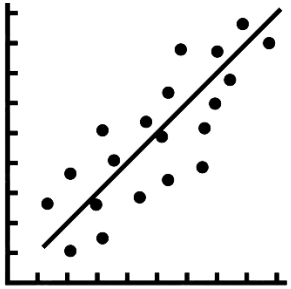
Experiment meets analysis

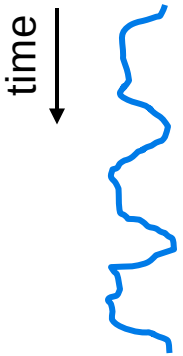
General Linear Model – Recap



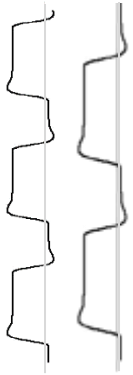
General Linear Model (GLM)

Data | Design Matrix | Beta estimates | Residuals/Error

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



Voxel time course



Regressors

One per regressor

How much does each regressor contribute to the explanation of the data?

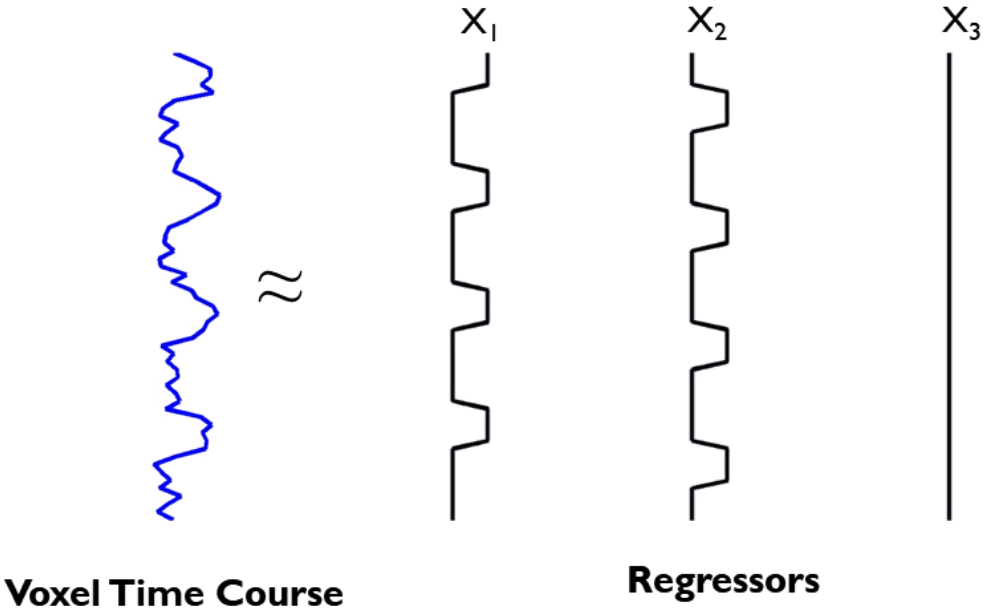
Error term

Variance that cannot be explained by the design matrix

Needs to be estimated!

General Linear Model (GLM)

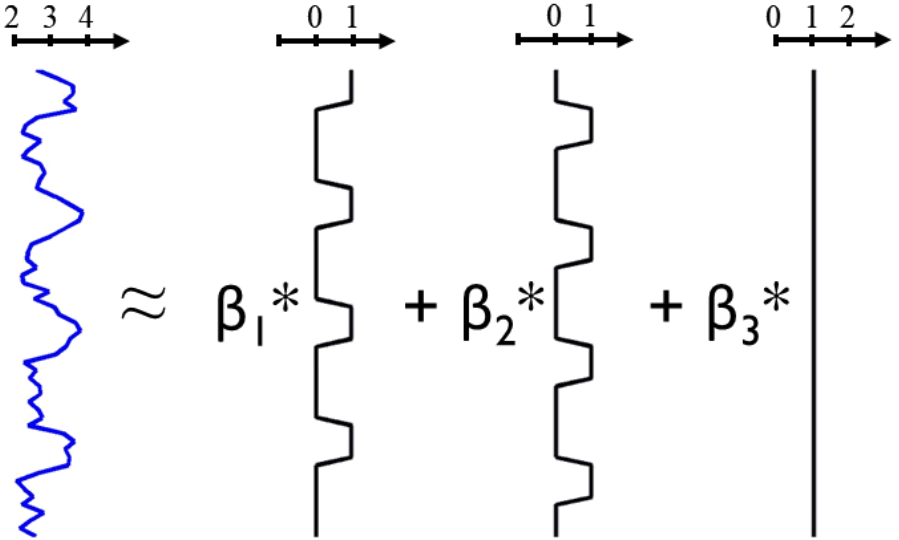
The estimation entails finding the beta estimates (β) such that the linear combination of the regressors 'best' matches the data.



Adapted from Chumbley (2008)

General Linear Model (GLM)

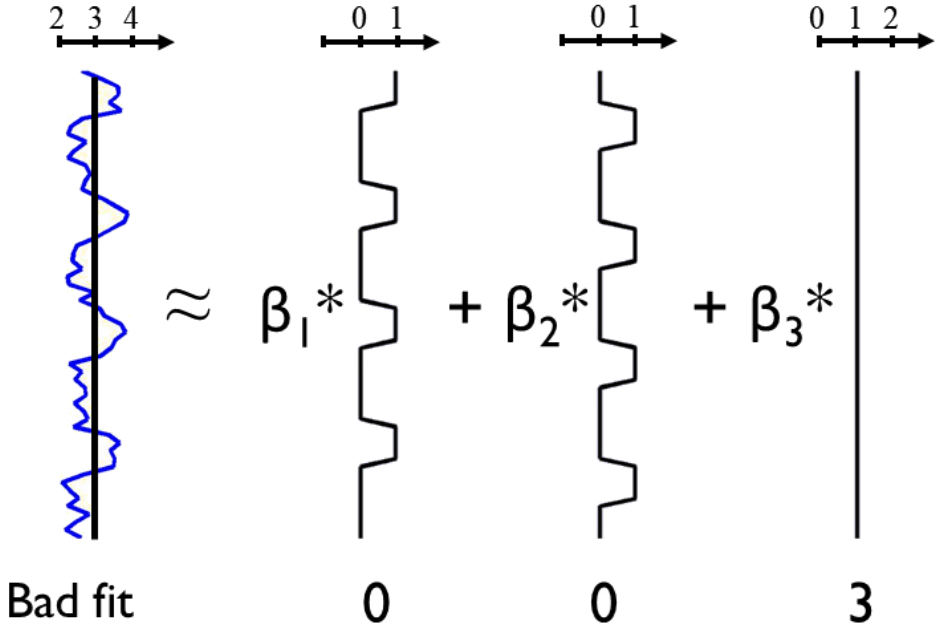
Modeling voxel time courses with a linear combination of hypothetical time-series (regressors).
Same model for each voxel → One beta estimate per regressor per voxel.



Adapted from Chumbley (2008)

General Linear Model (GLM)

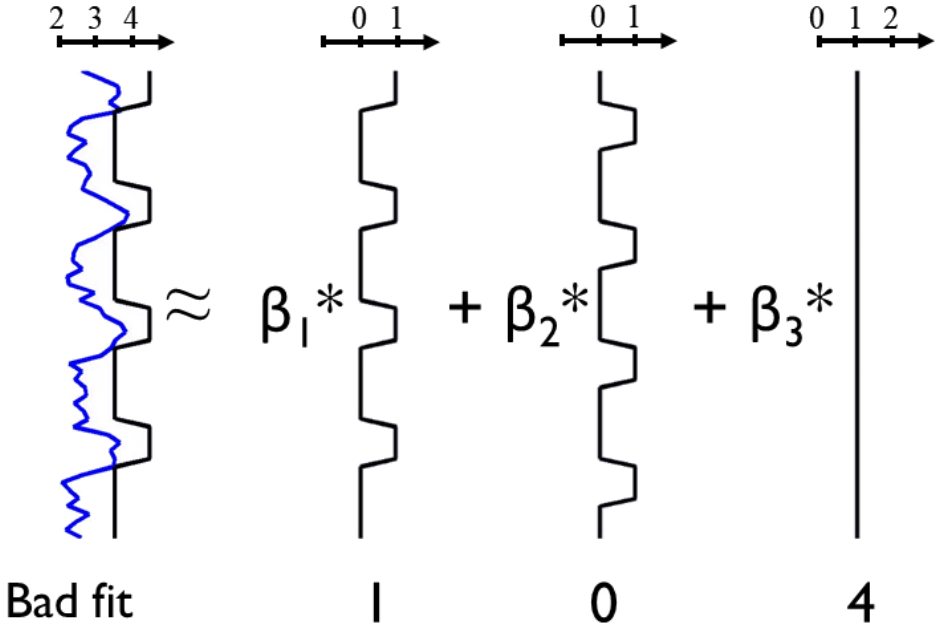
Modeling voxel time courses with a linear combination of hypothetical time-series (regressors).
Same model for each voxel → One beta estimate per regressor per voxel.



Adapted from Chumbley (2008)

General Linear Model (GLM)

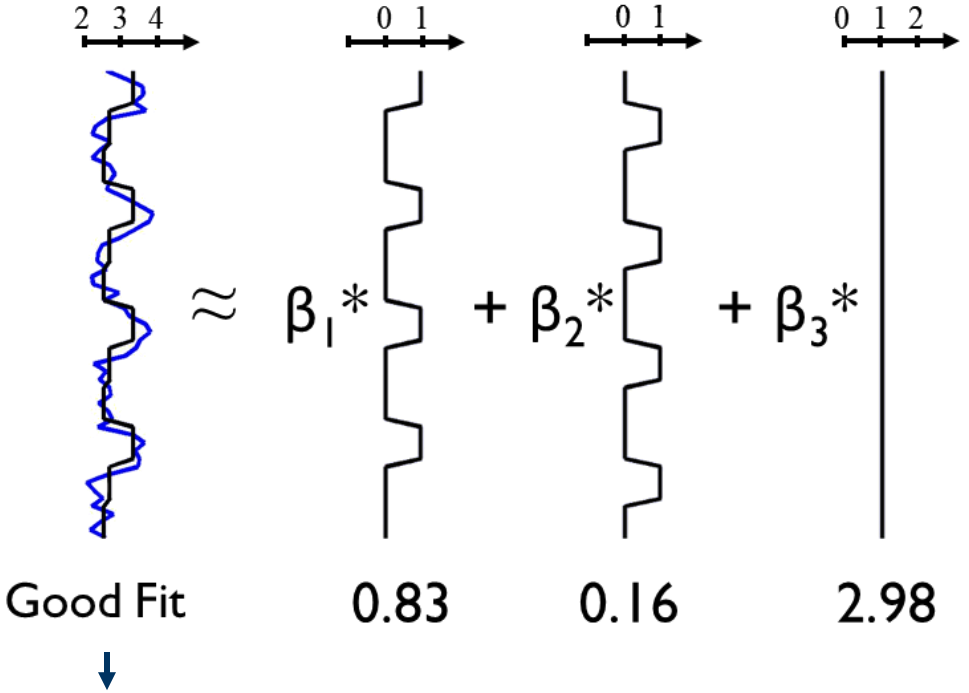
Modeling voxel time courses with a linear combination of hypothetical time-series (regressors).
Same model for each voxel → One beta estimate per regressor per voxel.



Adapted from Chumbley (2008)

General Linear Model (GLM)

Modeling voxel time courses with a linear combination of hypothetical time-series (regressors).
Same model for each voxel → One beta estimate per regressor per voxel.



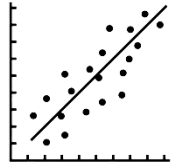
Minimal sum of squared errors

Adapted from Chumbley (2008)

GLM assumptions

Linearity

Any change in the regressor is associated with a proportional change in the data (i.e., there is a linear relationship between regressor and data)



✓ **Assumption typically met (but non-linear effects exist)**

Normality

Residuals are normally distributed

✓ **Assumption typically met (with large enough sample size)**

No multicollinearity

Regressors are independent of each other (i.e., they are uncorrelated)

✗ **Assumption often violated (e.g., HRF-convolution makes regressors more similar)**

Independence

Observations are independent of each other (e.g., different time points), and so are the residuals

Homoscedasticity

The variance of the residuals is constant across all levels of the data (e.g., all time points)

Multicollinearity

Two or more predictors in the model are highly correlated, or predictors are linear combinations of other predictors

Multicollinearity – What is the problem?

Correlated regressors explain overlapping variance in the signal

- Model coefficients (β) become unstable (i.e., small changes in the data lead to large changes in the coefficients)
- In case of perfect collinearity, there are infinite solutions to the regression (β_1=1, β_2=-1 vs. β_1=5, β_2=-5, …)
- ”Bouncing beta” effect: Model coefficients for the same regressor can be strongly positive OR strongly negative depending on the coefficients of other regressors
- **Coefficients are not reliable, and the resulting model does not generalize to new data**

Multicollinearity – How can we quantify the problem?

Look at your data!

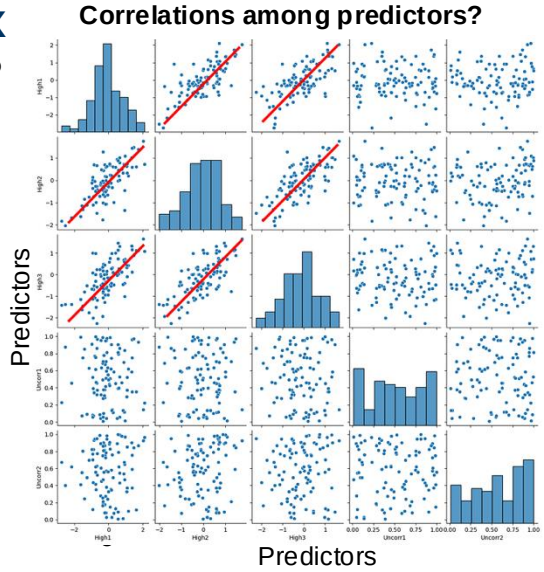
For example, are different stimulus features correlated? Are behavioral variables correlated?
These variables will cause problems in every voxel of the brain!

Look at the covariance structure of your design matrix

Are there high correlations among predictors (after HRF convolution)?
If so, ask yourself, do you need these predictors in the model?
Are important comparisons affected?

Compute variance inflation factors (VIF)

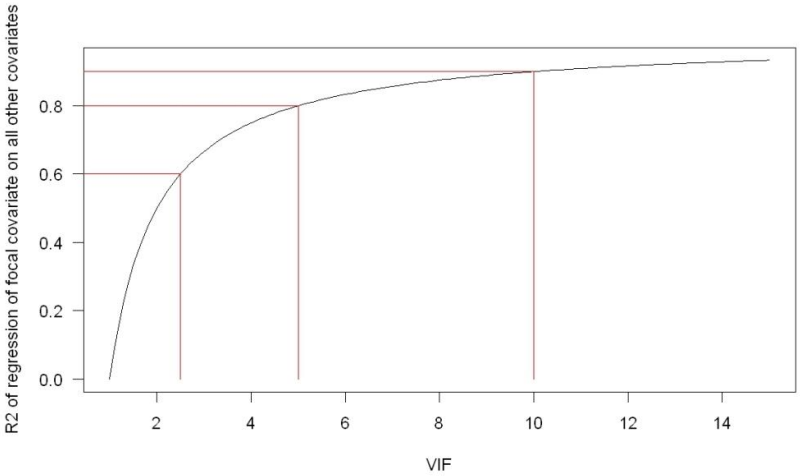
Quantify how much the variance of a regression coefficient increases due to multicollinearity



Multicollinearity – How can we quantify the problem?

Variance inflation factor (VIF):

- Quantifies how much the variance of a regression coefficient increases due to multicollinearity
- R^2 = Variance explained in a predictor by all other predictors in the model
- VIF = $\frac{1}{1-R^2}$
- VIF = 1 means no collinearity. You are good!
- VIF = 5-10 means you are in trouble (80-90% of your predictor is explained by other predictors)
- VIF > 20, close the laptop, take a holiday, think again how you want to analyze the data



Quiz:

Which of the following design matrices exhibits multicollinearity & why?

Answer: **Both!**

Design matrix 1

$$\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 2 & 0 \\ 1 & 2 & 0 \end{pmatrix}$$

Regressor2 = Regressor1 x 2

Design matrix 2

$$\begin{pmatrix} 0 & 3 & 1 \\ 0 & 3 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix}$$

Regressor2 = Regressor1 + Regressor3 x 3

Multicollinearity – Can we solve the problem?

Short answer is NO, but we can:

1) Avoid the problem before it occurs through **experimental design**

2) Compensate for the problem through **analytical strategies**

Multicollinearity – Can we solve the problem?

Experimental considerations

- **Think of the analysis before designing the experiment**
  Determine a priori which factors need to be independent

- **Orthogonal task designs**
  (e.g., vary each experimental component independently from all others, balance their combination)

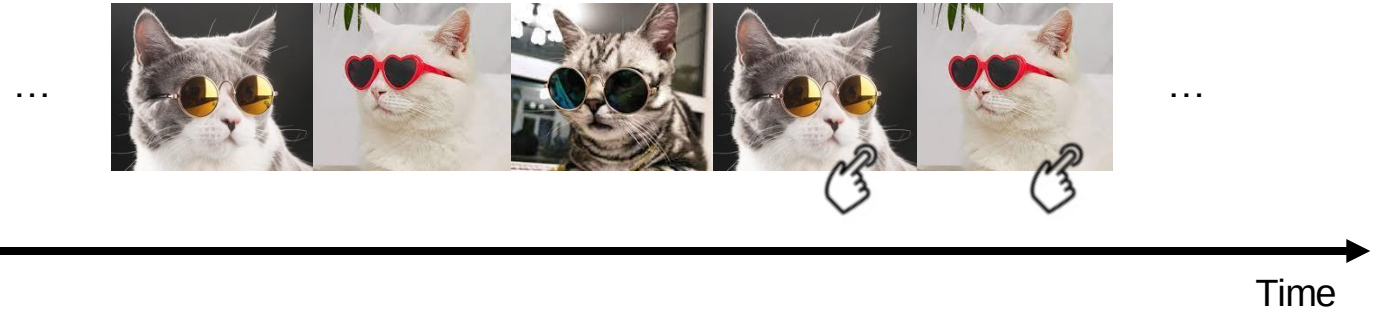
Multicollinearity – Can we solve the problem?

Experimental considerations

- **Think of the analysis before designing the experiment**
  Determine a priori which factors need to be independent

- **Orthogonal task designs**
  (e.g., vary each experimental component independently
  from all others, balance their combination)

Recognition-memory task: Click a button when an image repeats



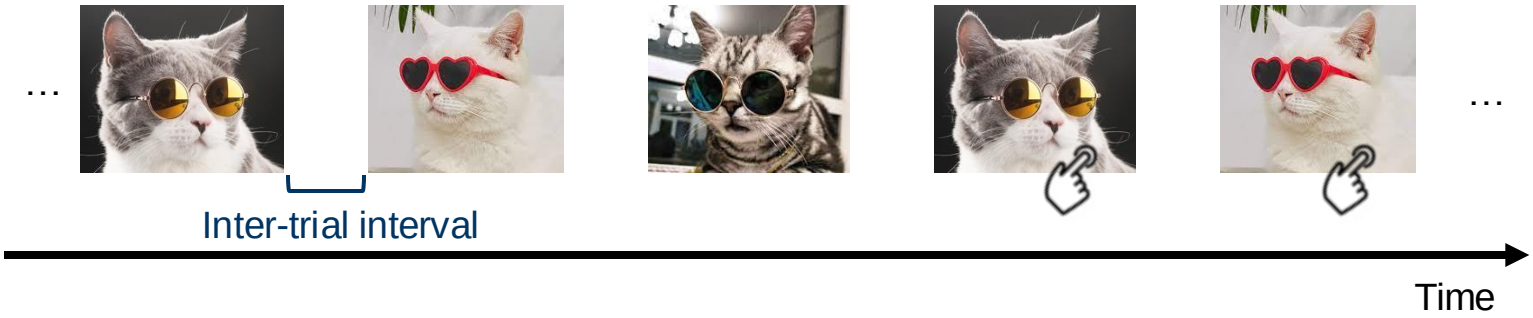
Multicollinearity – Can we solve the problem?

Experimental considerations

- **Think of the analysis before designing the experiment**
  Determine a priori which factors need to be independent

- **Orthogonal task designs**
  (e.g., vary each experimental component independently from all others, balance their combination)

Recognition-memory task: Click a button when an image repeats



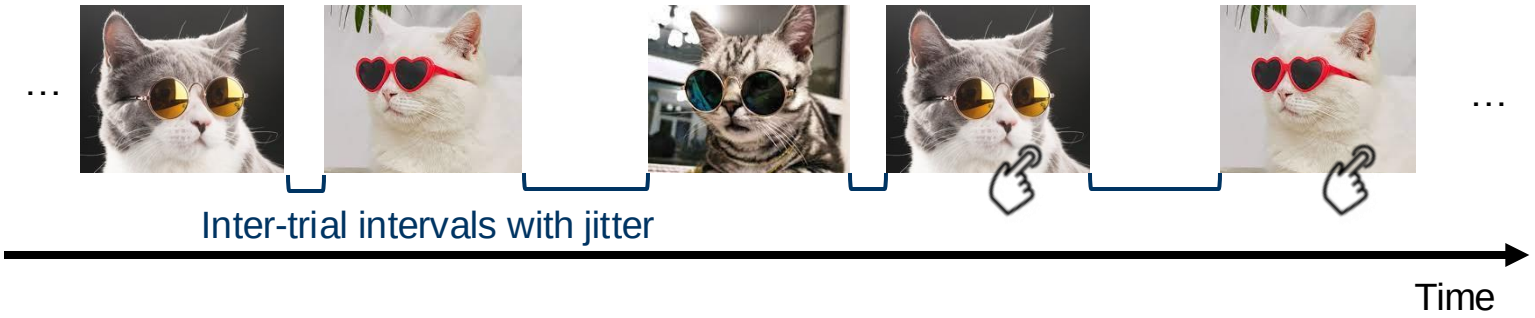
Multicollinearity – Can we solve the problem?

Experimental considerations

- **Think of the analysis before designing the experiment**
  Determine a priori which factors need to be independent

- **Orthogonal task designs**
  (e.g., vary each experimental component independently
  from all others, balance their combination)

Recognition-memory task: Click a button when an image repeats



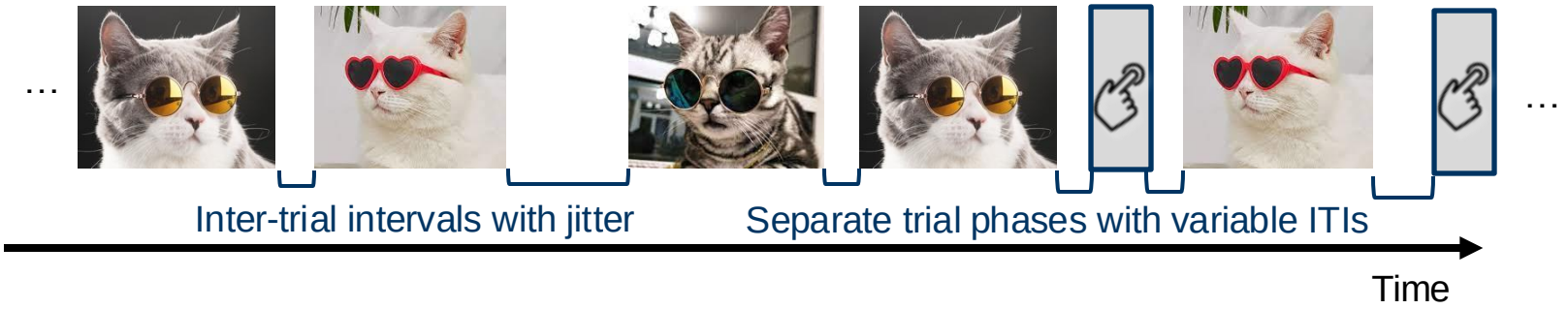
Multicollinearity – Can we solve the problem?

Experimental considerations

- **Think of the analysis before designing the experiment**
  Determine a priori which factors need to be independent

- **Orthogonal task designs**
  (e.g., vary each experimental component independently from all others, balance their combination)

Recognition-memory task: Click a button when an image repeats



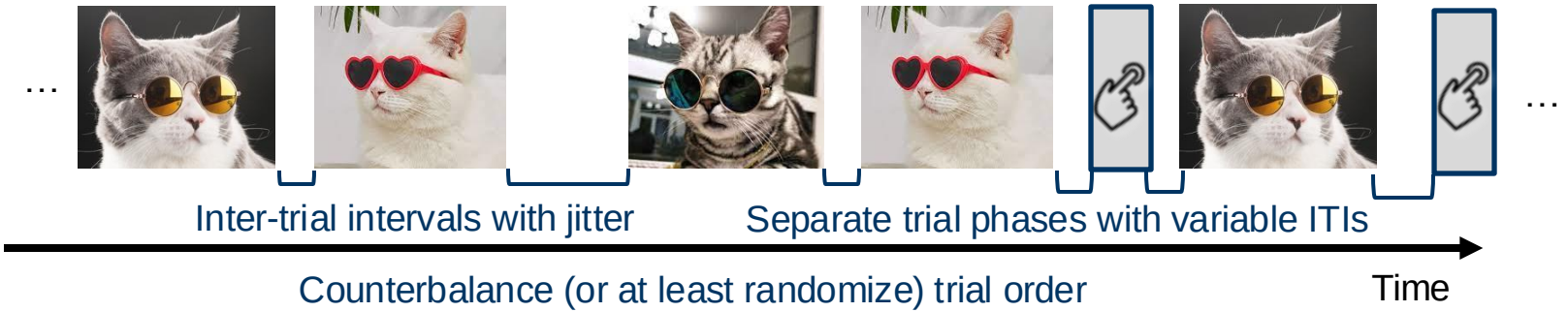
Multicollinearity – Can we solve the problem?

Experimental considerations

- **Think of the analysis before designing the experiment**
  Determine a priori which factors need to be independent

- **Orthogonal task designs**
  (e.g., vary each experimental component independently from all others, balance their combination)

Recognition-memory task: Click a button when an image repeats

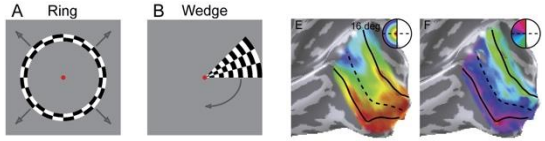


Multicollinearity – Can we solve the problem?

Experimental considerations

- **Think of the analysis before designing the experiment**
  Determine a priori which factors need to be independent
- **Orthogonal task designs**
  (e.g., vary each experimental component independently from all others, balance their combination)
- **Separate conditions in time**
  Add inter-trial intervals with jitter, separate task phases (e.g., stimuli & button clicks)
- **Counterbalance trial order**
  Ensure that each condition precedes each other condition equally often (at least randomize order)
- **Block designs**
  Group together trials of a certain condition to separate them from trials of another condition (unlike event-related designs)

Example experiment: Visual field mapping



Dumoulin et al. 2008

Correlations among predictors minimal by design

Only applicable to certain research questions & highly controlled tasks
(e.g., naturalistic tasks often lack control, behavioral variables often covary…)

See Lecture 2

Multicollinearity – Can we solve the problem?

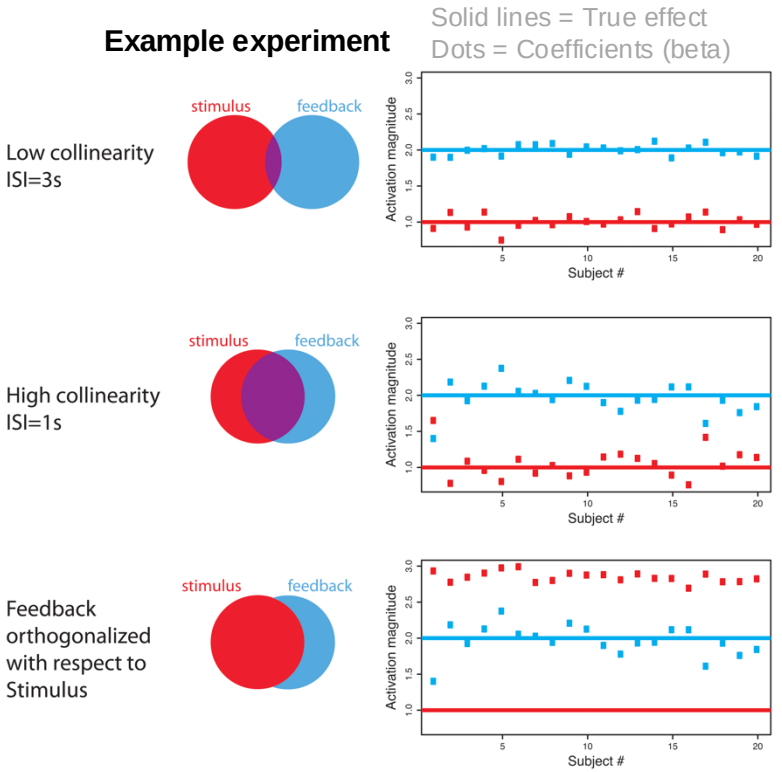
Analytical considerations

- **Reduce model complexity**
  Remove predictors that are not needed.
  Rule of thumb: n_regressors < n_datapoints/20

- **Orthogonalization of regressors**
  Decide which predictor gets credit for explaining overlapping variance

  Orthogonalization <u>can be appropriate</u>,
  (i.e., for covariate regressors of a main regressor)

  Orthogonlization <u>can be misleading</u>
  (e.g., difference between model coefficients is ”not real” but reflects your decision)



<u>Read this:</u> Orthogonlization of regressors in fMRI models. Mumford et al. 2015, PLOS ONE

Multicollinearity – Can we solve the problem?

Analytical considerations

- **Reduce model complexity**
  Remove predictors that are not needed.
  Rule of thumb: n_regressors < n_datapoints/20

- **Orthogonalization of regressors**
  Decide which predictor gets credit for explaining overlapping variance

- **Regularized regression (e.g., Ridge regression)**
  ”Penalty term” (λ) added to the GLM that ”shrinks” coefficients, with larger coefficients being compressed more

  Penalty term needs to be estimated (λ) through cross-validation (i.e., data needs to be split into Training vs. Test set)

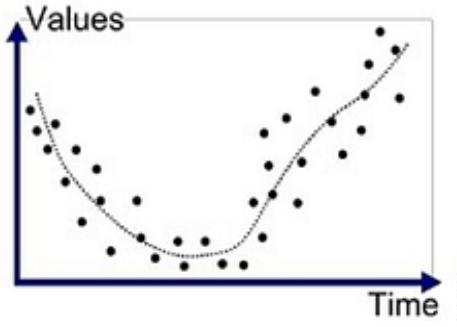
  Model fits the training data less well, but it generalizes better to new data (Popular in machine learning to reduce overfitting)

Read more: Dupré la Tour et al. 2022

Residual sum of squares (RSS)

$$RSS_{L2} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \lambda \sum_{j=1}^{P} B_j^2$$

“Regular” GLM Penalty

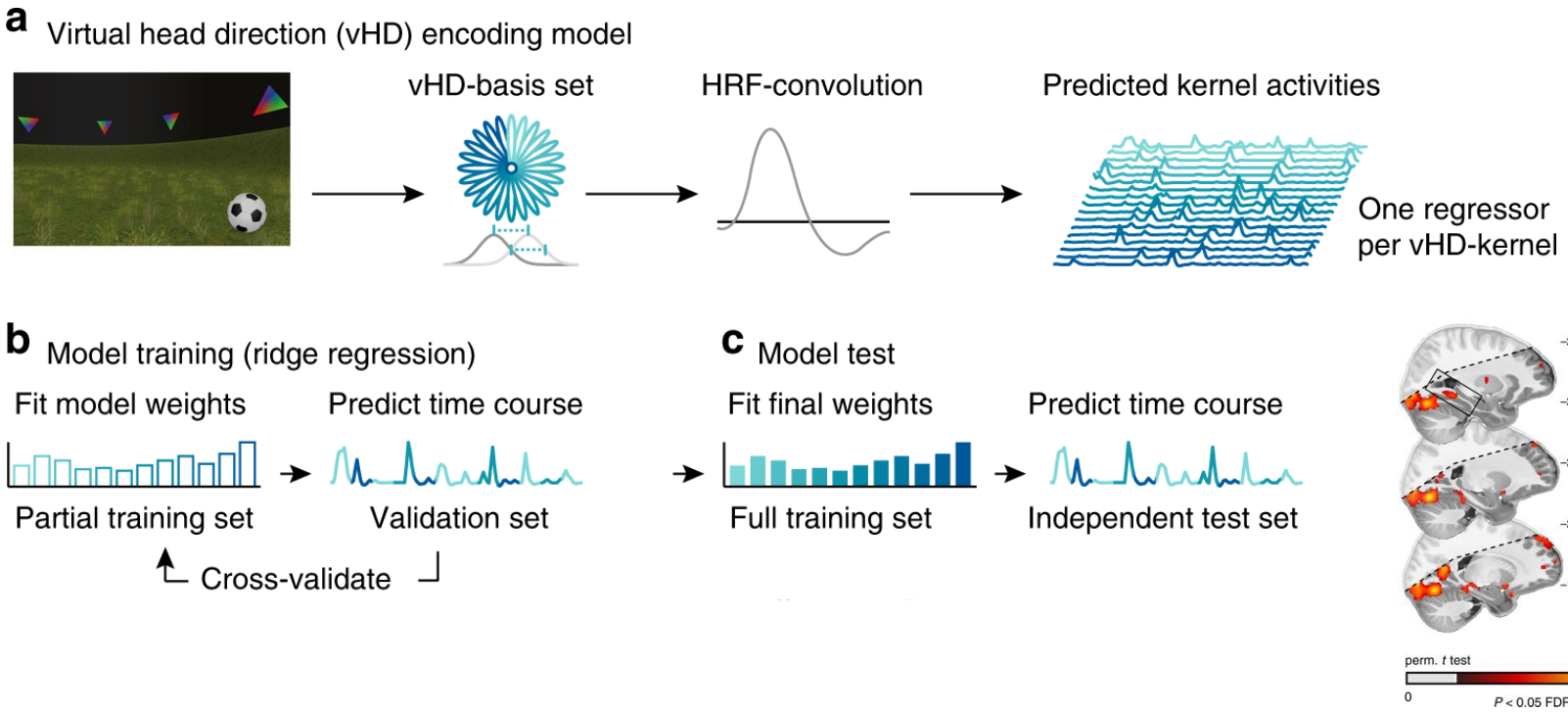


Good Fit/Robust

Multicollinearity – Can we solve the problem?

Regularization – Example: Ridge regression links fMRI & navigation behavior

Nau et al. 2020



Multicollinearity – Can we solve the problem?

Analytical considerations

- **Reduce model complexity**
  Remove predictors that are not needed.
  Rule of thumb: n_regressors < n_datapoints/20

- **Orthogonalization of regressors**
  Decide which predictor gets credit for explaining overlapping variance

- **Regularized regression (e.g., Ridge regression)**
  ”Penalty term” (λ) added to the GLM that ”shrinks” coefficients, with larger coefficients being compressed more

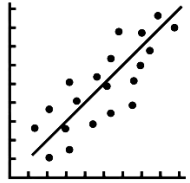
- **Dimensionality reduction**
  Find principal components of design matrix & fit those to the data

”There is no free lunch”
Each experimental design choice & method against multicollinearity has pros & cons!

GLM assumptions

Linearity

Any change in the regressor is associated with a proportional change in the data (i.e., there is a linear relationship between regressor and data)



✓ **Assumption typically met (but non-linear effects exist)**

Normality

Residuals are normally distributed

✓ **Assumption typically met (with large enough sample size)**

No multicollinearity

Regressors are independent of each other (i.e., they are uncorrelated)

✗ **Assumption often violated** **But experimental and/or analytical techniques can help** ✓

Independence

Observations are independent of each other (e.g., different time points), and so are the residuals

✗ **Assumption often violated (e.g., due to temporal autocorrelations)**

Homoscedasticity

The variance of the residuals is constant across all levels of the data (e.g., all time points)

Temporal autocorrelation

The signal is correlated with a delayed version of itself, meaning that each value in the time series can be predicted based on the values that came before. Also known as **serial dependence**.

Temporal autocorrelation – What is the problem?

Observations are not independent

- Samples acquired close in time are very similar (e.g., because of the HRF)
- The amount of independent information in the data is reduced
- Degrees of freedom are overestimated, leading standard errors to be underestimated
- Autocorrelation leads to inflated t-Statistic and to an increase in False Positive results

Degrees of freedom

$$df = N - P$$

Standard error

$$\hat{\sigma} = \sqrt{\frac{\Sigma_{i=1}^{N}(\hat{y}_i - y_i)^2}{df}}$$

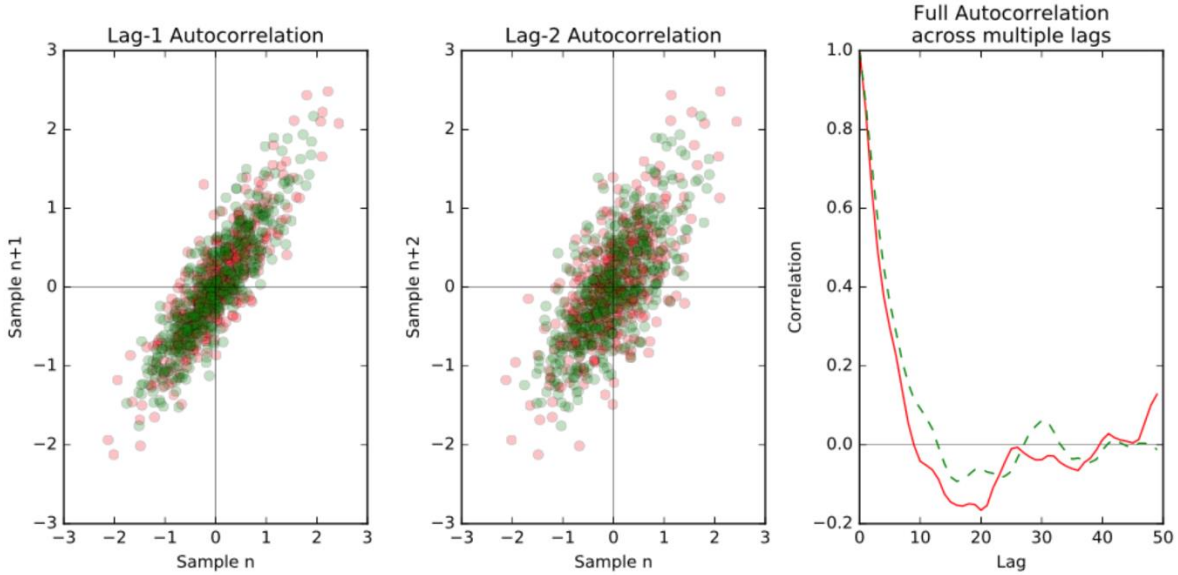
t-Statistic

$$t = \frac{\hat{\beta}}{\hat{\sigma}}$$

Temporal autocorrelation – How can we quantify the problem?

Compute Autocorrelogram

- Correlate time series with a delayed version of itself
- Do this for all possible delays
- Inspect the resulting curve (i.e., the autocorrelogram for all delays)



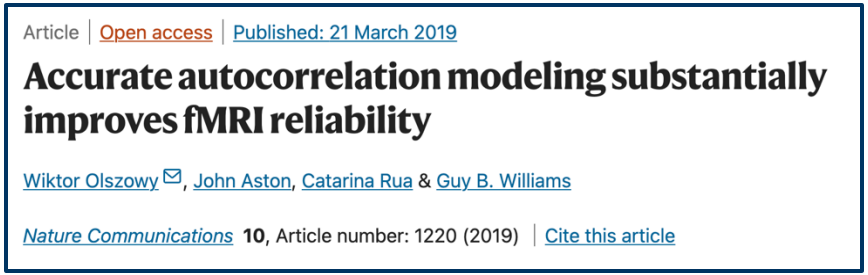
Temporal autocorrelation – Can we solve the problem?

Prewhitening

- Remove autocorrelation by transforming the data such that the residuals resemble white noise (thus the name “prewhitening”)

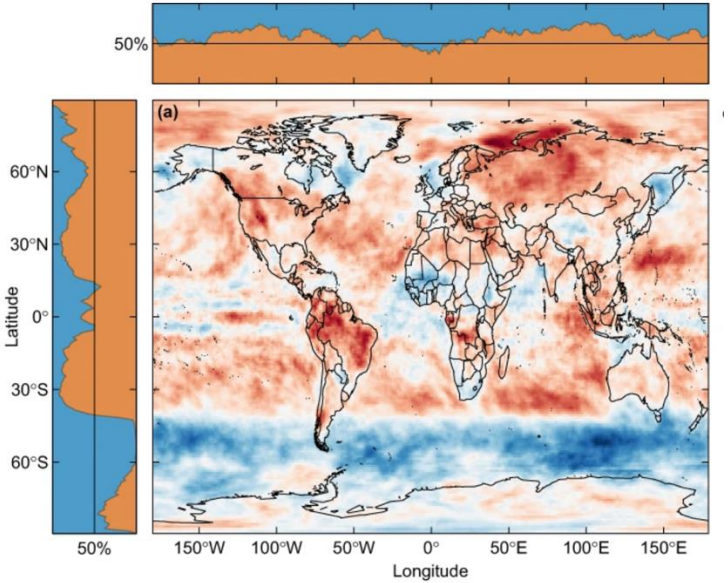
- Different software packages (FSL, SPM, AFNI…) use different algorithmic solutions

- Standard recipe
  1) Fit a GLM model
  2) Compute residual autocorrelation
  3) Correct residual autocorrelation (e.g., through filtering)
  4) Add the uncorrected residuals to the 'explained (fitted) signal’
  5) Re-run the GLM on corrected data



Temporal autocorrelation as a feature, not a bug

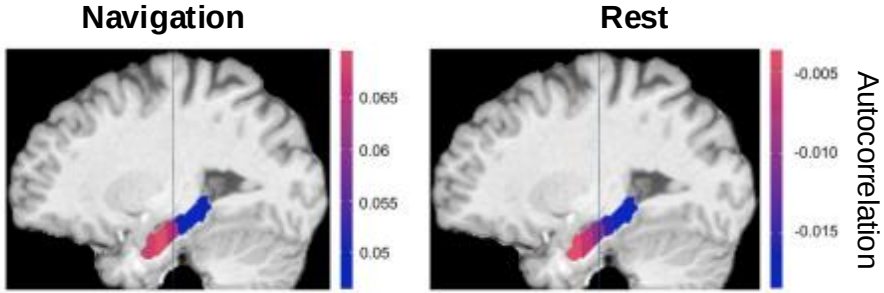
Example 1:
Autocorrelation in global temperature is changing (Climate change?)



Autocorrelation **Increase** / **Decrease** over the years

Cecco & Gouhier 2018

Example 2:
Autocorrelation in fMRI reflects task demands



Brunec et al. 2018

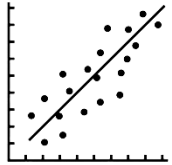
Take home
Check for autocorrelations in your data!

They might speak towards your research question or cause problems (e.g., violating assumptions)

GLM assumptions

Linearity

Any change in the regressor is associated with a proportional change in the data (i.e., there is a linear relationship between regressor and data)



✓ **Assumption typically met (but non-linear effects exist)**

Normality

Residuals are normally distributed

✓ **Assumption typically met (with large enough sample size)**

No multicollinearity

Regressors are independent of each other (i.e., they are uncorrelated)

✗ **Assumption often violated** **But experimental and/or analytical techniques can help** ✓

Independence

Observations are independent of each other (e.g., different time points), and so are the residuals

✗ **Assumption often violated** **But correcting for autocorrelations in the data can help** ✓

Homoscedasticity

The variance of the residuals is constant across all levels of the data (e.g., all time points)

✗ **Assumption often violated (e.g., noise & physiological signals such as breathing can vary over scan duration)**

Heteroscedasticity

Physiological or thermal noise can vary over scan duration (e.g., head motion), leading to variations in the variance of residuals

Heteroscedasticity

What is the problem?

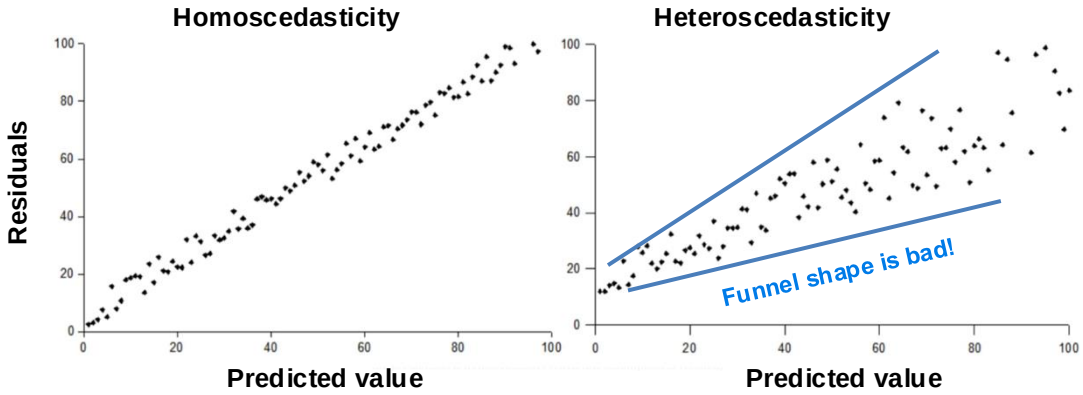
- Variance in residuals may change over time
- Standard errors differing between conditions etc., leading to biased t-statistic

How can we detect the problem?

- Plot residuals over predicted values of regression model

Can we solve the problem?

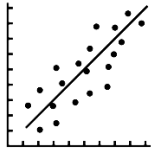
- Correct metrics for residual variance (e.g., Weighted least squares, Robust standard errors)



GLM assumptions

Linearity

Any change in the regressor is associated with a proportional change in the data (i.e., there is a linear relationship between regressor and data)



✓ **Assumption typically met (but non-linear effects exist)**

Normality

Residuals are normally distributed

✓ **Assumption typically met (with large enough sample size)**

No multicollinearity

Regressors are independent of each other (i.e., they are uncorrelated)

✗ **Assumption often violated** **But experimental and/or analytical techniques can help** ✓

Independence

Observations are independent of each other (e.g., different time points), and so are the residuals

✗ **Assumption often violated** **But correcting for autocorrelations in the data can help** ✓

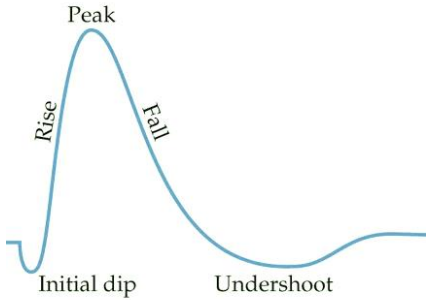
Homoscedasticity

The variance of the residuals is constant across all levels of the data (e.g., all time points)

✗ **Assumption often violated** **But robust regression and robust standard errors help** ✓

General take aways

- The general linear model makes certain assumptions
- When assumptions are violated, statistical metrics can be over/underestimated
- Experimental design and analyses techniques can help reduce these problems
- fMRI researchers need to be extra aware of multicollinearity and autocorrelations
- The hemodynamic response effectively “smooths” the time series, bluring lines between trials and making regressors similar



Key terms to remember

- General linear model
- Design matrix
- Beta estimates
- Error term
- Linearity
- Normality
- Multicollinearity
- «Bouncing beta effect»
- Variance inflation factor
- Orthogonalization of regressors
- Counterbalancing
- Regularization
- Ridge regression
- Independence of observations
- Temporal autocorrelation
- Prewhitening
- Homoscedascticity
- Heteroscedascitic

Image: amberusa.com

Happy scanning!