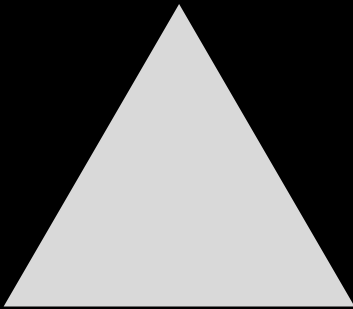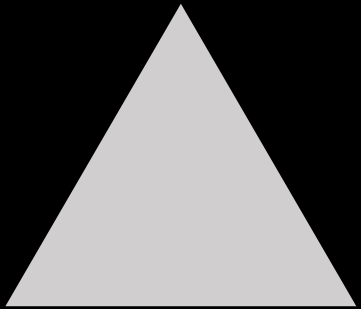# Building a Model

Covariance Matrices, Assumptions, and Model Construction
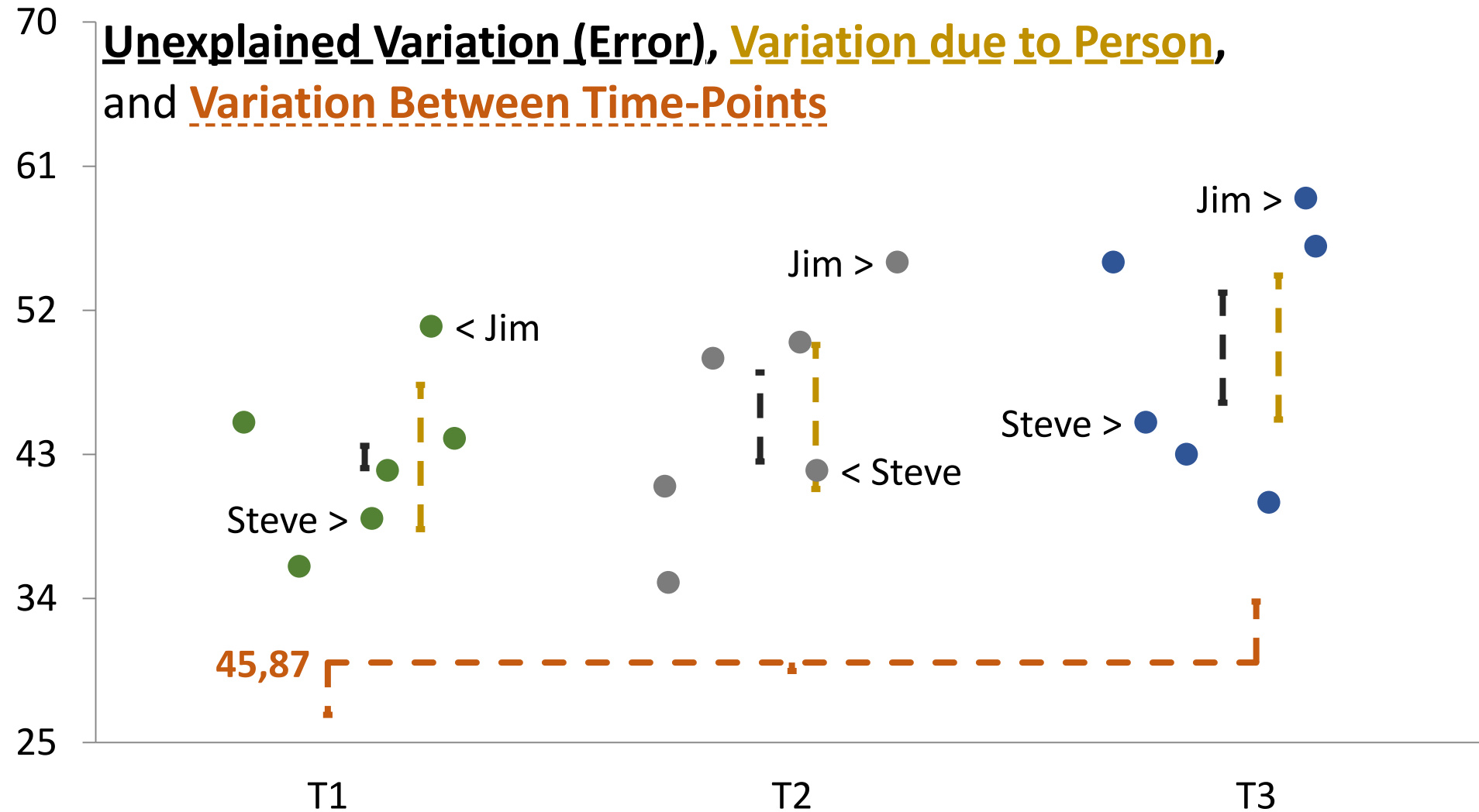
Welcome

This Session

- What Covariance Matrices are
- Which matrices are important to check
- The assumptions of Mixed Models
- Building Models and check for the best fit
- How to describe your model and results

01 Covariance Matrix

The RM ANOVA

Unexplained Variation (Error), Variation due to Person, and Variation Between Time-Points

Sphericity

The effect of person is always the same

It's not about variation between individuals but about the correlation between replications

Compound Symmetry

All Variances and Covariance as the same

The **covariance between T1 and T2** is the same as the **covariance between T1 and T3** which is the same as the **covariance between T2 and T3**

Compound Symmetry

$$Covariance\ Matrix: \begin{bmatrix} Var_1 & Cov_{1,2} & Cov_{1,3} \\ Cov_{1,2} & Var_2 & Cov_{2,3} \\ Cov_{1,3} & Cov_{2,3} & Var_3 \end{bmatrix}$$

$$Compound\ Symmetry: \sigma^2 \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

Sphericity

Repeated Measures ANOVA uses sphericity, which is Compound Symmetry Light. If you have Compound Symmetry you have Sphericity, but not the other way around

$$VarCovar1 = \begin{matrix} & T1 & T2 & T3 \\ T1 & \\ T2 & \\ T3 & \end{matrix} \begin{bmatrix} 20 & 10 & 10 \\ 10 & 20 & 10 \\ 10 & 10 & 20 \end{bmatrix} \qquad VarCovar2 = \begin{matrix} & T1 & T2 & T3 \\ T1 & \\ T2 & \\ T3 & \end{matrix} \begin{bmatrix} 10 & 5 & 10 \\ 5 & 20 & 15 \\ 10 & 15 & 30 \end{bmatrix}$$
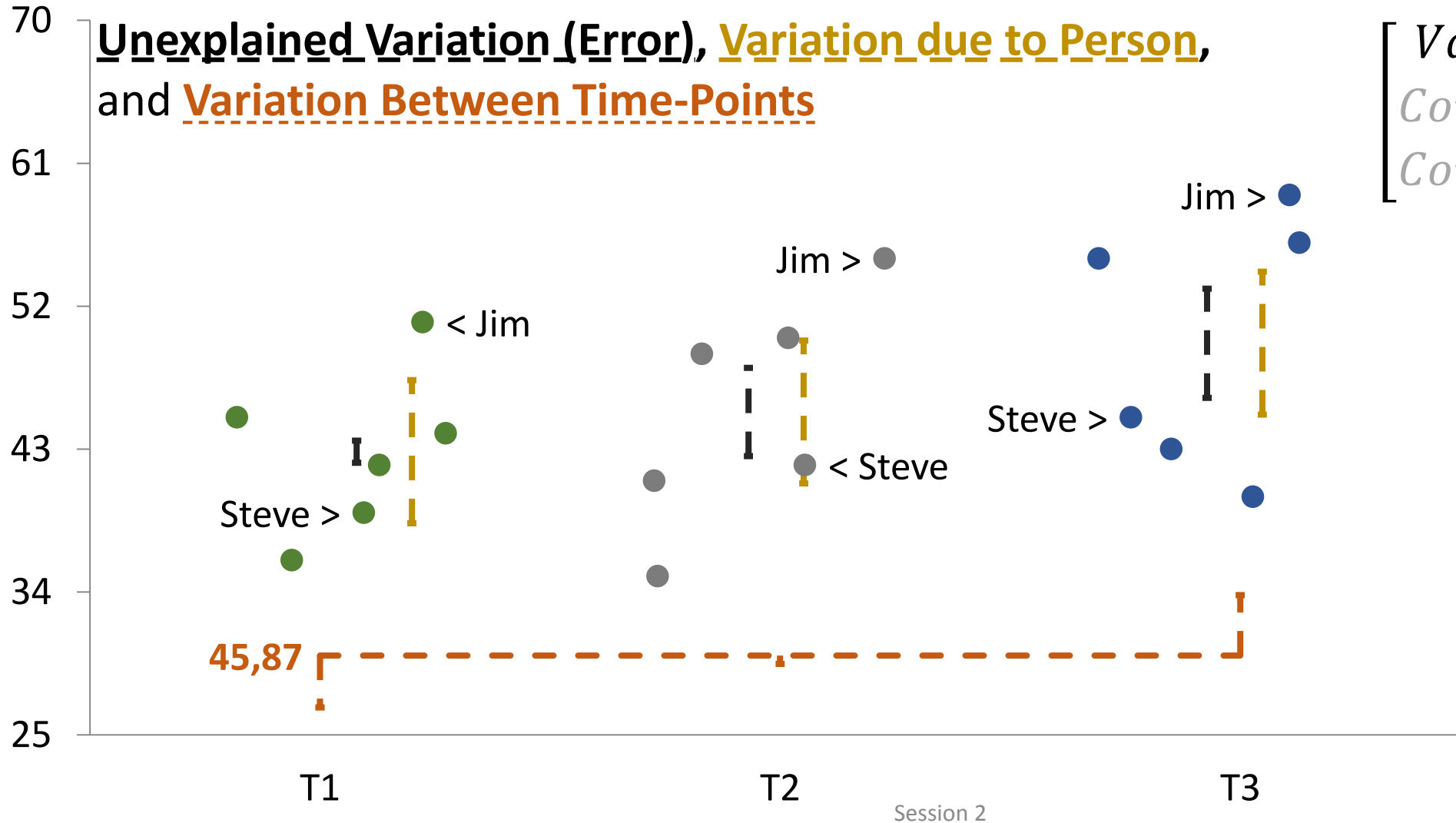
$$S_{x-y}^2 = S_x^2 + S_y^2 - 2(S_{xy})$$

$$Var_{diff} = Var_{T1} + Var_{T2} - 2(Cov_{T1;T2})$$

$$Var_{T1-T2} = 20 + 20 - (2 * 10) = 20$$

$$Var_{T1-T2} = 10 + 20 - (2 * 5) = 20$$

The RM ANOVA

The Multivariate Test (MANOVA) had no assumptions about the dependence between replications. It uses an **Unstructured Covariance Matrix**

The Univariate Test assumes the same people do the same test, it assumes a high dependence between replications. It uses a **Sphericity Structure (Compound Symmetry Light)**

$$Covariance\ Matrix: \begin{bmatrix} Var_1 & Cov_{1,2} & Cov_{1,3} \\ Cov_{1,2} & Var_2 & Cov_{2,3} \\ Cov_{1,3} & Cov_{2,3} & Var_3 \end{bmatrix}$$

Mixed Models

Mixed Models don't assume anything, they hand it over to you, make you decide what structure is best

$$\begin{bmatrix} Var_1 & Cov_{1,2} & Cov_{1,3} \\ Cov_{1,2} & Var_2 & Cov_{2,3} \\ Cov_{1,3} & Cov_{2,3} & Var_3 \end{bmatrix}$$

- **Compound Symmetry (Exchangeable):** assumes equal covariances meaning any order is fine
- **Autoregressive (AR1):** covariance is strong in adjacent points, decreasing with distance (such as equally spaced time)
- **Toeplitz:** covariance depends on the distance, this works well for ordinal data (such as time)
- **Unstructured:** has no constraints, most complex, but will always provide the best estimate

- **Heterogenous Structures:** have the same constraints on covariances, but let the variances vary
- **Correlation Structures:** transform the covariances in correlations for easier interpretation

$$CS = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \rho \\ & 1 & \rho & \rho \\ & & 1 & \rho \\ & & & 1 \end{bmatrix}$$

$$AR1 = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ & 1 & \rho & \rho^2 \\ & & 1 & \rho \\ & & & 1 \end{bmatrix}$$

$$TP = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ & 1 & \rho_1 & \rho_2 \\ & & 1 & \rho_1 \\ & & & 1 \end{bmatrix}$$

$$UN = \begin{bmatrix} \sigma_1^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} \\ & \sigma_2^2 & \sigma_{32} & \sigma_{42} \\ & & \sigma_3^2 & \sigma_{43} \\ & & & \sigma_4^2 \end{bmatrix}$$

Why not always use Unstructured?

Unstructured works well if there's **no clear pattern** in your matrix. If there is, a different structure will save you **degrees of freedom**.

-or-

If you have **a lot of replications** you won't have enough data to estimate everything
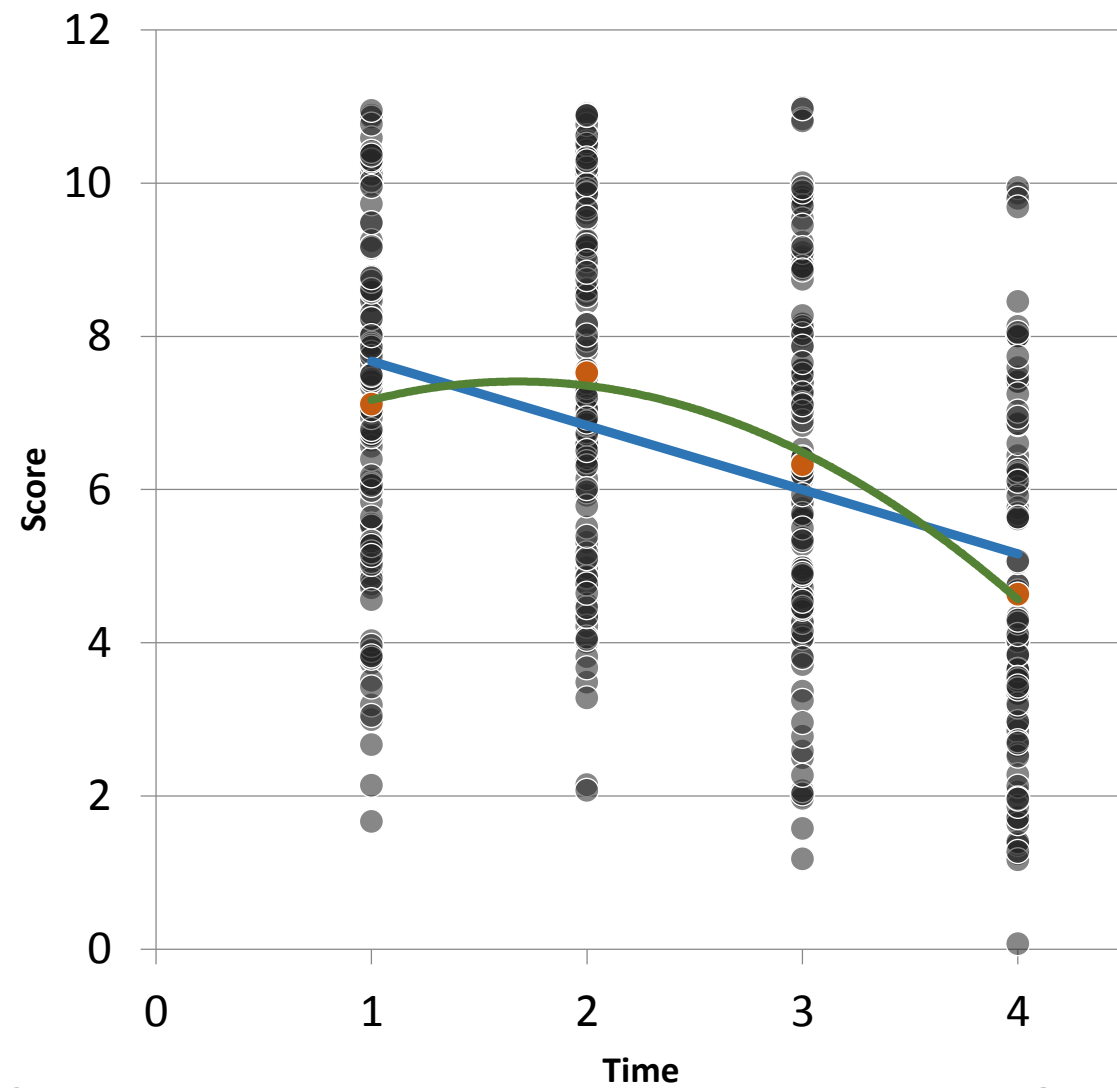
02

Model Assumptions

Assumption 1: Linearity

All Covariates have a Linear Relationship with the dependent
(or at least not a non-linear relationship)

Linearity
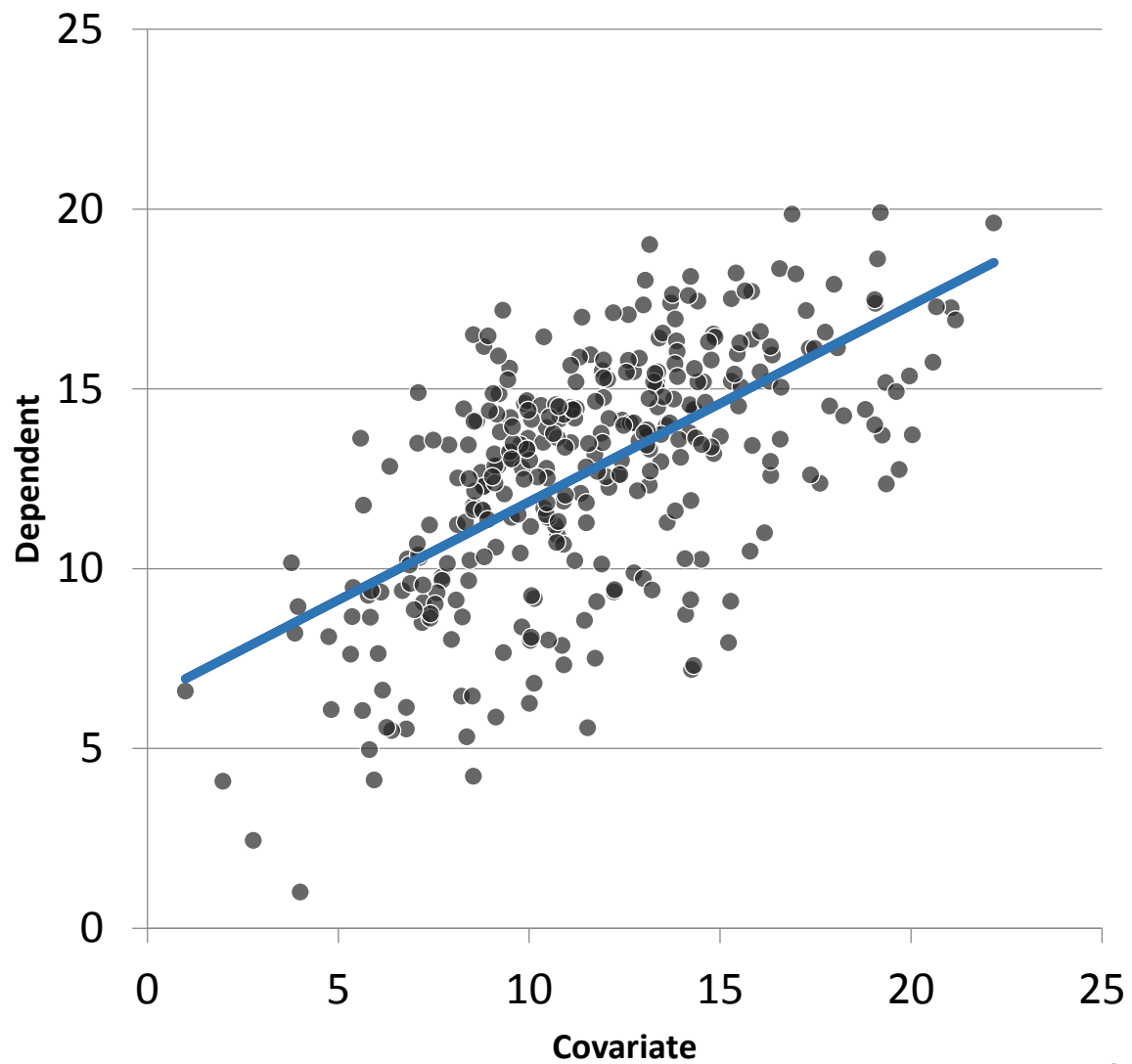
Assumption 2: Homoscedasticity

Residuals are equally distributed across predicted values
(the model is not better or worse at specific value ranges)

Homoscedasticity
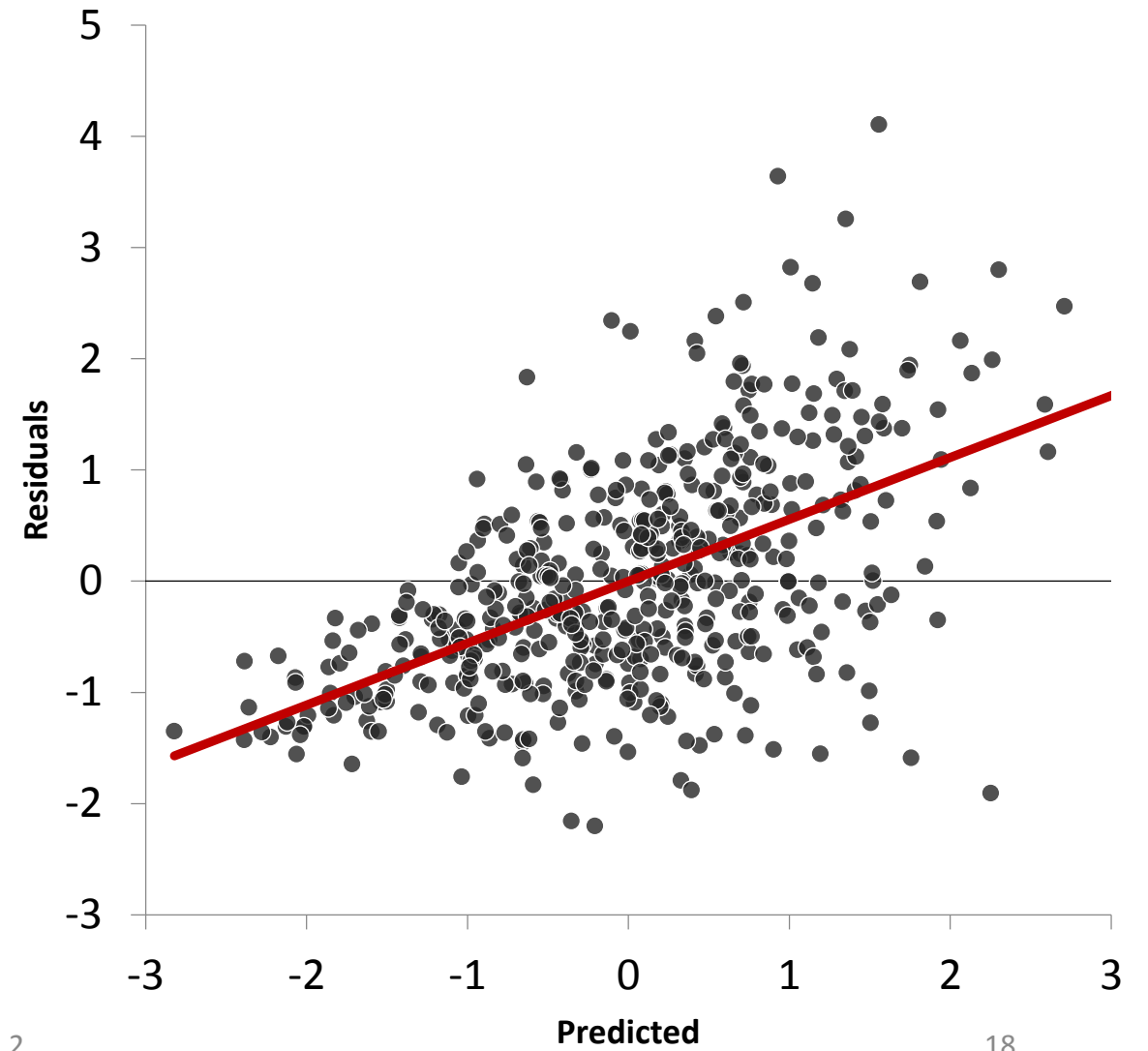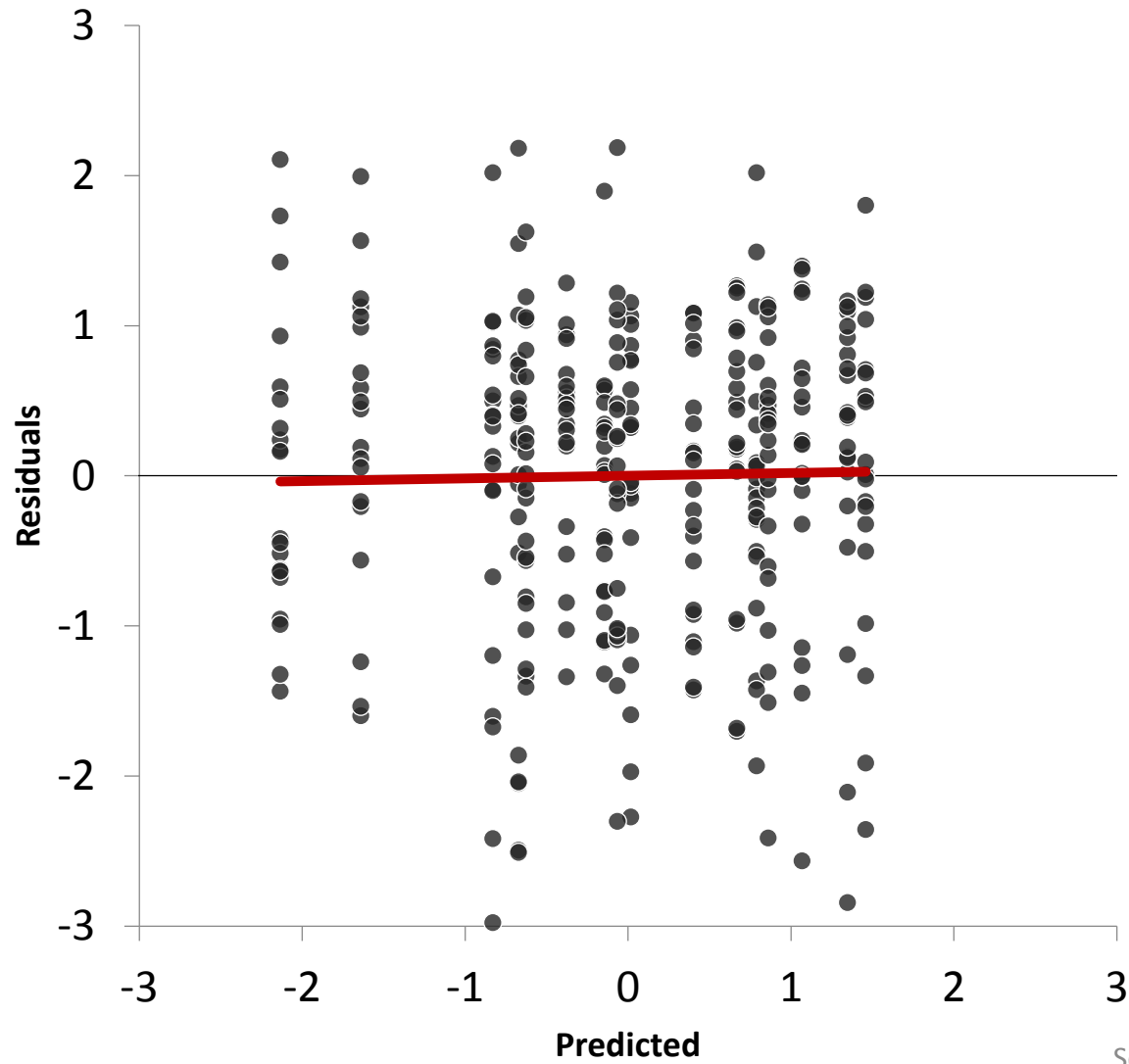
Assumption 3: Multicollinearity

Predictors are not correlated too much
(you're not predicting with the same data)

Multicolinearity

Correlations

	Predictor 1	Predictor 2	Predictor 3
Predictor 1	1	.622	.330
Predictor 2	.622	1	.264
Predictor 3	.330	.264	1

Correlations

	Predictor 1	Predictor 2	Predictor 3
Predictor 1	1	.967	.456
Predictor 2	.967	1	.345
Predictor 3	.456	.345	1

Rule of Thumb: Correlation Should be <.70

Assumption 4: Outliers

No observations have extremely Low or High residuals
(the model is not affected by strange data)

Outliers

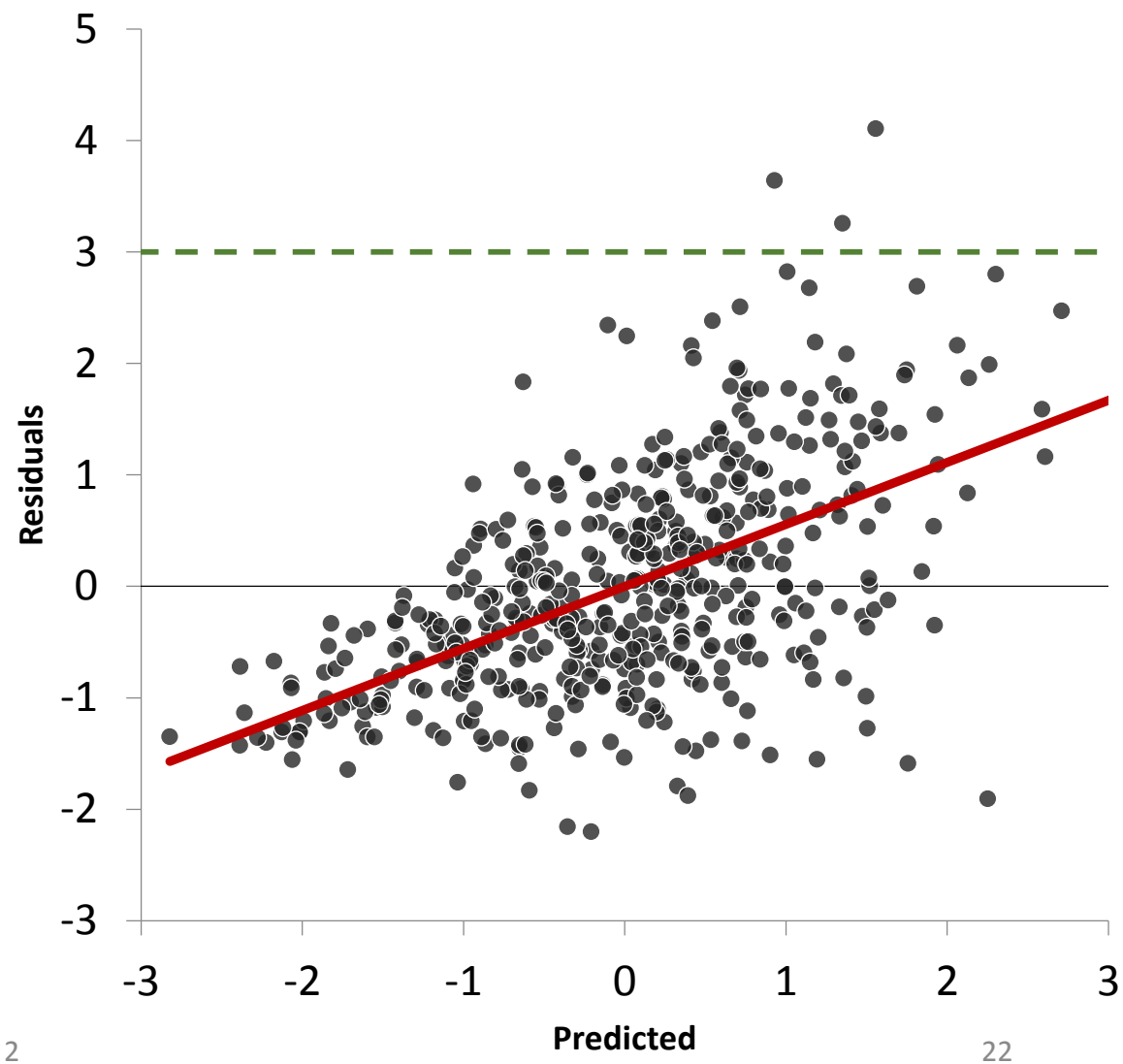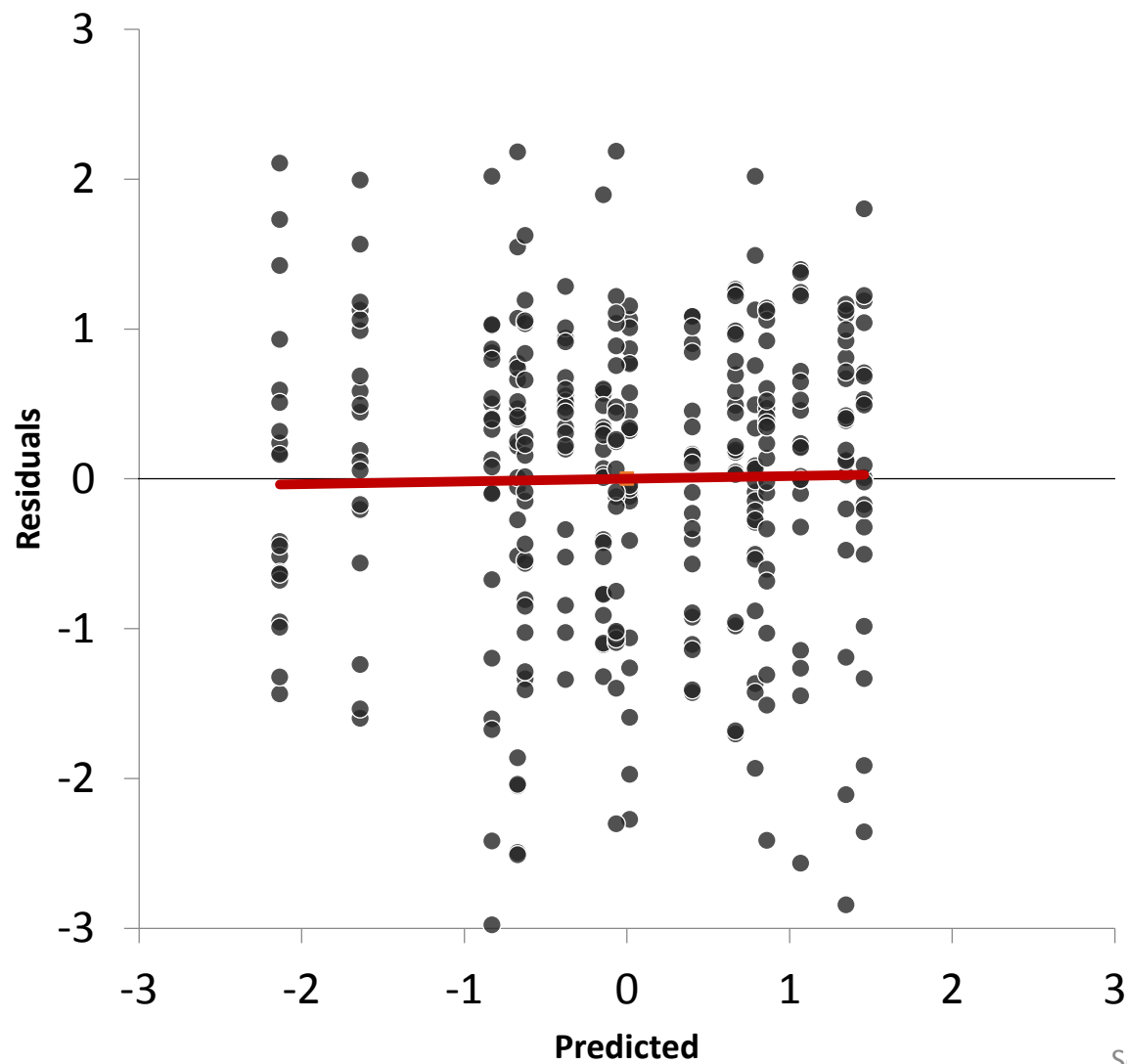# Assumption 5: Normality of Residuals

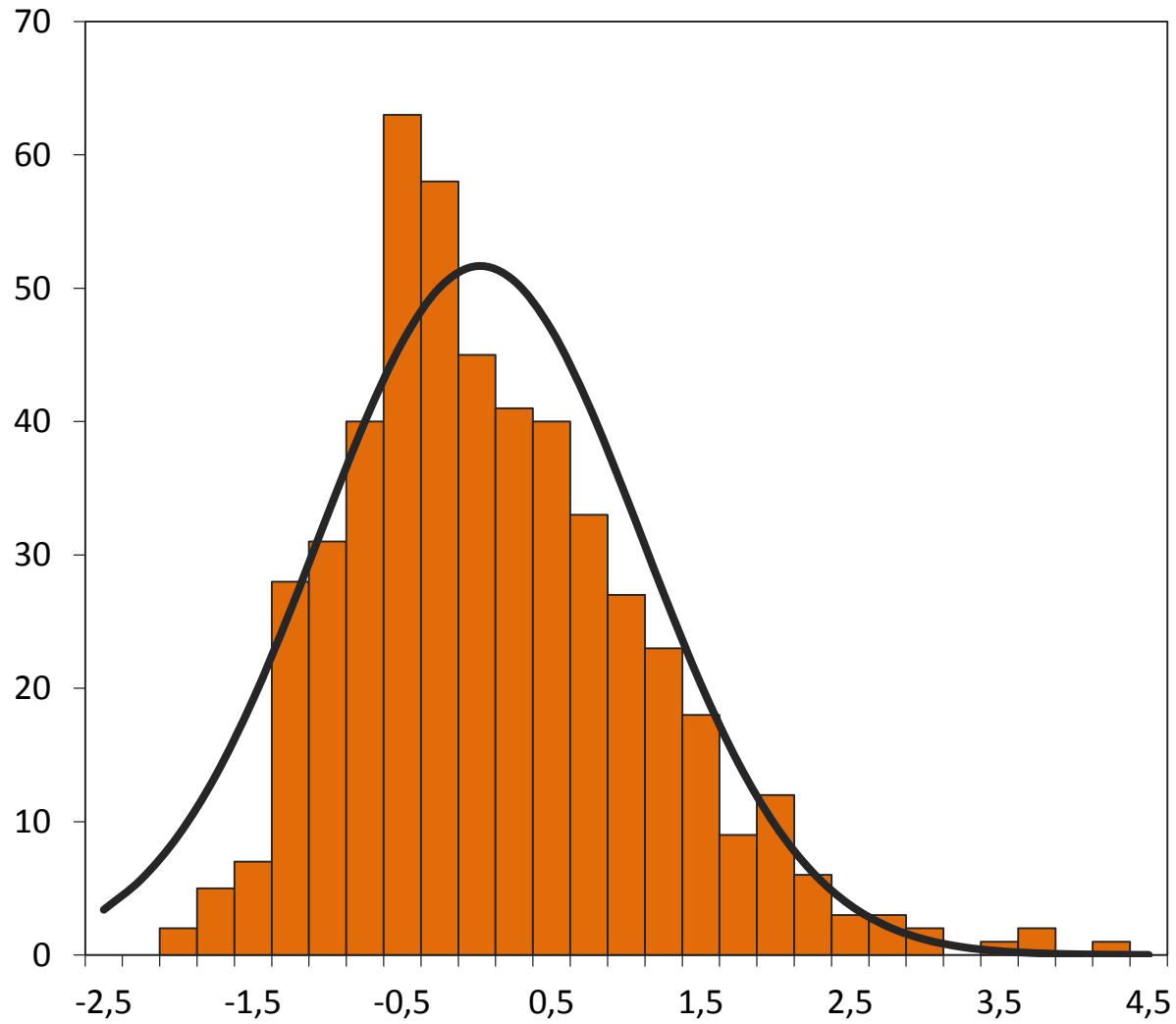Residuals are normally distributed
(the model should not over or underestimate)

Normality
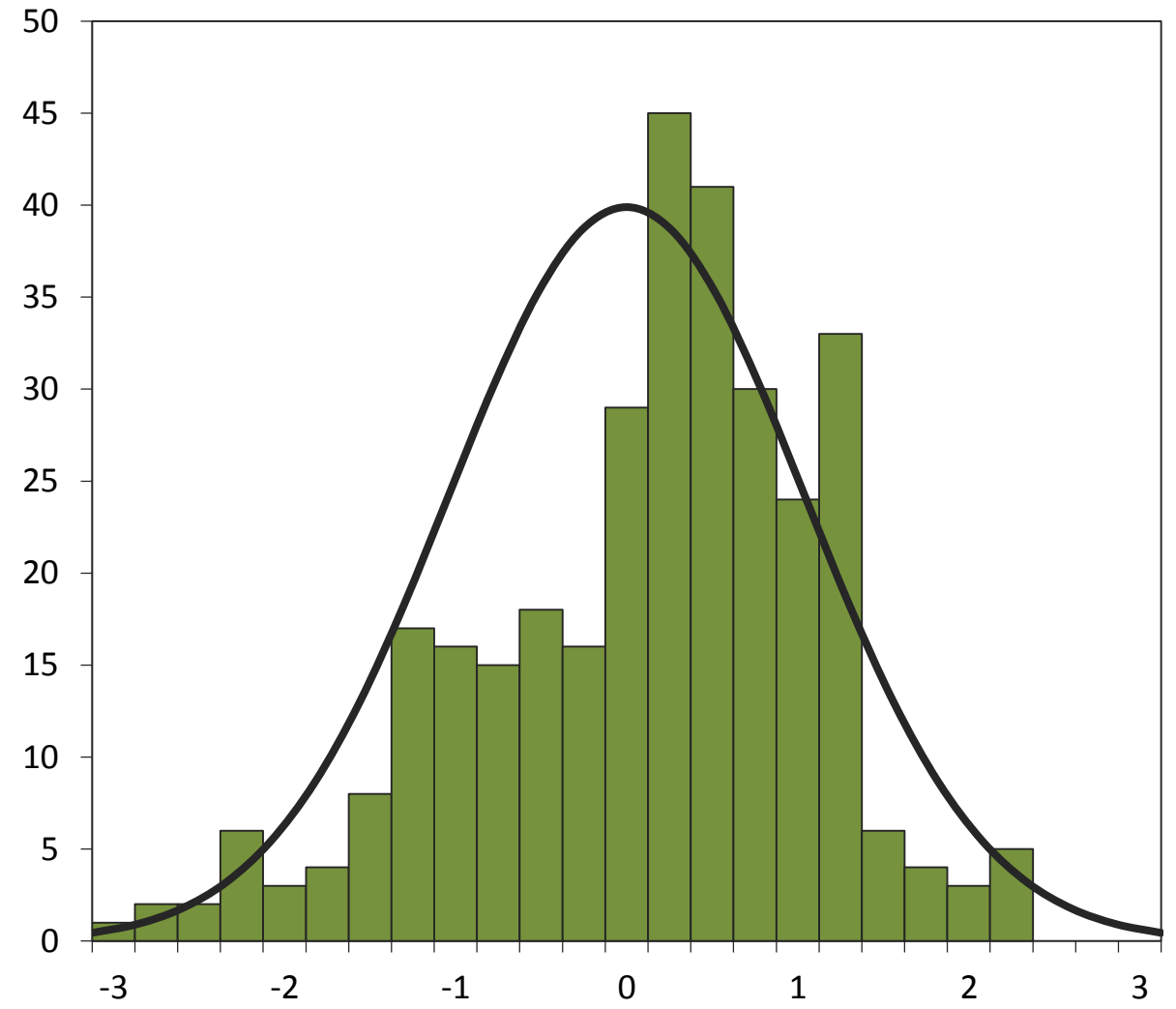
Short Break

03

Model Building

Comparing Models

Before building a model we quickly need to cover the Information Criteria table. Models can be compared by using these values. On their own they mean nothing, but they are invaluable when comparing one model to another with slight changes (different terms or different covariance patterns).

Information Criteria

-2 Restricted Log Likelihood	2402,637
Akaike's Information Criterion (AIC)	2444,637
Hurvich and Tsai's Criterion (AICC)	2448,762
Bozdogan's Criterion (CAIC)	2539,249
Schwarz's Bayesian Criterion (BIC)	2518,249

-2LL is a good start, but it can only compare nested models. AIC is more common and can compare models with equal number of parameters.

These values represent "Fit" and the rule is: Lower is better. SPSS can show you negative values as well, rule still applies.

If the difference between the smallest AIC and the second smallest AIC is smaller than 2 you can safely use the second smallest (it's a valid simplification). Between 2 and 4 is still strong support for the simpler model, while more than 10 means you can better stick with the minimum AIC model.

A

1 2 3 4   6 7 8 9

B

700/900/1100ms
Fixation

500ms
Prime

700/900ms ISI

500ms
Target

500ms
Response

250ms
Feedback
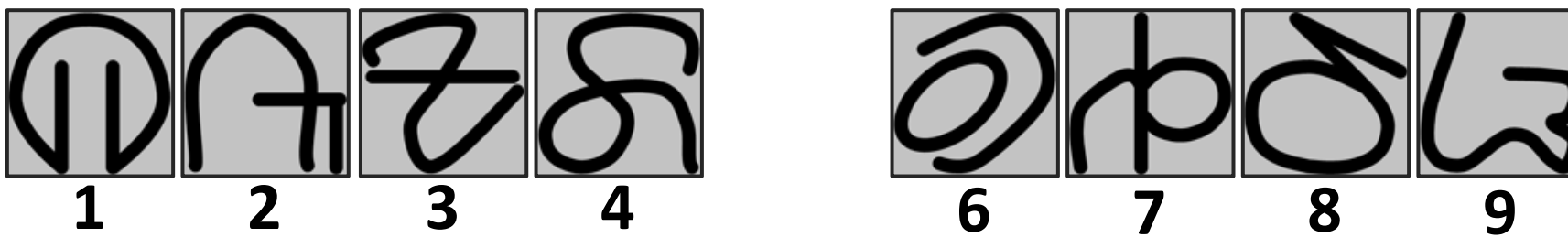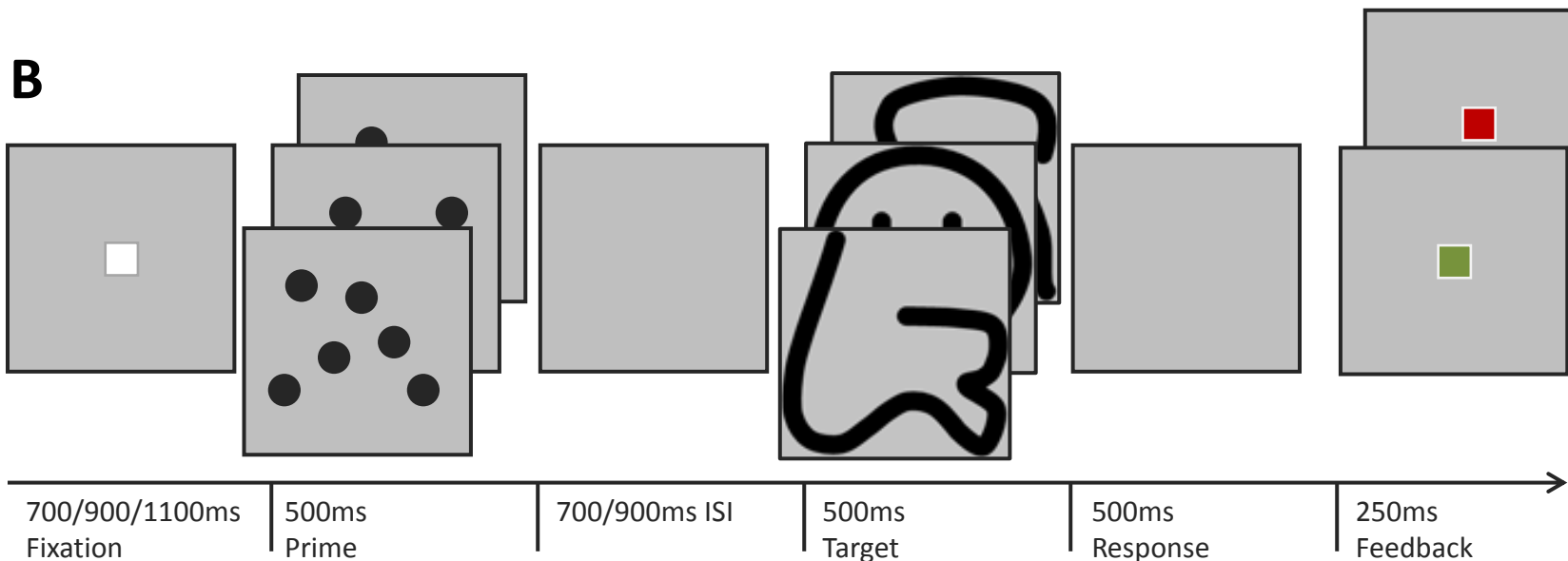
Step 1: Factor Selection

Add all Factors, Covariates, and Interactions
(overfitting is preferred)

Factor Selection

Type III Tests of Fixed Effects

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	25	19576.679	.000
Session	1	375	73.391	.000
Hemisphere	1	375	65.703	.000
Prime	1	375	667.198	.000
Target	1	375	427.971	.000
Session * Hemisphere	1	375	0.181	.670
Session * Prime	**1**	**375**	**188.867**	**.000**
Session * Target	1	375	0.005	.941
Hemisphere * Prime	1	375	0.067	.796
Hemisphere * Target	1	375	3.436	.065
Prime * Target	**1**	**375**	**21.917**	**.000**
Session * Hemisphere * Prime	1	375	0.945	.332
Session * Hemisphere * Target	**1**	**375**	**17.486**	**.000**
Session * Prime * Target	1	375	3.759	.053
Hemisphere * Prime * Target	1	375	0.175	.676
Session * Hemisphere * Prime * Target	1	375	0.394	.531

Assumption Checks
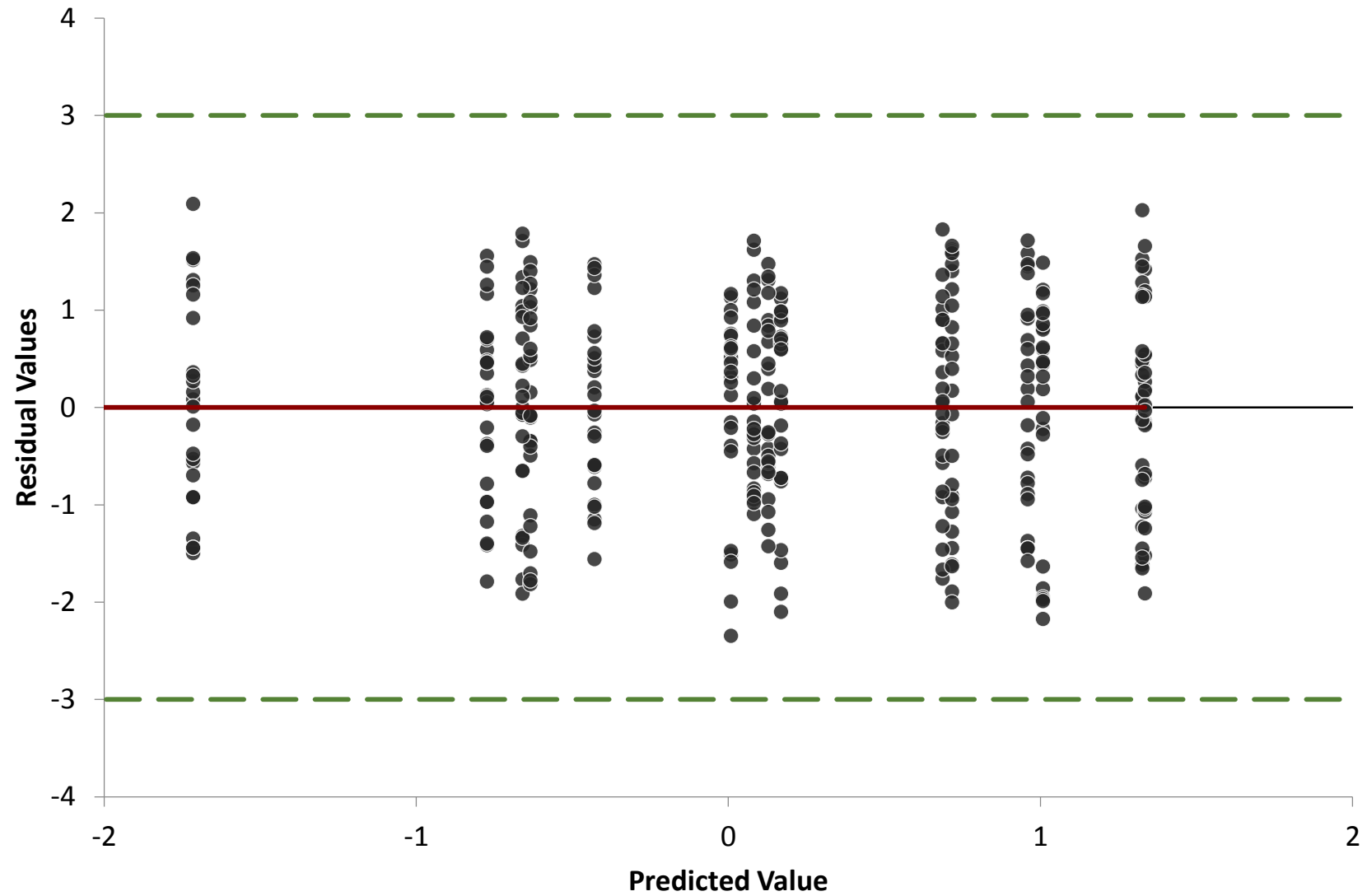
Assumption Checks
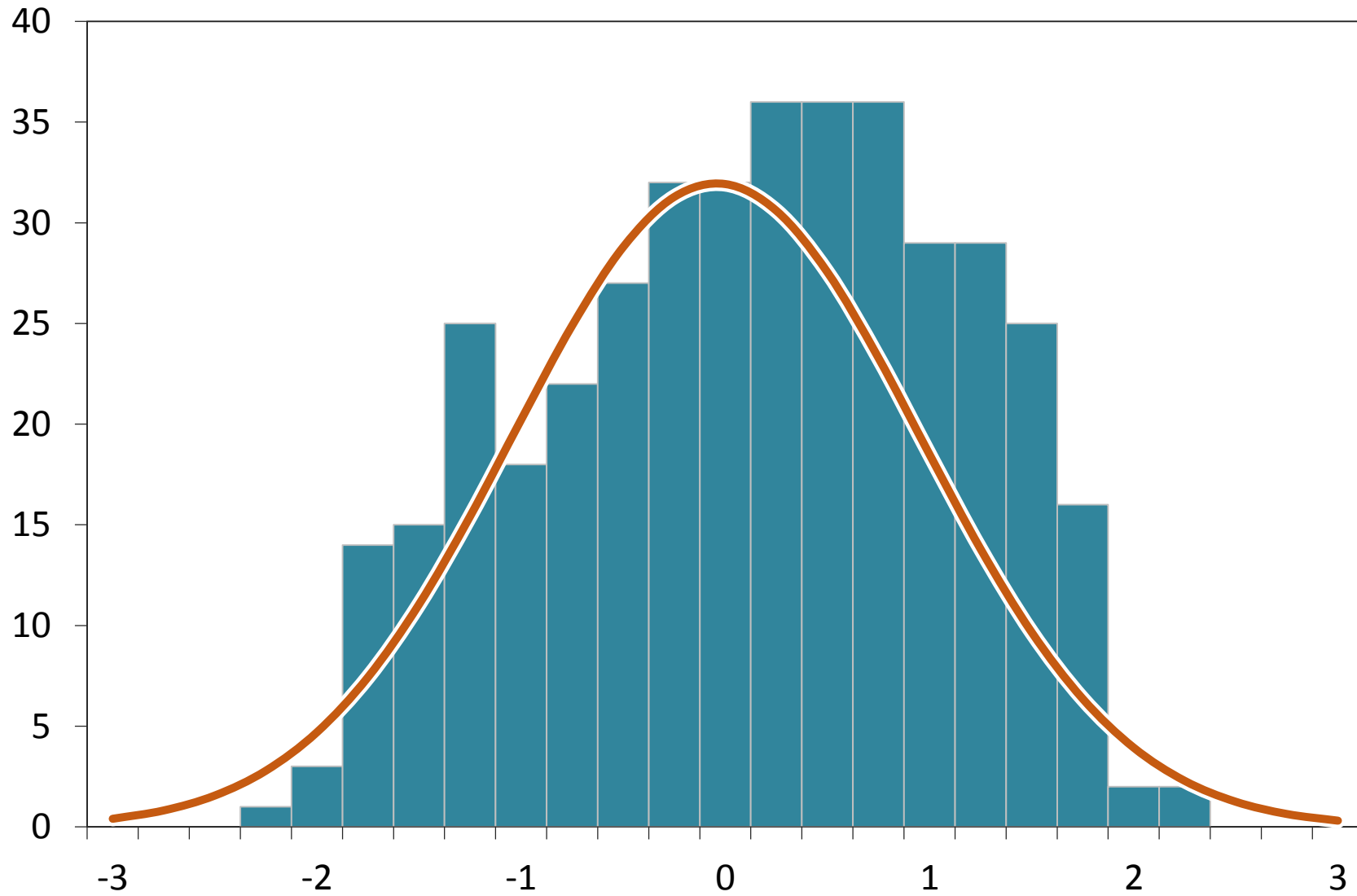
Step 2: Covariance Structure Selection

Compare Covariance Structures
(change to REML)

Covariance Matrix Selection

Structure	Parameters	-2LL	AIC	BIC
UN	152	519.580	791.580	1328.867
TPH	47	657.066	719.066	841.536
ARH(1)	33	667.526	701.526	768.687
CSH	**33**	**667.780**	**701.780**	**768.941**
TP	32	667.986	699.986	763.197
AR(1)	18	679.227	683.227	691.128
CS	18	679.392	683.392	691.293

Covariance Matrix Selection

Type III Tests of Fixed Effects

Source	UN (AIC=791.580)			CSH (AIC=701.780)		
	DF	F	Sig.	DF	F	Sig.
Intercept	24	18793.612	.000	24.229	18907.997	.000
Session	24	66.220	.000	339.600	70.414	.000
Hemisphere	24	97.786	.000	340.036	63.044	.000
Prime	24	786.394	.000	340.476	640.264	.000
Target	24	568.393	.000	339.337	410.590	.000
Session * Hemisphere	24	0.181	.674	339.413	0.174	.677
Session * Prime	**24**	**117.673**	**.000**	340.134	181.229	.000
Session * Target	24	0.004	.950	339.376	0.005	.942
Hemisphere * Prime	24	0.052	.821	339.382	0.064	.800
Hemisphere * Target	24	4.525	.044	340.046	3.297	.070
Prime * Target	**24**	**21.716**	**.000**	339.251	21.027	.000
Session * Hemisphere * Prime	24	1.044	.317	339.560	0.907	.342
Session * Hemisphere * Target	**24**	**17.576**	**.000**	**344.641**	**16.796**	**.000**
Session * Prime * Target	24	4.698	.040	339.907	3.607	.058
Hemisphere * Prime * Target	24	0.214	.648	339.358	0.168	.682
Session * Hemisphere * Prime * Target	24	0.234	.633	339.503	0.378	.539

Step 3: Model Reduction

Remove non-significant effects
(switch back to ML)

Model Reduction

Type III Tests of Fixed Effects

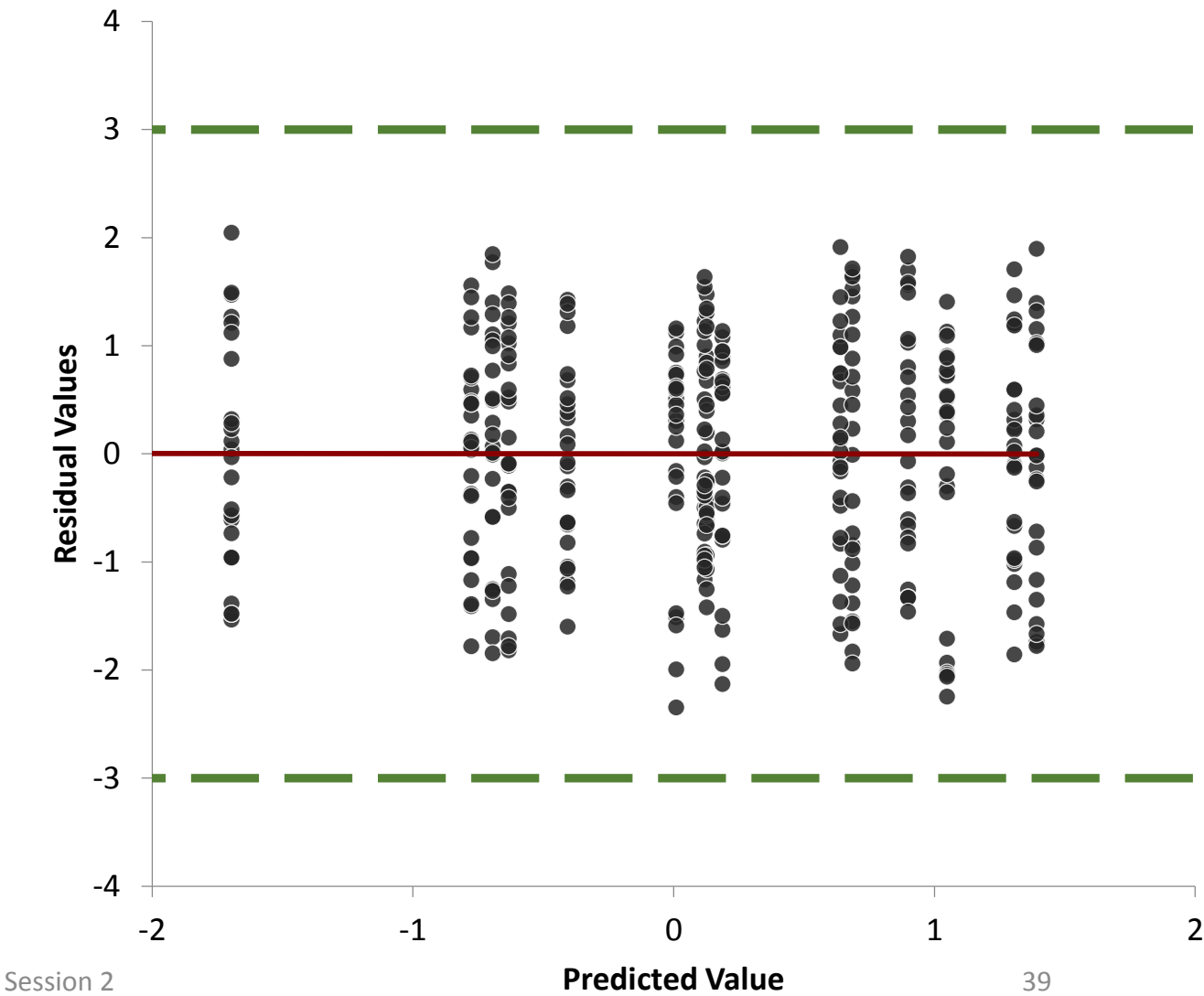| Source	Sig.	Sig.	Sig.	Sig.	Source
Intercept	.000	.000	.000	.000	Intercept
Session	.000	.000	.000	.000	Session
Hemisphere	.000	.000	.000	.000	Hemisphere
Prime	.000	.000	.000	.000	Prime
Target	.000	.000	.000	.000	Target
Session * Hemisphere	.671	.670	.651	.596	Session * Hemisphere
Session * Prime	.000	.000	.000	.000	Session * Prime
Session * Target	.941	.954	.969	.956	Session * Target
Hemisphere * Prime	.796	.809	.816	.784	Hemisphere * Prime
Hemisphere * Target	.065	.071	.065	.057	Hemisphere * Target
Prime * Target	.000	.000	.000	.000	Prime * Target
Session * Hemisphere * Prime	.332	.338	.344		
Session * Hemisphere * Target	.000	.000	.000	.000	Session * Hemisphere * Target
Session * Prime * Target	.053	.048	.047	.054	Session * Prime * Target
Hemisphere * Prime * Target	.676	.655			
Session * Hemisphere * Prime * Target	.531				

Step 4: The Final Model

Re-Run the final model with extra Post-Hoc testing
(switch to REML again for inference)

The Final Model

Type III Tests of Fixed Effects

Source	Denominator df	F	Sig.
Intercept	24.132	18995.535	.000
Session	343.185	70.955	.000
Hemisphere	343.756	63.420	.000
Prime	344.281	671.031	.000
Target	343.559	414.250	.000
Session * Hemisphere	342.950	0.272	.603
Session * Prime	344.428	186.437	.000
Session * Target	340.544	0.003	.956
Hemisphere * Prime	345.203	0.073	.788
Hemisphere * Target	343.152	3.531	.061
Prime * Target	344.165	19.956	.000
Session * Hemisphere * Target	347.960	17.249	.000
Session * Prime * Target	344.887	3.624	.058

Step 5: Writing it Down

There's no template to follow and APA will not help you
(check similar papers and keep the audience in mind)

*For primary analyses, **we used a linear mixed model for repeated measures over time** by type of diabetes (SAS Proc Mixed) to analyze the impact of the 3 education interventions on HbA$_{1c}$ at baseline and follow-up **with fixed effects of time, group, type of diabetes, the interactions between time and group, and between time and type of diabetes. This procedure prevented list-wise deletion due to missing data.***

Weinger K, Beverly EA, Lee Y, Sitnokov L, Ganda OP, Caballero AE. The Effect of a Structured Behavioral Intervention on Poorly Controlled Diabetes: A Randomized Controlled Trial. *Arch Intern Med.*2011;171(22):1990–1999. doi:10.1001/archinternmed.2011.502

*The primary analysis, to determine the effects of age and condition on step latency, APA onset, and APA errors, **used a linear mixed model, with age group (younger vs older) and condition (SRT vs CRT) as fixed factors and participant as a random factor**. APA errors occurred in 83 trials (24 among 12 of the 13 young participants and 59 among 12 of the 13 older participant) for the combined CRT and SRT conditions.*

Rajal G. Cohen, John G. Nutt, Fay B. Horak; Errors in Postural Preparation Lead to Increased Choice Reaction Times for Step Initiation in Older Adults, *The Journals of Gerontology: Series A*, Volume 66A, Issue 6, 1 June 2011, Pages 705–713, https://doi.org/10.1093/gerona/glr054

***We used a linear mixed model (LMM)** to test for statistical differences between power values across action sequence categories. **Since the experimental design had multiple conditions within a single subject, correlation within subject data had to be accounted for**, and since the beta-ERD and transient rebound values had large variance between subjects, we accommodated the mixed model in order to consider these factors when comparing relative power values across categories across subjects. **A variance components structure was used as a covariance structure, with action category as the fixed effect and subjects as the random effect.** The category-wise maximum values of beta-ERD and transient rebound amplitudes were the dependent variables respectively. All datasets were tested and confirmed graphically for normality in advance. Results at p < 0.05 were considered to be statistically significant for all tests. Statistical tests were performed with MATLAB and IBM SPSS® Statistics 17.0.0 (SPSS Inc., IBM Company, Chicago, IL, USA).*

Park, H., Kim, J. S., & Chung, C. K. (2013). Differential beta-band event-related desynchronization during categorical action sequence planning. *PLoS One*, *8*(3), e59544.

Make sure your analysis can be reproduced

Can someone, with the same data, reproduce your results based on your description?

- What were the Fixed Effects in the model?
    - were they centered, coded, scaled?
- What was the covariance structure?
- How were the Fixed effect and Covariance structure determined?
- How were different models evaluated?

- What software was used?
- Which method of estimation was used (REML/ML)?
- Were assumptions violated and/or corrected?
- Was there missing data (did it impact the model)?

Statistical Description

A Linear Mixed Model was constructed (IBM SPSS 24) using within-subject factors Session (2 Levels; Session 1 & Session 3), Hemisphere (2 Levels; Left & Right), Prime (2 Levels; Low & High), and Target (2 Levels; Low & High). Covariance Structures were compared on Restricted Maximum Likelihood (REML) models using Akaike's Information Criterion (AIC).

Non-significant higher order interactions were removed on Maximum Likelihood (ML) models in a step-wise manner starting with the four-way interaction Session*Hemisphere*Prime*Target (p=.531), followed by the three-way interaction between Hemisphere*Prime*Target (p=.655) and Session*Hemisphere*Prime (p=.344). The Session*Prime*Target interaction was kept based on theoretical expectations and all two-way interactions also remained in the model.

The final model used for the analyses (reported below) was a fixed effects model (REML) using a Heterogeneous Compound Symmetry Covariance Matrix for repeated measures. Residuals were normally distributed, showed no heteroscedasticity, and no observations were removed as outliers.

Results

Type III Tests of Fixed Effects

Source	Denominator df	F	Sig.
Intercept	24.132	18995.535	.000
Session	**343.185**	**70.955**	**.000**
Hemisphere	**343.756**	**63.420**	**.000**
Prime	**344.281**	**671.031**	**.000**
Target	**343.559**	**414.250**	**.000**
Session * Hemisphere	342.950	0.272	.603
Session * Prime	**344.428**	**186.437**	**.000**
Session * Target	340.544	0.003	.956
Hemisphere * Prime	345.203	0.073	.788
Hemisphere * Target	343.152	3.531	.061
Prime * Target	**344.165**	**19.956**	**.000**
Session * Hemisphere * Target	**347.960**	**17.249**	**.000**
Session * Prime * Target	344.887	3.624	.058

The Linear Mixed Procedure on [DATA] showed main effects for Session ($F_{(1,343.18)}$=70.955, $p<.001$), Hemisphere ($F_{(1,343.756)}$=63.42, $p<.001$), Prime ($F_{(1,344.28)}$=671.03, $p<.001$), and Target ($F_{(1,343.56)}$=414.25, $p<.001$).

There were two-way interactions of Session*Prime ($F_{(1,344.43)}$=186.44, $p<.001$) and Prime*Target ($F_{(1,344.17)}$=19.96, $p<.001$).

Lastly there was a three-way interaction between Session, Target, and Hemisphere ($F_{(1,347.96)}$=17.25, $p<.001$).

Step 1: Factor Selection

Step 2: Covariance Structure Selection

Step 3: Model Reduction

Step 4: The Final Model

Step 5: Writing it Down

This Session

- What Covariance Matrices are
- Which matrices are important to check
- The assumptions of Mixed Models
- Building Models and check for the best fit
- How to describe your model and results

The Next Sessions

~~How to build Models and decide on covariance structures~~
~~(plus talk about what covariance structures are)~~

Adding covariates, random intercepts, and random slopes
(Making the Mixed Models truly Mixed)

Adding Planned and Polynomial Contrasts
(Thinking ahead like a good data scientist)

MAASTRICHT UNIVERSITY

Session Evaluation