Methods in Psycholinguistics — Linear regression —

Judith Degen

Acknowledgments

As always, it takes a village. These slides contain my own material as well as bits and pieces contributed by:

Maureen Gillespie (UNH)
Peter Graff (Vienna)
T. Florian Jaeger (Rochester)
Dave Kleinschmidt (Princeton)
Victor Kuperman (McMaster)
Roger Levy (MIT)

… with Florian's permission

What are some of the questions we’ve encountered?

What was the evidence brought to bear on answering those questions?

What will we cover?

- introduction to Generalized Linear Models (GLMs) and Generalized Linear Mixed Models (GLMMs)
  - mathematical background
  - intuition / conceptualization
  - geometric interpretation
  - common issues & solutions for GLM/GLMMs
  - relation to ANOVA
- we'll learn how to
  - conduct, interpret, and report GLM/GLMM analyses in R
  - visualize data in R

Part lecture, part learning by doing, part asking questions!

What kind of data can you analyze with GLMs?

- continuous (nominal)
  response/reading times, slider ratings, speech onset times,…
- categorical (binary)
  truth value judgments,
  any binary choice prediction…
- ordered discrete (ordinal)
  Likert scale ratings…

- unordered discrete
  any choice between more than two options

What kind of data can you analyze with GLMs?

- continuous (nominal) ....linear regression
  response/reading times, slider ratings, speech onset times,…
- categorical (binary) ....logistic regression
  truth value judgments, any binary choice prediction…
- ordered discrete (ordinal) ....ordinal regression
  Likert scale ratings…
- unordered discrete .....multinomial regression
  any choice between more than two options

What kind of data can you analyze with GLMs?

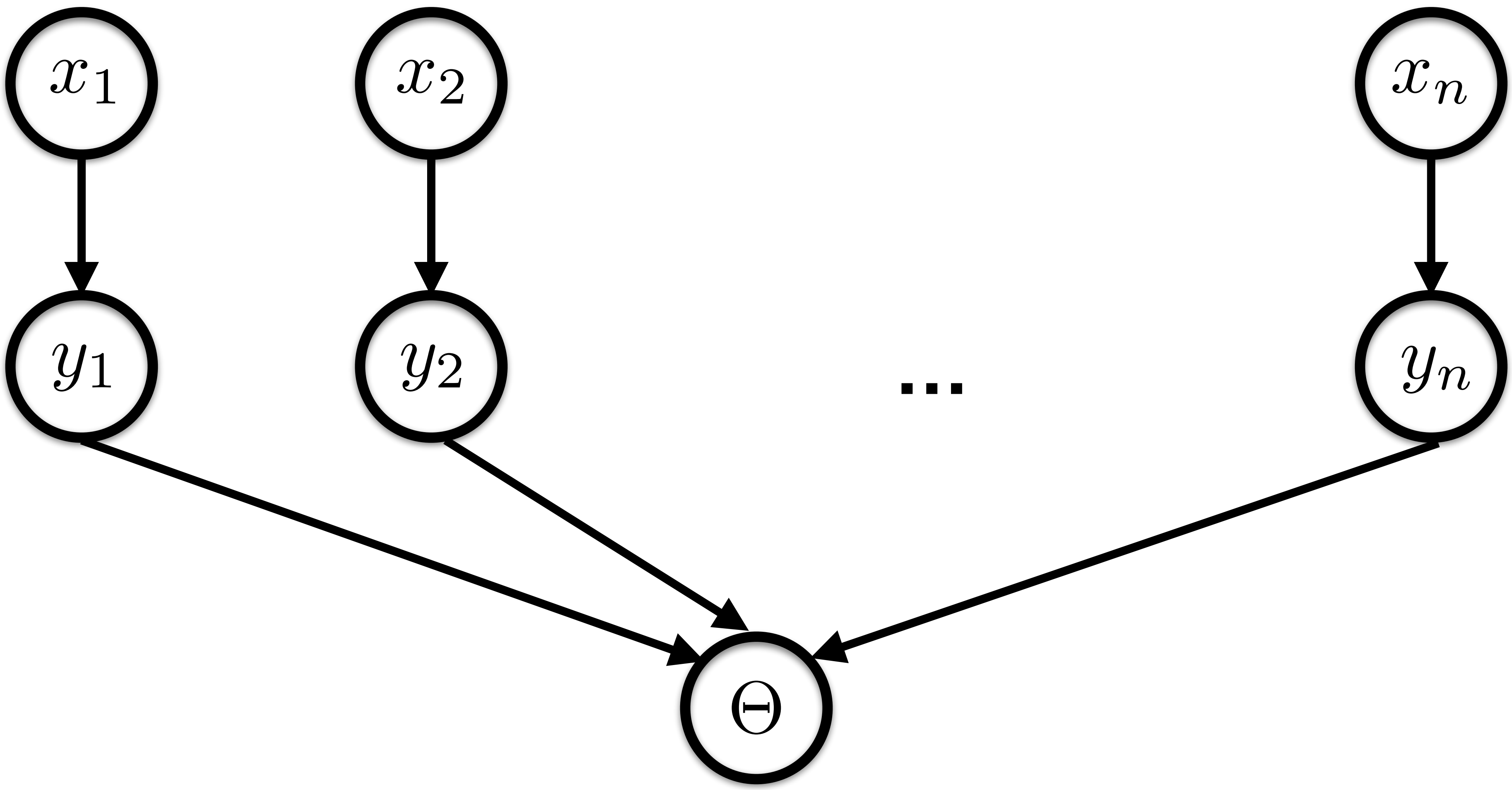
- continuous (nominal) ....linear regression
  response/reading times, slider ratings, speech onset times,…
- categorical (binary) ....logistic regression
  truth value judgments, any binary choice prediction…
- ordered discrete (ordinal) ....ordinal regression
  Likert scale ratings…
- unordered discrete .....multinomial regression
  any choice between more than two options

What kind of data can you analyze with GLMs?

- continuous (nominal) ....linear regression
  response/reading times, slider ratings, speech onset times,…
- categorical (binary) ....logistic regression
  truth value judgments, any binary choice prediction…
- ordered discrete (ordinal) ....ordinal regression
  Likert scale ratings…
- unordered discrete .....multinomial regression
  any choice between more than two options

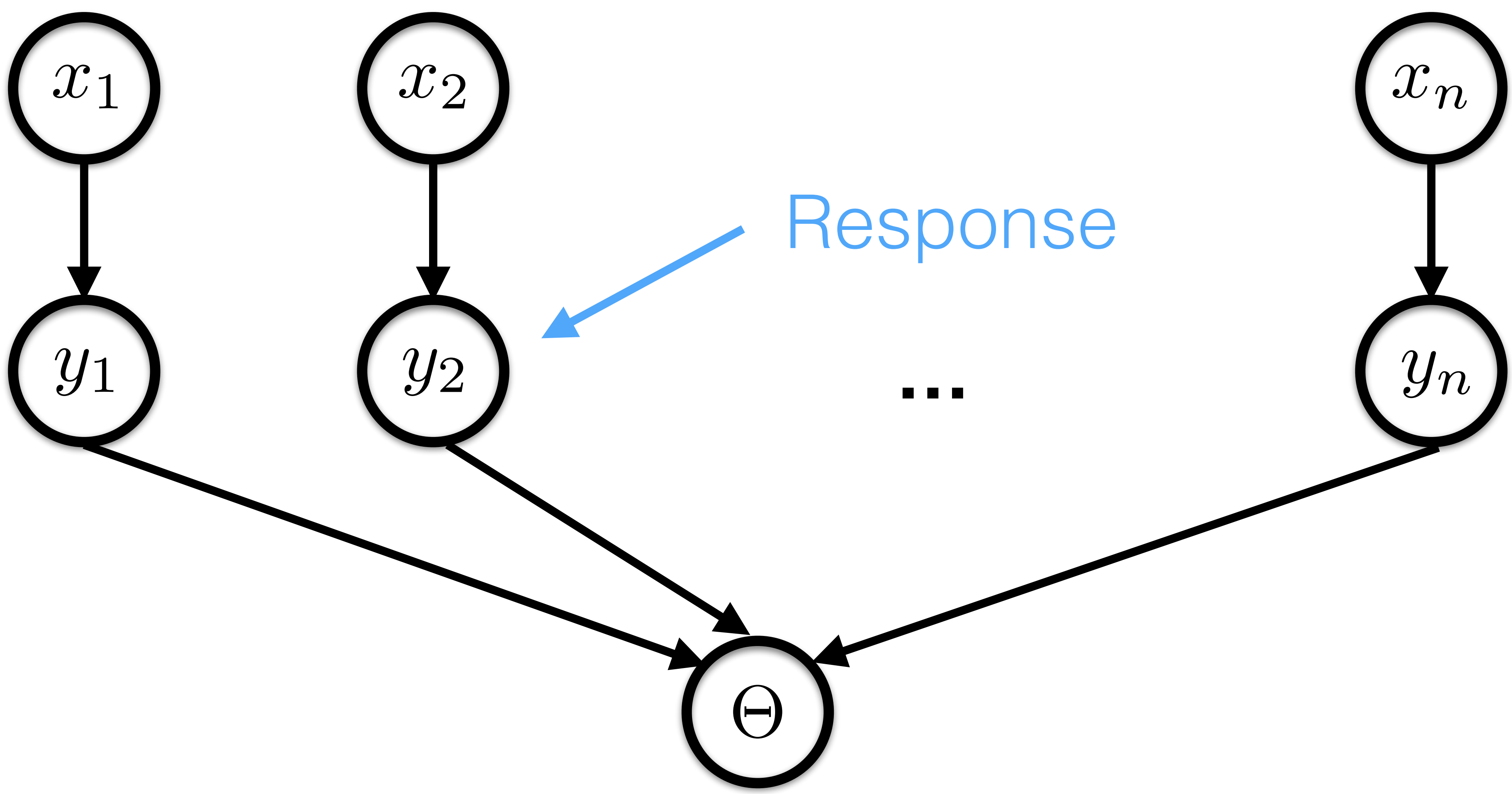
Generalized Linear Models

Goal: model effects of predictors (independent variables) X on a response (dependent variable) Y



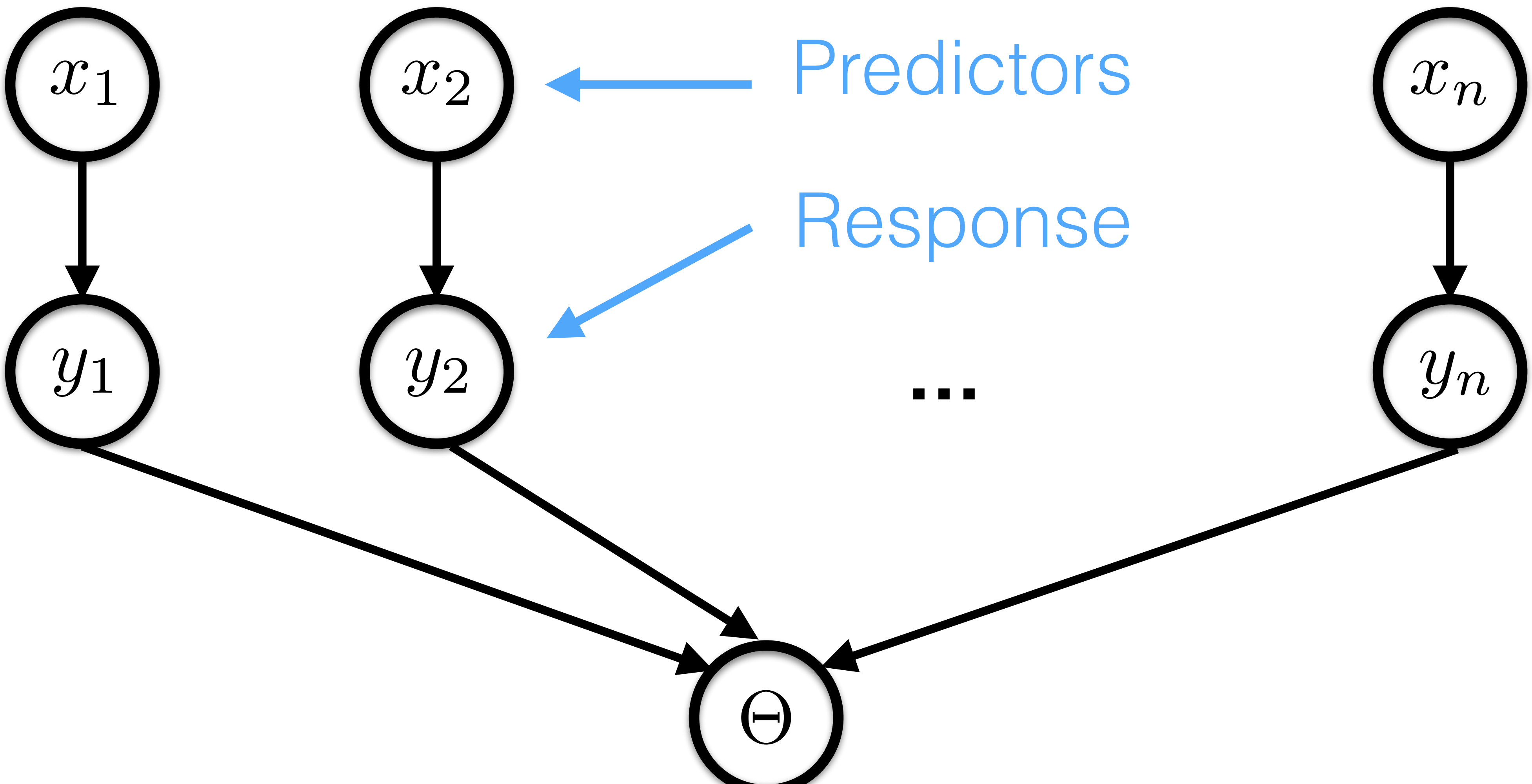
Generalized Linear Models

Goal: model effects of predictors (independent variables) X on a response (dependent variable) Y



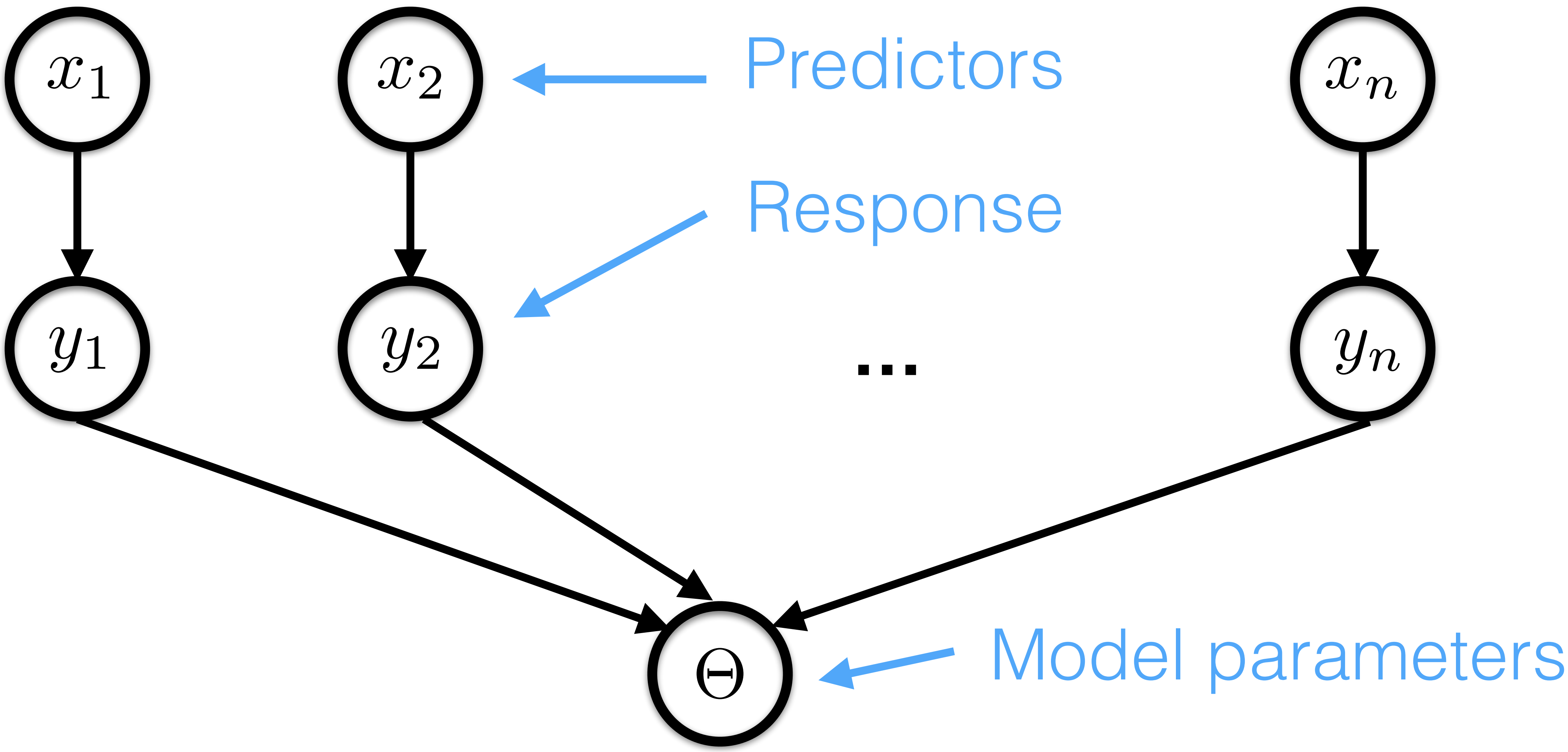
Generalized Linear Models

Goal: model effects of predictors (independent variables) X on a response (dependent variable) Y



Generalized Linear Models

Goal: model effects of predictors (independent variables) X on a response (dependent variable) Y



Reviewing GLMs

Assumptions of the generalized linear model:

1. Predictors X_i influence Y through the mediation of a linear predictor η
2. η is a linear combination of the X_i

$$\eta = \alpha + \beta_1 X_1 + \cdots + \beta_N X_N$$

3. η determines the predicted mean μ of Y

$$\eta = g(\mu) \qquad \text{(link function)}$$

4. There is some noise distribution P around the predicted mean μ of Y:

$$P(Y = y; \mu)$$

Linear regression

Linear regression (which underlies ANOVA) is a kind of generalized linear model.

The predicted mean is simply the linear predictor:

$$\eta = l(\mu) = \mu$$

Noise is normally (=Gaussian) distributed around 0 with standard deviation σ :

$$\epsilon \sim N(0, \sigma)$$

This results in the traditional linear regression equation:

Predicted mean $\mu = \eta$ Noise $\sim N(0, \sigma)$

$$Y = \underbrace{\alpha + \beta_1 X_1 + \cdots + \beta_n X_n}_{\text{Predicted mean } \mu = \eta} + \underbrace{\epsilon}_{\text{Noise } \sim N(0,\sigma)}$$

An example: lexical decision

Baayen, Feldman, & Schreuder (2006)

tpozt *Word or non-word?*

house *Word or non-word?*

Measure response times (RT)

Question: which factors predict RTs?

Let’s analyze…
open RStudio!

The dataset

- lexical decisions from 79 concrete nouns, each seen by 21 participants (1,659 observations)
- **Outcome/response:** log-transformed lexical decision times
- **Inputs:**
  - continuous: e.g. frequency
  - categorical: e.g., native language (English vs other)

The basic model

Let's assume that frequency has a *linear* effect on average log RT, and trial-level noise is *normally distributed.*

If x_i is frequency, this simple model is:

$$RT_{ij} = \alpha + \beta x_{ij} + \overbrace{\epsilon_{ij}}^{\text{Noise} \sim N(0, \sigma_\epsilon)}$$

The basic model

Let's assume that frequency has a *linear* effect on average log RT, and trial-level noise is *normally distributed.*

If x_i is frequency, this simple model is:

Given

$$\boxed{RT_{ij}} = \alpha + \beta \boxed{x_{ij}} + \overbrace{\epsilon_{ij}}^{\text{Noise} \sim N(0, \sigma_\epsilon)}$$

The basic model

Let's assume that frequency has a *linear* effect on average log RT, and trial-level noise is *normally distributed.*

If x_i is frequency, this simple model is:

Given

$$\boxed{RT_{ij}} = \boxed{\alpha} + \boxed{\beta}\boxed{x_{ij}} + \overbrace{\epsilon_{ij}}^{\text{Noise} \sim N(0,\boxed{\sigma_\epsilon})}$$

Inferences

The basic model

Let's assume that frequency has a *linear* effect on average log RT, and trial-level noise is *normally distributed.*

If x_i is frequency, this simple model is:

Given

$$\boxed{RT_{ij}} = \boxed{\alpha} + \boxed{\beta}\boxed{x_{ij}} + \overbrace{\epsilon_{ij}}^{\text{Noise} \sim N(0,\boxed{\sigma_\epsilon})}$$

Inferences

E.g. "Does frequency affect RT?"

The basic model

Let's assume that frequency has a *linear* effect on average log RT, and trial-level noise is *normally distributed.*

If x_i is frequency, this simple model is:

Given

$$\boxed{RT_{ij}} = \boxed{\alpha} + \boxed{\beta}\boxed{x_{ij}} + \overbrace{\epsilon_{ij}}^{\text{Noise} \sim N(0,\boxed{\sigma_\epsilon})}$$

Inferences

E.g. "Does frequency affect RT?"—> is β reliably non-zero?

Let’s translate this into R

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.588778   0.022296 295.515   <2e-16 ***
Frequency   -0.042872   0.004533  -9.459   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 0.2353 on 1657 degrees of freedom
Multiple R-squared:  0.05123,	Adjusted R-squared:  0.05066
```

$$RT_{ij} = \alpha + \beta x_{ij} + \overbrace{\epsilon_{ij}}^{\text{Noise} \sim N(0, \sigma_\epsilon)}$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.588778   0.022296 295.515   <2e-16 ***
Frequency   -0.042872   0.004533  -9.459   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 0.2353 on 1657 degrees of freedom
Multiple R-squared:  0.05123,    Adjusted R-squared:  0.05066
```

$$RT_{ij} = \boxed{\alpha} + \beta x_{ij} + \overbrace{\epsilon_{ij}}^{\text{Noise} \sim N(0, \sigma_\epsilon)}$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.588778   0.022296 295.515   <2e-16 ***
Frequency   -0.042872   0.004533  -9.459   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 0.2353 on 1657 degrees of freedom
Multiple R-squared:  0.05123,	Adjusted R-squared:  0.05066
```

$$RT_{ij} = \boxed{\alpha} + \boxed{\beta} x_{ij} + \overbrace{\epsilon_{ij}}^{\text{Noise} \sim N(0, \sigma_\epsilon)}$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.588778   0.022296 295.515   <2e-16 ***
Frequency   -0.042872   0.004533  -9.459   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 0.2353 on 1657 degrees of freedom
Multiple R-squared:  0.05123,   Adjusted R-squared:  0.05066
```

$$RT_{ij} = \boxed{\alpha} + \boxed{\beta} x_{ij} + \overbrace{\epsilon_{ij}}^{\text{Noise} \sim N(0, \boxed{\sigma_\epsilon})}$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.588778   0.022296 295.515   <2e-16 ***
Frequency   -0.042872   0.004533  -9.459   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 0.2353 on 1657 degrees of freedom
Multiple R-squared:  0.05123,	Adjusted R-squared:  0.05066
```

$$RT_{ij} = \boxed{\alpha} + \boxed{\beta} x_{ij} + \overbrace{\epsilon_{ij}}^{\text{Noise} \sim N(0, \boxed{\sigma_\epsilon})}$$

"There was a significant main effect of frequency such that more frequent words were responded to more quickly $(\beta = -0.04, SE = 0.004, t = -9.46, p < .0001)$*."*

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.588778   0.022296 295.515   <2e-16 ***
Frequency   -0.042872   0.004533  -9.459   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 0.2353 on 1657 degrees of freedom
Multiple R-squared:  0.05123,	Adjusted R-squared:  0.05066
```

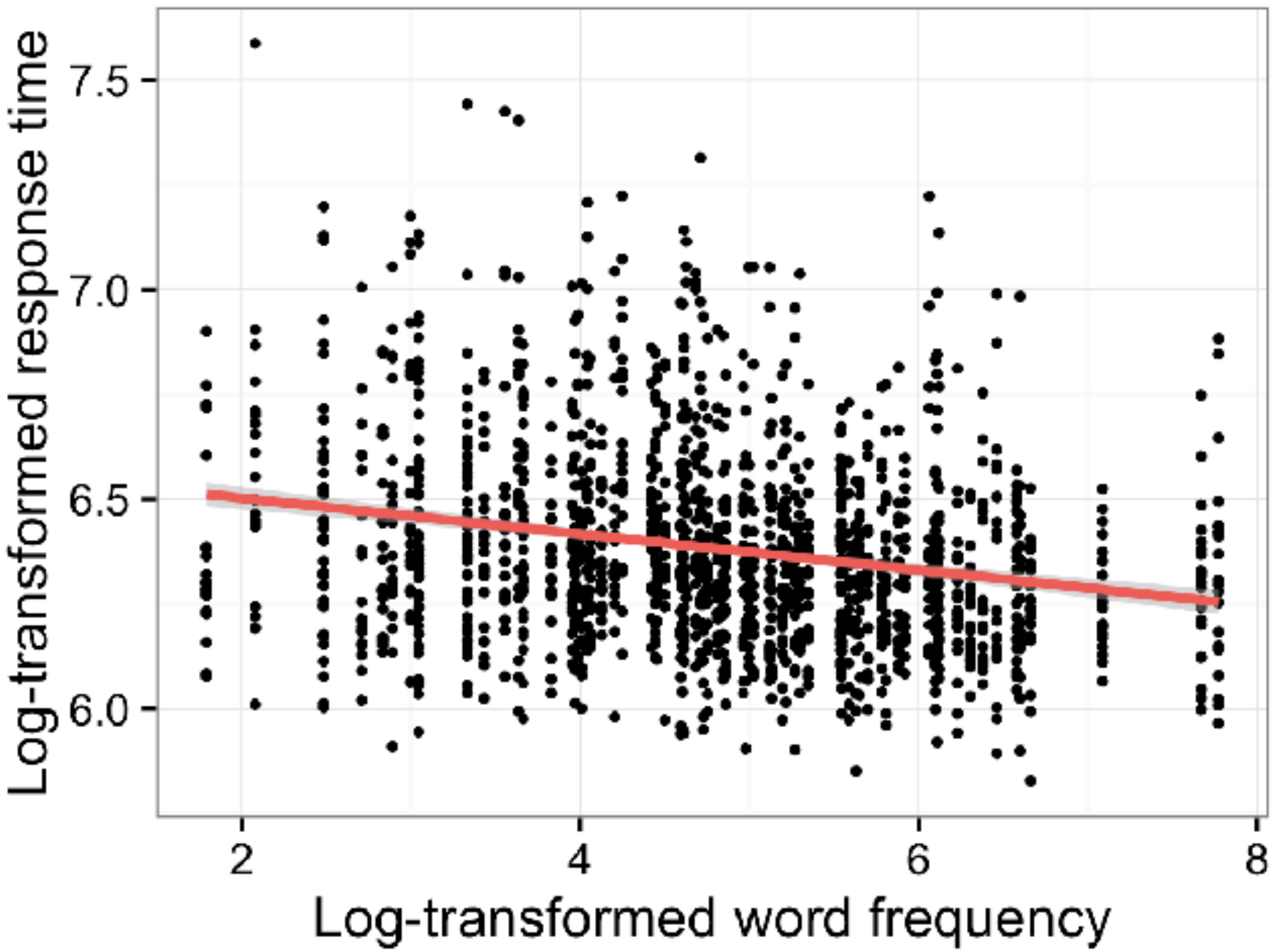
$$RT_{ij} = \boxed{\alpha} + \boxed{\beta} x_{ij} + \overbrace{\epsilon_{ij}}^{\text{Noise} \sim N(0, \boxed{\sigma_\epsilon})}$$

"There was a significant main effect of frequency such that more frequent words were responded to more quickly $(\beta = -0.04, SE = 0.004, t = -9.46, p < .0001)$*."*

Why is R^2 so low even though frequency has tiny p-value?

Geometric intuitions

-.04



Geometric interpretation of linear regression: find slopes for predictors that minimize squared error

Linearity assumption

Like ANOVA, the linear model assumes the outcome is linear in the *coefficients* (**linearity assumption**).

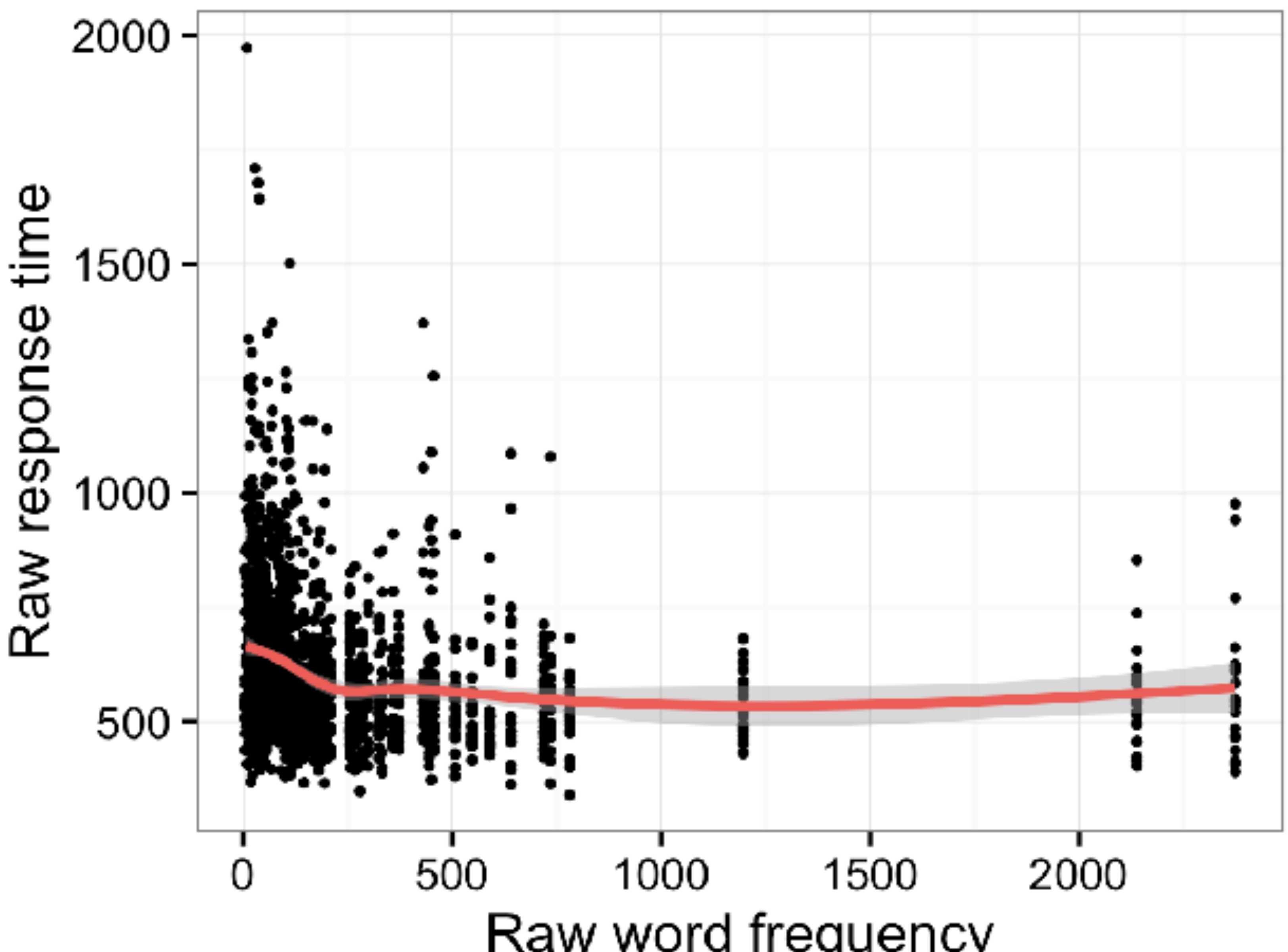
Linearity assumption

Like ANOVA, the linear model assumes the outcome is linear in the *coefficients* (**linearity assumption**).

This doesn't mean that outcome and input *variables* need to be linearly related!

Linearity assumption

Like ANOVA, the linear model assumes the outcome is linear in the *coefficients* (**linearity assumption**).



This doesn't mean that outcome and input *variables* need to be linearly related!

Adding predictors (multiple regression)

Extend the simple model to include an additional predictor for **morphological family size** (number of words in the morphological family of the target word).

Adding predictors (multiple regression)

Extend the simple model to include an additional predictor for **morphological family size** (number of words in the morphological family of the target word).

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.563853   0.026826 244.685  < 2e-16 ***
Frequency   -0.035310   0.006407  -5.511 4.13e-08 ***
FamilySize  -0.015655   0.009380  -1.669   0.0953 .
```

1. Is the interpretation of the output clear?
2. What is the interpretation of the intercept?
3. How much faster is the most frequent word expected to be read compared to the least frequent word?

Categorical predictors

Extend the model to include a predictor for participants' **native language** (English vs other).

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         6.497073   0.025784 251.977  < 2e-16 ***
Frequency          -0.035310   0.006054  -5.832 6.56e-09 ***
FamilySize         -0.015655   0.008863  -1.766   0.0775 .
NativeLanguageOther 0.155821   0.011025  14.133  < 2e-16 ***
```

The output is a linear combination of predictors, so categorical predictors need to be coded numerically
—> Default in R: dummy/treatment coding (more tomorrow)

Categorical predictors

Extend the model to include a predictor for participants' **native language** (English vs other).

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           6.497073   0.025784 251.977  < 2e-16 ***
Frequency            -0.035310   0.006054  -5.832 6.56e-09 ***
FamilySize           -0.015655   0.008863  -1.766   0.0775 .
NativeLanguageOther   0.155821   0.011025  14.133  < 2e-16 ***
```

The output is a linear combination of predictors, so categorical predictors need to be coded numerically
—> Default in R: dummy/treatment coding (more tomorrow)

What is the "mean" that is being predicted in this model?

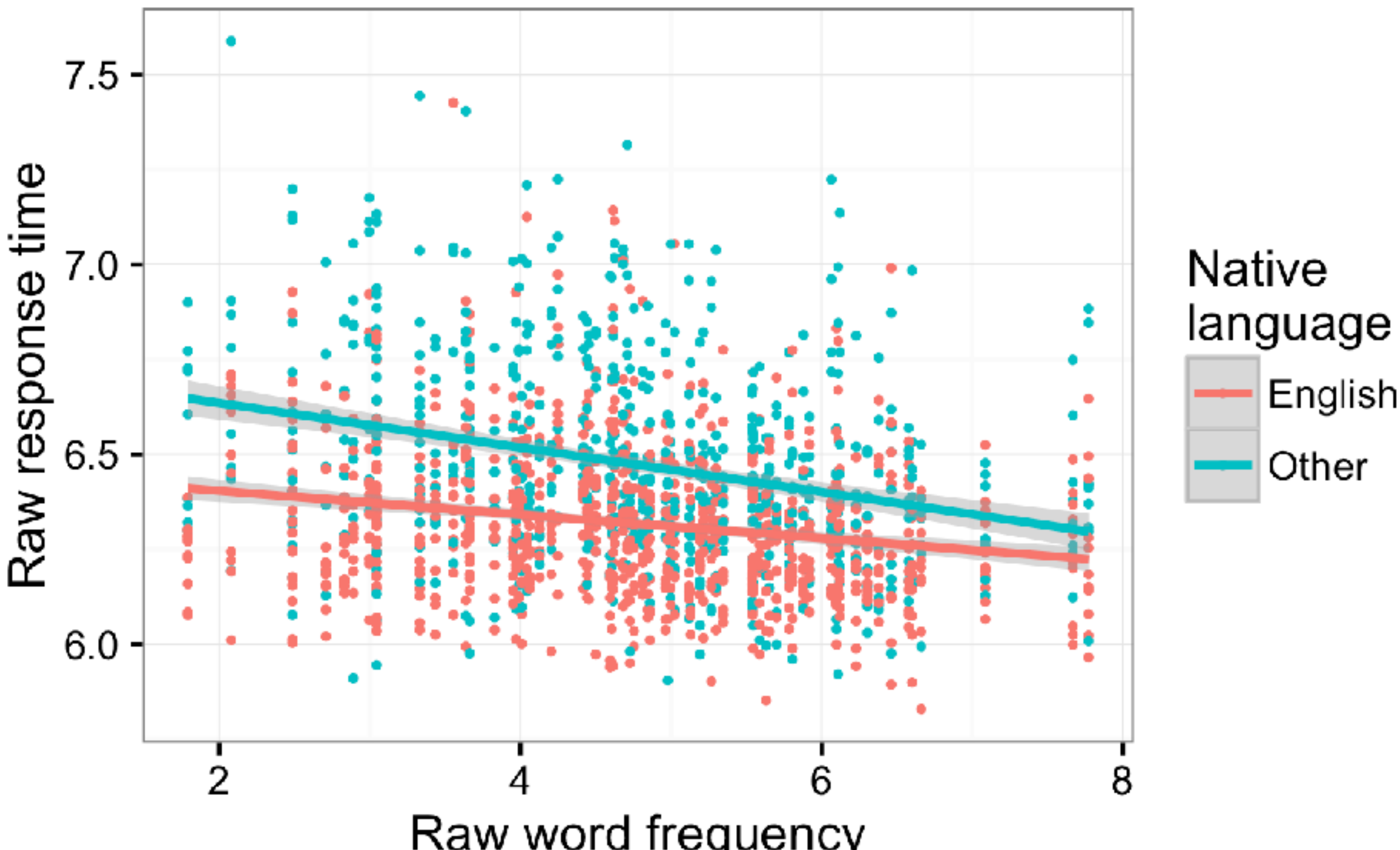
Interactions

Interactions are products of predictors.
Interpretation of significant interactions: the slope of one predictor differs for different values of the other predictor.

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   6.441135   0.031140 206.847  < 2e-16 ***
FamilySize                   -0.015655   0.008839  -1.771 0.076726 .
Frequency                    -0.023536   0.007079  -3.325 0.000905 ***
NativeLanguageOther           0.286343   0.042432   6.748 2.06e-11 ***
Frequency:NativeLanguageOther -0.027472   0.008626  -3.185 0.001475 **
```

How should we interpret the interaction between frequency and native language?

Plotting the interaction



Linear regression vs. ANOVA

- **shared**
  - linearity assumption (though investigation of non-linearities easily possible in regression)
  - assumption of normality
  - assumption of independence (of noise)
  - ANOVA is basically linear regression with only categorical predictors
- **different**
  - Generalized Linear Model
  - consistent and transparent way of treating continuous and categorical predictors
  - regression encourages a priori explicit coding of hypothesis (reducing post-hoc tests)

Hypothesis testing in psycholinguistic research

- often, we make predictions not just about the **existence**, but also about the **direction** of the effect
- sometimes, we're also interested in effect **shapes** (e.g., non-linearities)
- unlike ANOVA, regression analyses test hypotheses about effect **direction**, **shape**, and **size** without requiring post-hoc analyses
  - if predictors are coded appropriately (see next time)
  - if the model can be trusted (see next time)

Determining parameters

How do we choose parameters (model coefficients) β_i and σ?

Find the best ones. (cf Andrew Ng's videos)

Two major approaches:

1. Maximum Likelihood Estimation (ML): pick parameter values that maximize the (log) probability of data, i.e., maximize $P(Y|\beta_i, \sigma)$
2. Bayesian inference: infer best model parameters via Bayes' rule, given a prior distribution over model parameters

$$P(\beta_i, \sigma|Y) = \frac{\overbrace{P(Y|\beta_i, \sigma)}^{\text{Likelihood}} \cdot \overbrace{P(\beta_i, \sigma)}^{\text{Prior}}}{P(Y)}$$