

Rutgers Linguistics Workshop on Mixed Effects Models — Linear regression —

09/16/2016

Judith Degen
Stanford University

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:

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

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:

Given

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

Inferences

The diagram illustrates the flow of information in a linear model. The word 'Given' is at the top, with three green arrows pointing down to the terms RT_{ij} , α , and x_{ij} in the equation. The word 'Inferences' is at the bottom, with three arrows pointing up to the same terms: a blue arrow to α , a red arrow to β , and a purple arrow to σ_ϵ . The equation itself is $RT_{ij} = \alpha + \beta x_{ij} + \epsilon_{ij}$, where ϵ_{ij} is under a brace labeled 'Noise $\sim N(0, \sigma_\epsilon)$ '. The terms RT_{ij} , α , β , and x_{ij} are each enclosed in a colored box: green for RT_{ij} and x_{ij} , blue for α , and red for β .

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	0.0222296	295.515	<2e-16 ***
Frequency	-0.042872	0.004533	-9.459	<2e-16 ***

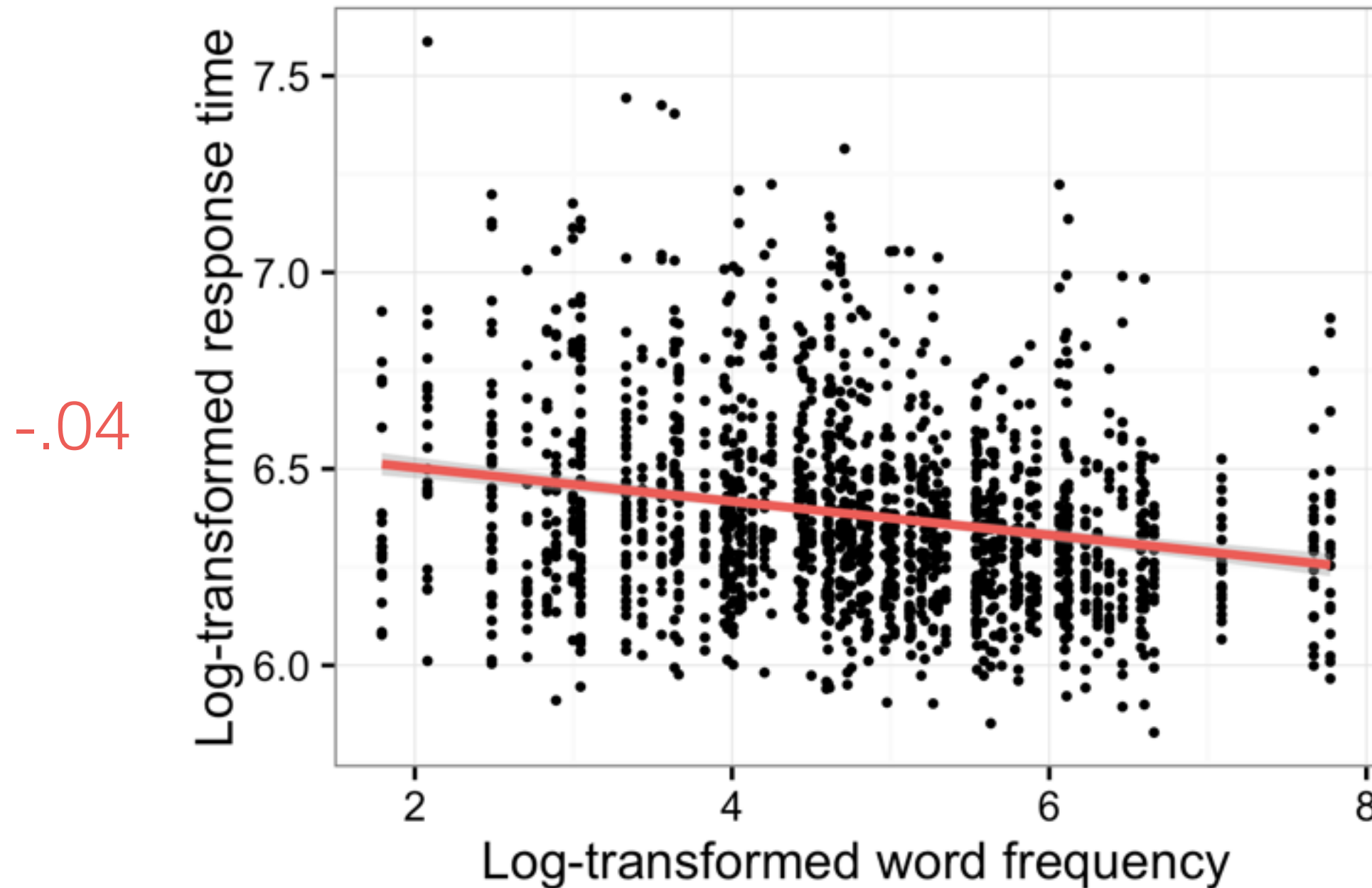
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.2353 on 1657 degrees of freedom
Multiple R-squared: 0.05123, Adjusted R-squared: 0.05066

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

“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?

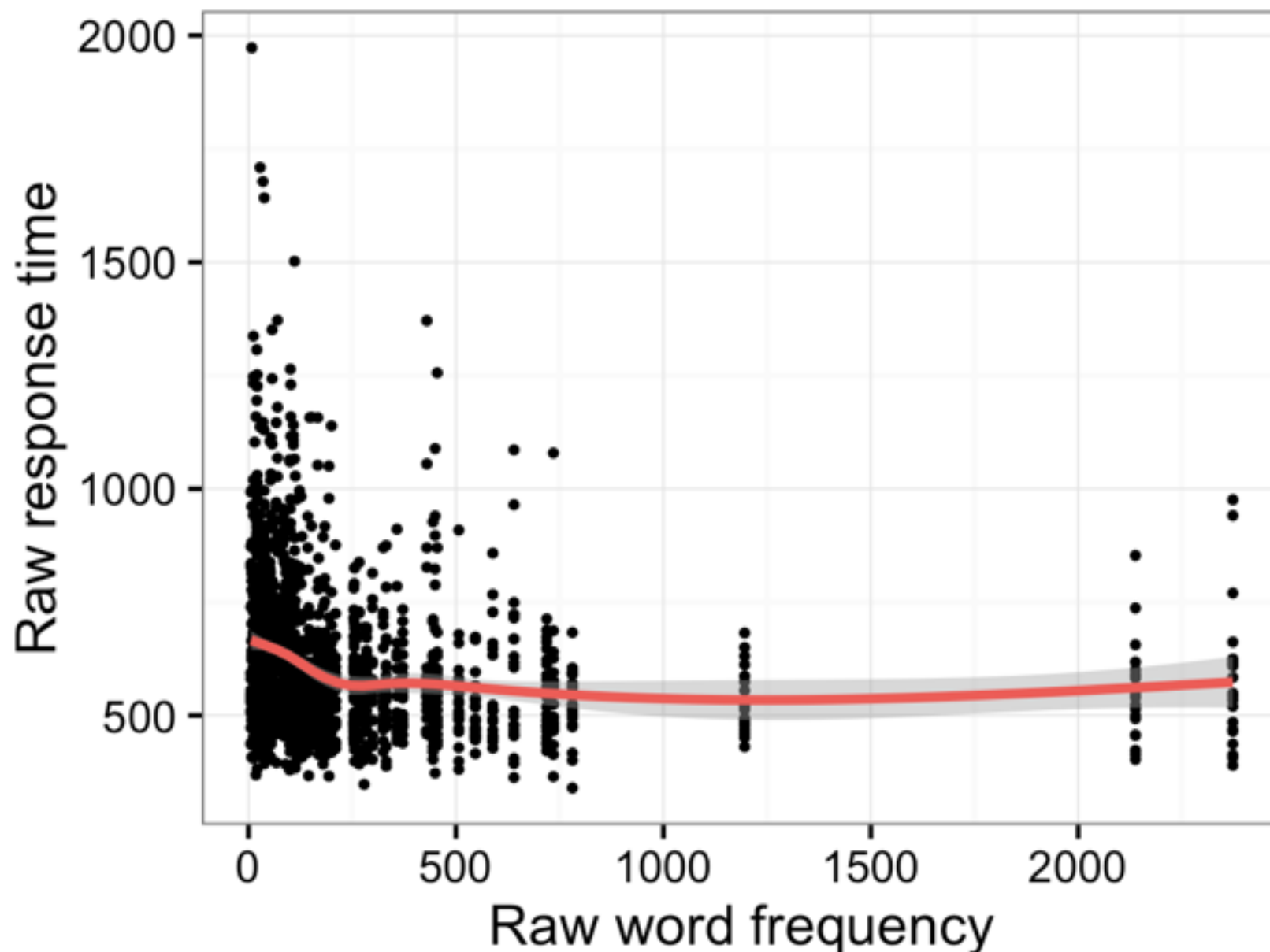
Geometric intuitions



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**).



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).

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)

What is the “mean” that is being predicted in this model?

Interactions

Interactions are products of predictors.

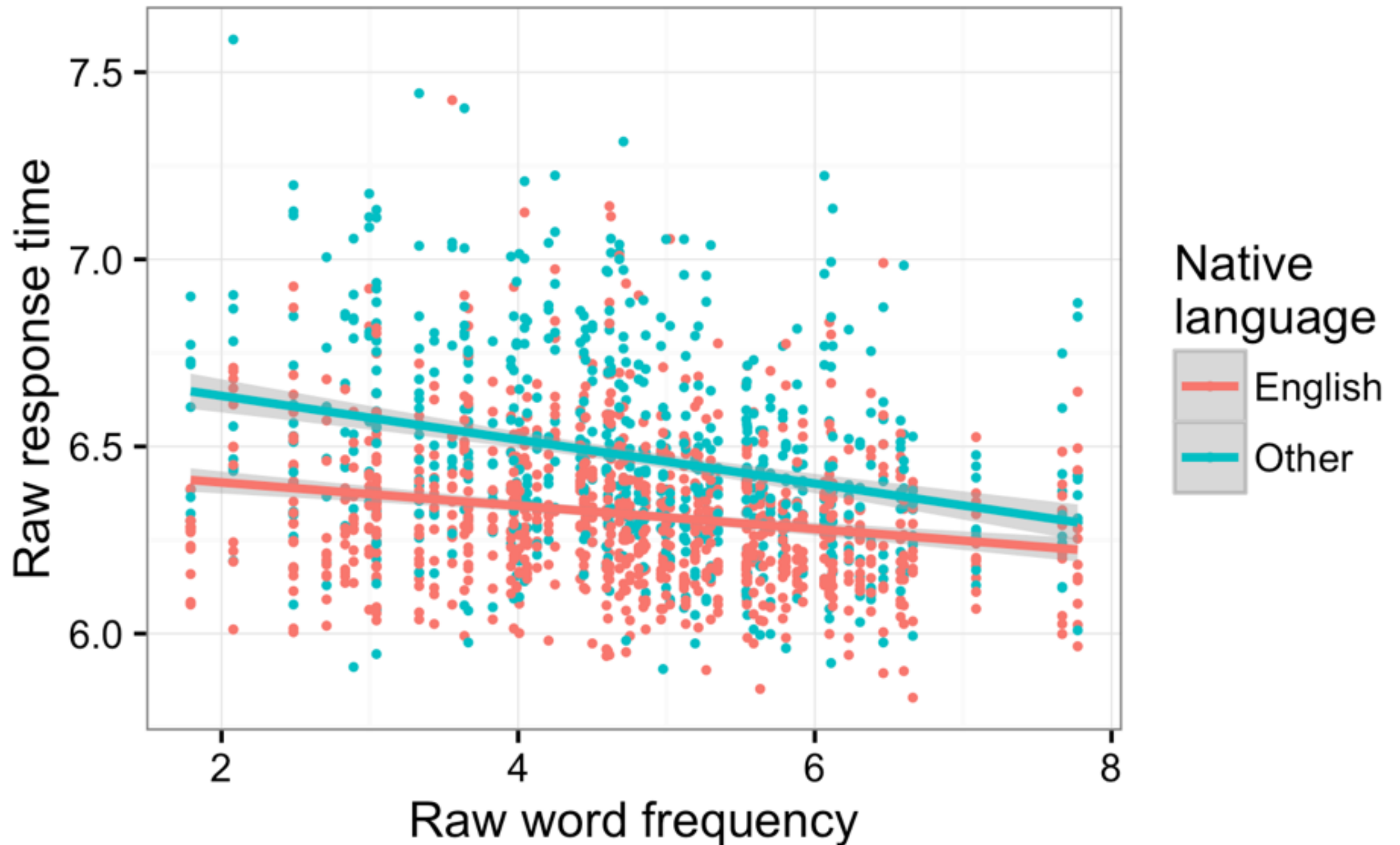
Interpretation of significant interactions: the slope of one predictor differs for different values of the other predictor.

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 tomorrow)
 - if the model can be trusted (see tomorrow)