Interpretable Linear Modeling

Jeffrey M. Girard

Carnegie Mellon University

Overview

Linear Modeling

Generalized Linear Modeling

Multilevel Modeling

Linear Modeling

What is linear modeling?

- Linear models (LM) are statistical prediction models with high interpretability
- They are best combined with a *small* number of *interpretable* features
- They allow you to quantify the *size* and *direction* of each feature's effect
- They allow you to isolate the *unique* effect of each feature

- LM incorporates regression, ANOVA, t-tests, and F-tests
- LM allows for multiple X (predictor or independent) variables
- LM allows for multiple Y (explained or dependent) variables

- LM assumes the Y variables are normally distributed (to avoid → GLM)
- LM assumes the data points are independent (to avoid → MLM)

Simple linear regression

- Linear regression is often presented with the following notation:

$$y_i = \alpha + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i$$

  - y is a vector of observations on the explained variable
  - x is a vector of observations on the predictor variable
  - α is the intercept
  - β are the slopes
  - ϵ is a vector of normally distributed and independent residuals

Simple linear regression

- Instead, we will use the following (equivalent) notation
- This notation will make it easier to generalize LM

$$y_i \sim \text{Normal}(\mu_i, \sigma) \qquad \text{Likelihood}$$

$$\mu_i = \alpha + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} \qquad \text{Linear model}$$

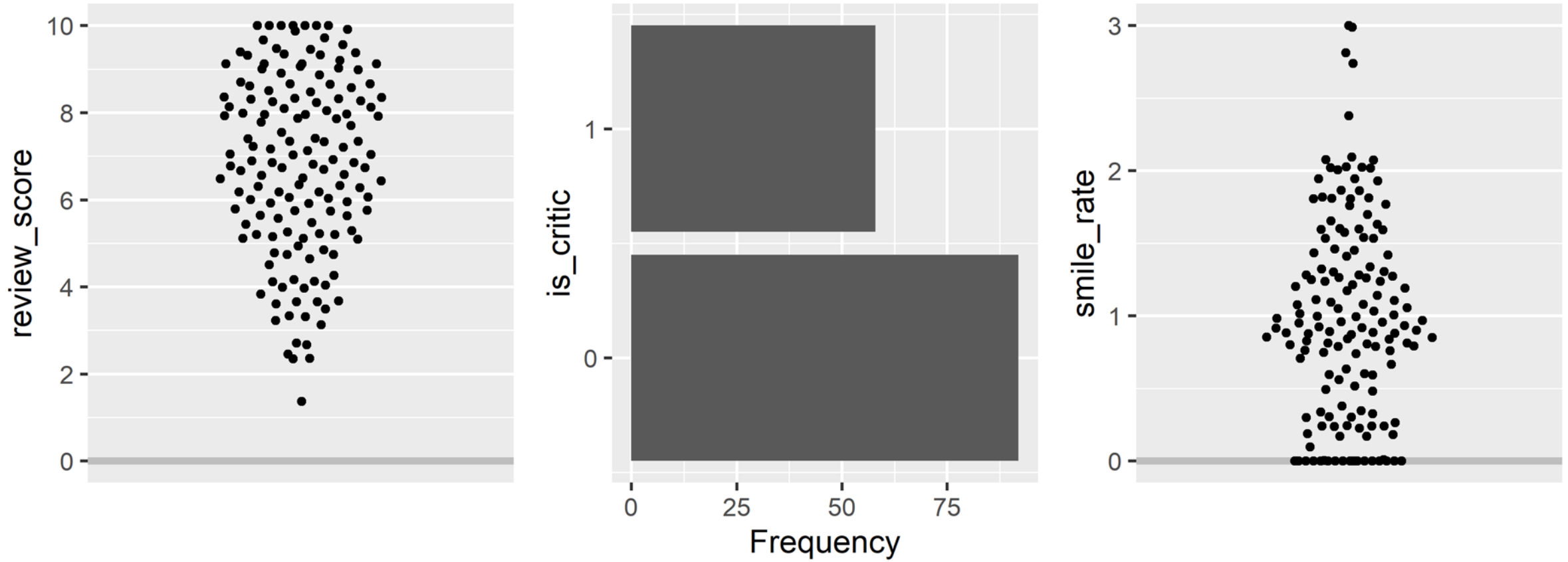
- This can be read as y is normally distributed around μ_i with SD σ
- μ_i relates y_i to x_i and defines the "best fit" prediction line
- σ captures the residuals or errors when $y_i \neq \mu_i$

Estimating the regression coefficients

- The goal is to find the parameter values that minimize the residuals
  - In linear modeling, this means estimating α, β, and σ
- This can be accomplished in several different ways
  - Maximum likelihood (i.e., minimize the residual sum of squares)
  - Bayesian approximation (e.g., maximum a posteriori or MCMC)
- I will provide code for use in R's *lme4* and in Python's *statsmodels*
  - These both use maximum likelihood estimation by default
  - I also highly recommend MCMC (especially for MLM)

Simulated LM example dataset

Simulated Data of 150 YouTube Book Review Videos



Example of LM with intercept-only

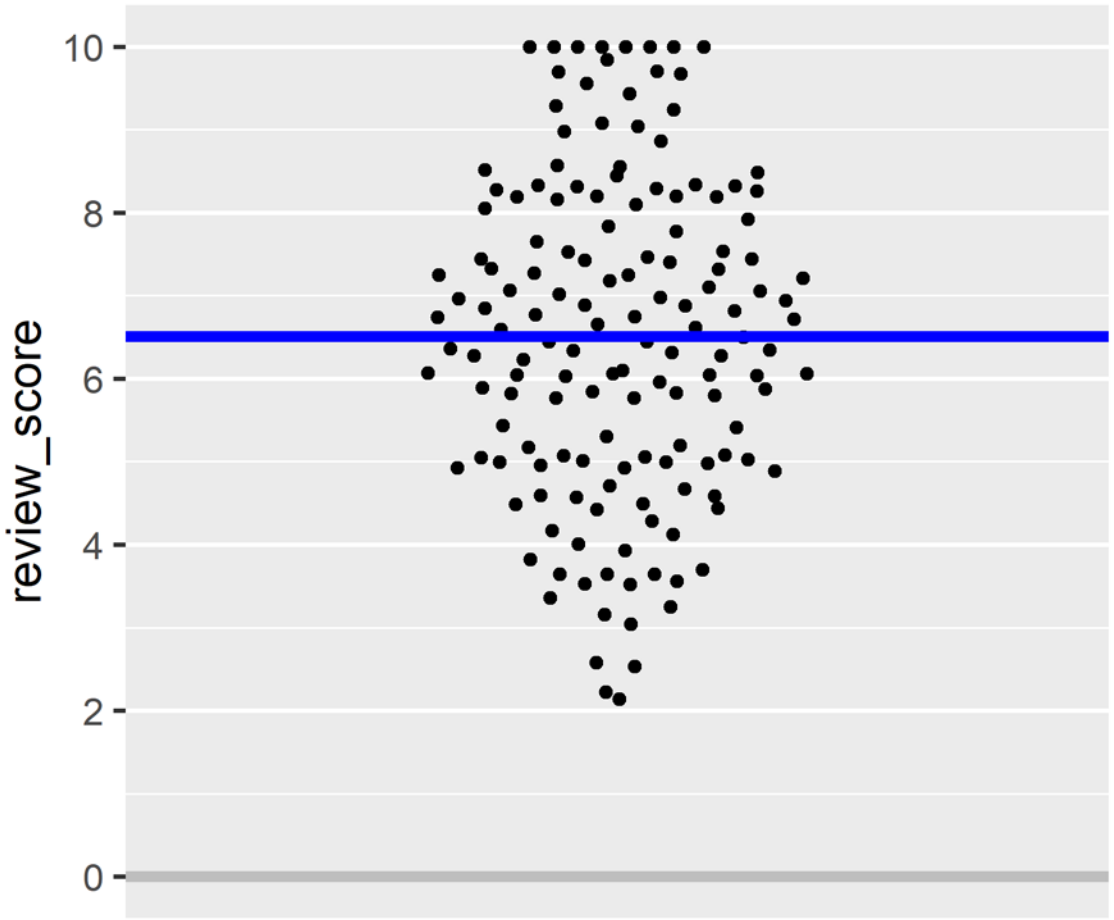
$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha$$

```
review_score ~ 1
```

	Variable	Estimate	95% CI
α	(Intercept)	6.51	[6.20, 6.83]
σ	Residual SD	1.94	

Deviance $= 563.16$, Adjusted $R^2 = 0.000$



Example of LM with binary predictor

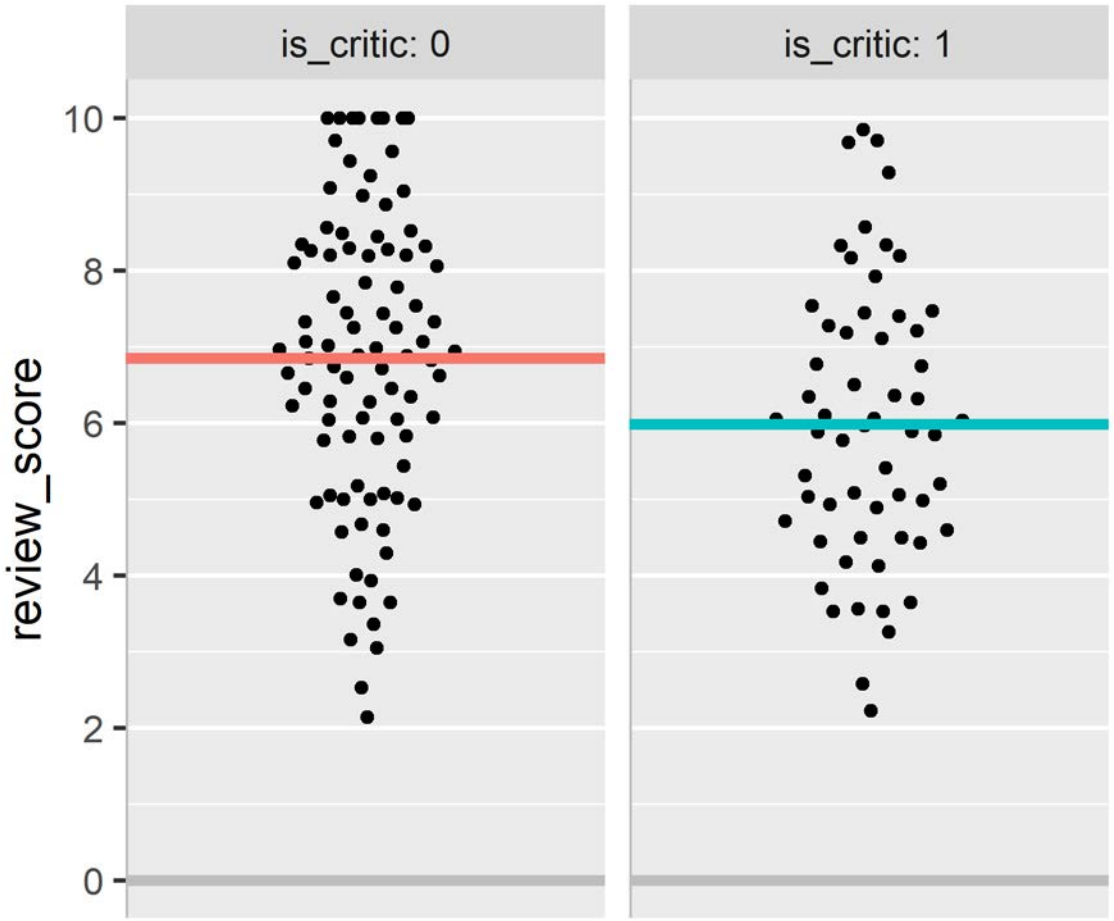
$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta x_i$$

```
review_score ~ 1 + is_critic
```

	Variable	Estimate	95% CI
α	(Intercept)	6.85	$[6.46, 7.24]$
β	is_critic	-0.87	$[-1.50, -0.24]$
σ	Residual SD	1.90	

Deviance $= 536.15$, Adjusted $R^2 = 0.042$



Example of LM with continuous predictor

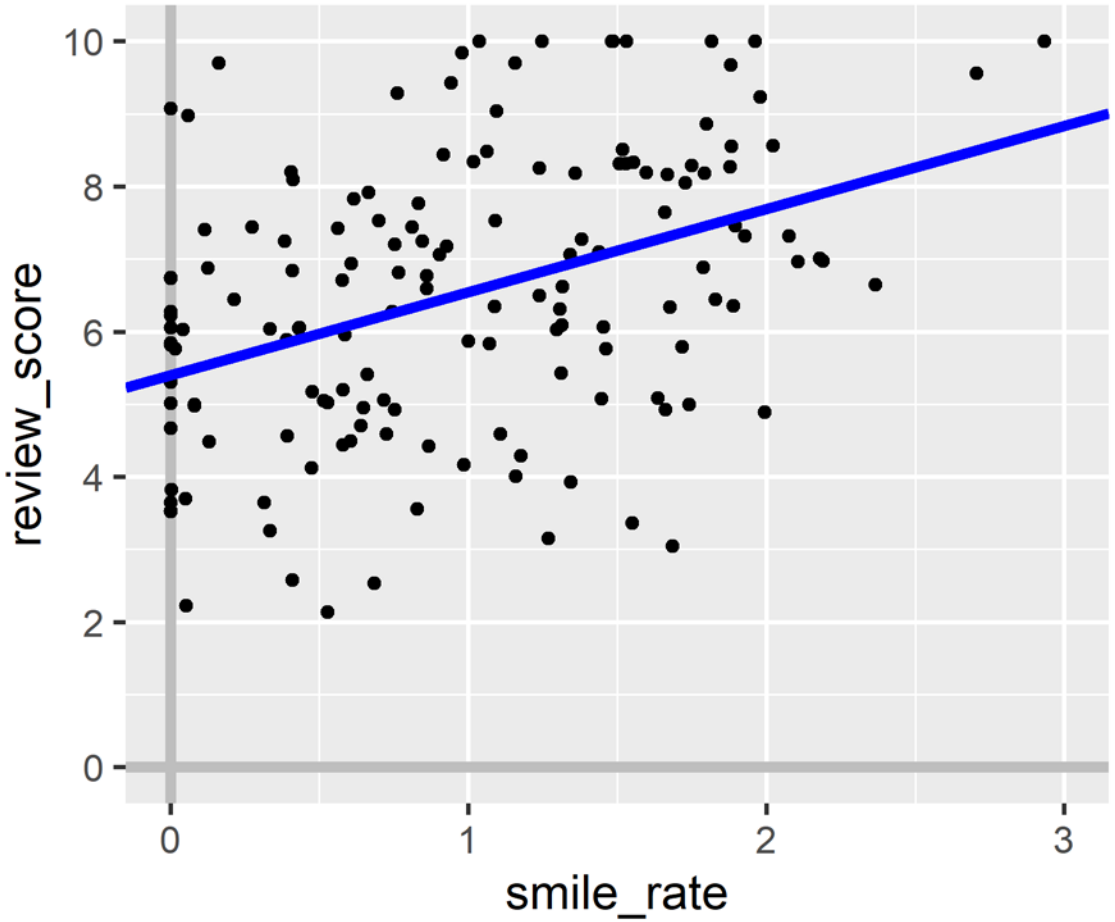
$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta x_i$$

```
review_score ~ 1 + smile_rate
```

	Variable	Estimate	95% CI
α	(Intercept)	5.40	$[4.90, 5.90]$
β	smile_rate	1.15	$[0.72, 1.57]$
σ	Residual SD	1.78	

Deviance $= 471.43$, Adjusted $R^2 = 0.157$



Example of LM with two predictors

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i}$$

```
review_score ~
1 + smile_rate + is_critic
```

	Variable	Estimate	95% CI
α	(Intercept)	5.71	$[5.13, 6.29]$
β_1	smile_rate	1.07	$[0.65, 1.49]$
β_2	is_critic	-0.61	$[-1.20, -0.01]$
σ	Residual SD	1.77	

Deviance $= 458.68$, Adjusted $R^2 = 0.174$

Example of LM with interaction effect

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i}$$

```
review_score ~ 1 +
smile_rate * is_critic
```

	Variable	Estimate	95% CI
α	(Intercept)	5.96	$[5.31, 6.61]$
β_1	smile_rate	0.83	$[0.33, 1.34]$
β_2	is_critic	-1.31	$[-2.32, -0.30]$
β_3	Interaction	0.79	$[-0.13, 1.70]$
σ	Residual SD	1.76	

Deviance $= 449.88$, Adjusted $R^2 = 0.185$

Generalized Linear Modeling

What is generalized linear modeling?

- Generalized linear modeling (GLM) is an extension of LM
- It allows LM to accommodate non-normally distributed *Y* variables
- Examples include variables that are: binary, discrete, and bounded

- This is accomplished using GLM families and link functions
- Families model the likelihood function using a specific distribution
- Link functions connect the linear model to the mean of the likelihood
- Link function also ensure that the mean is the "right" kind of number

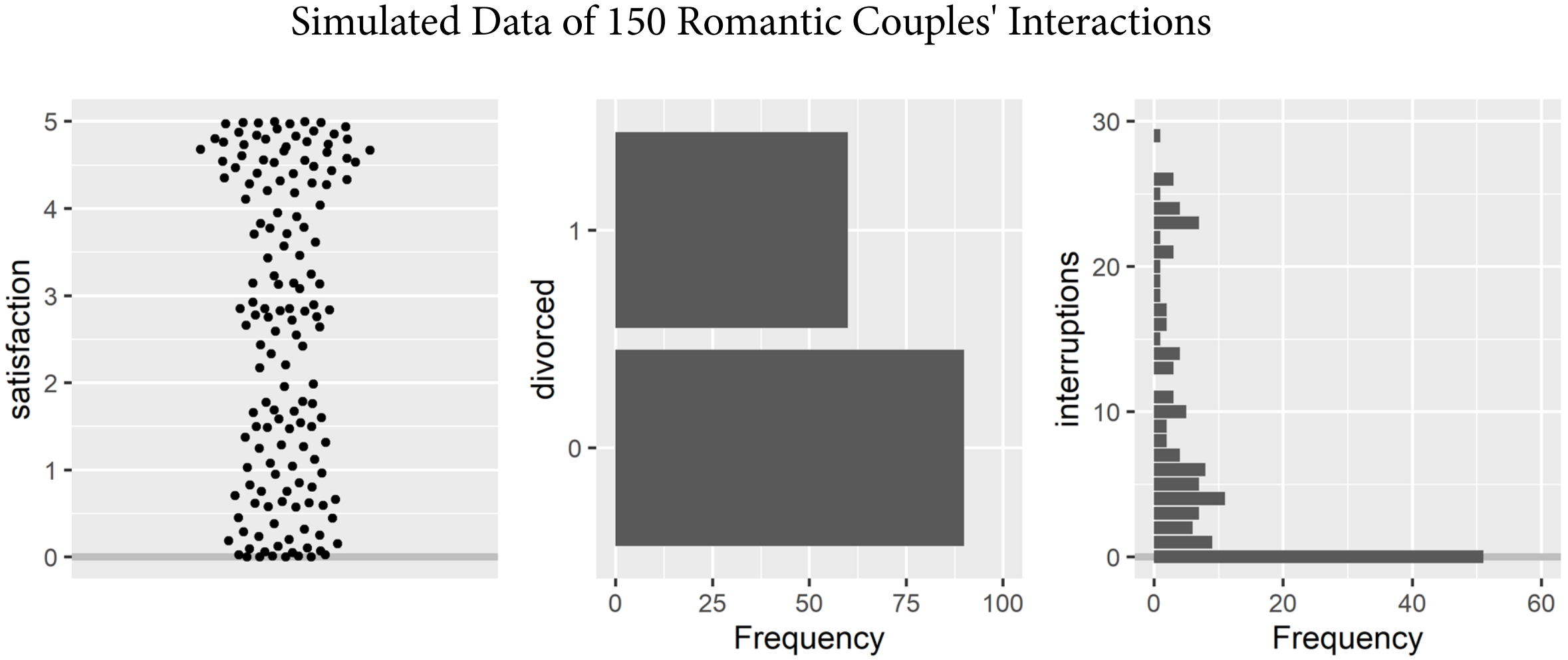
Common GLM families and link functions

$$y_i \sim \text{Family}(\mu_i, \dots)$$

$$\text{link}(\mu_i) = \alpha + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$$

Family	Uses	Supports	Link Function	
Normal	Linear data	$\mathbb{R}$: $(-\infty, \infty)$	Identity	μ
Gamma	Non-negative data (reaction times)	$\mathbb{R}$: $(0, \infty)$	Inverse or Power	μ^{-1}
Poisson	Count data	$\mathbb{Z}$: $0, 1, 2, \dots$	Log	$\log(\mu)$
Binomial	Binary data Categorical data	$\mathbb{Z}$: $\{0, 1\}$ $\mathbb{Z}$: $[0, K)$	Logit	$\log\left(\frac{\mu}{1-\mu}\right)$

Simulated GLM example dataset



Example of LM for binary data

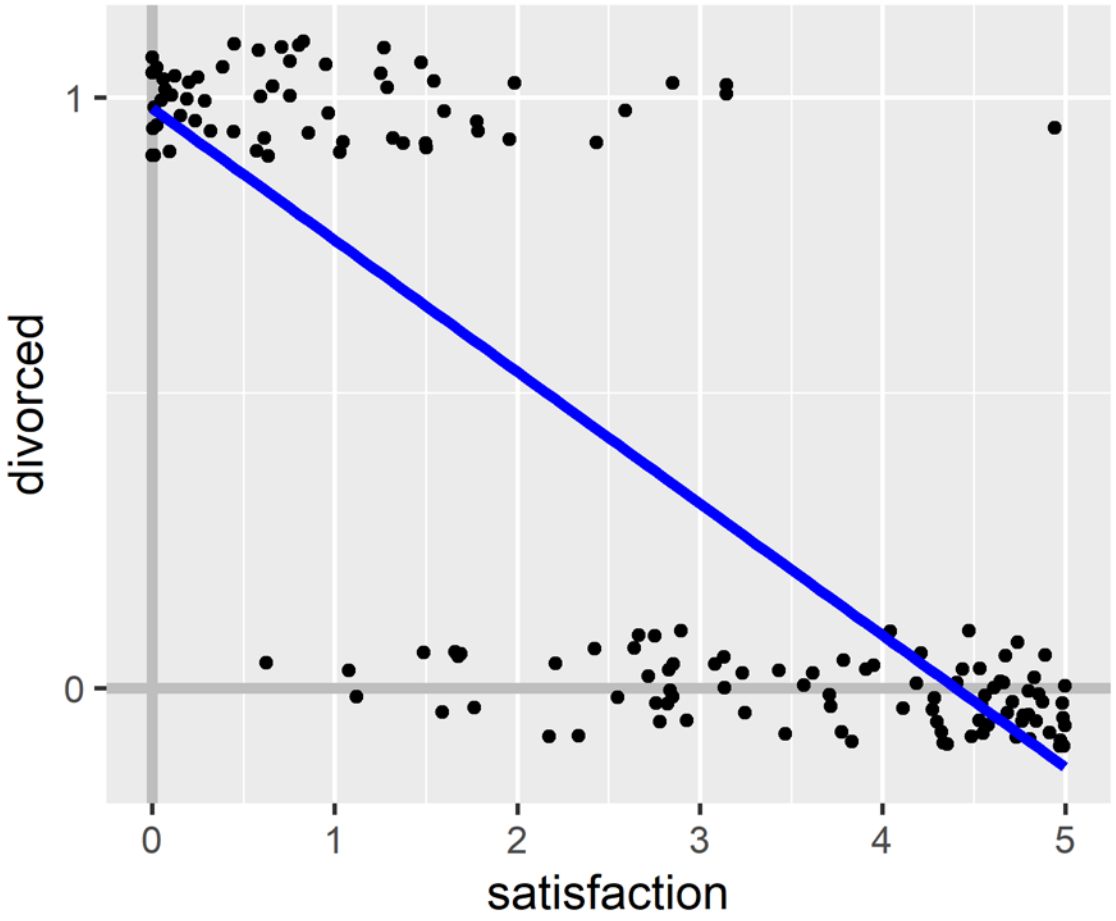
$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta x_i$$

```
divorced ~ 1 + satisfaction
```

	Variable	Estimate	95% CI
α	(Intercept)	0.98	$[0.89, 1.07]$
β	satisfaction	-0.22	$[-0.25, -0.19]$
σ	Residual SD	0.304	

Deviance $= 13.70$, Adjusted $R^2 = 0.617$



Example of GLM for binary data

$$y_i \sim \text{Binomial}(1, p_i)$$

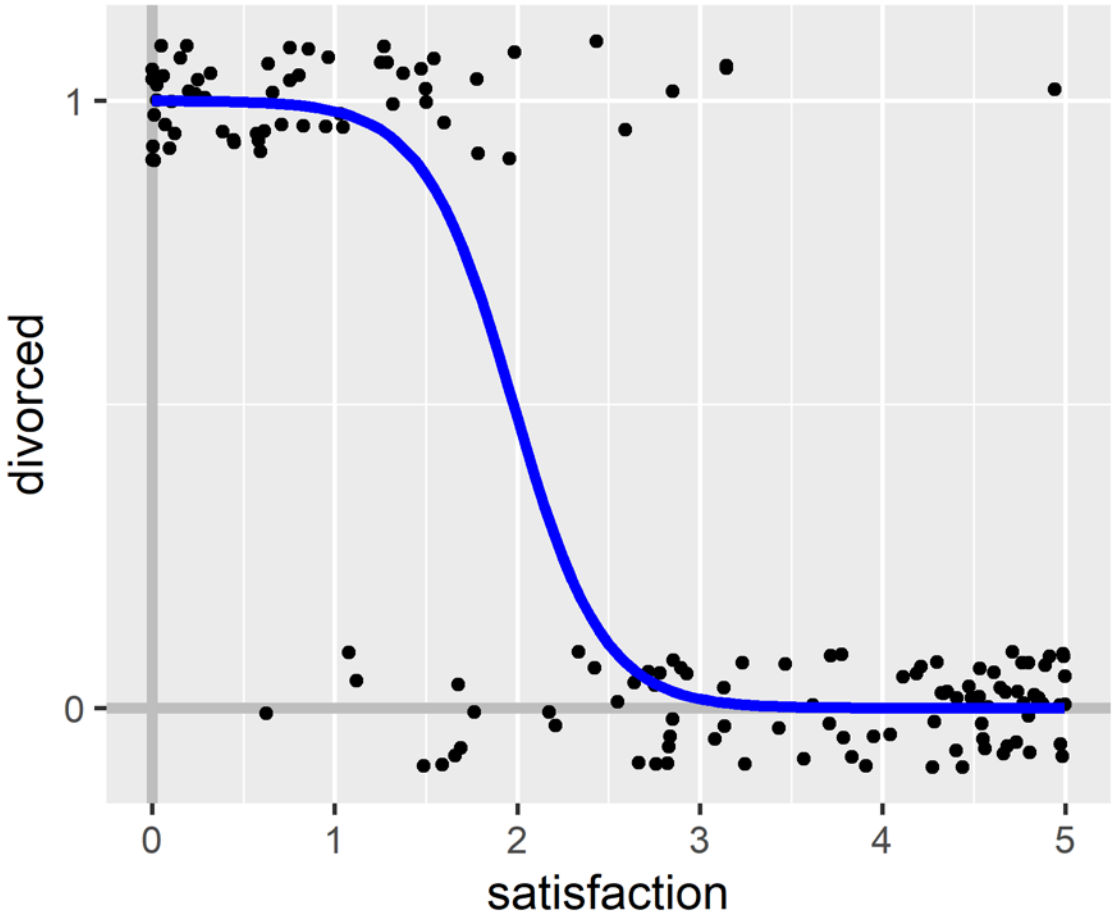
$$\text{logit}(p_i) = \alpha + \beta x_i$$

```
divorced ~ 1 + satisfaction
family = binomial(link = "logit")
```

	Variable	Estimate	95% CI
α	(Intercept)	3.52	$[2.45, 4.87]$
β	satisfaction	-1.78	$[-2.40, -1.30]$

$\text{Deviance} = 82.04, \qquad \text{Pseudo } R^2 = 0.664$

Note: Results are in transformed (i.e., logit) units



Example of LM for count data

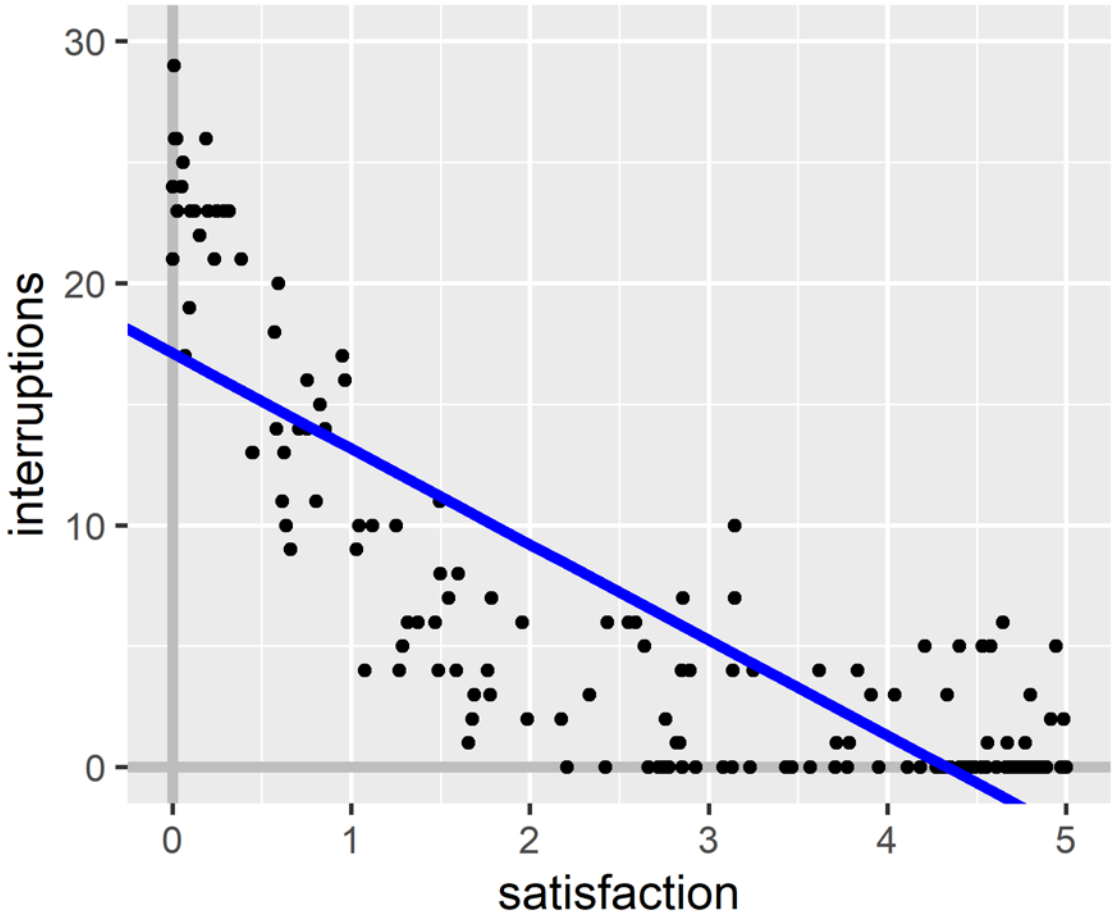
$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta x_i$$

```
interruptions ~ 1 + satisfaction
```

	Variable	Estimate	95% CI
α	(Intercept)	17.11	$[15.76, 18.45]$
β	satisfaction	-3.95	$[-4.38, -3.52]$
σ	Residual SD	4.61	

Deviance $= 3143.62$, Adjusted $R^2 = 0.688$



Example of GLM for count data

$$y_i \sim \text{Poisson}(\lambda_i)$$

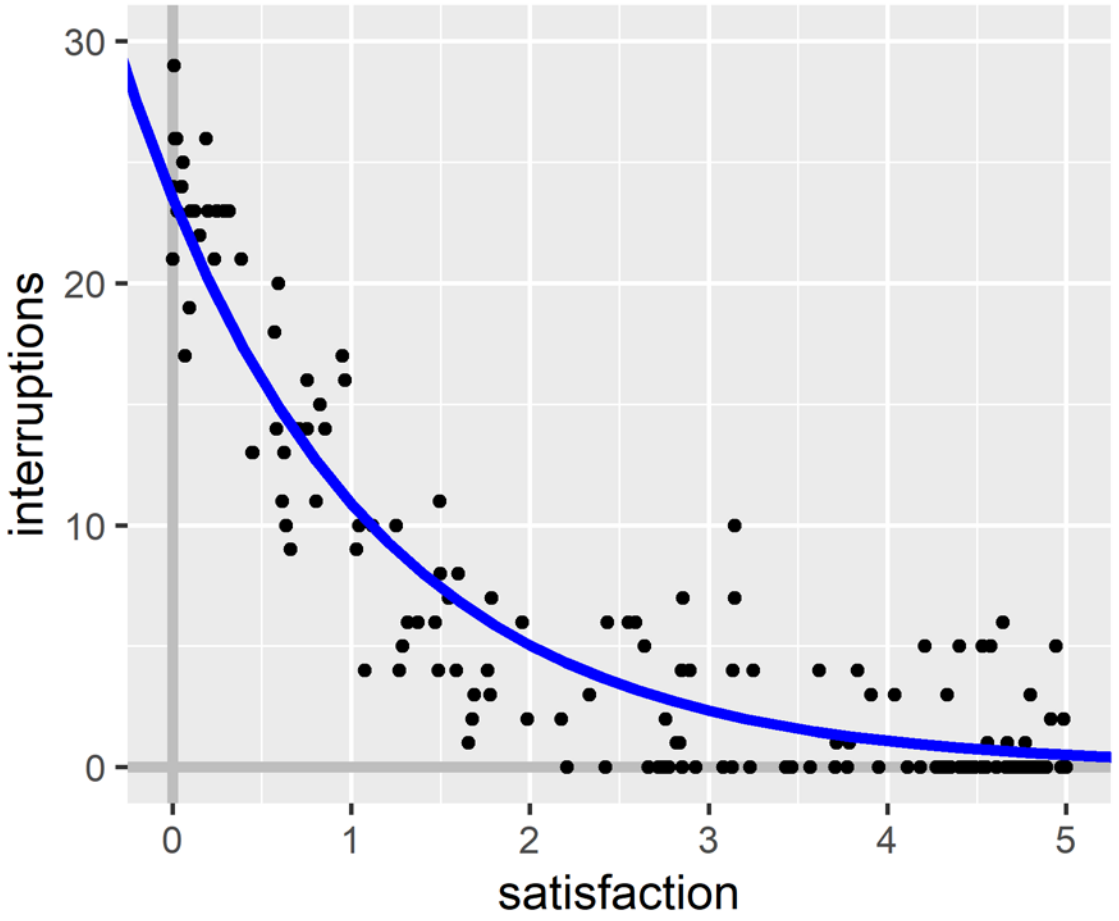
$$\log(\lambda_i) = \alpha + \beta x_i$$

```
interruptions ~ 1 + satisfaction
 family = poisson(link = "log")
```

	Variable	Estimate	95% CI
α	(Intercept)	3.16	$[3.08, 3.24]$
β	satisfaction	-0.77	$[-0.83, -0.72]$

$\text{Deviance} = 325.24, \qquad \text{Pseudo } R^2 = 0.895$

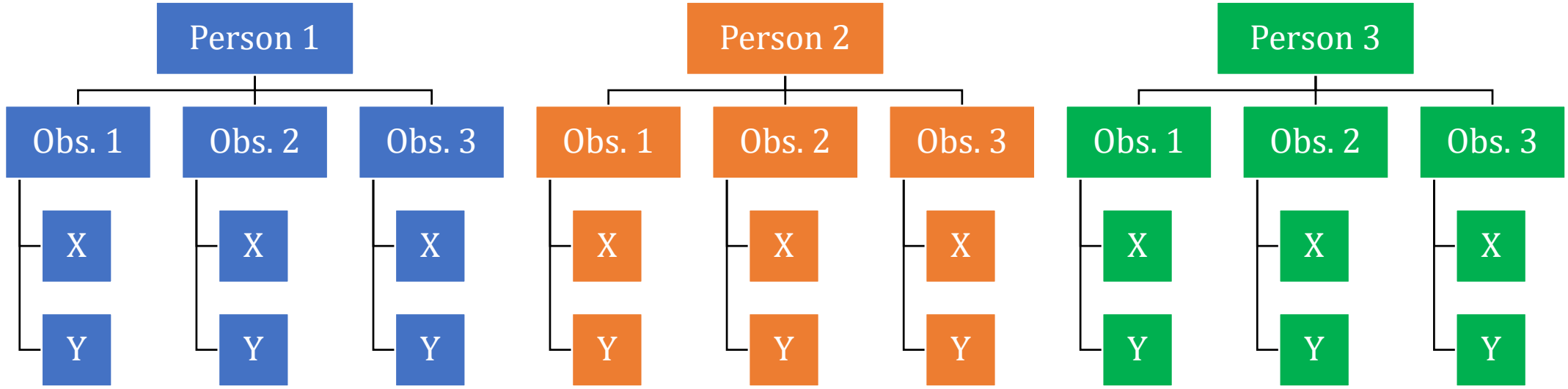
Note: Results are in transformed (i.e., log) units



Multilevel Modeling

What is multilevel modeling?

- MLM allows (G)LM to handle data points that are dependent/clustered
- This is problematic because data within a cluster are likely to be similar
- Effects (e.g., intercepts and slopes) may also differ between clusters



What is multilevel modeling?

- One approach would be to estimate effects across all clusters (complete pooling)
  - *But this would require all clusters to be nearly identical, which they may not be*
- Another would be to estimate effects for each cluster separately (no pooling)
  - *But this would ignore similarities between clusters and may overfit the data*
- The MLM approach tries to have the best of both approaches (partial pooling)
  - *MLM tries to learn the similarities and the differences between clusters*
  - It estimates effects for each individual cluster separately
  - But these effects are assumed to be drawn from a population of clusters
  - MLM explicitly models this population and the cluster-specific variation

What is multilevel modeling?

- MLM comes with several important benefits over single-level models
  - Improved estimates for repeat sampling
  - Improved estimates for imbalanced sampling
  - Estimates of variation across clusters
  - Avoid averaging and retain variation
  - Better accuracy and inference!
- Many have argued that "multilevel regression deserves to be the default approach"

Varying intercepts

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

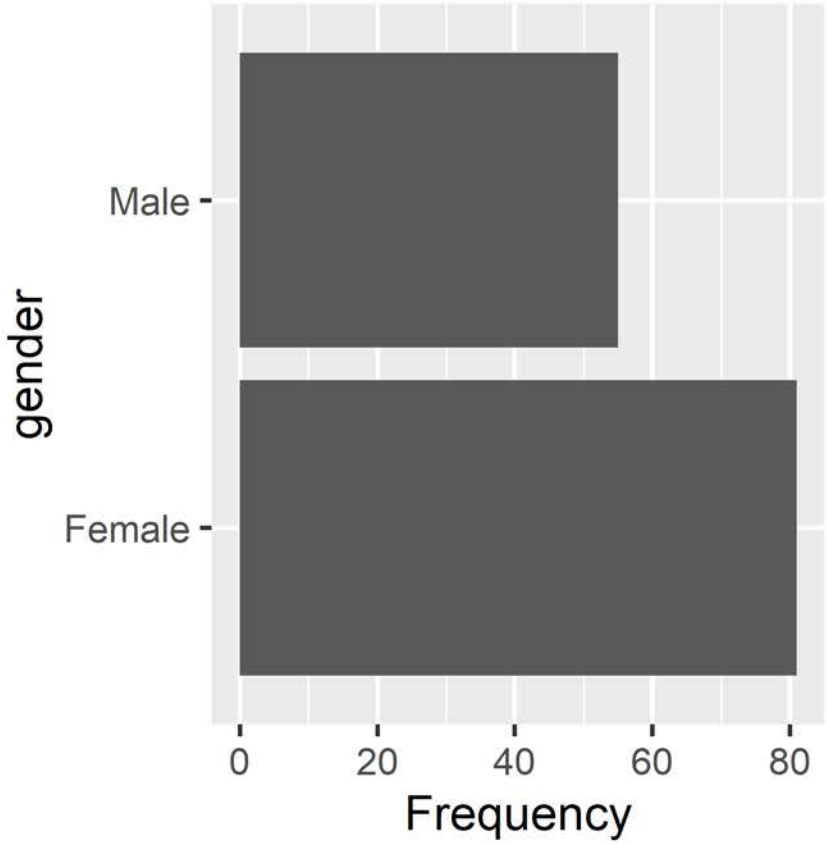
$$\mu_i = \alpha_{\text{CLUSTER}[i]}$$

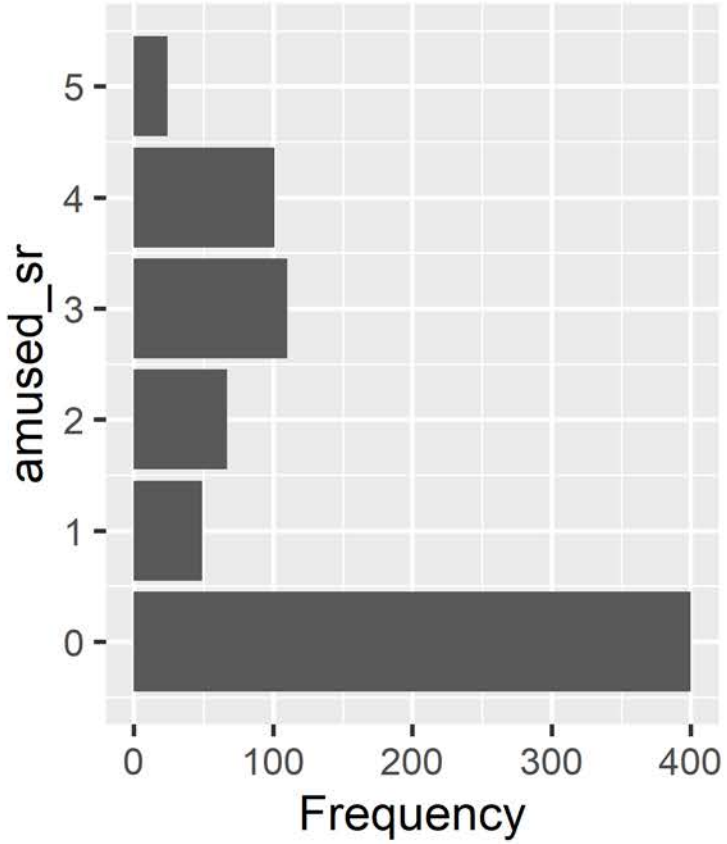
$$\alpha_{\text{CLUSTER}} \sim \text{Normal}(\alpha, \sigma_{\text{CLUSTER}})$$

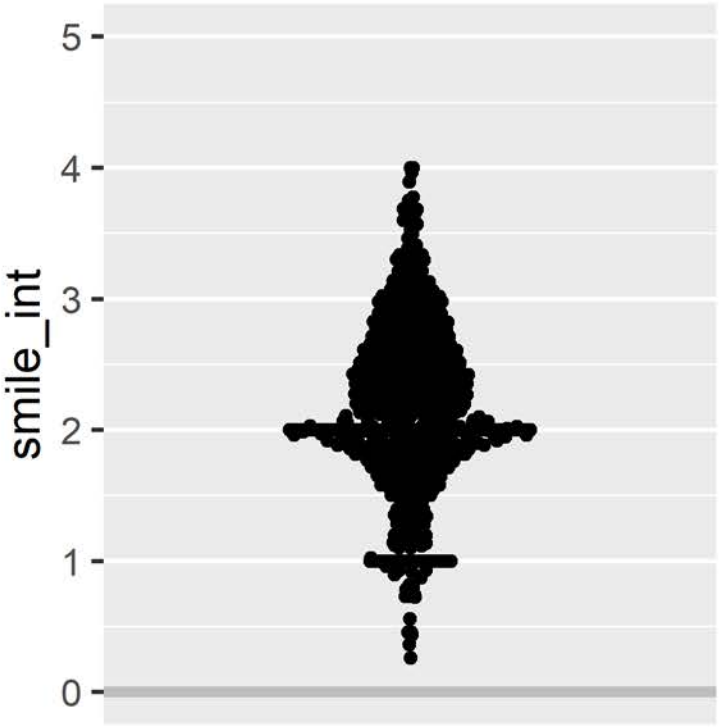
- α captures the overall intercept (i.e., the mean of all clusters' intercepts)
- $\alpha_{\text{CLUSTER}[i]}$ captures each cluster's deviation from the overall intercept
- σ_{CLUSTER} captures how much variability there is across clusters' intercepts
- We now have an intercept for every cluster and a "population" of intercepts
- The model learns about each cluster's intercept from the population of intercepts
  - This pooling *within* parameters is very helpful for imbalanced sampling

Actual MLM example dataset

Real Data of 751 Smiles from 136 Subjects (Girard, Shandar, et al. 2019)







Complete pooling (underfitting)

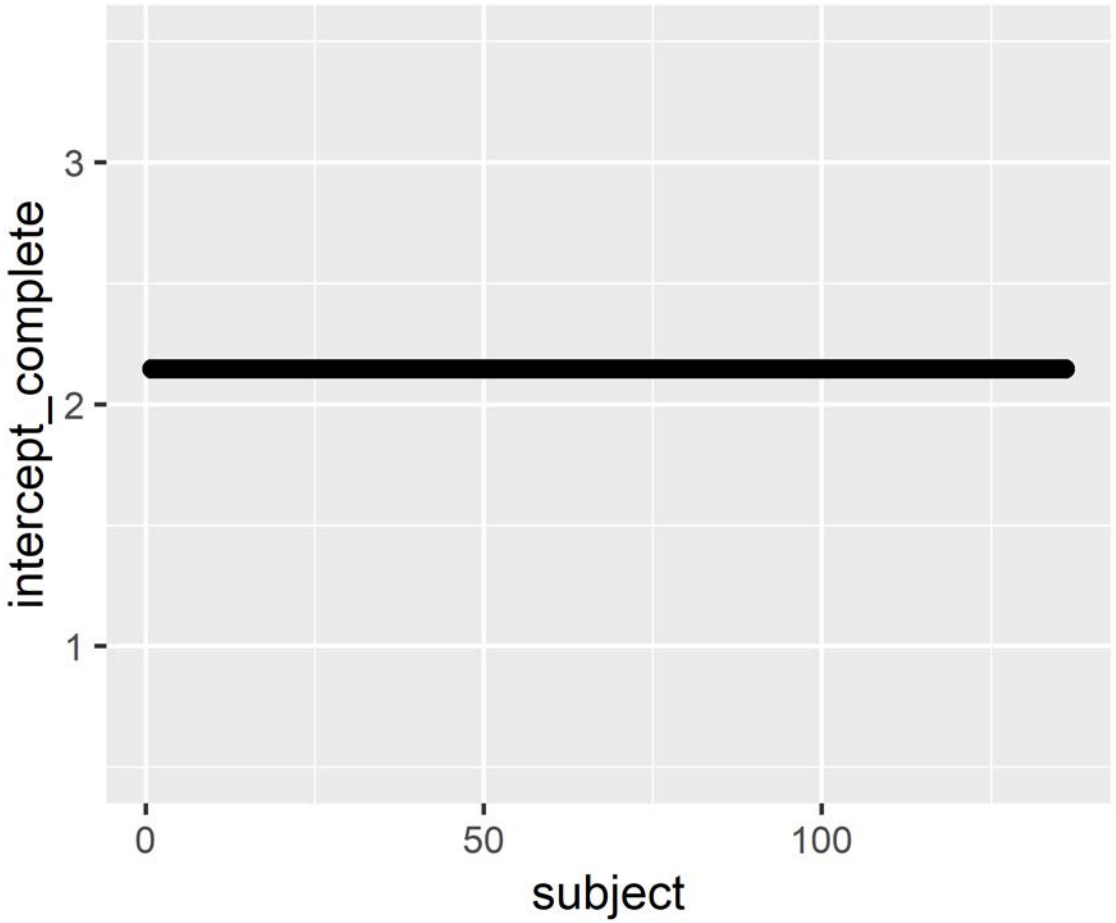
$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha$$

```
smile_int ~ 1
```

	Variable	Estimate	95% CI
α	(Intercept)	2.14	[2.10, 2.19]
σ	Residual SD	0.66	

Deviance = 1495.8



No pooling (overfitting)

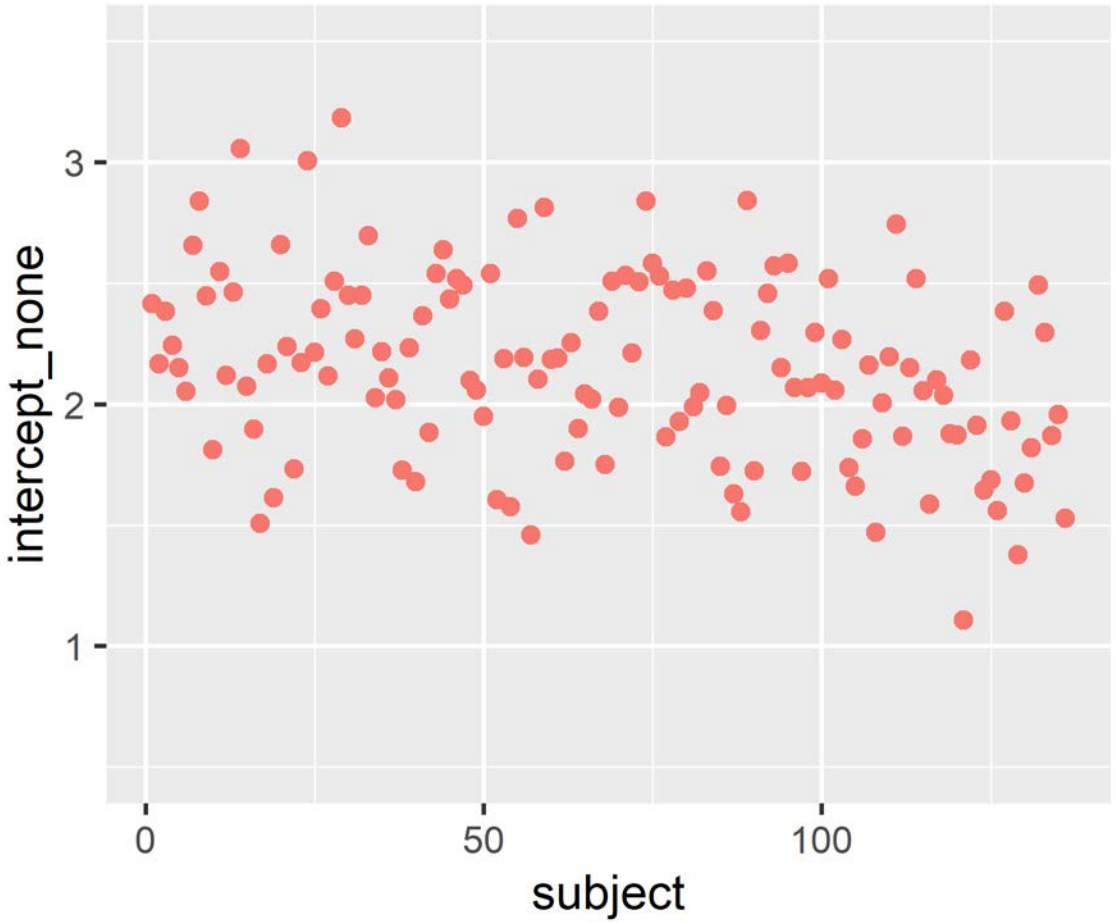
$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha_{\text{PERSON}[i]}$$

```
smile_int ~ 1 + factor(subject)
```

	Variable	Estimate	95% CI
$\alpha_{[1]}$	Intercept for P1	2.14	[1.93, 2.89]
$\alpha_{[2]}$	Intercept for P2	2.17	[1.49, 2.85]
…	…	…	…
σ	Residual SD	0.60	

$$\text{Deviance} = 1210.10$$



Partial pooling (ideal fit)

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

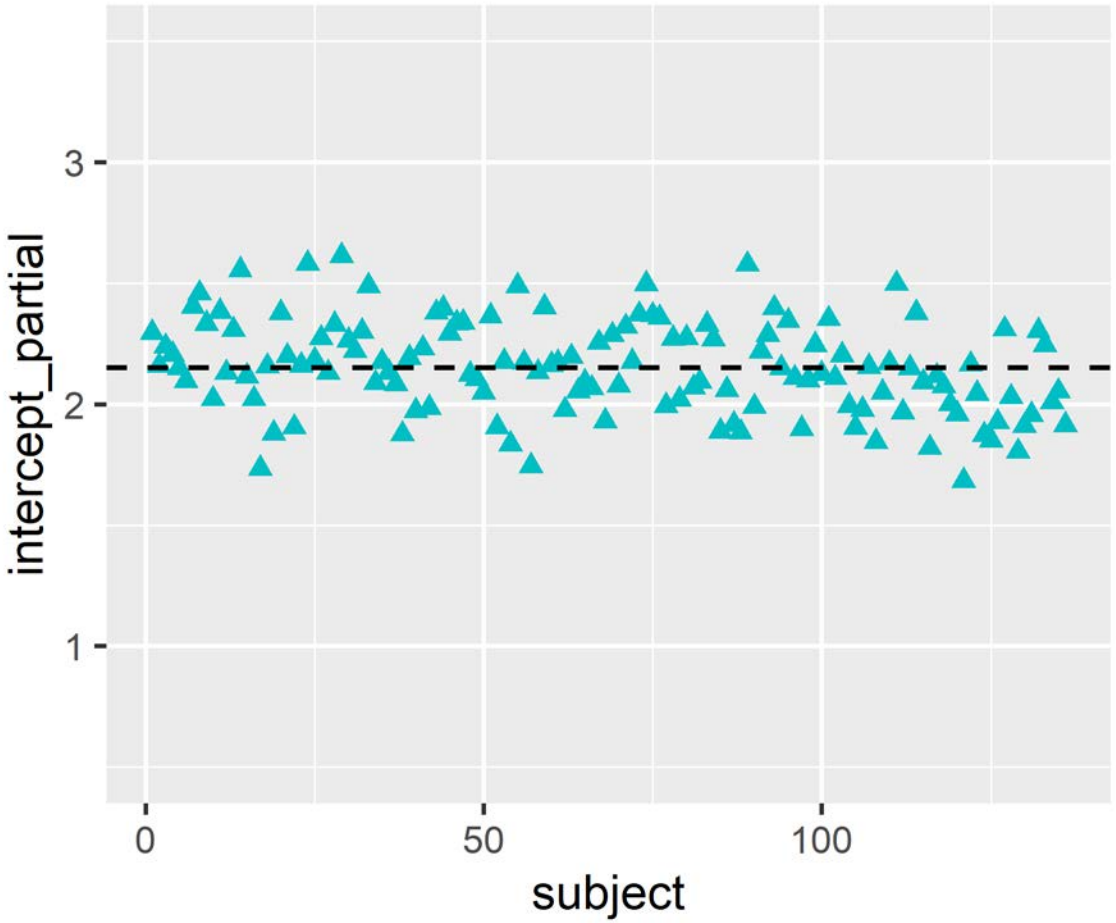
$$\mu_i = \alpha_{PERSON[i]}$$

$$\alpha_{PERSON} \sim \text{Normal}(\alpha, \sigma_{PERSON})$$

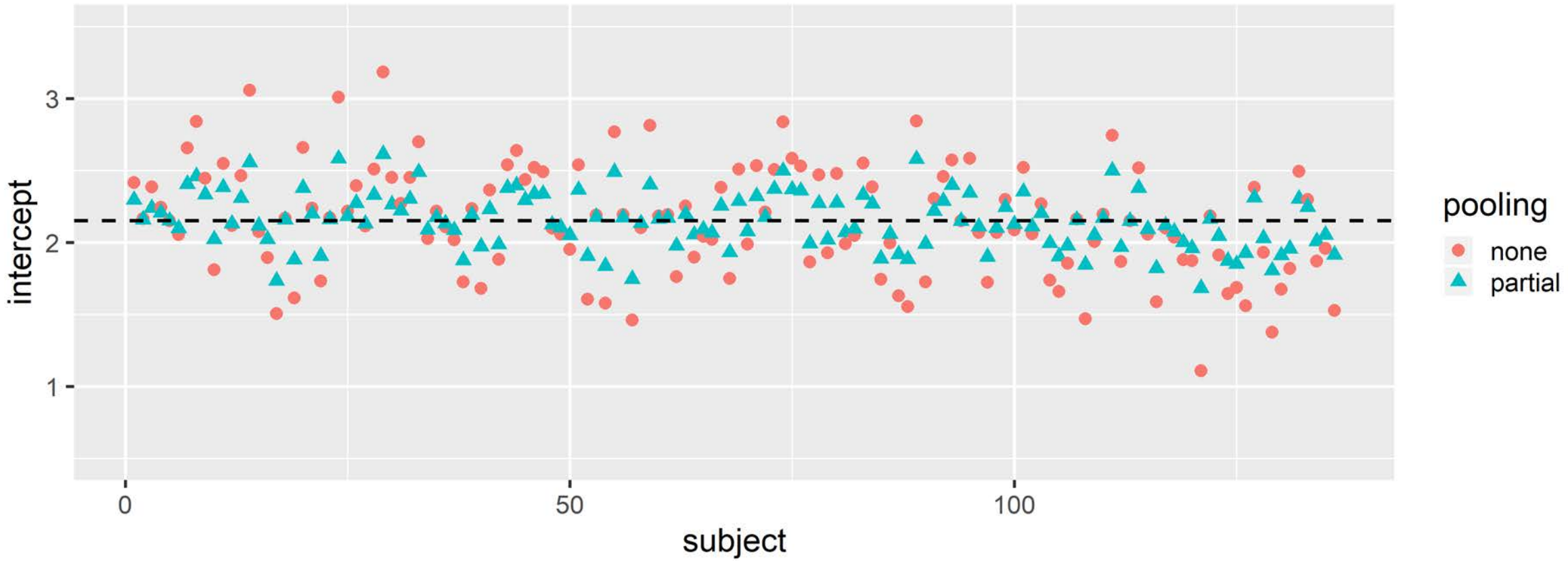
```
smile_int ~ 1 + (1 | subject)
```

	Variable	Estimate	95% CI
α	(Intercept)	2.15	[2.09, 2.21]
σ_P	Intercept SD	0.27	
σ	Residual SD	0.60	

$$\text{Deviance} = 1458.81$$



Shrinkage in action



Varying intercepts and slopes

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha_{\text{CLUSTER}[i]} + \beta_{\text{CLUSTER}[i]} x_i$$

$$\begin{bmatrix} \alpha_{\text{CLUSTER}} \\ \beta_{\text{CLUSTER}} \end{bmatrix} \sim \text{MVNormal}\left(\begin{matrix} \alpha \\ \beta \end{matrix}, \mathbf{S}\right)$$

$$\mathbf{S} = \begin{pmatrix} \sigma_\alpha & 0 \\ 0 & \sigma_\beta \end{pmatrix} \mathbf{R} \begin{pmatrix} \sigma_\alpha & 0 \\ 0 & \sigma_\beta \end{pmatrix}$$

- We can pool and learn from the covariance between varying intercepts and slopes
- We do this with a 2D normal distribution with means α and β and covariance matrix **S**
- We build **S** through matrix multiplication of the variances and correlation matrix **R**
- This is more complex but results in even better pooling: within and *across* parameters

MLM with varying intercepts and slopes

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha_{\text{PERSON}[i]} + \beta_{\text{PERSON}[i]} x_i$$

$$\begin{bmatrix} \alpha_{\text{PERSON}} \\ \beta_{\text{PERSON}} \end{bmatrix} \sim \text{MVNormal}\left(\begin{matrix} \alpha \\ \beta \end{matrix}, \mathbf{S}\right)$$

$$\mathbf{S} = \begin{pmatrix} \sigma_\alpha & 0 \\ 0 & \sigma_\beta \end{pmatrix} \mathbf{R} \begin{pmatrix} \sigma_\alpha & 0 \\ 0 & \sigma_\beta \end{pmatrix}$$

```
smile_int ~ 1 + amused_sr +
  (1 + amused_sr| subject)
```

	Variable	Estimate	95% CI
α	(Intercept)	1.98	$[1.91, 2.06]$
β	amused_sr	0.12	$[0.09, 0.15]$
σ_α	Intercept SD	0.24	$[0.17, 0.32]$
σ_β	Slope SD	0.04	$[0.00, 0.08]$
$\mathbf{R}$	$\text{corr}(\sigma_\alpha, \sigma_\beta)$	0.30	$[-0.69, 0.96]$
σ	Residual SD	0.57	$[0.54, 0.60]$

Depiction of varying intercepts and slopes