Language Technologies Institute

Carnegie Mellon University

Multimodal Affective Computing

Lecture 8: Statistical Modeling

Jeffrey Girard
Louis-Philippe Morency

Outline of this week's lecture

1. The general linear model (LM)
2. The generalized linear model (GLM)
3. Preview of advanced frameworks
   - Multilevel modeling (MLM)
   - Structural equation modeling (SEM)
   - Regularization and prediction (GLMNET)

What is a model?

- Models tell a story of how the observed data came to be
- This story is translated into a formal probability model
- We can then quantify and compare models' fit and stability
- Models are neither true nor false, but they can be useful
- Models can be used for description and for prediction
  - In science, we are usually more interested in interpretable description
  - In engineering, we are usually more interested in accurate prediction
  - Depending on our priorities, we can choose different types of model

Types of models

- Models often get packaged into specific "tests"
  - ANOVA, ANCOVA, MANOVA, MANCOVA, t-tests, F-tests, etc.
  - Linear regression, multiple regression, multivariate regression, etc.
  - Logistic regression, Poisson regression, polynomial regression, etc.
  - Multilevel models, factor analytic models, mixture models, etc.
- These tests are often treated as if they are totally different
- However, there is an underlying uniformity to these models
- It is often possible to collapse them into broader frameworks
- Today we will discuss several such modeling frameworks

The general linear model (LM)

What is the general linear model?

- The general linear model (LM) is flexible and expandable
  - It incorporates linear regression^{+}, ANOVA^{+}, t-tests, and F-tests
  - It allows for multiple X (predictor or independent) variables
  - It allows for multiple Y (explained or dependent) variables
  - It can be further expanded into GLM, MLM, SEM, GLMNET, etc.
- Situations when LM is a particularly good choice
  - You have relatively few predictor variables
  - Your predictor variables are meaningful/interpretable
  - You want to know the direction and size of every relationship
  - You have reason to believe the LM assumptions are met

A refresher on linear regression

- Linear regression is often presented as:

$$y = \beta_0 + \beta_1 x + \epsilon$$

  - y is a vector of observations on the explained variable
  - x is a vector of observations on the predictor variable
  - β are the model parameters
    - β_0 is the "intercept"
    - β_1 is the "slope"
  - ϵ is an error or residual term

A refresher on linear regression

- We want to find the β values that minimize the residuals
- One approach is to minimize the residual sum of squares

$$RSS(\beta) = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \sum_{i=1}^{n} (y_i - \beta^T x_i)^2$$

- This calculates the "maximum likelihood" estimates of β

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- This approach is called Ordinary Least Squares (OLS)

What is the general linear model?

- To expand linear regression to LM, we rewrite it as:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U}$$

  - $\mathbf{Y}$ is a matrix of observations on explained variables
  - $\mathbf{X}$ is a matrix of observations on predictor variables
  - $\mathbf{B}$ is a matrix of model parameters to be estimated
  - $\mathbf{U}$ is a matrix containing errors/residuals

$$RSS(\mathbf{B}) = \sum_{i=1}^{n} (\mathbf{Y}_i - \mathbf{B}^T \mathbf{X}_i)^2$$

How to implement the general linear model

- Most statistical software includes a function for LM using OLS
- We will focus on **statsmodels** in Python and **stats** in R
- Both take in a data frame and a "design formula"
- Both will provide parametric confidence intervals by default
- ```
  import statsmodels.formula.api as sm
  res = sm.ols(formula='y~1+x1+x2+x1*x2', data=df).fit()
  print(res.summary())
  ```
- ```
  res <- stats::lm(formula='y~1+x1+x2+x1*x2', data=df)
  summary(res)
  confint(res)
  ```

A simple example

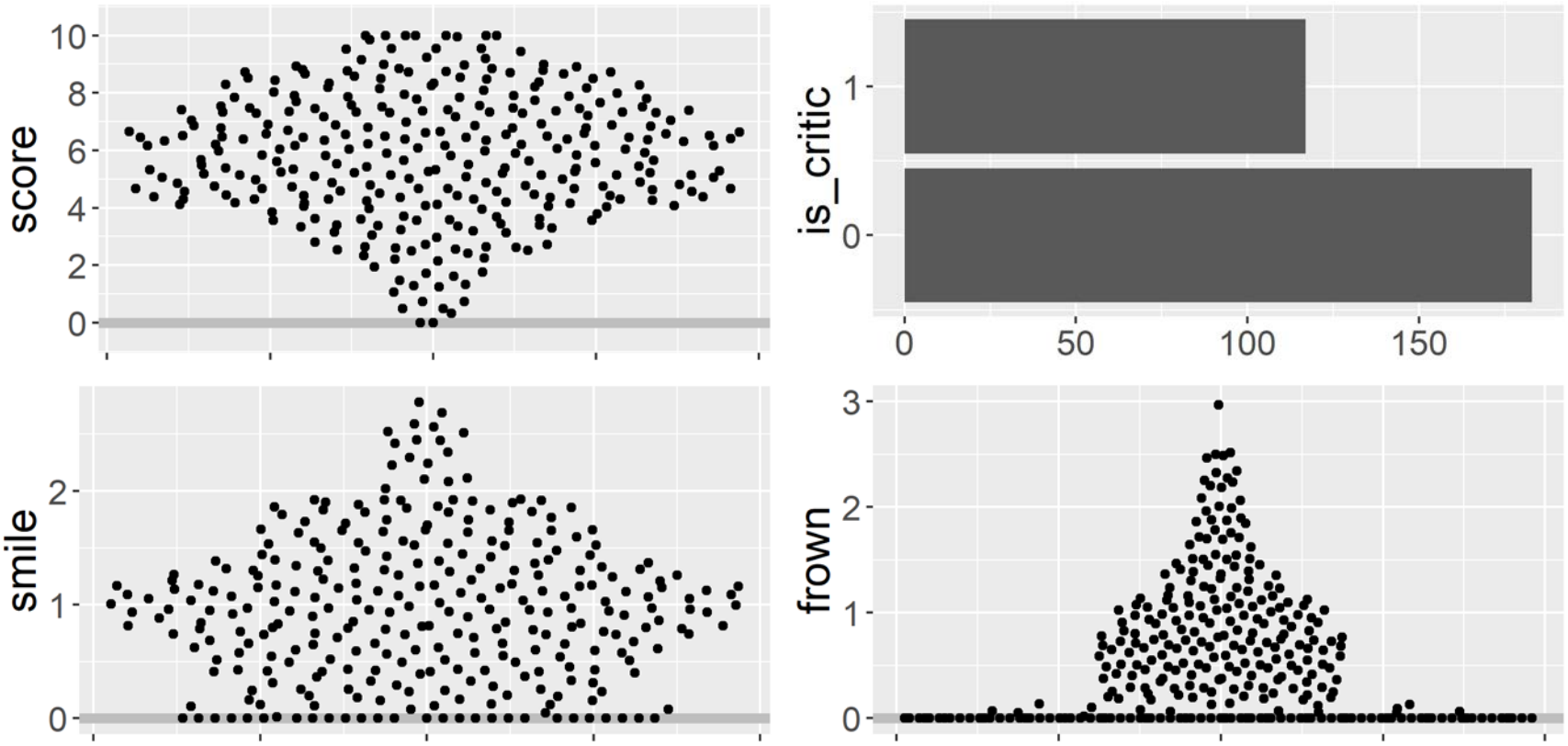
Background

- We download 300 book review videos from YouTube
- We measure reviewers' rates of smiling and frowning
- We record reviewers' review scores and professional status

Questions

- Does facial behavior reveal what the review score was?
- Do smiling and frowning provide unique information?
- Does the effect of smiling depend on professional status?

Visualize the distributions



Intercept-only

$$y_i = \beta_0 + \epsilon_i$$

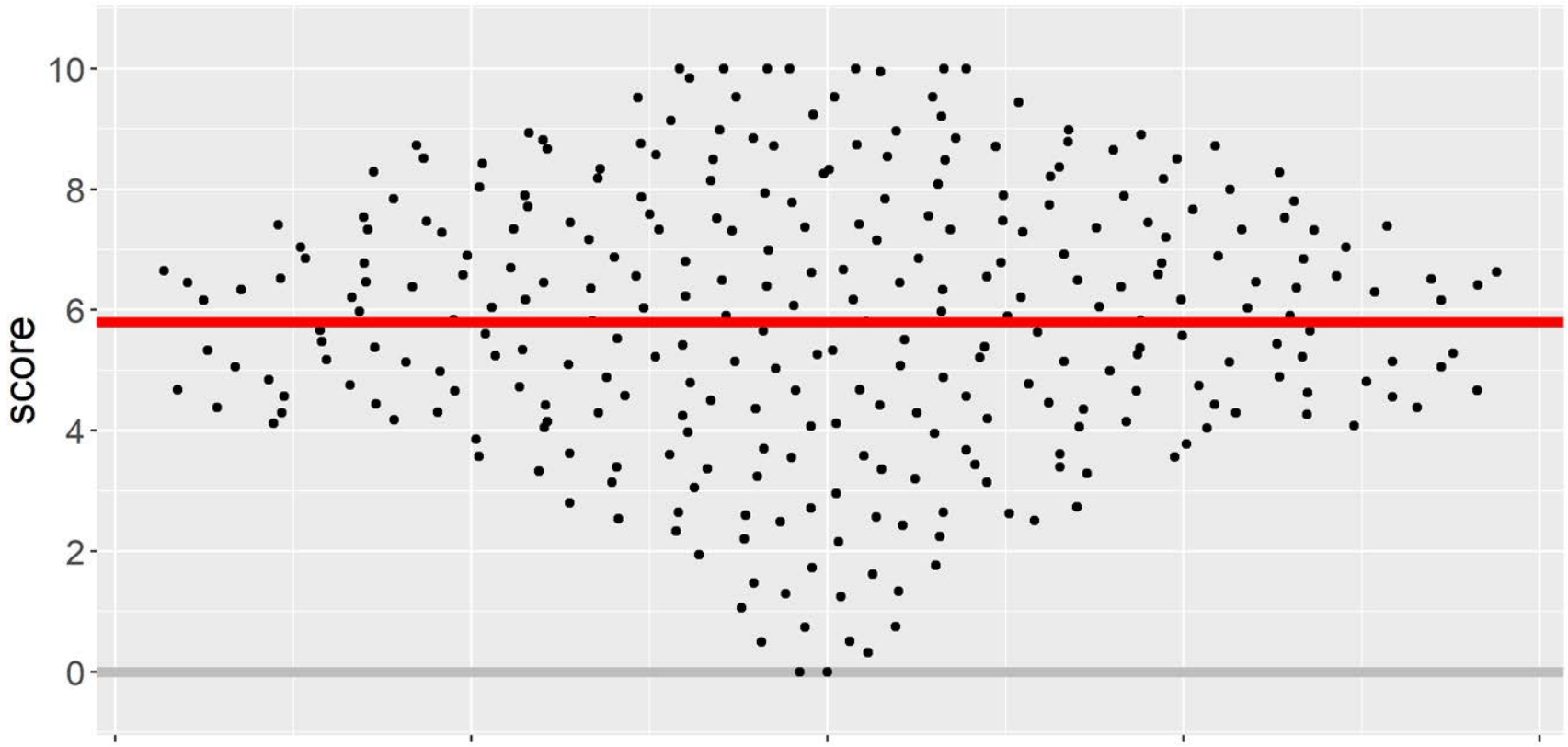
```
score ~ 1
```

	Variable	Est.	95% CI
β_0	(Intercept)	5.80	$[5.54, 6.05]$

The intercept (β_0) is the value of y when all $x = 0$

"The average review score in the sample was 5.80."

Intercept-only



Single continuous predictor

$$y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i$$

```
score ~ 1 + smile
```

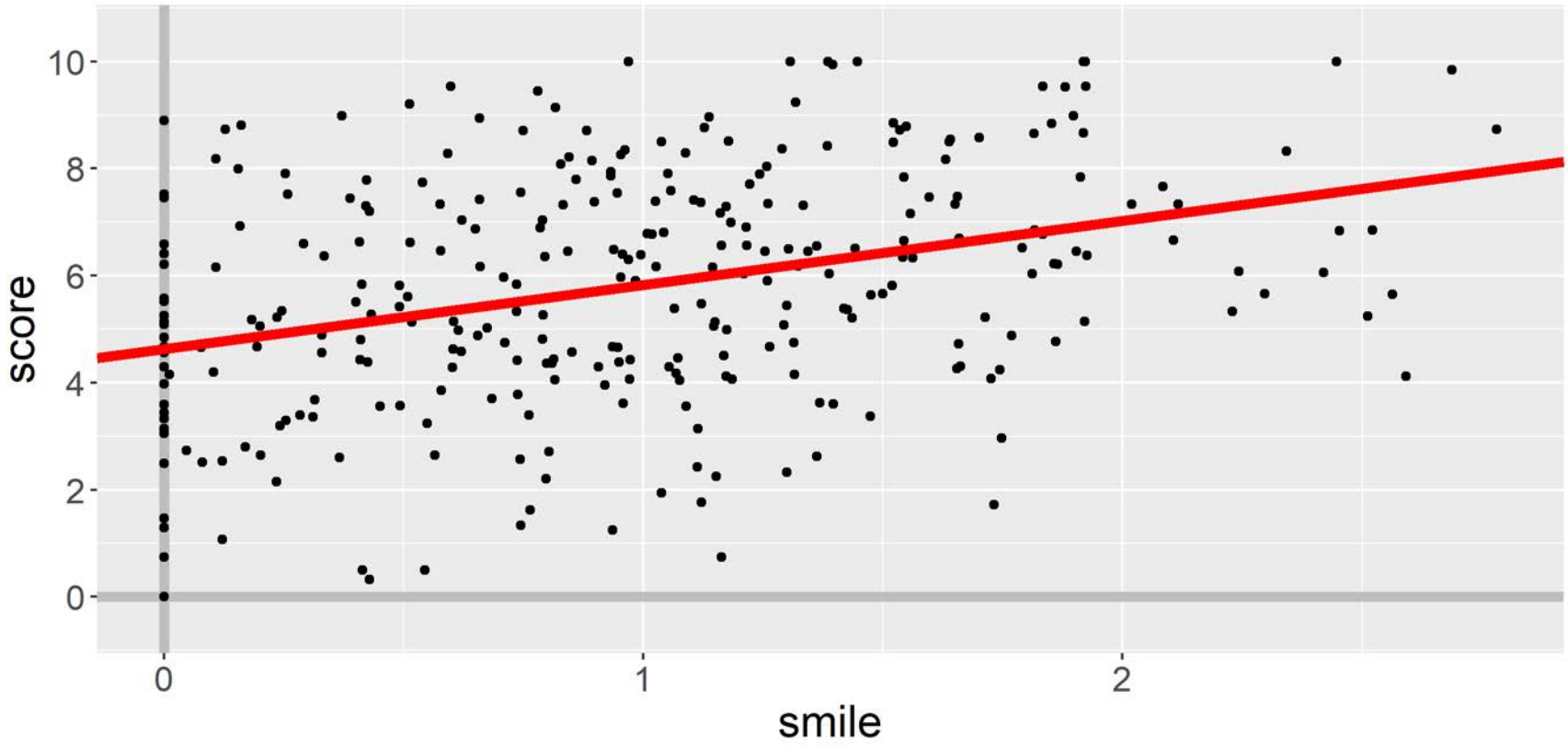
	Variable	Est.	95% CI
β_0	(Intercept)	4.63	$[4.20, 5.06]$
β_1	Smile	1.20	$[0.83, 1.57]$
R^2	Var. Explained	0.12	$[0.05, 0.19]$

The slope (β_j) is the change in y for each one-unit increase in x_j

"For each one-unit increase in smiling rate, the predicted review score will be 1.20 points higher."

"The predicted review score for a video with a smiling rate of 0 is 4.63."

Single continuous predictor



Single continuous predictor

$$y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i$$

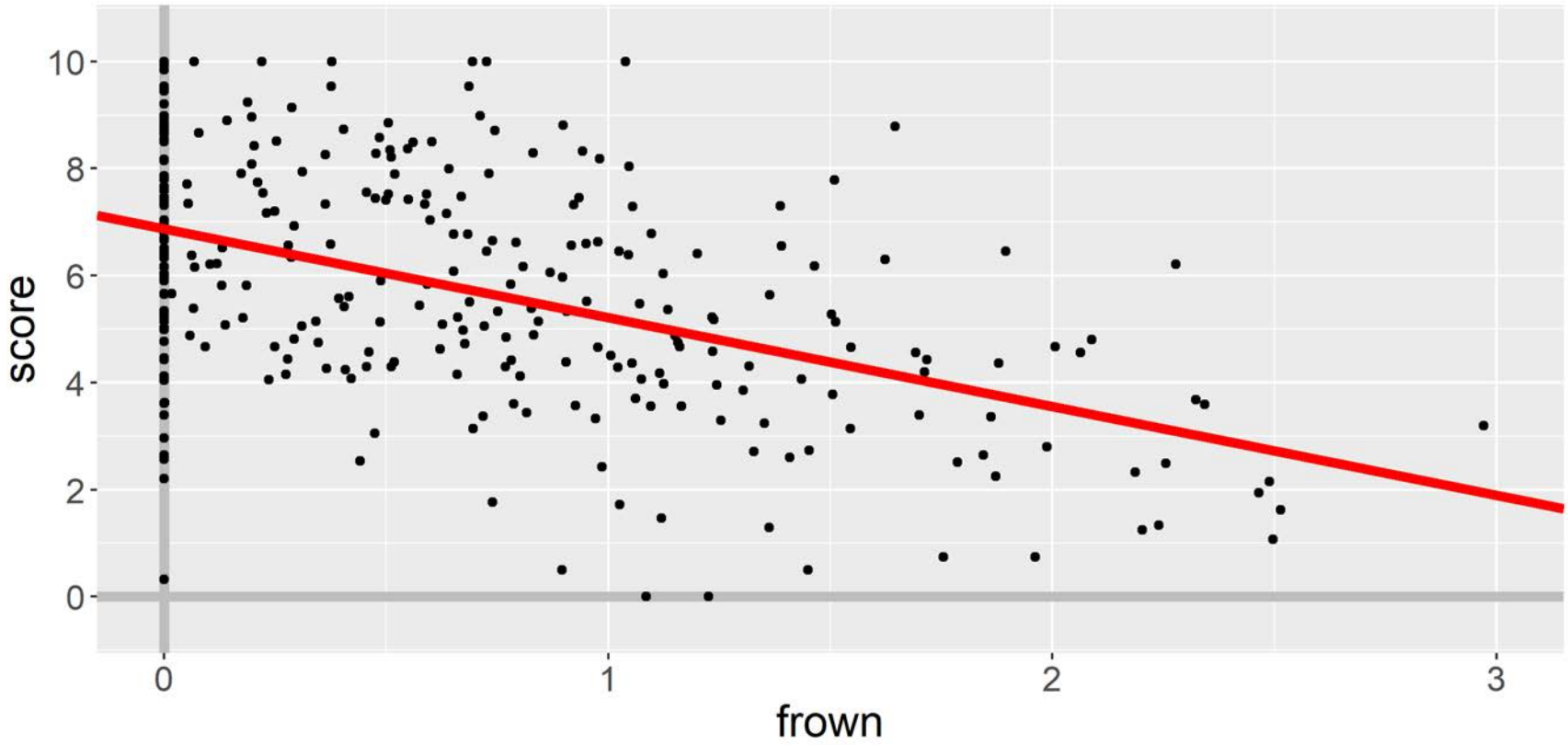
```
score ~ 1 + frown
```

	Variable	Est.	95% CI
β_0	(Intercept)	6.88	$[6.57, 7.19]$
β_1	Frown	-1.66	$[-2.00, -1.32]$
R^2	Var. Explained	0.24	$[0.16, 0.32]$

"For each one-unit increase in frowning rate, the predicted review score will be 1.66 points lower."

"The predicted review score for a video with a frowning rate of 0 is 6.88."

Single continuous predictor



Two continuous predictors

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$

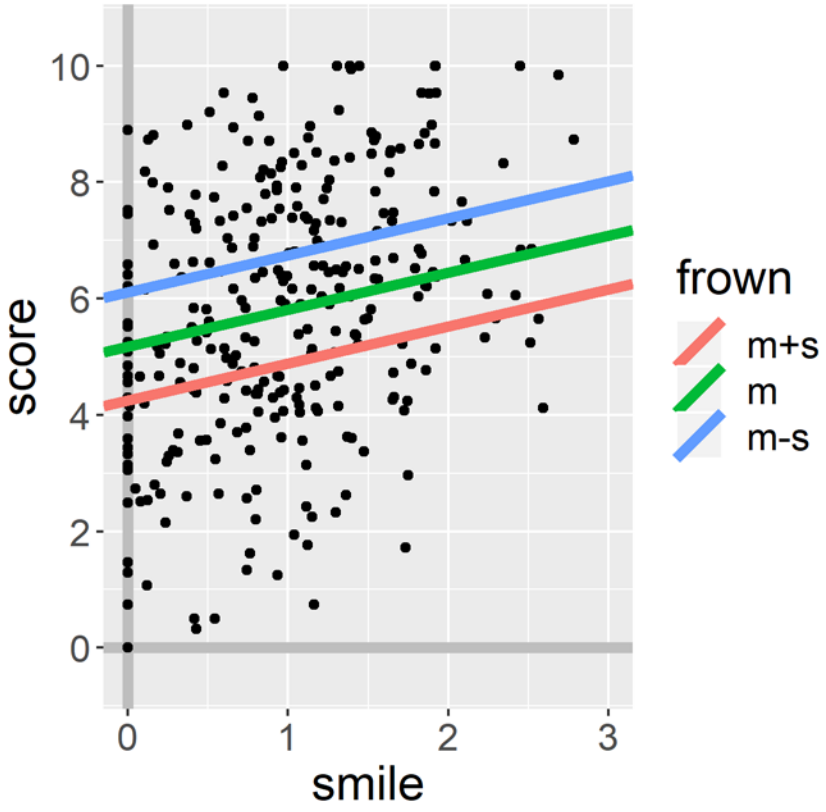
```
score ~ 1 + smile + frown
```

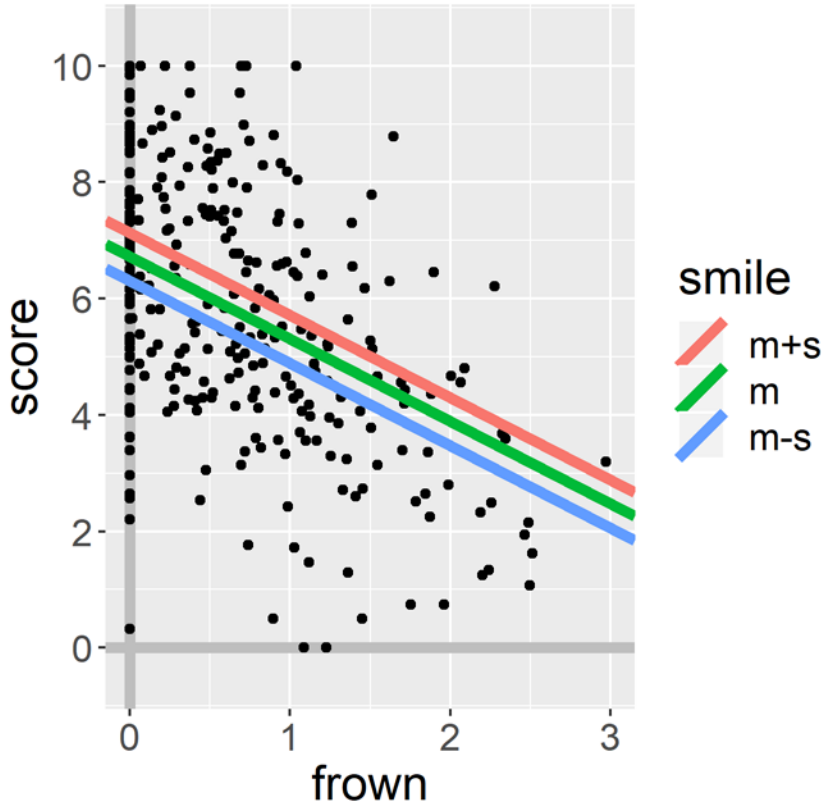
	Variable	Est.	95% CI
β_0	(Intercept)	6.10	$[5.56, 6.64]$
β_1	Smile	0.64	$[0.27, 1.00]$
β_2	Frown	-1.42	$[-1.78, -1.06]$
R^2	Var. Explained	0.27	$[0.18, 0.35]$

With multiple x variables, β_j becomes the *unique* contribution of x_j

"When controlling for frowning rate, for each one-unit increase in smiling rate, the predicted review score will be 0.64 points higher."

Two continuous predictors





Comparing models and model parameters

Model	(Intercept)	Smile	Frown	R^2
I	5.80			
I + S	4.63	1.20		0.12
I + F	6.88		−1.66	0.24
I + S + F	6.10	0.64	−1.42	0.27

- Why did the intercept change so much in value between models?
- Why did the smile and frown coefficients change in the last model?
- Why isn't the final model's R^2 the sum of the other two?
- Would you rather know the smiling rate or the frowning rate?

Single binary predictor (dummy code)

$$y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i$$

```
score ~ 1 + is_critic
```

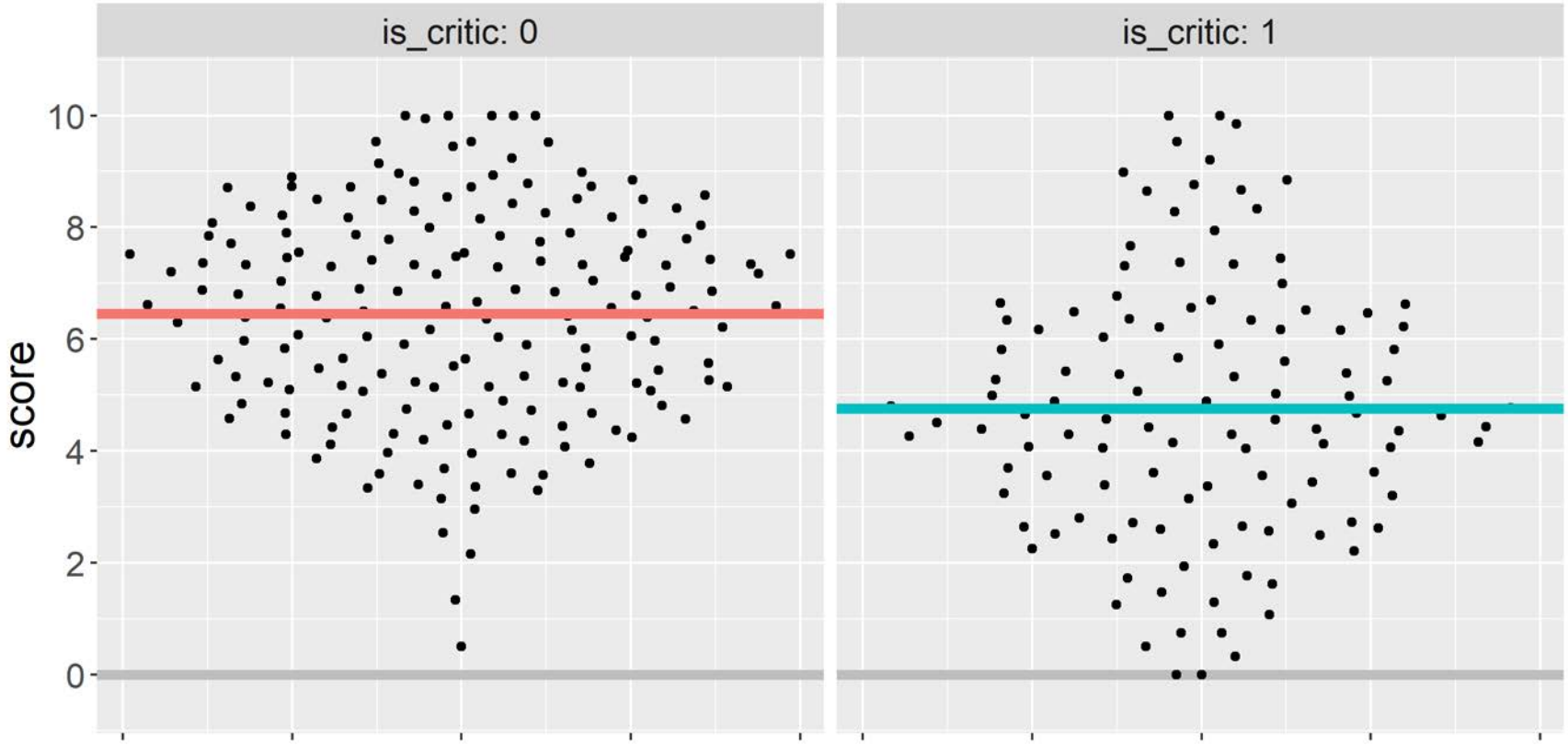
	Variable	Est.	95% CI
β_0	(Intercept)	6.46	$[6.16, 6.76]$
β_1	Critic	-1.70	$[-2.18, -1.22]$
R^2	Var. Explained	0.14	$[0.07, 0.21]$

With binary dummy codes, the intercept is the value of y in the "reference" group.

"The average review score for non-critics was 6.46."

"The average review score for critics was 1.70 points lower than that for non-critics."

Single binary predictor (dummy code)



Binary and continuous predictors (main effects)

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$

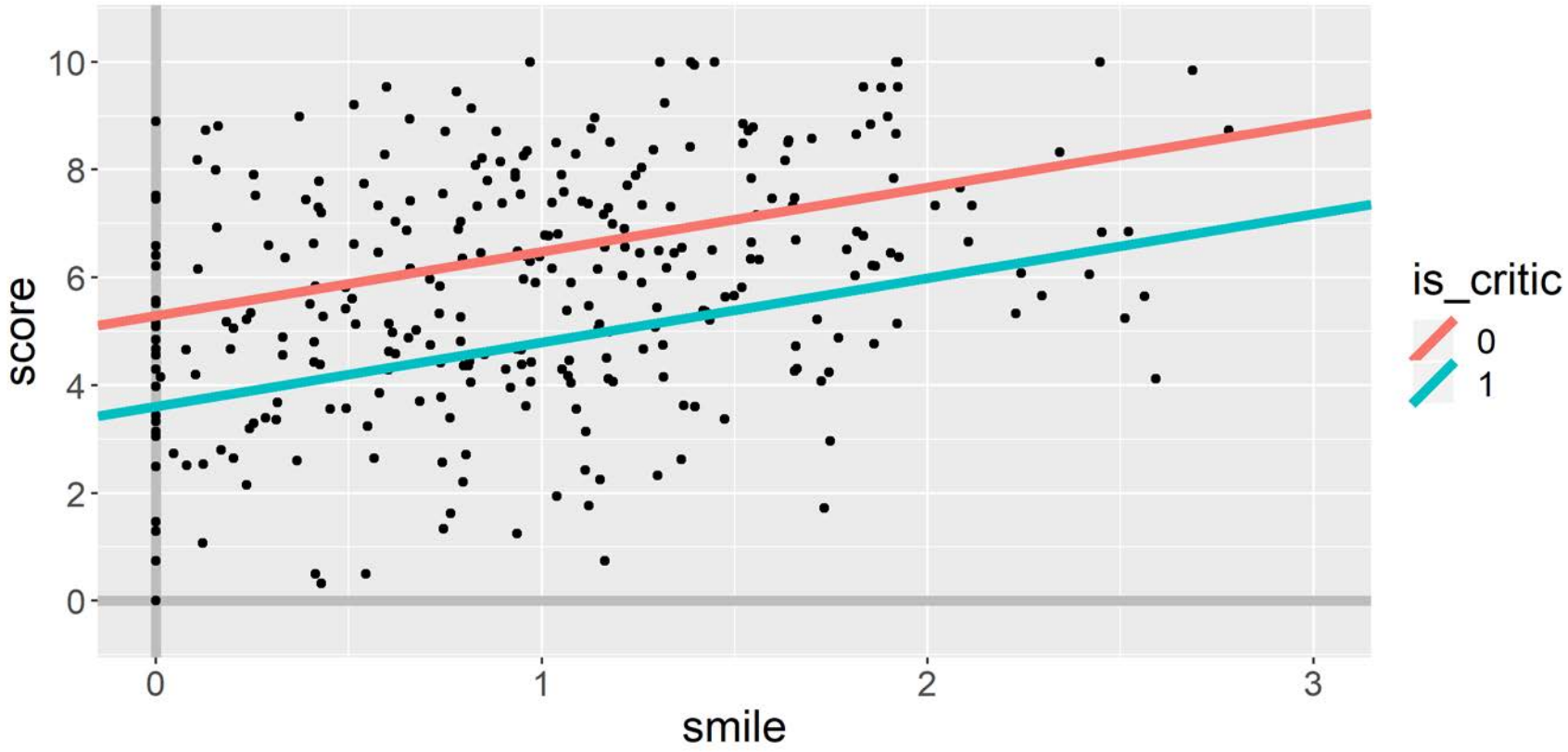
```
score ~ 1 + smile + is_critic
```

	Variable	Est.	95% CI
β_0	(Intercept)	5.29	$[4.86, 5.73]$
β_1	Smile	1.19	$[0.85, 1.53]$
β_2	Critic	-1.69	$[-2.14, -1.24]$
R^2	Var. Explained	0.26	[0.17,0.34]

"For a non-critic with a smiling rate of zero, the average review score was 5.29."

"Controlling for professional status, for each one-unit increase in smiling rate, the predicted review score was 1.19 points higher."

Binary and continuous predictors (main effects)



Binary and continuous predictors (interaction)

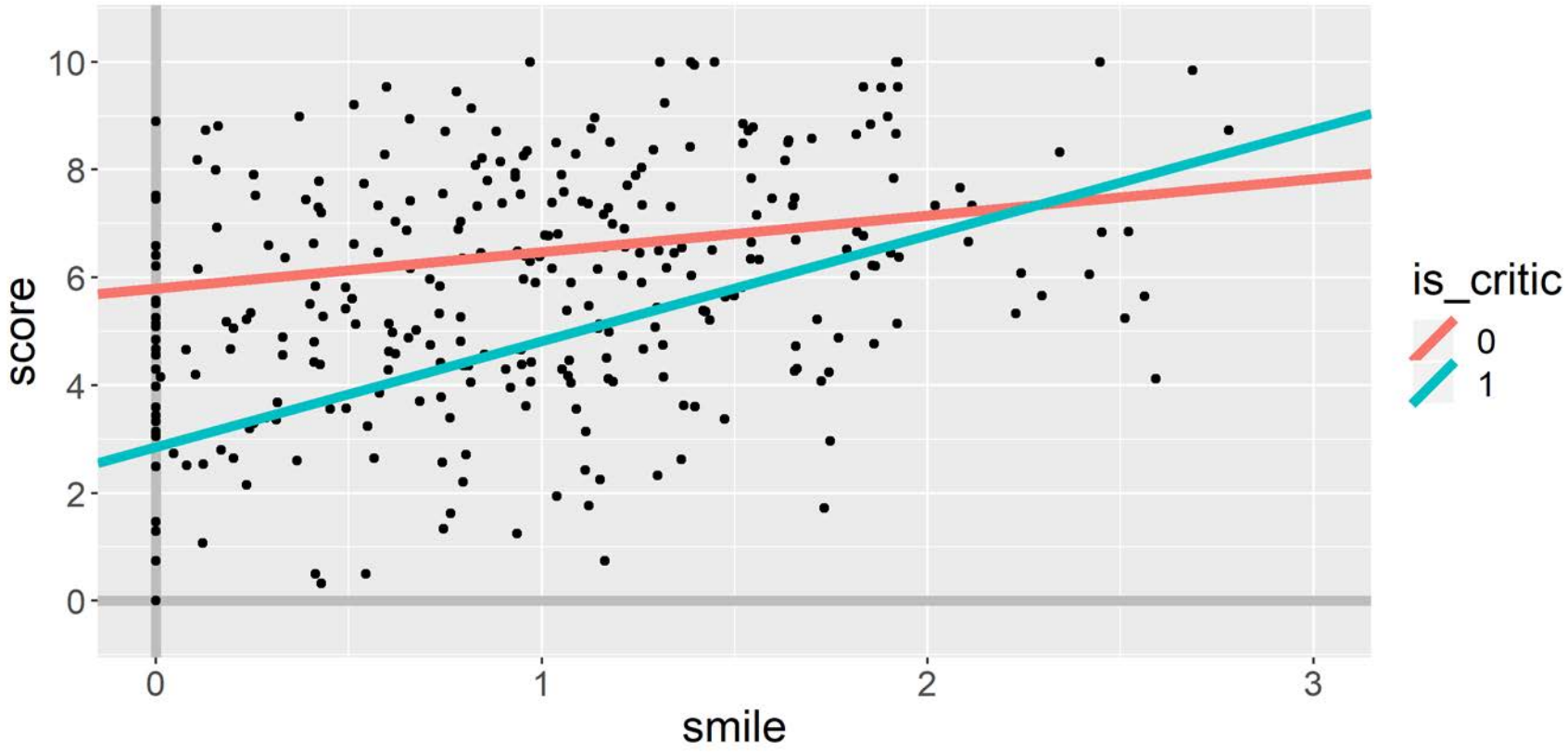
- With many predictors, the slope is the *unique* contribution of x

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i$$

```
score ~ 1 + smile + is_critic + smile*is_critic
```

	Variable	Est.	95% CI
β_0	(Intercept)	5.79	$[5.29, 6.29]$
β_1	Smile	0.68	$[0.25, 1.11]$
β_2	Critic	-2.94	$[-3.73, -2.15]$
β_3	Interaction (SxC)	1.29	$[0.61, 1.97]$
R^2	Var. Explained	0.29	$[0.21, 0.38]$

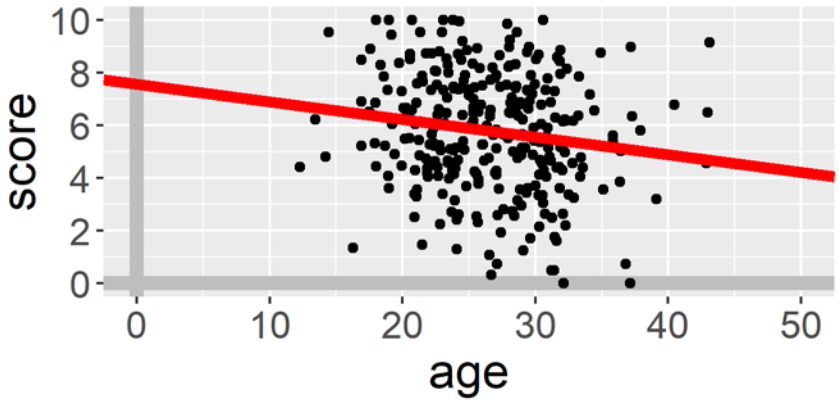
Binary and continuous predictors (interaction)

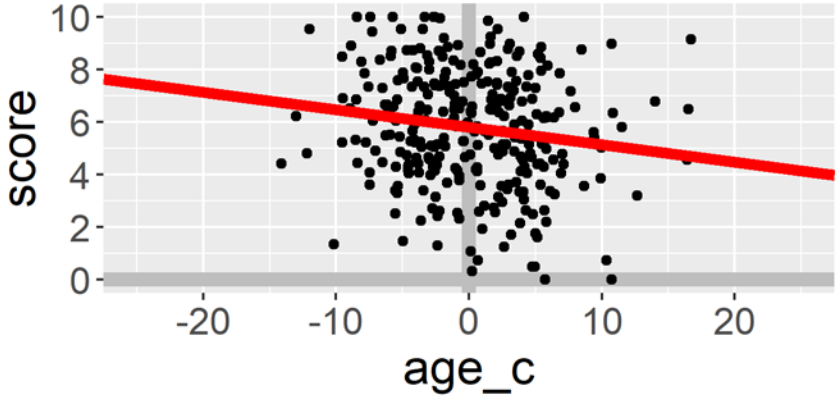


Centering and standardization

- The intercept (β_0) is the value of y when $x = 0$
- This is not very informative when 0 is not in the sample
- It is helpful to "center" the predictor by subtracting its mean ($x_i^c = x_i - \bar{x}$)
- Now the intercept is the value of y when $x = \bar{x}$

$\beta_0 = 7.55$ $\beta_1 = -0.07$





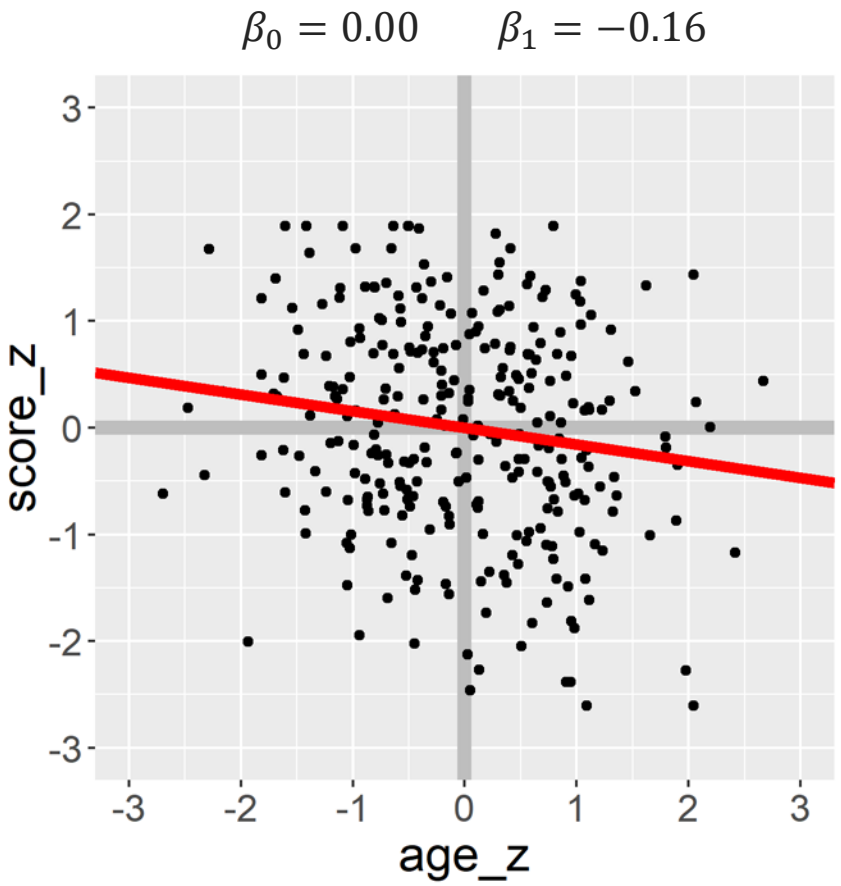
$\beta_0 = 5.80$ $\beta_1 = -0.07$

Centering and standardization

- It can help to standardize each variable by centering it and then dividing by its SD

$$\left(x_i^z = \frac{x_i - \bar{x}}{s_x}\right) \qquad \left(y_i^z = \frac{y_i - \bar{y}}{s_y}\right)$$

- This puts everything on the same scale, which eases comparison/interpretation
- The β are now standardized coefficients and in SD units



Assumptions of the linear model

Assumptions of LM using the OLS approach

1. Correct specification of the form of the relationships (i.e., linear)
2. All important predictors are included and perfectly measured
3. The residuals have constant variance around the regression line
4. The residuals are normally distributed around the regression line
5. The residuals of the observations are independent of one another

Consequences of violating these assumptions

- Estimates of the regression coefficients may be biased
- Standard errors (and thus hypothesis testing) may be biased

Assumptions of the linear model

Assumption	Diagnosis	Remedies
Correct specification	Scatterplots, F-test	Transformations, Power polynomial terms
No omitted predictors	Theory/literature review, Added variable plots	Adding predictors, Regularization
Perfect measurement	Reliability analysis	Corrections, SEM
Constant variance	Residual plots, Levene test, Breusch-Pagan test	Transformations, Weighted least squares
Normality of residuals	Plot residual distribution, q-q plots, Shapiro-Wilk test	Transformations, GLM
Independence	Index plots, ACF plots, ICC, Durbin-Watson test	Transformations, Dummy variables, MLM

Practical issues with the linear model

Other Issues

- Outliers can influence your results (run with and without outliers)
- Missing data can bias your results (use FIML or MI procedures)
- Correlated predictors can cause problems (use regularization)
- High-order interaction terms are hard to interpret (use carefully)
- LM models can overfit the sample (use cross-validation)

Resources for LM

- Cohen, Cohen, West, & Aiken (2002) Applied multiple regression…
- Gelman & Hill (2007) Data analysis using regression and multilevel…
- McElreath (2015) Statistical rethinking: A Bayesian course with examples…

The generalized linear model (GLM)

The generalized linear model (GLM)

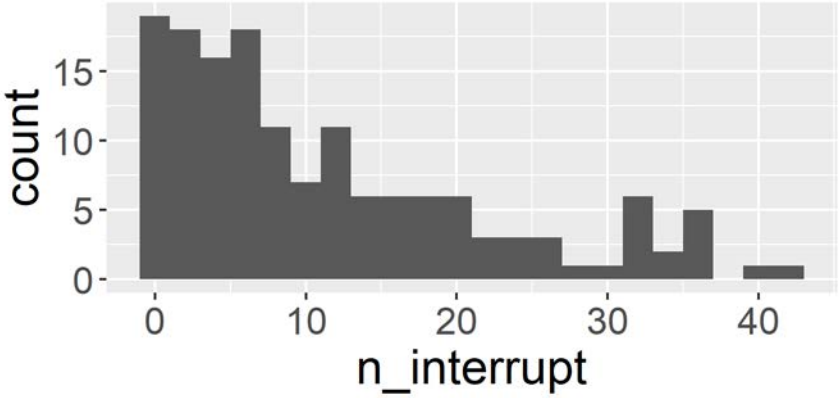
- LM assumes that y variables are normally distributed
- GLM handles y variables with specified distributions
- GLM is also implemented in **statsmodels** and **stats**
  - You need to specify a "family" describing the y variable's distribution
  - This will transform y using a "link function" appropriate to that family
- ```
  sm.glm(formula='y~1+x1', data=df,
    family=sm.families.Poisson)
  ```
- ```
  stats::glm(formula='y~1+x1', data=df,
    family=stats::poisson)
  ```

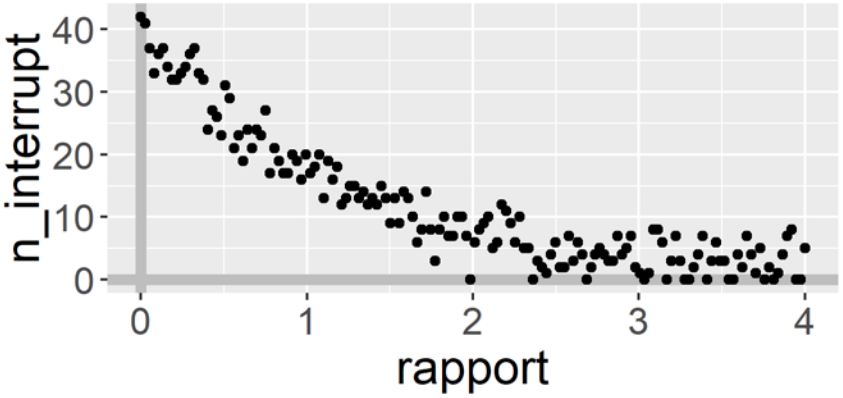
Common GLM families and link functions

Family	Uses		Link Function	
Gaussian	Linear data	real: $(-\infty, +\infty)$	Identity	μ
Gamma	Exponential data	real: $(0, +\infty)$	Inverse or Power	μ^{-1}
Poisson	Count data	integer: 0,1,2, …	Log	$\log(\mu)$
Binomial	Binary data Categorical data	integer: $\{0,1\}$ integer: $[0, K)$	Logit	$\log\left(\frac{\mu}{1-\mu}\right)$

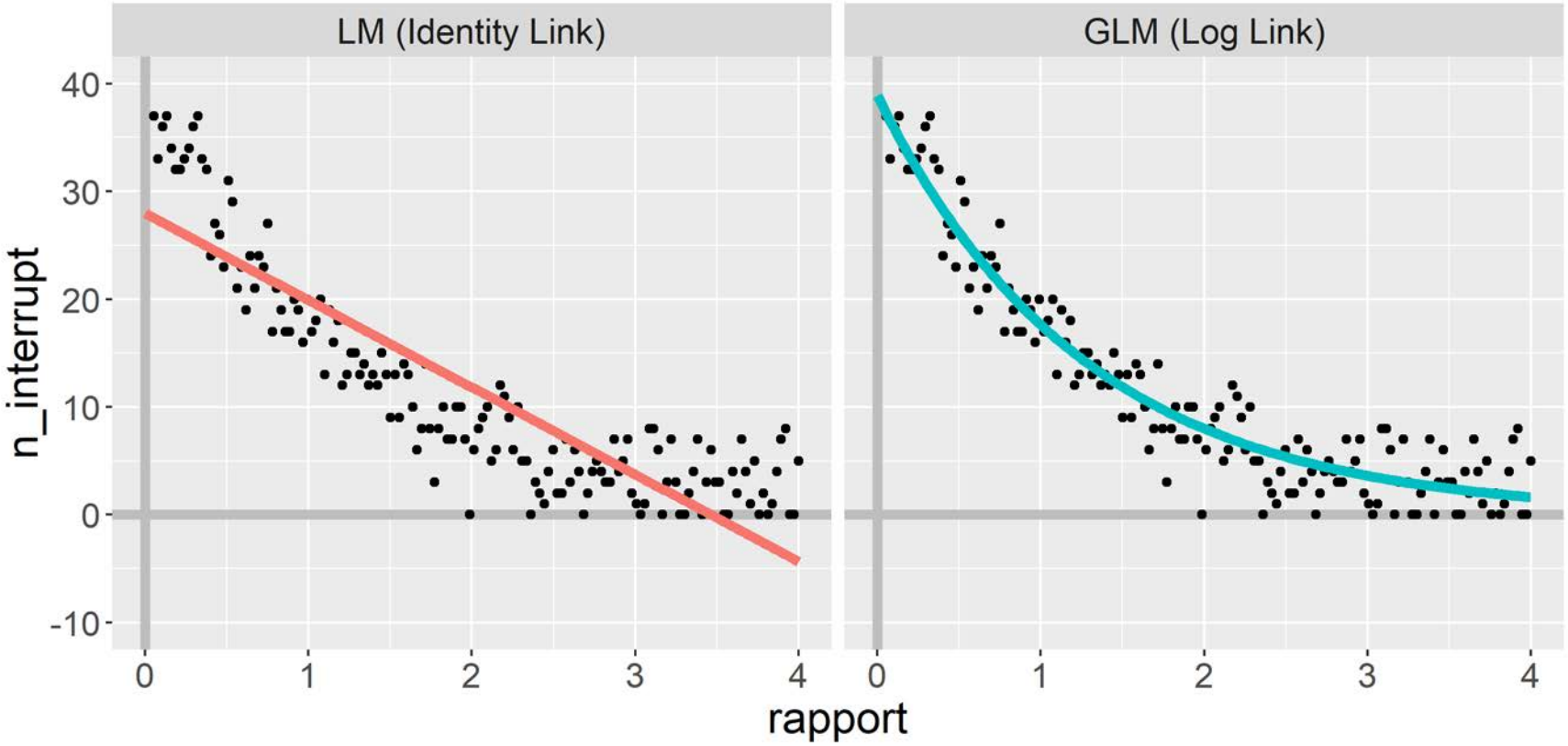
Applied GLM example

- Let's say we want to predict the count of interruptions during a 5 minute social interaction using ratings of rapport
- LM assumes y is normally distributed, but counts are not
- A straight line will do a poor job modeling this relationship

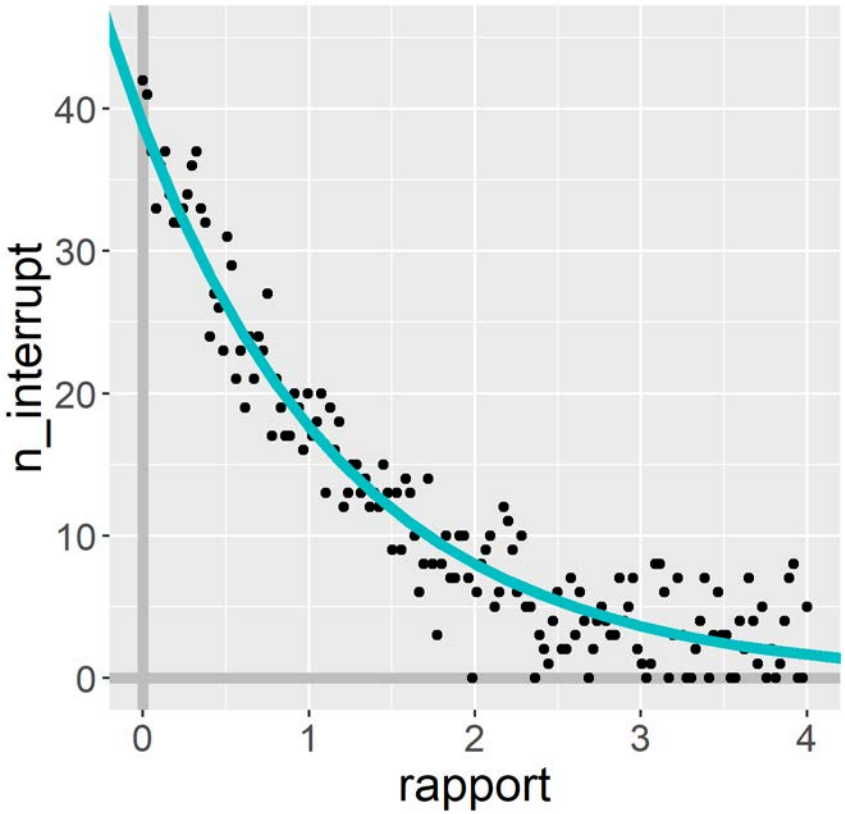


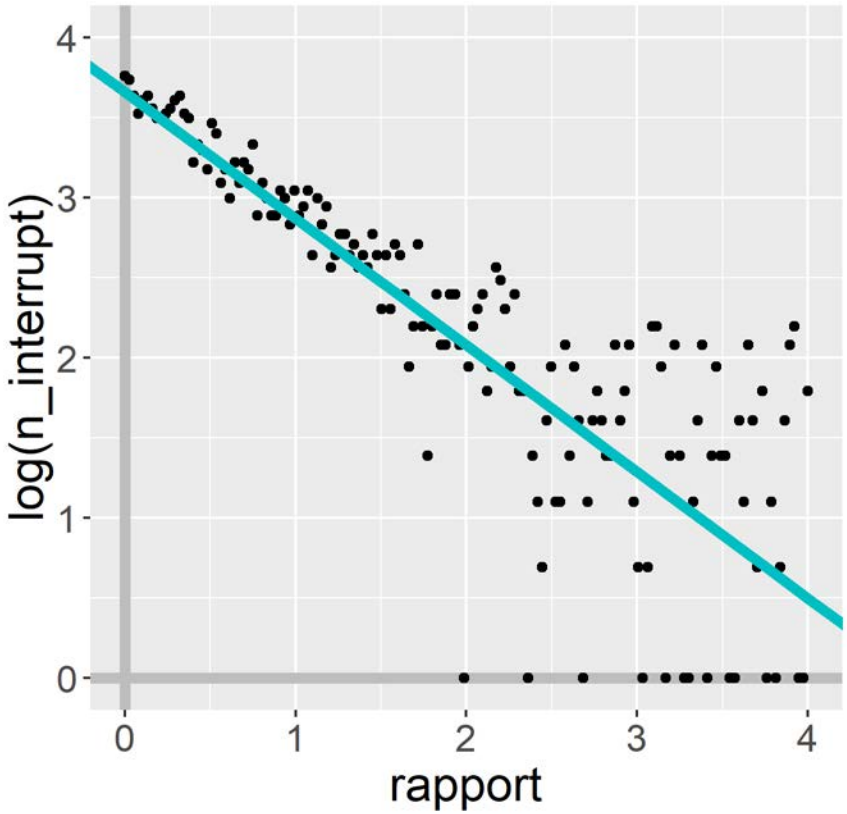


Applied GLM example



Applied GLM example





Applied GLM example

- GLM estimates will, by default, be given in transformed units

$$\log(y_i) = \beta_0 + \beta_1 x_{1i} + \epsilon_i$$

```
n_interrupts ~ 1 + rapport, family = poisson
```

	Variable	Est. (in log units)	95% CI (in log units)
β_0	(Intercept)	3.66	$[3.59, 3.73]$
β_1	Rapport	-0.79	$[-0.84, -0.74]$

GLM coefficients can often be re-transformed to enhance their interpretability. We can transform β_1 into an incidence rate ratio ($IRR = 0.45$), which means that a one-unit increase in rapport cuts the expected number of interruptions roughly in half.

Preview of advanced frameworks

Multilevel modeling (MLM)

- LM and GLM assume that the observations are independent
- In practice, observations are frequently "nested" or "clustered"
  - Multiple observations drawn from each participant, object, task, etc.
  - Multiple participants drawn from each group, location, population, etc.
- To accommodate this, we can model each "level" separately
  - A 2-level model: multiple tests (L1) within multiple students (L2)
  - A 2-level model: multiple students (L1) within multiple schools (L2)
  - A 3-level model: tests (L1) within students (L2) within schools (L3)

Multilevel modeling (MLM)

- MLM gives more accurate representations of the higher levels
- We can capture each cluster's central tendency and variability
  - e.g., we know each student's average test score and how variable
- MLM enables us to answer questions across levels
  - e.g., does a school's location predict its students' average test scores?
- MLM enables model parameters to vary by cluster
  - e.g., is studying more beneficial for some students than for others?
- Most implementations of MLM incorporate GLM's features

Multilevel modeling (MLM)

- MLM has many different instantiations and names
  - Multilevel modeling (MLM)
  - Linear mixed effects modeling (LME)
  - Hierarchical liner modeling (HLM)
  - Random effects/random coefficients modeling
- Resources for MLM
  - Gelman & Hill (2007) Data analysis using regression and multilevel…
  - Snijders & Bosker (2011) Multilevel analysis: An introduction…
  - McElreath (2015) Statistical rethinking: A Bayesian course with…

Structural equation modeling (SEM)

- LM, GLM, and MLM assume that all variables are observed
- In practice, we are often interested in latent variables
  - We often can't measure y directly and instead measure its indicators
  - We often want to measure multi-faceted and hierarchical constructs
- LM, GLM, and MLM generally assume simple relationships
- In practice, we are often interested in complex relationships
  - We often have several sets of distinct x and y variables
  - We often want variables to play the role of both x and y variables
  - We often want to understand systems or networks of relationships

Structural equation modeling (SEM)

- SEM incorporates many related techniques
  - Path analysis, factor analysis, latent growth modeling, etc.
  - Most SEM implementations incorporate LM and GLM features
  - Advanced implementations even incorporate MLM features (MSEM)
- SEM confers many benefits over LM, GLM, and MLM
  - We can account for measurement error in our latent variables
  - We can model complex relationships between many different variables
  - We can generate diagrams to represent our models and results
- Resources for SEM
  - Kline (2010) Principles and practice of structural equation modeling

Regularization and prediction (GLMNET)

- LM and GLM are susceptible to overfitting and multicollinearity
  - Multicollinearity is when two or more x variables are highly related
  - In this case, small changes in the data can dramatically change β
  - This also makes it difficult to include large numbers of predictors
- Regularization addresses these issues by adding information
  - A Bayesian interpretation is that we are introducing informative priors
- There are several common regularization approaches
  - The ridge penalty shrinks the β of the predictors toward each other
  - The lasso tends to pick one of the predictors and discard the others
  - The elastic-net penalty combines/bridges the ridge penalty and lasso

Regularization and prediction (GLMNET)

- GLMNET is an approach for prediction using regularized GLM
  - It uses the elastic-net penalty and is extremely fast to estimate/train
  - It can handle many more predictors than non-regularized GLM
  - Because it is based on GLM, it is still extremely interpretable (e.g., β)
- GLMNET is like a "missing link" between statistics and ML
  - I often estimate GLMNET models as linear baselines for ML models
  - You can use the exact same cross-validation scheme for all models
- Resources for GLMNET
  - https://glmnet-python.readthedocs.io/en/latest/glmnet_vignette.html
  - https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html