Finding ‘good enough’: Statistical model building for real-world data

April 4, 2024

Jessica R. Shaw


All models are wrong
but some are useful
George E.P. Box

Linear models allow us to make valid inference about the relationship between two variables, after adjustment for confounding factors

For example, if we wish to understand the relationship between **Age** and **Height** during childhood, we might consider the linear model:

Height ~ Age + Sex

This model will estimate of the mean change in Height per one-year increase in Age, after adjustment for Sex.

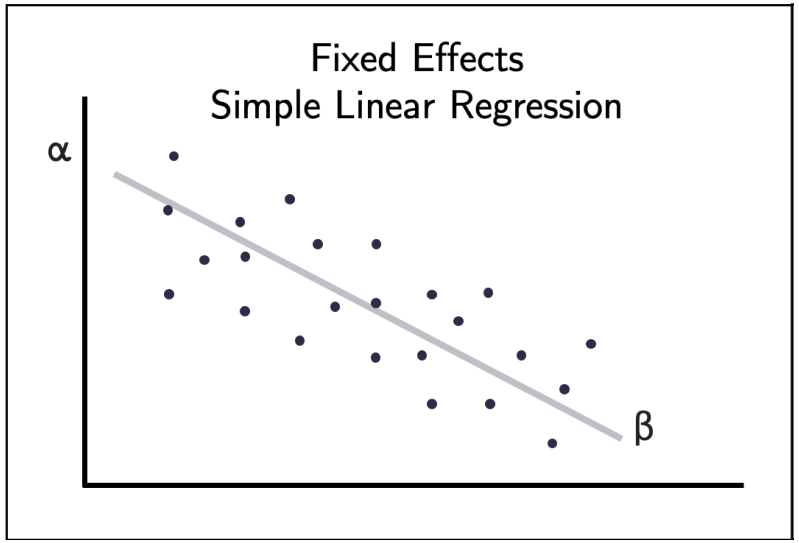
With a **linear mixed model**, we can also adjust for correlated observations

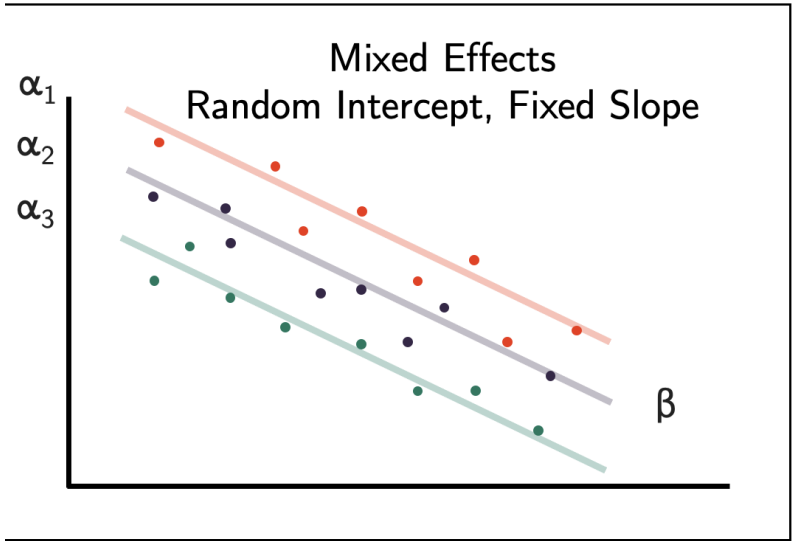
For example, say we want to understand the relationship between **Age** and **Height** during childhood, and we now have *two measurements* from each participant.

We might consider the linear mixed model:

$$\mathbf{Height_i \sim Age_i + Sex + (1|Participant\ ID)},$$

- i indexes each measurement of Height and Age, with possible values i=1 or i=2.
- (1|Participant ID) is a random intercept that accounts for intra-individual correlation of measurements.





For linear mixed models to enable robust inference, **model assumptions** must be met

- Linear mixed model assumptions:
  - Independent variables are linearly related to the dependent variable.
  - Model residuals have approximately constant variance (‘homoscedasticity’)
  - Model residuals are mutually independent.
  - Model residuals are approximately normally distributed.

The **Akaike Information Criterion (AIC)** is a statistic used to compare goodness of fit of candidate models

- AIC can be used to compare the goodness of fit of non-nested models (ie, models without overlapping independent variables).
- For example, using AIC we can compare:

Model A: $Y \sim Age + Sex$

vs.

Model B: $Y \sim BMI$

When using **AIC** for model selection, remember "***lower is better***"

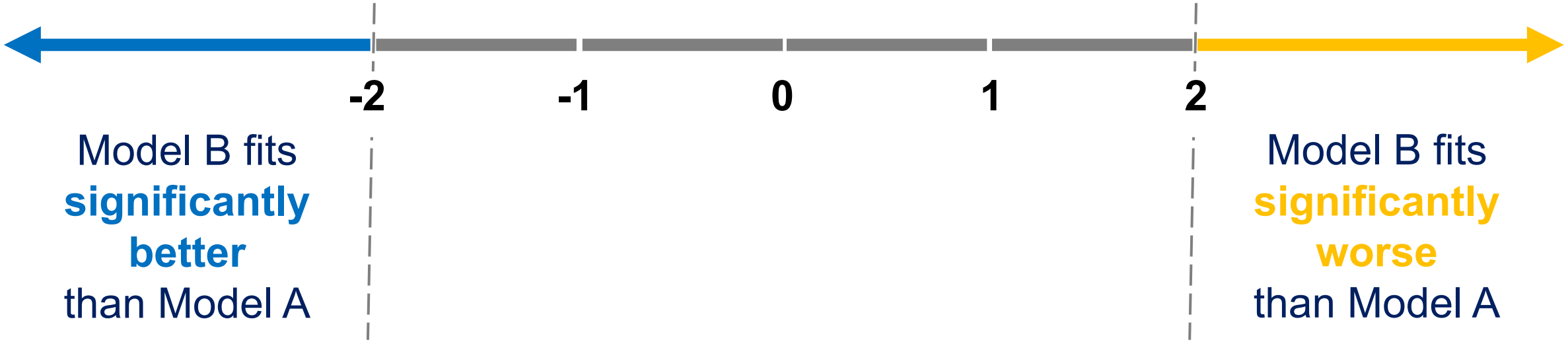
- A lower AIC value indicates 'better' model fit, eg:
  - If AIC(Model A) < AIC(Model B), then model A is a better fit for the data than Model B.
  - If AIC(Model B) < AIC(Model A), then model B is a better fit for the data than Model A.

We calculate **delta AIC (Δ_{AIC})** to quantify the relative goodness of fit of two candidate models

Calculation of $\Delta_{\mathbf{AIC}}$

$$\Delta_{\mathbf{AIC}} = \mathbf{AIC(Model\ B)} - \mathbf{AIC(Model\ A)}$$

Qualitative interpretation of $\Delta_{\mathbf{AIC}}$



Multicollinearity occurs when the independent variables in a regression model are correlated with one another

- Linear regression models assume that independent variables are mutually independent predictors of the dependent variable.
- When independent variables are correlated with one another, this assumption is violated.
- When multicollinearity is present, model-produced estimates may be misleading, and the validity of inference can be compromised.

The **Variance Inflation Factor (VIF)** is a measure of multicollinearity in a regression model

- Can be calculated for any model with two or more fixed-effect predictor variables.
- One VIF value is calculated for each fixed-effect predictor in the model.
- Variables correlated with one another will yield higher VIF values.

Qualitative interpretation of VIF

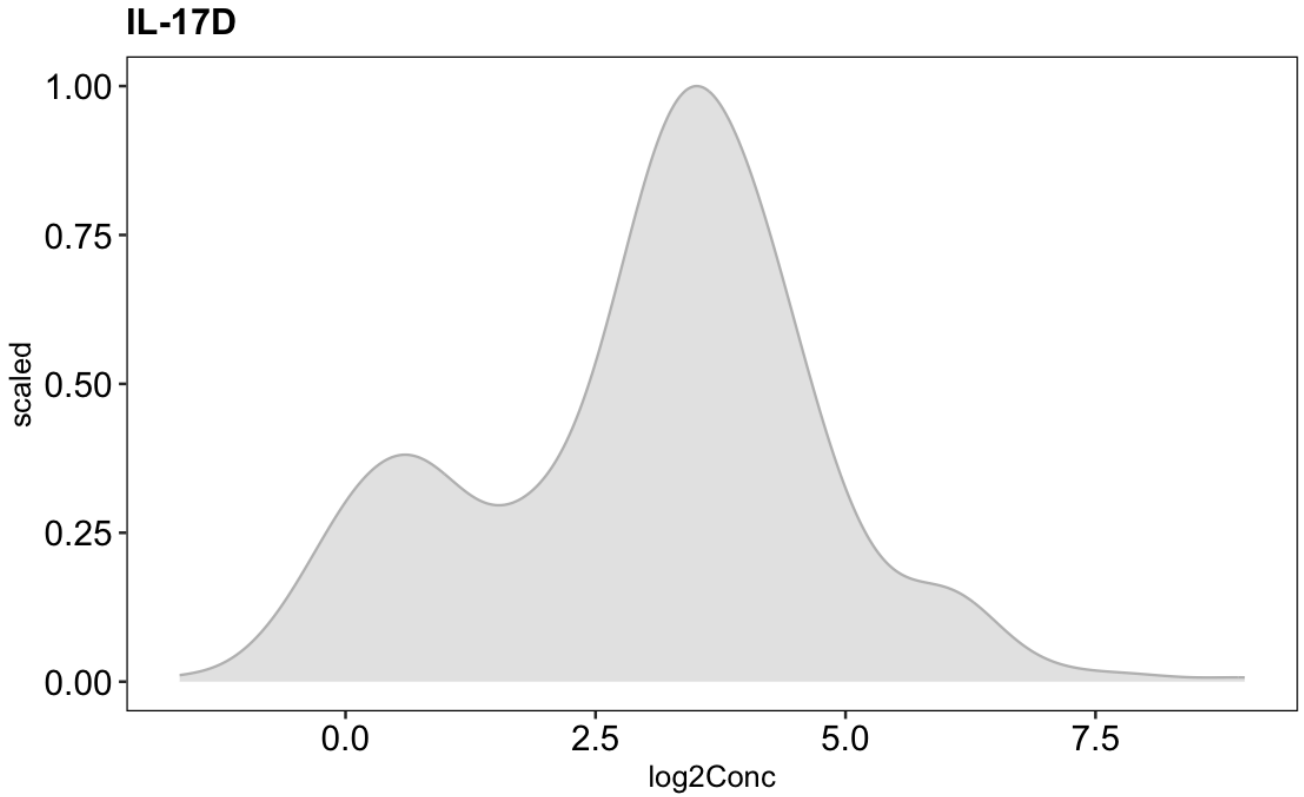
- $VIF \approx 1.0$ Indicates minimal correlation of the predictor variables (ideal).
- $4.0 < VIF \leq 10.0$ Indicates some correlation of predictor variables – requires further investigation – consider model revision.
- $VIF > 10.0$ Indicates serious multicollinearity – requires model revision to correct.

Real-world data example: Circulating IL-17D in individuals with vs. without trisomy 21

The research question

- **Goal:** Understand the relationship between trisomy 21 and circulating levels of inflammatory cytokine IL-17D.
- We must find the best statistical model that will:
  - Optimize our power to detect a difference between groups, if a difference truly exists.
  - Minimize our risk of detecting a difference between groups when one does not truly exist.
  - Yield robust and unbiased estimates of the difference between individuals with and without trisomy 21.

Can we use a linear (mixed) model?

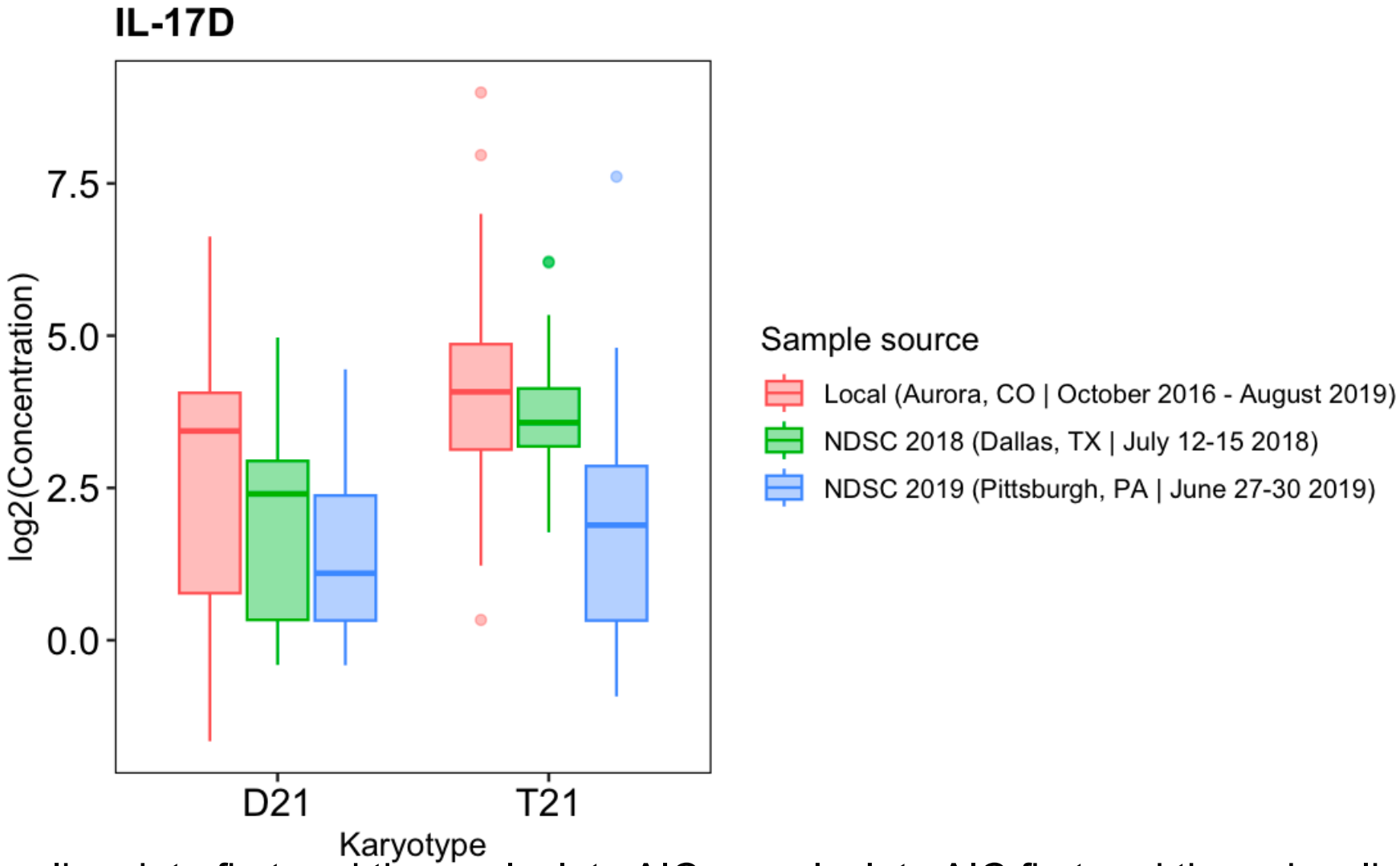


- We wish to understand the relationship between trisomy 21 and circulating levels of IL-17D, after adjustment for key confounders (age, sex).
- We plot the log2-transformed raw data to establish whether it is approximately normally distributed before adjustment for confounders.
- Because the data is *approximately* normally distributed after log2 transformation, we determine that a linear model or linear mixed model for log2(Concentration) is *likely* to work.

Are there any issues with collinearity if we include the variables T21, Age, and Sex in the same linear (mixed) model?

Analyte <chr>	T21 <dbl>	Age <dbl>	Female <dbl>
IL-17D	1.037588	1.028858	1.039235

Could any batch effects be present?



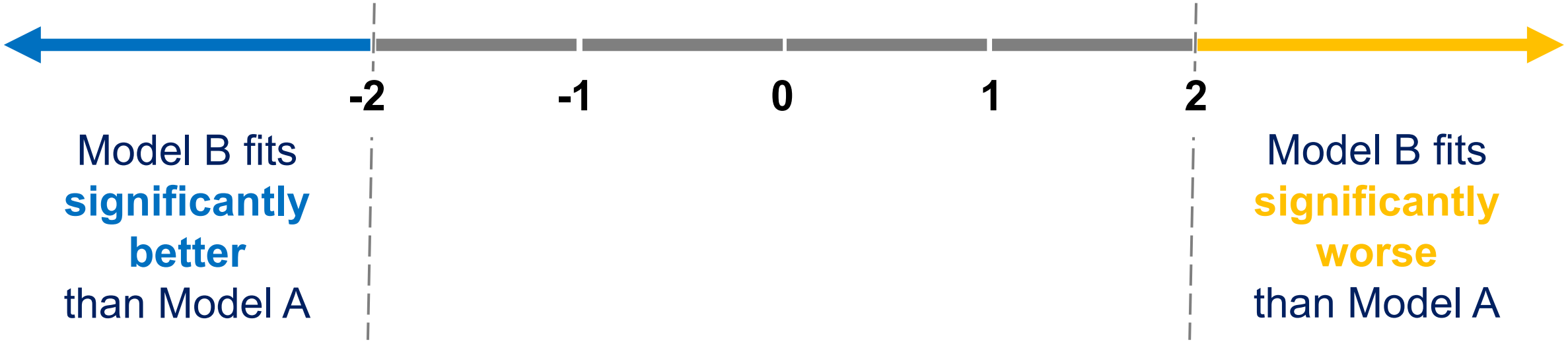
Note: OK to visualize data first and then calculate AIC, or calculate AIC first and then visualize the data, but I recommend ALWAYS doing both!

Does the model fit better with or without a random intercept for sample source?

Analyte <chr>	AIC(lm(log2Conc ~ T21 + Age + Female)) <dbl>	AIC <dbl>	model <chr>
IL-17D	1812.064	1669.58	log2Conc ~ T21 + Age + Female + (1|Source), REML = FALSE

YES IT DOES.

Qualitative interpretation of Δ_{AIC}



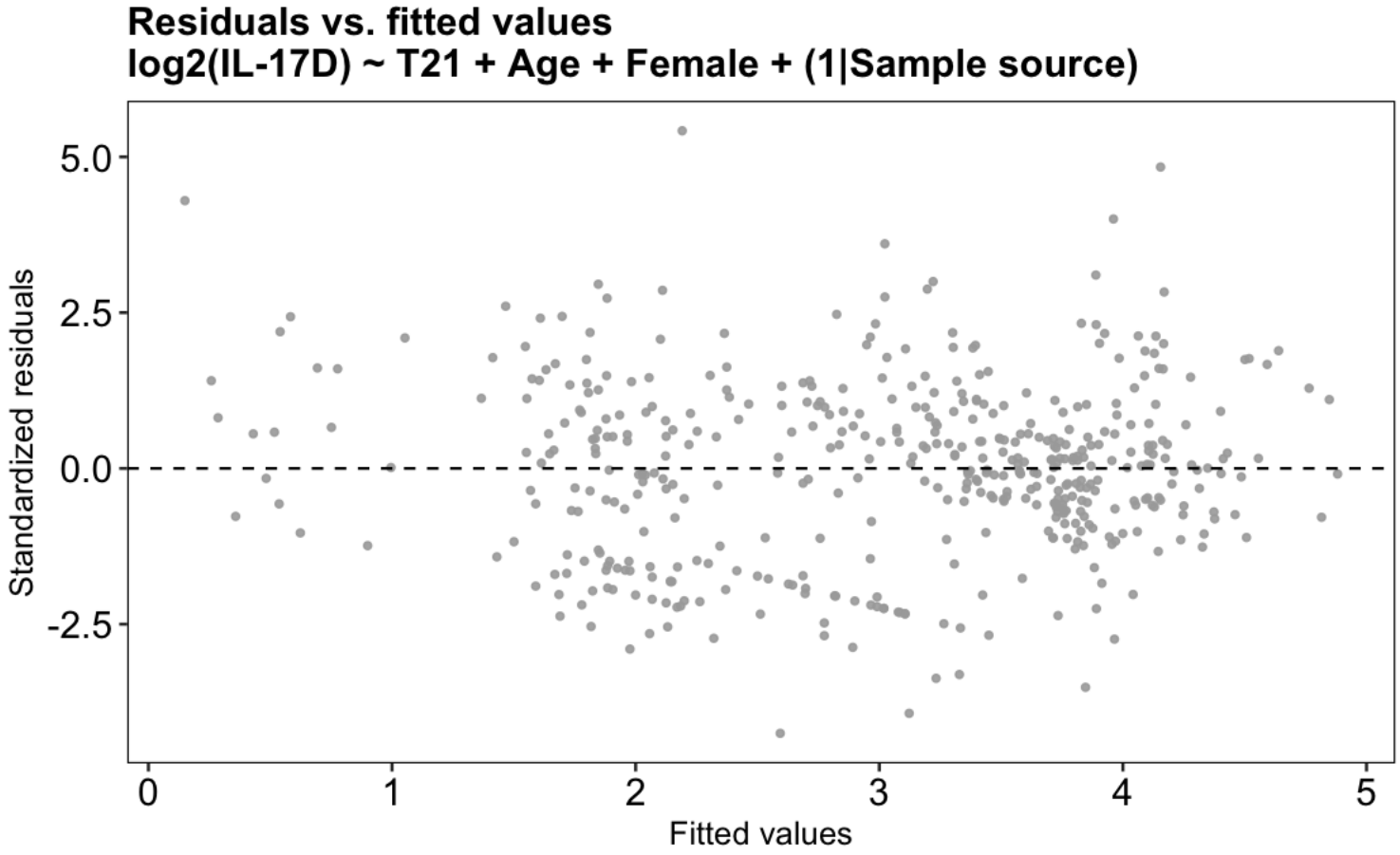
Can we verify the model assumptions?

Eg, Do we have homoscedasticity of residuals?
Yep!

Is it perfect?

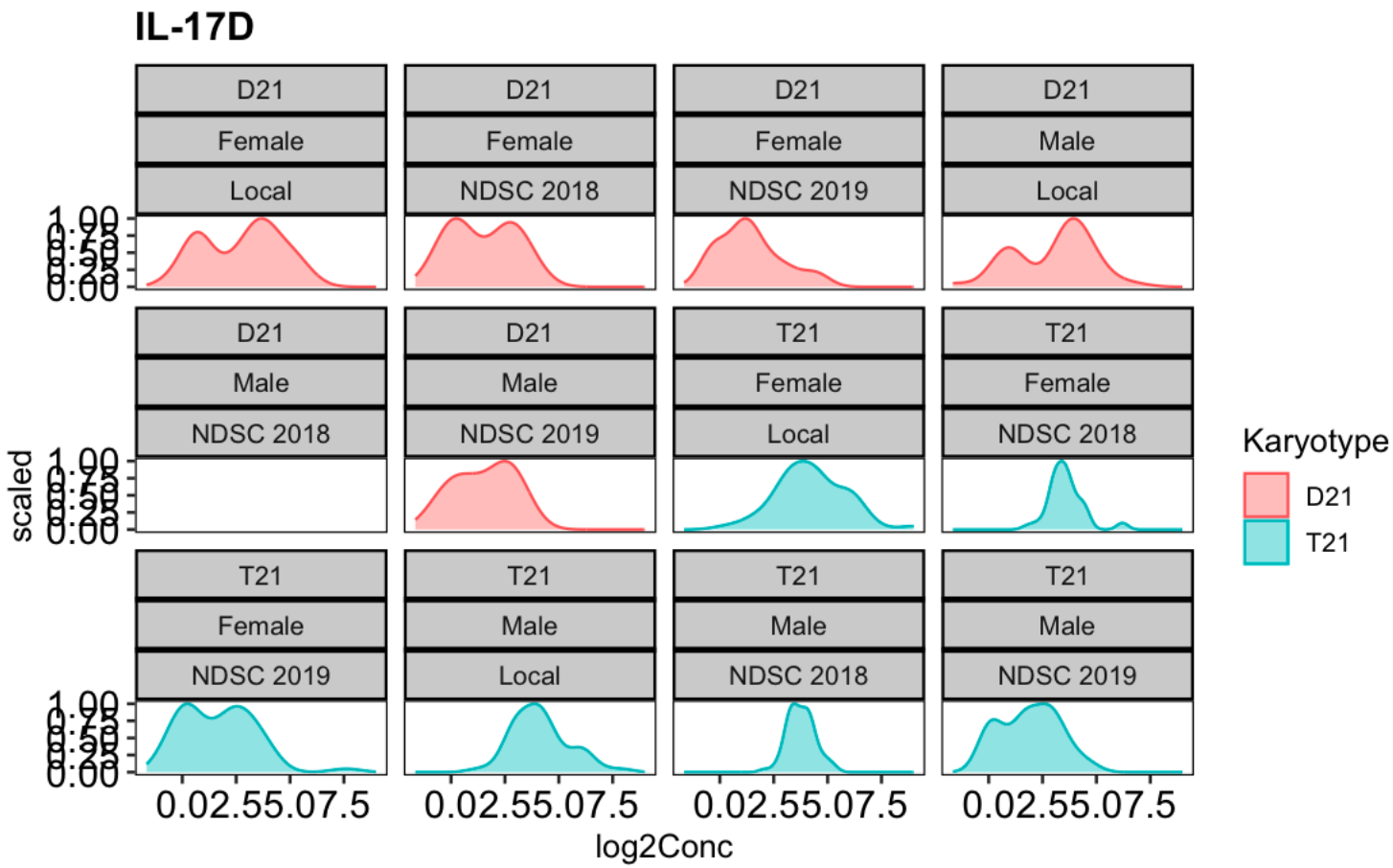
Nope.

Is it good enough to be useful?
Yep.



Thank you!

Questions?


IL-17D
D21
Female
Local
D21
Female
NDSC 2018
D21
Female
NDSC 2019
D21
Male
Local
D21
Male
NDSC 2018
D21
Male
NDSC 2019
T21
Female
Local
T21
Female
NDSC 2018
T21
Female
NDSC 2019
T21
Male
Local
T21
Male
NDSC 2018
T21
Male
NDSC 2019
scaled
0.02.55.07.5
log2Conc
Karyotype
D21
T21