9.S918: Quantitative Inference in Brain and Cognitive Sciences

Penultimate class!

- **"by-cluster" analysis versus multilevel models**
- **high-level statistical modeling (esp. regression) do's & don'ts**

Roger Levy
Dept. of Brain & Cognitive Sciences
Massachusetts Institute of Technology

May 7, 2025

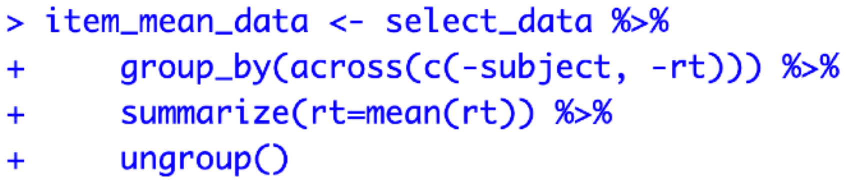
"By-cluster" analysis versus multi-level models

"By-cluster" analysis versus multi-level models

- Remember Boyce & Levy 2023 data some classes ago?

"By-cluster" analysis versus multi-level models

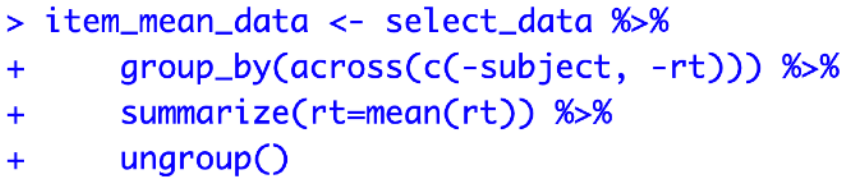
- Remember Boyce & Levy 2023 data some classes ago?
- I presented you with analysis of *item means*



```
> item_mean_data <- select_data %>%
+     group_by(across(c(-subject, -rt))) %>%
+     summarize(rt=mean(rt)) %>%
+     ungroup()
```

"By-cluster" analysis versus multi-level models

- Remember Boyce & Levy 2023 data some classes ago?
- I presented you with analysis of *item means*

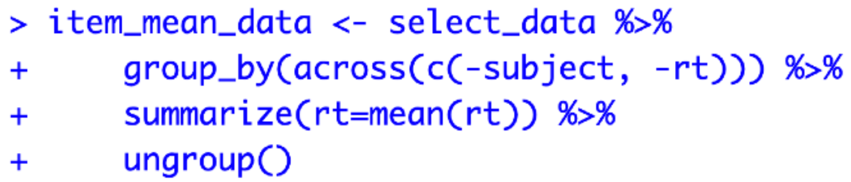


```
> item_mean_data <- select_data %>%
+     group_by(across(c(-subject, -rt))) %>%
+     summarize(rt=mean(rt)) %>%
+     ungroup()
```

- This is what is sometimes called a by-"group" analysis, where "group" might be item, participant, ...

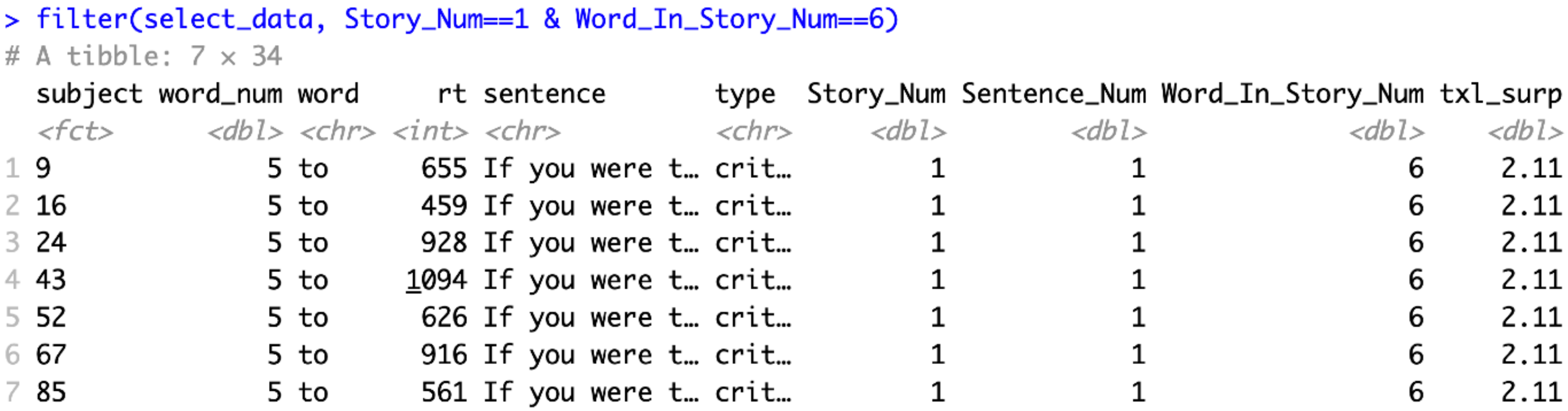
"By-cluster" analysis versus multi-level models

- Remember Boyce & Levy 2023 data some classes ago?
- I presented you with analysis of *item means*



```
> item_mean_data <- select_data %>%
+     group_by(across(c(-subject, -rt))) %>%
+     summarize(rt=mean(rt)) %>%
+     ungroup()
```

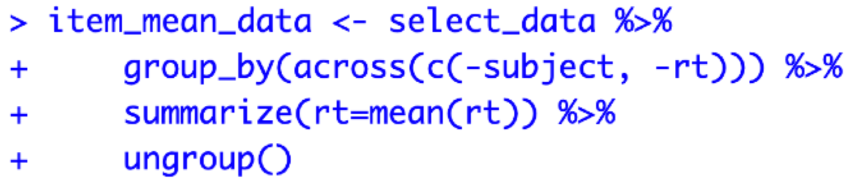
- This is what is sometimes called a by-"group" analysis, where "group" might be item, participant, ...
- But, we actually have the *raw* data too!



```
> filter(select_data, Story_Num==1 & Word_In_Story_Num==6)
# A tibble: 7 × 34
  subject word_num word     rt sentence       type  Story_Num Sentence_Num Word_In_Story_Num txl_surp
  <fct>      <dbl> <chr> <int> <chr>          <chr>     <dbl>        <dbl>             <dbl>    <dbl>
1 9              5 to      655 If you were t… crit…         1            1                 6     2.11
2 16             5 to      459 If you were t… crit…         1            1                 6     2.11
3 24             5 to      928 If you were t… crit…         1            1                 6     2.11
4 43             5 to     1094 If you were t… crit…         1            1                 6     2.11
5 52             5 to      626 If you were t… crit…         1            1                 6     2.11
6 67             5 to      916 If you were t… crit…         1            1                 6     2.11
7 85             5 to      561 If you were t… crit…         1            1                 6     2.11
```

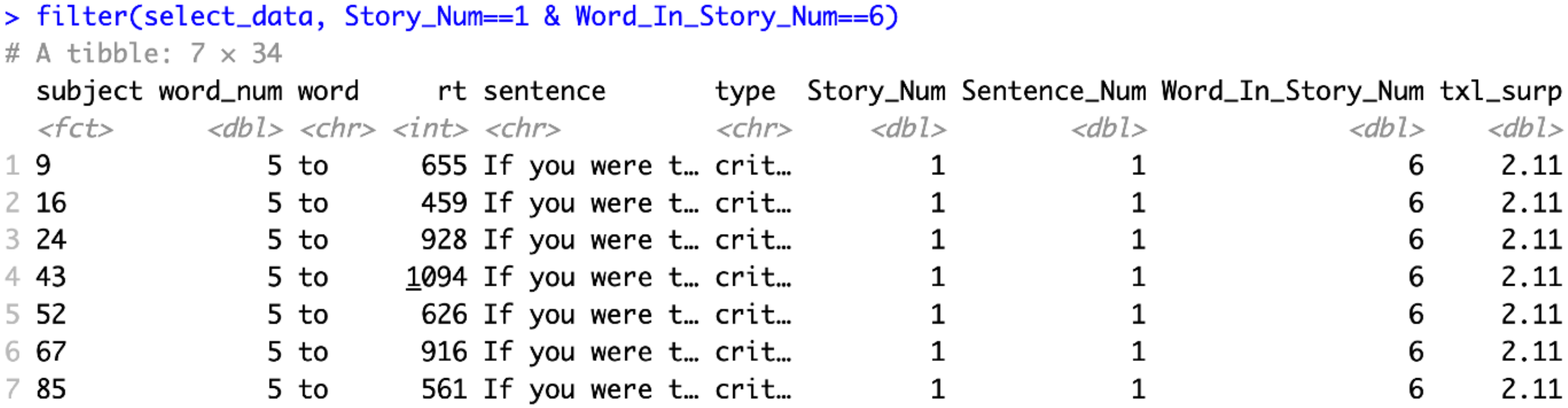
"By-cluster" analysis versus multi-level models

- Remember Boyce & Levy 2023 data some classes ago?
- I presented you with analysis of *item means*



```
> item_mean_data <- select_data %>%
+     group_by(across(c(-subject, -rt))) %>%
+     summarize(rt=mean(rt)) %>%
+     ungroup()
```

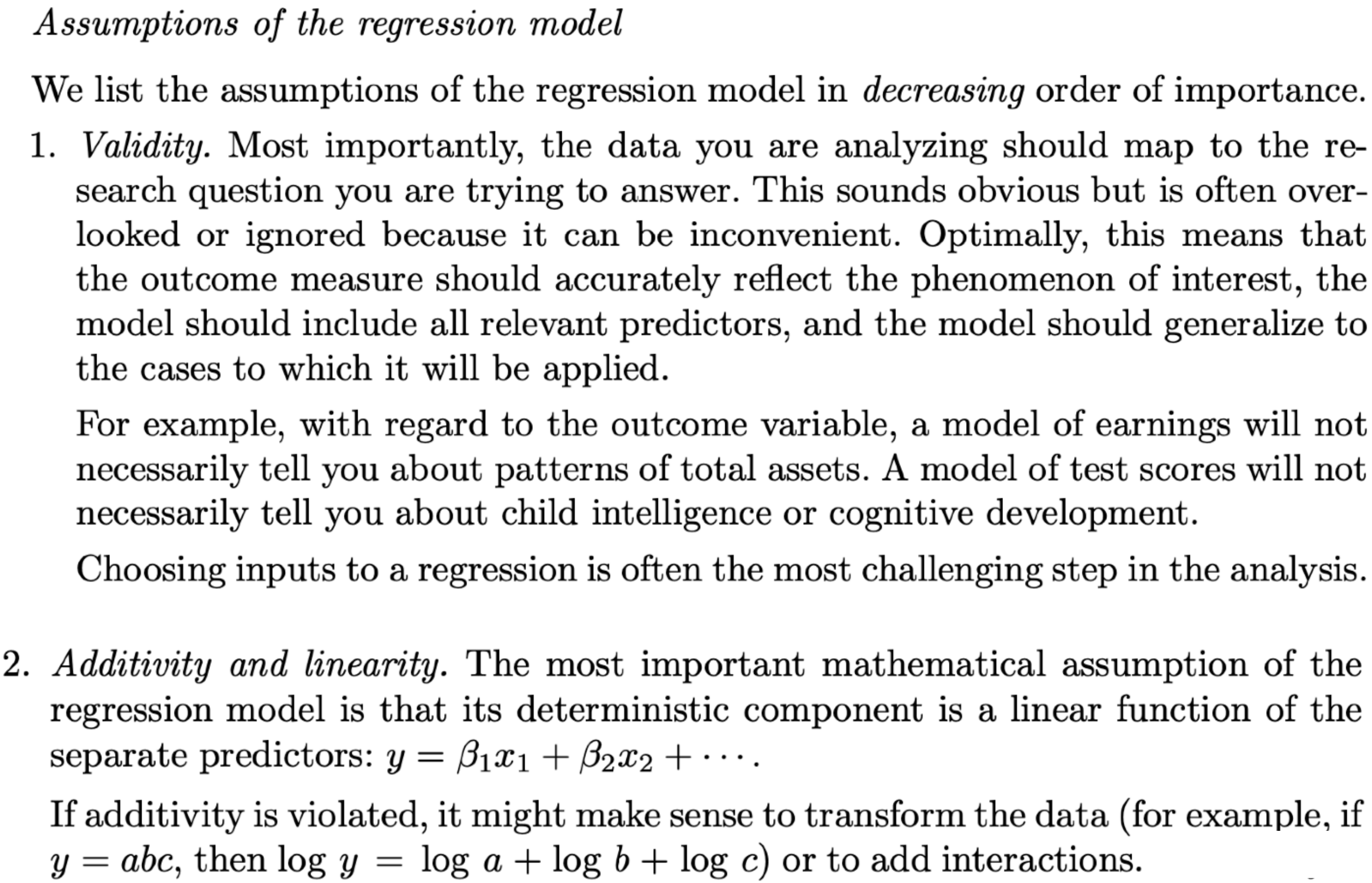
- This is what is sometimes called a by-"group" analysis, where "group" might be item, participant, ...
- But, we actually have the *raw* data too!



```
> filter(select_data, Story_Num==1 & Word_In_Story_Num==6)
# A tibble: 7 × 34
  subject word_num word     rt sentence       type  Story_Num Sentence_Num Word_In_Story_Num txl_surp
  <fct>       <dbl> <chr> <int> <chr>          <chr>     <dbl>        <dbl>             <dbl>    <dbl>
1 9               5 to      655 If you were t… crit…         1            1                 6     2.11
2 16              5 to      459 If you were t… crit…         1            1                 6     2.11
3 24              5 to      928 If you were t… crit…         1            1                 6     2.11
4 43              5 to     1094 If you were t… crit…         1            1                 6     2.11
5 52              5 to      626 If you were t… crit…         1            1                 6     2.11
6 67              5 to      916 If you were t… crit…         1            1                 6     2.11
7 85              5 to      561 If you were t… crit…         1            1                 6     2.11
```

- What are the tradeoffs in by-group vs multi-level analysis?

Assumptions & diagnostics for regression models



Assumptions of the regression model

We list the assumptions of the regression model in *decreasing* order of importance.

1. *Validity.* Most importantly, the data you are analyzing should map to the research question you are trying to answer. This sounds obvious but is often overlooked or ignored because it can be inconvenient. Optimally, this means that the outcome measure should accurately reflect the phenomenon of interest, the model should include all relevant predictors, and the model should generalize to the cases to which it will be applied.

   For example, with regard to the outcome variable, a model of earnings will not necessarily tell you about patterns of total assets. A model of test scores will not necessarily tell you about child intelligence or cognitive development.

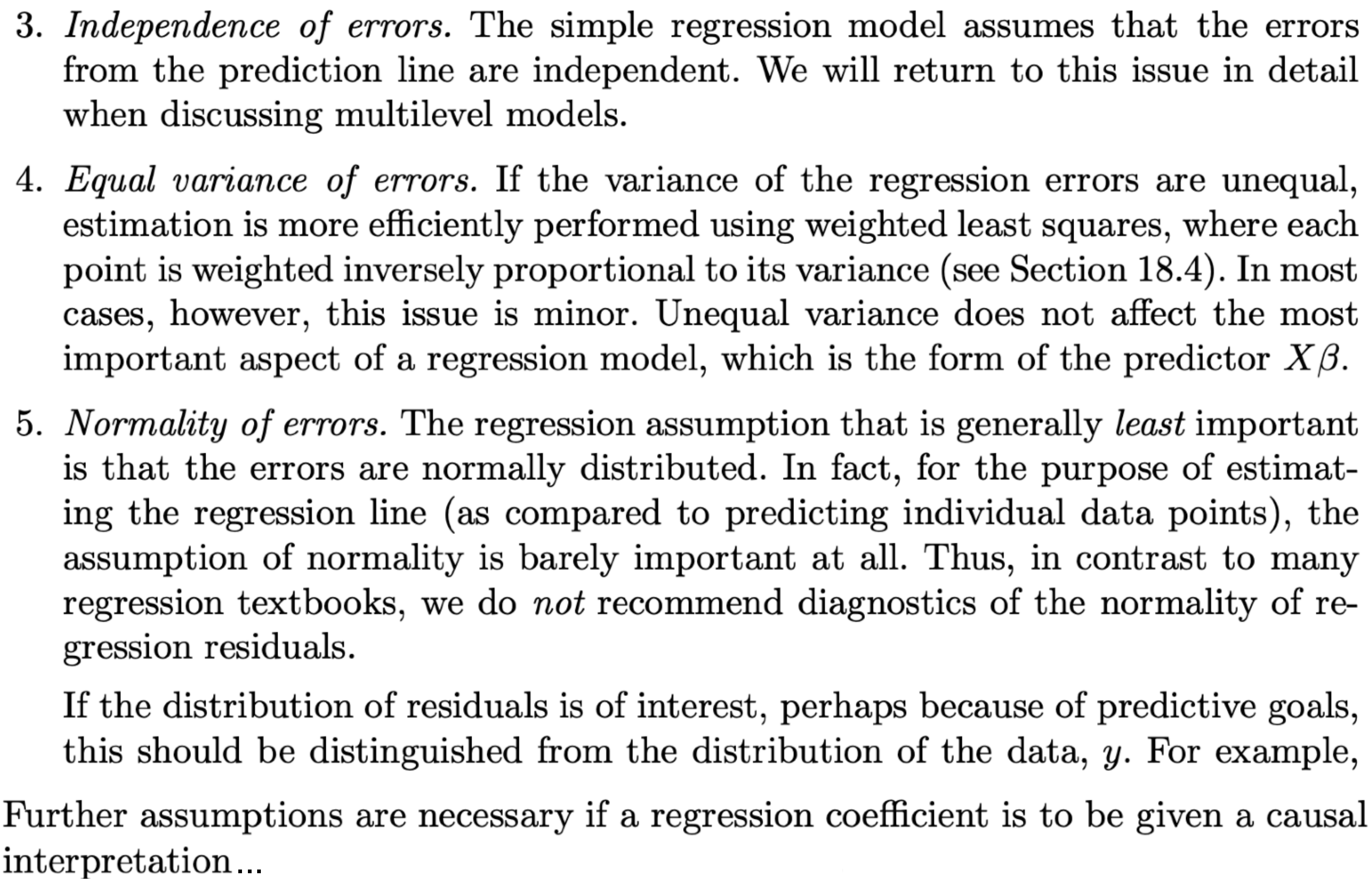
   Choosing inputs to a regression is often the most challenging step in the analysis.

2. *Additivity and linearity.* The most important mathematical assumption of the regression model is that its deterministic component is a linear function of the separate predictors: $y = \beta_1 x_1 + \beta_2 x_2 + \cdots$.

   If additivity is violated, it might make sense to transform the data (for example, if $y = abc$, then $\log y = \log a + \log b + \log c$) or to add interactions.

(Gelman & Hill 2007, drawn from the list in Section 3.6 on pp. 45–47; I encourage you to read the whole section carefully!!!)

Assumptions & diagnostics for regression models



3. *Independence of errors.* The simple regression model assumes that the errors from the prediction line are independent. We will return to this issue in detail when discussing multilevel models.

4. *Equal variance of errors.* If the variance of the regression errors are unequal, estimation is more efficiently performed using weighted least squares, where each point is weighted inversely proportional to its variance (see Section 18.4). In most cases, however, this issue is minor. Unequal variance does not affect the most important aspect of a regression model, which is the form of the predictor $X\beta$.

5. *Normality of errors.* The regression assumption that is generally *least* important is that the errors are normally distributed. In fact, for the purpose of estimating the regression line (as compared to predicting individual data points), the assumption of normality is barely important at all. Thus, in contrast to many regression textbooks, we do *not* recommend diagnostics of the normality of regression residuals.

   If the distribution of residuals is of interest, perhaps because of predictive goals, this should be distinguished from the distribution of the data, y. For example,

Further assumptions are necessary if a regression coefficient is to be given a causal interpretation...

(Gelman & Hill 2007, drawn from the list in Section 3.6 on pp. 45–47; I encourage you to read the whole section carefully!!!)

Model validation: plotting residuals

Plotting residuals to reveal aspects of the data not captured by the model

A good way to diagnose violations of some of the assumptions just considered (importantly, linearity) is to plot the residuals r_i versus fitted values $X_i\hat{\beta}$ or simply individual predictors x_i; Figure 3.12 illustrates for the test scores example where child's test score is regressed simply on mother's IQ. The plot looks fine; there do not appear to be any strong patterns. In other settings, residual plots can reveal systematic problems with model fit, as is illustrated, for example, in Chapter 6.

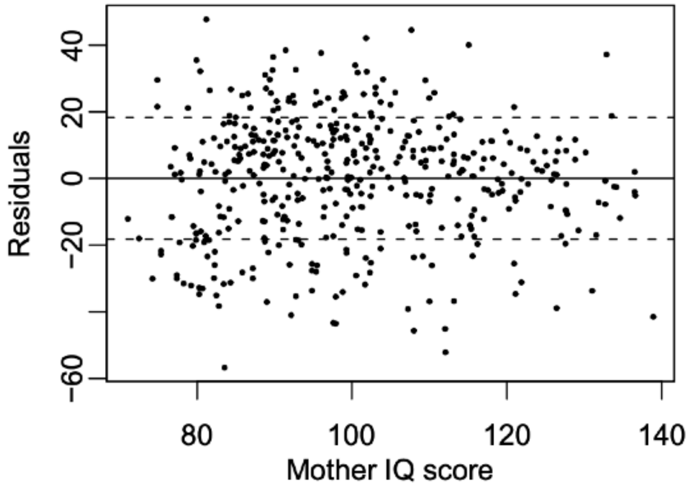


Figure 3.12 *Residual plot for child test score data when regressed on maternal IQ, with dotted lines showing ±1 standard-deviation bounds. The residuals show no striking patterns.*

(Gelman & Hill 2007, Section 3.6 on pp. 45–47; I encourage you to read the whole section carefully!!!)

High-level thoughts

- IMHO Gelman & Hill 2007's list is incredibly on point. This class's design reflects my generally aligned views. Reviewing the 5 assumptions:

1. *Validity.* "the data you are analyzing should map to the research question you are trying to answer" – we discussed this throughout both causal inference & regression
2. *Additivity and linearity.* "[the regression model's] deterministic component is a linear function of the separate predictors: $y = \beta_1 x_1 + \beta_2 x_2 + \ldots$" – discussed during regression
3. *Independence of errors.* "The simple regression model assumes that the errors from the prediction line are independent." – this was the single most important motivation for hierarchical/multi-level/mixed-effects models with **maximal design-critical random effects structure**

Retrospective

4. *Equal variance of errors.* "If the variance of the regression errors are unequal, estimation is more efficiently performed [by taking it into account in the regression model]. In most cases, however, this issue is minor. Unequal variance does not affect the most important aspect of a regression model, which is the form of the predictor $X\beta$." – we touched on this only lightly, as an aside in the context of mixed-effects models
5. *Normality of errors.* "The regression assumption that is generally *least* important is that the errors are normally distributed." – I agree, with the frequent exception of categorical response variables; hence we hardly discussed this at all

One last comment

- The one area where my approach in this class is different from Gelman & Hill 2007 involves causal inference
  - I've placed it front and center
  - We used causal graphical models quite a lot to formalize assumptions regarding the causal structure underlying data, and to identify what additional controls should be included in your analysis, and how

Statistical analysis **do's** and **don'ts**

Do:

- Translate each scientific question into a mathematical form, so that statistical inference on model parameter(s), and/or model comparison, answers the question
- Keep in mind the dimensions & units for all continuous variables in your data—both predictors and responses
- Consider potential confounding factors that would interfere with desired causal interpretations of parameter estimates, and include the right set of controls
- Keep in mind the potential limitations of any parametric assumptions you're making (e.g., that $y \sim x$ is linear)
- Carefully consider the grouping structure of your data; to account for it, use random effects that are maximal with respect to the analysis design for your scientific question

Statistical analysis **do's** and **don'ts**

Do:

- Ensure that your parameters of interest are identifiable
- Focus on the *precision* (CI size) of parameter estimates
- Keep in mind that, at least in the most popular software package(s) such as `lme4`, confidence intervals reported for a fitted mixed effects model are conditional on a point estimate of the random effects covariance matrix, $\hat{\Sigma}$
  - There may be considerable uncertainty regarding $\hat{\Sigma}$, but `lme4` doesn't tell you anything about that!
  - Because of this, **do** prefer likelihood ratio tests to t or z based confidence intervals
  - **Do** also consider Bayesian fitting of mixed models, e.g. with `brms`, to marginalize out uncertainty about Σ

Statistical analysis **do's** and **don'ts**

Do:

- Think about whether, when, and how you might convert your scientific question to a t test, with corresponding effect size estimate
- Use Monte Carlo simulation!
  - To improve your understanding of the models and algorithms you're using
  - For power analysis
  - ...for surely many other reasons that I haven't thought of!

Statistical analysis **do's** and **don'ts**

Don't:

- Transform your dependent variable just because its distribution doesn't perfectly match the residual noise assumptions of your data; consider the impact on:
  - the family of distributions your model can express; and
  - the interpretation of your model coefficients
- Include as "controls" predictors that are (plausibly) causally downstream of your response variable – it can bias the parameter estimates in your model relative to the causal effect(s) you want to estimate
- Be afraid of "singular fits" in point estimation of mixed effects models
- Drop design-critical random effects because of model convergence failure

Statistical analysis **do's** and **don'ts**

Don't:

- Mistake a large p-value (e.g., $p>0.2$) for strong evidence for the null hypothesis
- Do Bayesian hypothesis testing without principled and carefully justified choice of priors
  - Instead, **do** use domain knowledge to determine what might constitute a "practically equivalent to null" effect size, and estimate your model parameters of scientific interest precisely enough to determine how large your effect size is relative to that (this is basically Kruschke's *region of practical equivalence, or* "ROPE")