Developing the linear model

Rob Davies

2023-12-12

## 1 PSYC401: Classes weeks 6-10

* My name is Dr Rob Davies, I am an expert in communication, individual differences, and methods

|  |
| --- |
| Tip |
| **Ask me anything**:   * questions during class in person or anonymously through slido; * all other questions on discussion forum |

## 2 Week 10: Developing the linear model

|  |
| --- |
| Figure 1: Scatterplot showing the potential association between accuracy of comprehension and vocabulary scores: Data from eight studies |

## 3 Analyze + visualize + present

|  |  |
| --- | --- |
| |  | | --- | |  |   Figure 2: The data analysis pipeline or workflow: we focus on the linear model |

## 4 Develop the linear model: our aims

* We will learn how to:

1. Extend our capacity to code models so that we can incorporate **multiple predictors**
2. Develop the thought processes required to make **decisions about what predictors to include**
3. Develop the skills required to **critically evaluate results**

* Especially considering potential variation across samples

## 5 Develop the linear model: our aims

* We will revise how to:

1. Identify and interpret model statistics
2. Critically evaluate the results
3. Communicate the results

* We will learn how to: explore **extensions** of the linear model

## 6 We close the loop: Our *context*, the health comprehension project

1. Because public health impacts depend on giving people information they can understand
2. We want to know: **What makes it easy or difficult to understand written health information?**

|  |
| --- |
| flickr: Sasin Tipchair ‘Senior woman in wheelchair talking to a nurse in a hospital’ |

## 7 We close the loop: Health comprehension project, questions and analyses

1. We want to know: **What makes it easy or difficult to understand written health information?**
2. So our research questions are:

* What person attributes predict success in understanding?
* Can people accurately evaluate whether they correctly understand written health information?

## 8 Extensions to the linear model: Multiple predictors

* We need only a **limited change to R code**
* To specify a model with **multiple predictors**

## 9 How we *estimate* the association between two variables: One outcome and one predictor

model <- lm(mean.acc ~ SHIPLEY,   
 data = all.studies.subjects)  
summary(model)

1. Specify the lm function and the model mean.acc ~ ...
2. Specify what data we use data = all.studies.subjects
3. Get the results summary(model)

## 10 How we *estimate* the association between multiple variables: One outcome and *multiple* predictors

model <- lm(mean.acc ~ SHIPLEY + HLVA + FACTOR3 + AGE + NATIVE.LANGUAGE,   
 data = all.studies.subjects)  
summary(model)

1. Specify the lm function and the model:

* mean.acc ~ SHIPLEY + HLVA + FACTOR3 + AGE + NATIVE.LANGUAGE

## 11 The sentence structure of model code in R

Take a good look:

lm(mean.acc ~ SHIPLEY + HLVA + FACTOR3 + AGE + NATIVE.LANGUAGE, ...)

You will see this sentence structure in coding for *many* different analysis types

* method(outcome ~ predictors)
* predictors could be SHIPLEY + HLVA + FACTOR3 + AGE + NATIVE.LANGUAGE ...

## 12 Extensions to the linear model: Multiple predictors

* We assume that the outcome prediction errors *residuals* are normally distributed
* We do not assume that the distributions of *predictor variables* are normal

## 13 Revision: What differences between observed and predicted outcome values look like

* Differences between observed and predicted outcomes are shown by the vertical lines – outcome prediction errors: **residuals**
* Better models should show smaller differences between observed and predicted outcome values

|  |
| --- |
| Figure 3: The predicted change in mean comprehension accuracy, given variation in vocabulary scores. Observed values are shown in orange-red. Predicted values are shown in blue |

## 14 Revision: We typically assume that the residuals are normally distributed

* Some outcome prediction errors – **residuals** – are positive
* Some residuals are negative
* The average of the residuals will be zero overall

|  |
| --- |
| Figure 4: Plot showing the distribution of prediction errors – residuals – for the linear model of comprehension accuracy |

## 15 Multiple candidate predictor variables

|  |
| --- |
| Figure 5: Scatterplot showing the potential association between accuracy of comprehension and variation on each of a series of potential predictor variables. Data from 8 studies |

## 16 We do not assume normal *predictors*

|  |
| --- |
| Figure 6: Grid of plots showing the distribution of potential predictor variables. Data from 8 studies |

## 

## 17 Extensions to the linear model: Multiple predictors

|  |
| --- |
| Tip |
| We can try to model *anything* using linear models: that is the **real challenge we face**   * Any analysis you have learned can instead be done using a linear model: ANOVA, t-test, correlation, test, … * We can work with any kind of dependent or independent variable you can think of   **This is why we need to be careful** |

## 18 Analyses are done in context so when we conduct analyses we *must* use contextual information

**Closing the loop: The health comprehension project questions**

1. We want to know: *What makes it easy or difficult to understand written health information?*
2. So our research questions include:

* What person attributes predict success in understanding?

## 19 We *must* use contextual information: theory of comprehension

|  |  |
| --- | --- |
| |  | | --- | |  |   Figure 7: Understanding text depends on (1.) language experience and (2.) reasoning ability (Freed et al., 2017) |

## 20 Given theory, model of comprehension accuracy *should* include measures of

(1.) experience (HLVA, SHIPLEY) and (2.) reasoning ability (reading strategy)

|  |  |
| --- | --- |
| |  | | --- | |  |   Figure 8: Understanding text depends on (1.) language experience and (2.) reasoning ability (Freed et al., 2017) |

## 21 The flexibility and power of linear models requires us to be aware of the *garden of forking paths*

* Which variables *should be included* in an analysis?
* All of them; some of them; why?
* Will others disagree with reason?

|  |  |
| --- | --- |
| |  | | --- | |  |   Figure 9: Forking paths in data analysis |

## 22 Different researchers can reasonably make different choices

This is why we care about open science

* Theory- and evidence-based selection of critical variables for analysis *literature review*
* Share usable data and analysis code in open repositories *research report exercise*, PSYC403 data archiving

## 23 Let’s take a break

* End of part 1

## 24 Coding, thinking about, and reporting linear models with multiple predictors

lm(mean.acc ~ SHIPLEY + HLVA + FACTOR3 + AGE + NATIVE.LANGUAGE, ...)

## 25 Coding the linear model with multiple predictors

lm(mean.acc ~ SHIPLEY + HLVA + FACTOR3 + AGE + NATIVE.LANGUAGE, ...)

* The code represents a linear model with multiple predictors:

## 26 Thinking about the linear model with multiple predictors

Outcome is calculated as the sum of:

* The intercept plus
* The product of the coefficient of the effect of e.g. AGE multiplied by a person’s age +
* + any number of other variables +
* The error : mismatches between observed and predicted outcomes

## 27 Identifying key information in results

Call:  
lm(formula = mean.acc ~ SHIPLEY + HLVA + FACTOR3 + AGE + NATIVE.LANGUAGE,   
 data = all.studies.subjects)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.55939 -0.08115 0.02056 0.10633 0.41598   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.1873086 0.0472991 3.960 8.47e-05 \*\*\*  
SHIPLEY 0.0073947 0.0011144 6.635 7.70e-11 \*\*\*  
HLVA 0.0242787 0.0031769 7.642 9.44e-14 \*\*\*  
FACTOR3 0.0053455 0.0008947 5.975 4.12e-09 \*\*\*  
AGE -0.0026434 0.0004905 -5.390 1.05e-07 \*\*\*  
NATIVE.LANGUAGEOther -0.0900035 0.0141356 -6.367 4.04e-10 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.1612 on 555 degrees of freedom  
 (54 observations deleted due to missingness)  
Multiple R-squared: 0.4221, Adjusted R-squared: 0.4169   
F-statistic: 81.09 on 5 and 555 DF, p-value: < 2.2e-16

## 28 Identifying key information in results

1. The summary() of the linear model shows:
2. Estimates of the coefficients of the effects of the predictors we included, with null hypothesis significance tests of those estimates
3. Model fit statistics including R-squared and F-statistic estimates

## 29 For each predictor, e.g. HLVA, we see

Call:  
lm(formula = mean.acc ~ SHIPLEY + HLVA + FACTOR3 + AGE + NATIVE.LANGUAGE,   
 data = all.studies.subjects)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.55939 -0.08115 0.02056 0.10633 0.41598   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.1873086 0.0472991 3.960 8.47e-05 \*\*\*  
SHIPLEY 0.0073947 0.0011144 6.635 7.70e-11 \*\*\*  
HLVA 0.0242787 0.0031769 7.642 9.44e-14 \*\*\*  
FACTOR3 0.0053455 0.0008947 5.975 4.12e-09 \*\*\*  
AGE -0.0026434 0.0004905 -5.390 1.05e-07 \*\*\*  
NATIVE.LANGUAGEOther -0.0900035 0.0141356 -6.367 4.04e-10 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.1612 on 555 degrees of freedom  
 (54 observations deleted due to missingness)  
Multiple R-squared: 0.4221, Adjusted R-squared: 0.4169   
F-statistic: 81.09 on 5 and 555 DF, p-value: < 2.2e-16

1. The Coefficient Estimate: 0.0242787 for the slope of the effect of variation in HLVA scores
2. The Std. Error (standard error) 0.0031769 for the estimate
3. The t value of 7.642 and associated Pr(>|t|) p-value 9.44e-14 for the null hypothesis test of the coefficient

## 30 Identifying the key information in the linear model results: Coefficients

Call:  
lm(formula = mean.acc ~ SHIPLEY + HLVA + FACTOR3 + AGE + NATIVE.LANGUAGE,   
 data = all.studies.subjects)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.55939 -0.08115 0.02056 0.10633 0.41598   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.1873086 0.0472991 3.960 8.47e-05 \*\*\*  
SHIPLEY 0.0073947 0.0011144 6.635 7.70e-11 \*\*\*  
HLVA 0.0242787 0.0031769 7.642 9.44e-14 \*\*\*  
FACTOR3 0.0053455 0.0008947 5.975 4.12e-09 \*\*\*  
AGE -0.0026434 0.0004905 -5.390 1.05e-07 \*\*\*  
NATIVE.LANGUAGEOther -0.0900035 0.0141356 -6.367 4.04e-10 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.1612 on 555 degrees of freedom  
 (54 observations deleted due to missingness)  
Multiple R-squared: 0.4221, Adjusted R-squared: 0.4169   
F-statistic: 81.09 on 5 and 555 DF, p-value: < 2.2e-16

* Pay attention to **sign and the size** of coefficient estimate:
* Is the coefficient (e.g., HLVA 0.0242787) a positive or a negative number? is it relatively large or small?

## 31 Identifying the key information in the linear model results: R-squared

Call:  
lm(formula = mean.acc ~ SHIPLEY + HLVA + FACTOR3 + AGE + NATIVE.LANGUAGE,   
 data = all.studies.subjects)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.55939 -0.08115 0.02056 0.10633 0.41598   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.1873086 0.0472991 3.960 8.47e-05 \*\*\*  
SHIPLEY 0.0073947 0.0011144 6.635 7.70e-11 \*\*\*  
HLVA 0.0242787 0.0031769 7.642 9.44e-14 \*\*\*  
FACTOR3 0.0053455 0.0008947 5.975 4.12e-09 \*\*\*  
AGE -0.0026434 0.0004905 -5.390 1.05e-07 \*\*\*  
NATIVE.LANGUAGEOther -0.0900035 0.0141356 -6.367 4.04e-10 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.1612 on 555 degrees of freedom  
 (54 observations deleted due to missingness)  
Multiple R-squared: 0.4221, Adjusted R-squared: 0.4169   
F-statistic: 81.09 on 5 and 555 DF, p-value: < 2.2e-16

* Revision: Pay attention to R-squared
* R-squared indicates how much outcome variation we can predict, given our model
* Revision: we report Adjusted R-squared because it tends to be more accurate

## 32 Identifying the key information in the linear model results: F

Call:  
lm(formula = mean.acc ~ SHIPLEY + HLVA + FACTOR3 + AGE + NATIVE.LANGUAGE,   
 data = all.studies.subjects)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.55939 -0.08115 0.02056 0.10633 0.41598   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.1873086 0.0472991 3.960 8.47e-05 \*\*\*  
SHIPLEY 0.0073947 0.0011144 6.635 7.70e-11 \*\*\*  
HLVA 0.0242787 0.0031769 7.642 9.44e-14 \*\*\*  
FACTOR3 0.0053455 0.0008947 5.975 4.12e-09 \*\*\*  
AGE -0.0026434 0.0004905 -5.390 1.05e-07 \*\*\*  
NATIVE.LANGUAGEOther -0.0900035 0.0141356 -6.367 4.04e-10 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.1612 on 555 degrees of freedom  
 (54 observations deleted due to missingness)  
Multiple R-squared: 0.4221, Adjusted R-squared: 0.4169   
F-statistic: 81.09 on 5 and 555 DF, p-value: < 2.2e-16

* The model summary gives us the F-statistic:
* Revision: the F-test of the null hypothesis that the model *does not* predict the outcome

## 33 Plot predictions to interpret effects

|  |
| --- |
| Figure 10: A grid of plots showing model predictions, for outcome accuracy, given variation in (a.) age, (b.) vocabulary, (c.) health literacy, (d) reading strategy and (e.) native language. Data from eight studies |

## 34 Compare estimates with effects plots

|  |
| --- |
|  |

* Coefficients estimates in the summary match what we see
* Positive coefficients show upward slopes
* Larger coefficients show steeper slopes

## 35 The language and style of reporting linear model results

We fitted a linear model with mean comprehension accuracy as the outcome and, as predictors: vocabulary knowledge (Shipley), health literacy (HLVA), reading strategy (FACTOR3), age (years) and native language status. Our analysis indicated significant effects of all predictor variables. The model is significant overall, with , and explains 42% of variance (). The model estimates showed that the accuracy of comprehension increased with higher levels of participant vocabulary knowledge (), health literacy (), and reading strategy (). Younger participants () and native speakers of English as another language () tended to show lower levels of accuracy.

## 36 Look at what we do with the text

We fitted a linear model with mean comprehension accuracy as the outcome and, as predictors: vocabulary knowledge (Shipley), health literacy (HLVA), reading strategy (FACTOR3), age (years) and native language status. Our analysis indicated significant effects of all predictor variables. The model is significant overall, with , and explains 42% of variance (). The model estimates showed that the accuracy of comprehension increased with higher levels of participant vocabulary knowledge (), health literacy (), and reading strategy (). Younger participants () and native speakers of English as another language () tended to show lower levels of accuracy.

1. Explain: the method (linear model); the outcome (accuracy) and the predictors
2. Report the model fit statistics overall ()
3. Report the significant effects () and describe the nature of the effects

## 37 Let’s take a break

* End of part 2

## 38 Critically evaluating the results of analyses involving linear models

There are three levels of **uncertainty** when we look at sample data (McElreath, 2020) – uncertainty over:

1. The nature of the expected change in outcome
2. The ways that expected changes might vary between individual participants or between groups of participants
3. The random ways that specific responses can be produced

## 39 Critically evaluating the results of analyses involving linear models

* These uncertainties require us to carefully qualify the conclusions we draw from data analyses
* This does not mean we should avoid *causal language* when we think that psychological processes cause the behaviours we examine (Grosz et al., 2020)
* But it *does mean* we can be careful to identify the limits in the evidence we analyse

## 40 Revision: As we move into thinking about the data analysis, we need to identify our assumptions

1. **validity**: that differences in knowledge or ability cause differences in test scores
2. **measurement**: that this is equally true across the different kinds of people we tested
3. **generalizability**: that the sample of people we recruited resembles the population

## 41 How do *you* do this work?

1. **validity**

* We want to work with valid measures but *validity* requires explaining (Borsboom et al., 2004):

1. Does the thing exist in the world?
2. Is variation in that thing be reflected in variation in our measurement?

* What you can do: literature review to identify your reasoning in answer to these questions

## 42 How do *you* do this work?

1. **measurement**
2. **generalizability**

* It is most helpful to assume from the start that effects estimates will vary (Gelman, 2015; Vasishth & Gelman, 2021)
* So then we ask ourselves: will this test work in the same way in different groups?
* And we ask: how will these effects estimates vary across different groups

## 43 Why we need replication studies

|  |
| --- |
| Figure 11: Scatterplot showing the potential association between accuracy of comprehension and vocabulary scores: Data from eight studies. Effects will vary between different samples so: expect the variation (Gelman, 2015; Vasishth & Gelman, 2021) >>> important to evaluating claims in the literature, and to evaluation of your own results |

## 44 Why we need replication studies

|  |
| --- |
| Figure 12: Effects will vary between samples so expect the variation (Gelman, 2015; Vasishth & Gelman, 2021) >>> ask what variation may result from systematic differences between groups |

## 45 Why we need to consider the generalizability of sample data

|  |
| --- |
| Figure 13: Grid of plots showing the distribution of potential predictor variables |

## 46 Convenience samples are common in Psychology

* We test who we can – convenience sampling – and who we can test has an impact on the quality of evidence (Bornstein et al., 2013)
* If age, ethnicity or gender are not balanced does this matter to your research question?
* If samples are limited in size how does that affect our uncertainty over effects estimates?

## 47 Let’s take a break

* End of part 3

## 48 The linear model is very flexible, powerful and general

* Most introductory statistics classes teach each statistical test *as if* they are independent

|  |
| --- |
| Tip |
| Most common statistical tests are special cases of linear models, or are close approximations |

## 49 The t-test as linear model

* If you have two groups, with a variable X coding for group membership
* Then the mean outcome for one group
* The estimate of the slope tells about the average difference between groups
* And we can code the model like this: lm(y ~ group)

## 50 ANOVA as linear model

* If you have a 2 x 2 factorial design, with two factors factor.1, factor.2, and a dataset with variables X, Z coding for group membership
* Then the mean outcome for baseline conditions
* The estimates of the slopes tells about the average difference between groups
* The estimate of the slope tells us about the interaction
* And we can code the model like this: lm(y ~ factor.1\*factor.2)
* Or this Anova(aov(y ~ factor.1\*factor.2, data), type='II')

## 51 ANOVA as linear model

* In general, the psychological literature is full of ANOVA
* But the field is moving away from ANOVA towards mixed-effects models

|  |
| --- |
| Tip |
| We have to make choices in teaching and, here, we are choosing to focus on a powerful, flexible, and generally applicable method we can explain *in depth*: linear models   * Our aim is for students to better understand how to use a general approach |

## 52 Extensions to the linear model

* outcome can generalize to analyse data that are not metric, do not come from normal distributions
* predictors can be curvilinear, categorical, involve interactions
* error can be independent; can be non-independent

## 53 Look ahead: extensions to the linear model

* What if the outcome measurement data cannot be understood to be metric or to come from a normal probability distribution?

## 54 Extensions to the linear model – binary or dichotomous outcomes

1. Binary outcomes are very common in Psychology: yes or no; correct or incorrect; left or right visual field etc.
2. The change in coding is e.g. glm(ratings ~ predictors, family = "binomial")

## 55 Extensions to the linear model – ordinal outcomes

1. Likert scale or ratings data are best analysed using ordinal models (Liddell & Kruschke, 2018)
2. The change in coding (see Christensen, 2022) is e.g. clm(ratings ~ predictors)

## 56 Extensions to the linear model – non-independence of observations

1. Much – maybe most – psychological data are collected in ways that guarantee the non-independence of observations

* We test children in classes, patients in clinics, individuals in regions
* We test participants in multiple trials in an experiment, recording responses to multiple stimuli

1. These data should be analysed using **linear mixed-effects models** (Meteyard & Davies, 2020)

## 57 General advice

An old saying goes:

All models are wrong but some are useful

(attributed to George Box).

|  |
| --- |
| Tip |
| * Sometimes, it can be useful to adopt a simpler approach as a way to approximate *get closer to* better methods * Box also advises “Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad.” * Here, we focus on validity, measurement, generalizability and *critical thinking* |

## 58 Summary

* Linear models are a very general, flexible, and powerful analysis method
* We can use assuming that prediction outcomes (residuals) are normally distributed
* With potentially multiple predictor variables

## 59 Summary

* Closing the loop: when we plan an analysis we should try to use contextual information – theory and measurement understanding – to specify our model
* Closing the loop: when we critically evaluate our or others’ findings, we should consider validity, measurement, and generalizability

## 60 Summary

* When we report an analysis, we should report:

1. Explain what I did, specifying the method (linear model), the outcome variable (accuracy) and the predictor variables (health literacy, reading strategy, reading skill and vocabulary)
2. Report the model fit statistics overall ()
3. Report the significant effects () and describe the nature of the effects (does the outcome increase or decrease?)

## 61 End of week 10

## 62 References

Bornstein, M. H., Jager, J., & Putnick, D. L. (2013). Sampling in developmental science: Situations, shortcomings, solutions, and standards. *Developmental Review*, *33*(4), 357–370. <https://doi.org/10.1016/j.dr.2013.08.003>

Borsboom, D., Mellenbergh, G. J., & Heerden, J. van. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. <https://doi.org/10.1037/0033-295X.111.4.1061>

Christensen, R. H. B. (2022). *Ordinal: Regression models for ordinal data*. <https://CRAN.R-project.org/package=ordinal>

Freed, E. M., Hamilton, S. T., & Long, D. L. (2017). Comprehension in proficient readers: The nature of individual variation. *Journal of Memory and Language*, *97*, 135–153. <https://doi.org/10.1016/j.jml.2017.07.008>

Gelman, a. (2015). The connection between varying treatment effects and the crisis of unreplicable research: A bayesian perspective. *Journal of Management*, *41*(2), 632–643. <https://doi.org/10.1177/0149206314525208>

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The Taboo Against Explicit Causal Inference in Nonexperimental Psychology. *Perspectives on Psychological Science*, *15*(5), 1243–1255. <https://doi.org/10.1177/1745691620921521>

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*, 328–348.

McElreath, R. (2020). *Statistical rethinking*. Chapman; Hall/CRC. <https://doi.org/10.1201/9780429029608>

Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, 104092.

Vasishth, S., & Gelman, A. (2021). How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics*, *59*(5), 1311–1342. <https://doi.org/10.1515/ling-2019-0051>