Developing the linear model

Rob Davies

## 1 PSYC122: Classes in weeks 16-20

* 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 18: Developing the linear model

|  |
| --- |
| Figure 1: Scatterplot showing the potential association between accuracy of comprehension and health literacy |

## 3 Targets for weeks 16-19: Concepts

We are working together to develop concepts:

1. *Week 16* — Hypotheses, measurement and associations
2. *Week 17* — Predicting people using linear models
3. ***Week 18*** **— Everything is some kind of linear model**
4. *Week 19* — The real challenge in psychological science

## 4 Targets for weeks 16-19: Skills

We are working together to develop skills:

1. *Week 16* — Visualizing, estimating, and reporting associations
2. *Week 17* — Using data to *predict* people
3. ***Week 18*** **— Going deeper on linear models**
4. *Week 19* — Evaluating evidence across multiple studies

## 5 Learning targets for this week

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

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

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

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

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

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

1. Specify the lm function and the model:

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

## 8 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 ...

## 9 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

## 10 Revision: Differences between observed and predicted outcomes

* Differences between observed and predicted outcomes are outcome prediction errors: **residuals**
* These differences are shown by the vertical grey lines joining the points in [Figure 2](#fig-abline-predict-residuals-1)
* Better models should show smaller differences between observed and predicted outcome values

|  |
| --- |
| Figure 2: 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 |

## 11 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
* Because there is a balance of positive and negative prediction errors

|  |
| --- |
| Figure 3: Plot showing the distribution of prediction errors – residuals – for the linear model of comprehension accuracy |

## 12 Multiple candidate predictor variables

|  |
| --- |
| Figure 4: Scatterplots showing the potential association between accuracy of comprehension and variation on each of a series of potential predictor variables. Data are from two studies |

## 13 We do not assume normal *predictors*

|  |
| --- |
| Figure 5: Grid of plots showing the distribution of potential predictor variables |

## 14 Extensions to the linear model: Multiple predictors

|  |
| --- |
| Warning |
| We can try to model *anything* using linear models: that is the **real challenge we face**   * Any analysis you learn can instead be done using some form of (general) 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** |

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

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

|  |
| --- |
| Note |
| * What person attributes predict success in understanding? |

## 16 Given theory, a model of comprehension accuracy *should* include – as predictors:

(1.) experience HLVA, SHIPLEY and (2.) reasoning ability (FACTOR3, reading strategy) (Freed et al., 2017)

|  |  |
| --- | --- |
| |  | | --- | |  |   Figure 6: Understanding text depends on (1.) language experience and (2.) reasoning ability (Freed et al., 2017) |

## 17 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?
* Why?
* *Flexibility* in choosing predictors can impact results (A. Gelman & Loken, 2013)

|  |  |
| --- | --- |
| |  | | --- | |  |   Figure 7: Forking paths in data analysis |

## 18 Different researchers can reasonably make different choices

This is why we teach:

* *Evidence reviews* like meta-analysis theory- and evidence-based selection of critical variables for analysis
* *Open science*: we share usable data and analysis code in open repositories so others can critically evaluate our work

## 19 Let’s take a break

* End of part 1

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

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

## 21 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:

## 22 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

## 23 Identifying key information in results

Call:  
lm(formula = mean.acc ~ SHIPLEY + HLVA + FACTOR3 + AGE, data = clearly.both.subjects)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.38092 -0.05889 0.01296 0.06780 0.21677   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.2372834 0.0591436 4.012 7.43e-05 \*\*\*  
SHIPLEY 0.0067294 0.0015225 4.420 1.33e-05 \*\*\*  
HLVA 0.0179228 0.0026682 6.717 7.93e-11 \*\*\*  
FACTOR3 0.0032872 0.0008595 3.824 0.000156 \*\*\*  
AGE -0.0003125 0.0004374 -0.715 0.475391   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.101 on 336 degrees of freedom  
Multiple R-squared: 0.2938, Adjusted R-squared: 0.2854   
F-statistic: 34.94 on 4 and 336 DF, p-value: < 2.2e-16

## 24 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

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

Call:  
lm(formula = mean.acc ~ SHIPLEY + HLVA + FACTOR3 + AGE, data = clearly.both.subjects)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.38092 -0.05889 0.01296 0.06780 0.21677   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.2372834 0.0591436 4.012 7.43e-05 \*\*\*  
SHIPLEY 0.0067294 0.0015225 4.420 1.33e-05 \*\*\*  
HLVA 0.0179228 0.0026682 6.717 7.93e-11 \*\*\*  
FACTOR3 0.0032872 0.0008595 3.824 0.000156 \*\*\*  
AGE -0.0003125 0.0004374 -0.715 0.475391   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.101 on 336 degrees of freedom  
Multiple R-squared: 0.2938, Adjusted R-squared: 0.2854   
F-statistic: 34.94 on 4 and 336 DF, p-value: < 2.2e-16

1. The Coefficient Estimate: 0.0179228 for the slope of the effect of variation in HLVA scores
2. The Std. Error (standard error) 0.0026682 for the estimate
3. The t value of 6.717 and associated Pr(>|t|) p-value 7.93e-11 for the null hypothesis test of the coefficient

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

Call:  
lm(formula = mean.acc ~ SHIPLEY + HLVA + FACTOR3 + AGE, data = clearly.both.subjects)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.38092 -0.05889 0.01296 0.06780 0.21677   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.2372834 0.0591436 4.012 7.43e-05 \*\*\*  
SHIPLEY 0.0067294 0.0015225 4.420 1.33e-05 \*\*\*  
HLVA 0.0179228 0.0026682 6.717 7.93e-11 \*\*\*  
FACTOR3 0.0032872 0.0008595 3.824 0.000156 \*\*\*  
AGE -0.0003125 0.0004374 -0.715 0.475391   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.101 on 336 degrees of freedom  
Multiple R-squared: 0.2938, Adjusted R-squared: 0.2854   
F-statistic: 34.94 on 4 and 336 DF, p-value: < 2.2e-16

* Pay attention to **sign and the size** of coefficient estimate:
* Is the coefficient (e.g., HLVA 0.0179228) a positive or a negative number? is it relatively large or small?

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

Call:  
lm(formula = mean.acc ~ SHIPLEY + HLVA + FACTOR3 + AGE, data = clearly.both.subjects)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.38092 -0.05889 0.01296 0.06780 0.21677   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.2372834 0.0591436 4.012 7.43e-05 \*\*\*  
SHIPLEY 0.0067294 0.0015225 4.420 1.33e-05 \*\*\*  
HLVA 0.0179228 0.0026682 6.717 7.93e-11 \*\*\*  
FACTOR3 0.0032872 0.0008595 3.824 0.000156 \*\*\*  
AGE -0.0003125 0.0004374 -0.715 0.475391   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.101 on 336 degrees of freedom  
Multiple R-squared: 0.2938, Adjusted R-squared: 0.2854   
F-statistic: 34.94 on 4 and 336 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

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

Call:  
lm(formula = mean.acc ~ SHIPLEY + HLVA + FACTOR3 + AGE, data = clearly.both.subjects)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.38092 -0.05889 0.01296 0.06780 0.21677   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.2372834 0.0591436 4.012 7.43e-05 \*\*\*  
SHIPLEY 0.0067294 0.0015225 4.420 1.33e-05 \*\*\*  
HLVA 0.0179228 0.0026682 6.717 7.93e-11 \*\*\*  
FACTOR3 0.0032872 0.0008595 3.824 0.000156 \*\*\*  
AGE -0.0003125 0.0004374 -0.715 0.475391   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.101 on 336 degrees of freedom  
Multiple R-squared: 0.2938, Adjusted R-squared: 0.2854   
F-statistic: 34.94 on 4 and 336 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

## 29 Plot predictions to interpret effects

|  |
| --- |
| Figure 8: A grid of plots showing model predictions, for outcome accuracy, given variation in (a.) age, (b.) vocabulary, (c.) health literacy, and (d) reading strategy |

## 30 Compare estimates with effects plots

|  |
| --- |
|  |

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

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

We fitted a linear model with mean comprehension accuracy as the outcome and vocabulary knowledge (Shipley), health literacy (HLVA), reading strategy (FACTOR3), and age (years) as predictors. The model is significant overall, with , and explains 29% 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 () tended to show lower levels of accuracy but the age effect was not significant.

## 32 Look at what we do with the text

We fitted a linear model with mean comprehension accuracy as the outcome and vocabulary knowledge (Shipley), health literacy (HLVA), reading strategy (FACTOR3), and age (years) as predictors. The model is significant overall, with , and explains 29% 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 () tended to show lower levels of accuracy but the age effect was not significant.

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

## 33 Let’s take a break

* End of part 2

## 34 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

## 35 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

## 36 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

## 37 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

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

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

* It is most helpful to assume from the start that effects estimates will vary (a. 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

## 39 Why we need replication studies

|  |
| --- |
| Figure 9: 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 (a. Gelman, 2015; Vasishth & Gelman, 2021) >>> important to evaluating claims in the literature, and to evaluation of your own results |

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

|  |
| --- |
| Figure 10: Grid of plots showing the distribution of potential predictor variables |

## 41 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)

|  |
| --- |
| Tip |
| Practice **critical evaluation**:   * 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? |

## 42 Let’s take a break

* End of part 3

## 43 *Everything* is some kind of linear model

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

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

## 44 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)

## 45 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')

## 46 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

## 47 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")

## 48 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) which we study at *MSc*

## 49 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* |

## 50 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

## 51 Summary

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

## 52 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?)

## 53 End of week 18

## 54 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>

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>

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no ?fishing expedition? Or ?p-hacking? And the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*.

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>

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>