Better understanding 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 17: Better understanding the linear model

|  |
| --- |
|  |

## 3 Objectives: 1. Link together ideas on how to do psychological science

* You are learning about:

1. the scientific method
2. **measurement and hypothesis testing**
3. modern reproducible open science

* Our job now is to **connect these ideas together**

|  |
| --- |
| flickr, Khunal Gate ‘Web’ |

## 4 Objectives: 2. Strengthen your practice and build your independence

* In PSYC121 and PSYC122, you have learned about working with data
* In PSYC122, so far: you have been introduced to correlations and linear models
* Our job now is to **deepen and broaden your skills**

|  |
| --- |
| flickr, Magryciak ‘Great weekend’ |

## 5 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

## 6 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

## 7 Learning targets for this week

* *Skills* – Understand how to code: lm(mean.acc ~ SHIPLEY)
* *Concepts* – To answer questions like: **Is comprehension success influenced by vocabulary knowledge?**

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

## 8 Learning targets for this week

* We will learn how to do analysis in the *context* of a live research project: the **health comprehension project**

1. code linear models
2. identify and interpret model statistics
3. critically evaluate the results
4. communicate the results

## 9 Thinking about relationships in psychologial science

We often want to know about relationships

* Does variation in X predict variation in Y?
* What are the factors that influence outcome Y?
* Is a theoretical model consistent with observed behaviour?

## 10 Now consider our research aims in the *context* of 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’ |

## 11 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?

1. These kinds of research questions can be answered using methods like linear models

## 12 Context: Individual differences theory of comprehension success

* Understanding text depends on (1.) language experience and (2.) reasoning ability (Freed et al., 2017)

|  |  |
| --- | --- |
| |  | | --- | |  |   Figure 2: Hypothesized predictors of comprehension |

## 13 The measurement context: Where the data come from

* We measure reading comprehension: asking people to read text and then answer multiple choice questions
* We measure background knowledge: vocabulary knowledge (Shipley); health literacy (HLVA)

## 14 Reflect: The kinds of *critical evaluation* questions you can ask yourself

* Are multiple choice questions good ways to probe understanding? *What alternatives are there?*
* Are tests like the Shipley good measures of language knowledge? *What do we miss?*

## 15 Reflect: 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 looks like the population

## 16 We need to think about the derivation chain

|  |  |
| --- | --- |
| |  | | --- | |  |   Figure 3: The derivation chain, introduced in week 16 |

## 17 Questions, assumptions, predictions

Link: concepts, questions assumptions testable predictions

1. **concepts, questions**: Can people accurately understand health guidance?
2. **assumptions**: People who know more about language should also present more accurate understanding
3. **testable predictions**: Higher levels of vocabulary should be associated with higher levels of comprehension accuracy: we expect to estimate a positive coefficient

## 18 One way of thinking about the association is to visualize it

* For each value of the predictor **vocabulary**
* Does the the value of the outcome **accuracy**
* *Increase or decrease?*

|  |
| --- |
| Figure 4: The association between comprehension accuracy and vocabulary |

## 19 Let’s take a break

* End of part 1

## 20 Predicted association as *expected* change in average outcome

* [Figure 5](#fig-scatterplot-ridges-shipley-acc-1) shows the distribution curve of mean (comprehension) accuracy scores observed at each value of vocabulary
* You can see that the middle – the average – of each distribution increases
* as we go from left (low scores) to right (high scores) on vocabulary

|  |
| --- |
| Figure 5: Association between accuracy and vocabulary |

## 21 How do we *estimate* the association between two variables?

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

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

## 22 The sentence structure of models in R

Take a good look:

lm(mean.acc ~ SHIPLEY ...)

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

* method(outcome ~ predictors)
* method could be aov, brm, lm, glm, glmm, lmer, t.test, cor.test

## 23 Results: How does the outcome vary in relation to the predictor?

Call:  
lm(formula = mean.acc ~ SHIPLEY, data = clearly.one.subjects)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.42871 -0.04921 0.02079 0.07480 0.18430   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.44914 0.08053 5.577 9.67e-08 \*\*\*  
SHIPLEY 0.01050 0.00229 4.585 8.85e-06 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.1115 on 167 degrees of freedom  
Multiple R-squared: 0.1118, Adjusted R-squared: 0.1065   
F-statistic: 21.03 on 1 and 167 DF, p-value: 8.846e-06

* A model summary gives us estimates of:
* The coefficient for the intercept
* The coefficient for the slope of the SHIPLEY ‘effect’

## 24 These *coefficients* build a line

* The line represents:
* our prediction for how the outcome varies *on average*
* given change in the predictor

## 25 So now we need to think about straight lines

1. You may remember from school that to draw a straight line you need four numbers:
2. We calculate the height by adding

* the intercept, the value of y when
* to the product of the coefficient for the slope of the line
* multiplied by the value of the predictor variable

## 26 Let’s draw it

* Look at what we get if we draw the line using the linear model coefficients:
* for the intercept,
* for the slope,
* In the formula:
* (I round the numbers to three decimal places.)

|  |
| --- |
| Figure 6: The predicted association between comprehension accuracy and vocabulary |

## 27 We can understand the line as representing a set of predictions

* To see how — we use the coefficients to predict just one potential outcome:
* the expected accuracy for someone with a vocabulary score of 20
* We do this using the formula:

|  |
| --- |
| Figure 7: Predicted outcome in red |

## 28 We can understand the line as representing a set of predictions

* Let’s expand our predictions

1. Predict accuracy given a Shipley score of 20
2. Predict accuracy given a Shipley score of 30

|  |
| --- |
| Figure 8: Predicted outcome in red |

## 29 The linear model allows us to predict the average outcome we can expect given *any* value of the predictor

|  |
| --- |
| Figure 9: The predicted change in mean comprehension accuracy, given variation in vocabulary scores |

## 30 Let’s take a break

* End of part 2

## 31 We could draw a variety of different model prediction lines: how do we pick the *right* one?

|  |
| --- |
|  |

## 32 We could draw a variety of different model prediction lines: how do we pick the *right* one?

We need to go back to the prediction model

* To calculate a predicted outcome value, we calculated it as:
* Assuming the linear model
* But we missed a bit: **error**

## 33 Linear models are typically estimated given sample data

* Maybe you noticed that I talked about how the model allows us to predict
* how the outcome varies **on average** given different values of the predictor
* When we use a linear model to estimate the intercept and slope – to build the predictions – we fit a model to the sample data
* And no model will fit sample data perfectly

## 34 Linear models are typically estimated given sample data

Usually, this means there are differences between the expected outcomes that the model predicts and the observed outcomes

* So we often write the linear model like this:

1. The observed outcome equals
2. the intercept
3. plus the difference associated with a specific predictor value
4. plus some amount of mismatch or error , the difference between the observed outcome and the predicted outcome

## 35 We can derive the formulas used to calculate the estimates using calculus

* But we won’t
* Because the linear model calculations are done using matrix solution algorithms in R *so we don’t have to*

## 36 What do the prediction errors – the residuals – look like?

|  |
| --- |
| Figure 10: 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 |

## 37 What the prediction errors look like

* The model expectation: higher vocabulary predicts higher mean comprehension accuracy
* The predicted points are shown by the blue line
* The prediction line increases in height for higher values of vocabulary
* Look at the **differences in height** between the observed points (in orange-red) and predicted points

|  |
| --- |
| Figure 11: 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 |

## 38 What the prediction errors look like

* If the regression model were perfect then all the observed points would lie on the prediction line
* They do not

|  |
| --- |
| Figure 12: 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 |

## 39 What the prediction errors look like

* Differences between observed and predicted outcomes are shown by the vertical lines
* Better models should show smaller differences between observed and predicted outcome values
* Notice: some participants had same vocabulary scores but different outcomes

|  |
| --- |
| Figure 13: 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 |

## 40 We typically assume that the residuals are normally distributed

* Some are positive: observed outcome larger than predicted outcome
* Some are negative: observed outcome smaller than predicted outcome
* The average of the residuals will be zero overall

|  |
| --- |
| Figure 14: Plot showing the distribution of prediction errors – residuals – for the linear model of comprehension accuracy |

## 41 So: We pick the line that minimizes the residuals – the mismatch between predicted and observed outcomes

|  |
| --- |
|  |

## 42 Let’s take a break

* End of part 3

## 43 Identifying the key information in the linear model results

* The summary() of the linear model shows …
* The Estimate of the Coefficient of the effect of individual differences in vocabulary (SHIPLEY)
* how much the outcome mean.acc value changes, given differences in SHIPLEY score
* Associated t value and Pr(> |t|) statistics for the coefficient t-test
* Model fit statistics: R-squared and F-statistic

Call:  
lm(formula = mean.acc ~ SHIPLEY, data = clearly.one.subjects)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.42871 -0.04921 0.02079 0.07480 0.18430   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.44914 0.08053 5.577 9.67e-08 \*\*\*  
SHIPLEY 0.01050 0.00229 4.585 8.85e-06 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.1115 on 167 degrees of freedom  
Multiple R-squared: 0.1118, Adjusted R-squared: 0.1065   
F-statistic: 21.03 on 1 and 167 DF, p-value: 8.846e-06

## 44 For the effect of vocabulary (SHIPLEY), we have:

* The coefficient for the slope of the effect of variation in vocabulary scores: 0.01050
* The Std. Error (standard error) 0.00229 for that estimate
* The tvalue 4.585 and associated Pr(>|t|) p-value 8.85e-06 for the null hypothesis test of the coefficient

Call:  
lm(formula = mean.acc ~ SHIPLEY, data = clearly.one.subjects)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.42871 -0.04921 0.02079 0.07480 0.18430   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.44914 0.08053 5.577 9.67e-08 \*\*\*  
SHIPLEY 0.01050 0.00229 4.585 8.85e-06 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.1115 on 167 degrees of freedom  
Multiple R-squared: 0.1118, Adjusted R-squared: 0.1065   
F-statistic: 21.03 on 1 and 167 DF, p-value: 8.846e-06

## 45 Identifying the key information in the results

* Pay attention to the **sign and the size** of the coefficient estimate:
* Is the coefficient (e.g., SHIPLEY 0.01050) a positive or a negative number?
* Is it relatively large or small?
* We come back to this, shortly, in the context of interpretation and reporting

Call:  
lm(formula = mean.acc ~ SHIPLEY, data = clearly.one.subjects)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.42871 -0.04921 0.02079 0.07480 0.18430   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.44914 0.08053 5.577 9.67e-08 \*\*\*  
SHIPLEY 0.01050 0.00229 4.585 8.85e-06 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.1115 on 167 degrees of freedom  
Multiple R-squared: 0.1118, Adjusted R-squared: 0.1065   
F-statistic: 21.03 on 1 and 167 DF, p-value: 8.846e-06

## 46 The t-tests in the linear model

* For each coefficient, the t-test is used to evaluate if the coefficient is significantly different from zero
* We assume the **null hypothesis** that the coefficient is zero
* We do the test by comparing the estimated coefficient with the standard error of the estimate

## 47 The t-tests in the linear model

* The standard error indicates our **uncertainty** about the estimate
* Larger standard errors represent greater uncertainty

## 48 The t-tests in the linear model

* Standard errors can be calculated using information about:
* Error in the model — think of the distribution of residuals
* Variation of values in the predictor — how widely they range
* The sample size
* Standard errors will be smaller for the coefficients of effects that appear to have bigger impacts, in models that describe outcomes better, in larger samples

## 49 Identifying the key information in the results

* Pay attention to R-squared:
* The model summary gives us the Multiple R-squared and Adjusted R-squared
* These numbers represent how much of the variation in the outcome can be predicted by the model
* We usually report Adjusted R-squared because it tends to be more accurate

Call:  
lm(formula = mean.acc ~ SHIPLEY, data = clearly.one.subjects)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.42871 -0.04921 0.02079 0.07480 0.18430   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.44914 0.08053 5.577 9.67e-08 \*\*\*  
SHIPLEY 0.01050 0.00229 4.585 8.85e-06 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.1115 on 167 degrees of freedom  
Multiple R-squared: 0.1118, Adjusted R-squared: 0.1065   
F-statistic: 21.03 on 1 and 167 DF, p-value: 8.846e-06

## 50 R-squared – what is it? – an indicator of the proportion of outcome variation we can predict

* Better models should show smaller differences between observed and predicted outcomes
* R-squared () gives the proportion of outcome variance
* we can predict given information about differences in vocabulary

|  |
| --- |
| Figure 15: The difference between predicted and observed outcomes, given variation in vocabulary |

## 51 R-squared as an indicator of the proportion of the outcome variation we can predict

* To understand what this means, look at the scatterplot
* On average, values in outcome (accuracy) increase with increasing values in the predictor (vocabulary)
* But different people got different outcomes even with same vocabulary scores

|  |
| --- |
| Figure 16: The difference between predicted and observed outcomes, given variation in vocabulary |

## 52 R-squared as an indicator of the proportion of the outcome variation we can predict

* So: we have variation in the outcome that is related to variation in the predictor
* And: we have variation in the outcome that seems unrelated to the predictor
* tells us how much variation in the outcome is explained by the model
* gives us a proportion where

|  |
| --- |
| Figure 17: The difference between predicted and observed outcomes, given variation in vocabulary |

## 53 Identifying the key information in the results

* Pay attention to F:
* The model summary gives us the F-statistic:
* This is the test statistic for the test of the null hypothesis that the model does not predict the outcome

Call:  
lm(formula = mean.acc ~ SHIPLEY, data = clearly.one.subjects)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.42871 -0.04921 0.02079 0.07480 0.18430   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.44914 0.08053 5.577 9.67e-08 \*\*\*  
SHIPLEY 0.01050 0.00229 4.585 8.85e-06 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.1115 on 167 degrees of freedom  
Multiple R-squared: 0.1118, Adjusted R-squared: 0.1065   
F-statistic: 21.03 on 1 and 167 DF, p-value: 8.846e-06

## 54 Reporting the results of a linear model

* You will need to report three bits of information:

1. how much outcome variation is explained by the model
2. test for the null hypothesis that none of the predictors actually predict the outcome
3. Coefficient estimates with the t-tests for the null hypothesis for each coefficient

Call:  
lm(formula = mean.acc ~ SHIPLEY, data = clearly.one.subjects)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-0.42871 -0.04921 0.02079 0.07480 0.18430   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 0.44914 0.08053 5.577 9.67e-08 \*\*\*  
SHIPLEY 0.01050 0.00229 4.585 8.85e-06 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.1115 on 167 degrees of freedom  
Multiple R-squared: 0.1118, Adjusted R-squared: 0.1065   
F-statistic: 21.03 on 1 and 167 DF, p-value: 8.846e-06

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

Here is an example of results reporting text that is conventional:

We fitted a linear model with mean comprehension accuracy as the outcome and vocabulary (Shipley) as the predictor. Our analysis indicated a significant effect of vocabulary knowledge. The model is significant overall, with , and explains 11% of variance (). The model estimates showed that the accuracy of comprehension increased with increasing levels of participant vocabulary knowledge ().

## 56 Look at what we do with the text

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?)

We fitted a linear model with mean comprehension accuracy as the outcome and vocabulary (Shipley) as the predictor. Our analysis indicated a significant effect of vocabulary knowledge. The model is significant overall, with , and explains 11% of variance (). The model estimates showed that the accuracy of comprehension increased with increasing levels of participant vocabulary knowledge ().

## 57 Summary

* In psychological science, we often ask questions like:

1. Does variation in X predict variation in Y?
2. What are the factors that influence outcome Y?
3. Is a theoretical model consistent with observed behaviour?

* We can answer these questions using the linear model
* Given sample data, we can predict the average difference in outcome values, for different levels of a predictor variable
* We (or the math engine R uses) calculate the predictions so that they minimize the residuals, the errors of prediction or the mismatch between predicted and observed outcomes
* Our results report tells the reader about the model and the estimated effects

## 58 End of week 9

## 59 References

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>