categorical predictors, part 1: dichotomous predictors

January 29, 2023

1

things to know

- Problem Set 1 is due right about now
- The answer key should be done some time tomorrow, and (I hope) grading will be done by class time on Wednesday
- There will be drill this week, same place/time as "usual"

2

today's punchlines

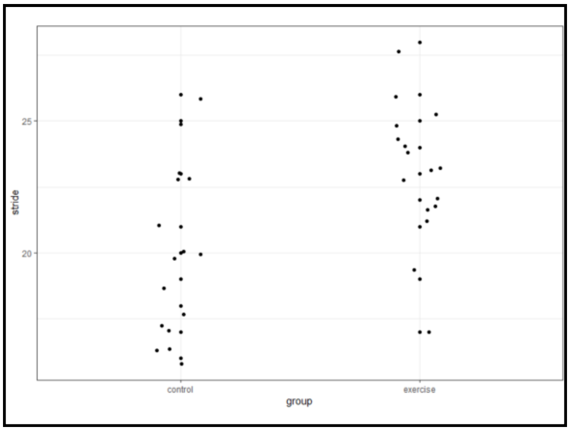
- not all designs will assess the relationship between *quantitative* predictors and outcomes
- categorical predictors can be easily handled using model comparison
- in a two-group design, *any* numbers can be assigned to each group
- we'll ultimately settle on (usually) using +0.5 and -0.5 as the numbers to indicate groups
- despite that it will look different, this *is* an independent-samples *t*-test

3

some hypothetical data (based on real research)

- music-based training might help elderly people improve balance, walking efficiency, and reduce the risk of falls
- a group of 32 senior citizens are randomly assigned to either (n = 16) walk in time to music (responding changing rhythms) for 6 months (once weekly) or to a delayed-intervention control group (n = 16)
- the data show that stride-length (the outcome) in the exercise group was M = 23" and in the control group was M = 20"
- note the group means and how far apart they are!

4



5

how to analyze?

- assign numbers to groups
- use the numbers as a predictor just like any other predictor
- what numbers?
- if all you care about is significance, it doesn't matter (but please don't be that person)
- let's start with exercise = 1 and control = 2 and fit the model

$$\text{stride} = b_0 + b_1 * X_{\text{group1and2}}$$

6

the results

- b_0 = 26
- b_1 = -3
- expressed as an equation

$$\widehat{stride} = 26 + (-3)X_{group1and2}$$

7

what are predicted stride values?

- plug in 1 for the exercise group

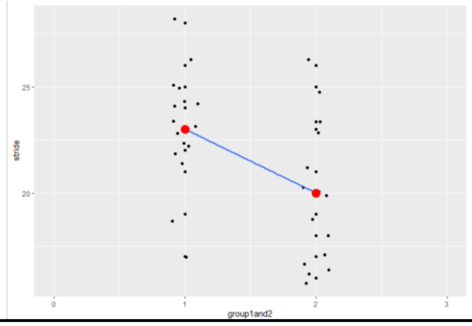
$$\widehat{stride} = 26 + (-3)(1) = 23$$

- plug in 2 for the control group

$$\widehat{stride} = 26 + (-3)(2) = 20$$

8

why are group means the predictions?



9

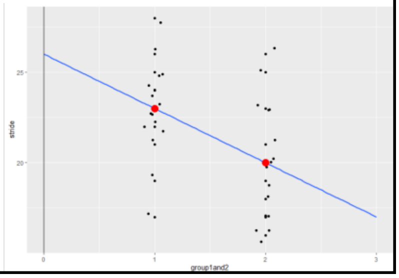
what is the slope?

- same as it ever was
- the predicted change as *X* increases by 1
- here, *increases by 1* means changing from *X* = 1 (the exercise group) to *X* = 2 (the control group)
- that is, if one is in the control group, their predicted stride length is -3 relative to the predicted stride length in the exercise group
- alternatively, we predict 23” for the exercise group and 20” for the control group

10

what is the intercept?

- it’s meaningless (in this case) because no group has a value of 0



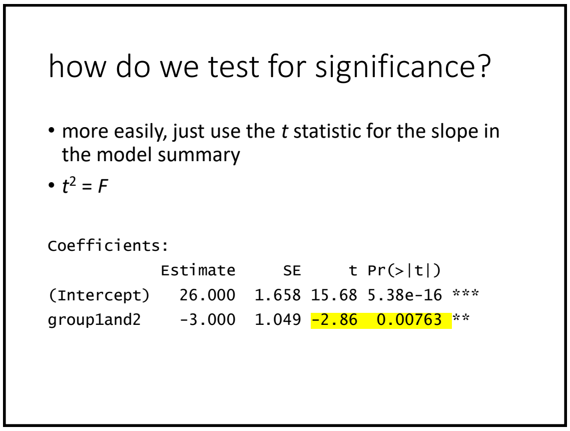
11

how do we test for significance?

- the usual: compare Model A to Model C
- Model A: stride = $b_0 + b_1X$
- Model C: stride = b_0
- H_0: $\beta_1 = 0$
- Model A SSE = 264
- Model C SSE = 336
- PRE = (336 – 264) / 336 = .214

$$F = \frac{.214/1}{(1 - .214)/(32 - 2)} = 8.18$$

12

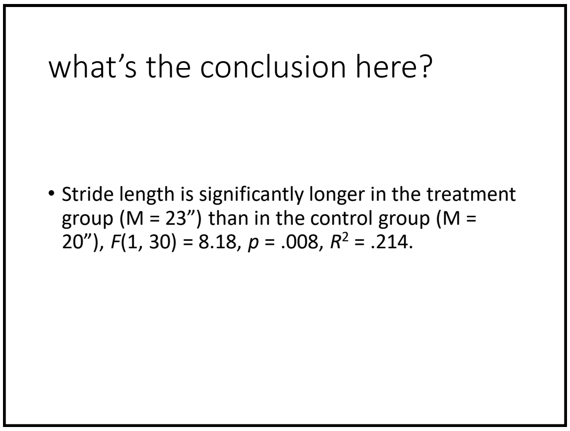


how do we test for significance?

- more easily, just use the *t* statistic for the slope in the model summary
- $t^2 = F$

```
Coefficients:
            Estimate     SE      t Pr(>|t|)
(Intercept)   26.000  1.658  15.68 5.38e-16 ***
group1and2    -3.000  1.049  -2.86  0.00763 **
```

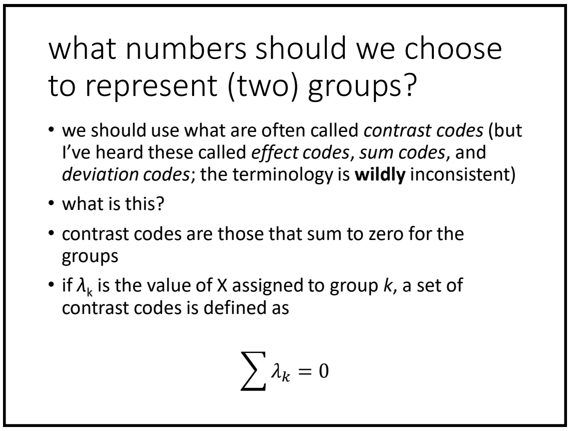
13



what’s the conclusion here?

- Stride length is significantly longer in the treatment group (M = 23”) than in the control group (M = 20”), *F*(1, 30) = 8.18, *p* = .008, R^2 = .214.

14



what numbers should we choose to represent (two) groups?

- we should use what are often called *contrast codes* (but I’ve heard these called *effect codes*, *sum codes*, and *deviation codes*; the terminology is **wildly** inconsistent)
- what is this?
- contrast codes are those that sum to zero for the groups
- if λ_k is the value of X assigned to group *k*, a set of contrast codes is defined as

$$\sum \lambda_k = 0$$

15

let’s try a -1 and +1

- if you dislike negative numbers, consider assigning +1 to the group with the higher mean
- what do we get in a model with ±1?

```
Coefficients:
              Estimate      SE      t Pr(>|t|)
(Intercept)    21.5000  0.5244 41.00  < 2e-16 ***
contrastCodes   1.5000  0.5244  2.86  0.00763 **
```

- *t* and *p* are the same!
- the slope is half the difference between group means
- the intercept is the overall mean (the *grand mean*)

16

what are predicted stride values?

- plug in +1 for the exercise group

$$\widehat{stride} = 21.5 + (1.5)(1) = 23$$

- plug in -1 for the control group

$$\widehat{stride} = 21.5 + (1.5)(-1) = 20$$

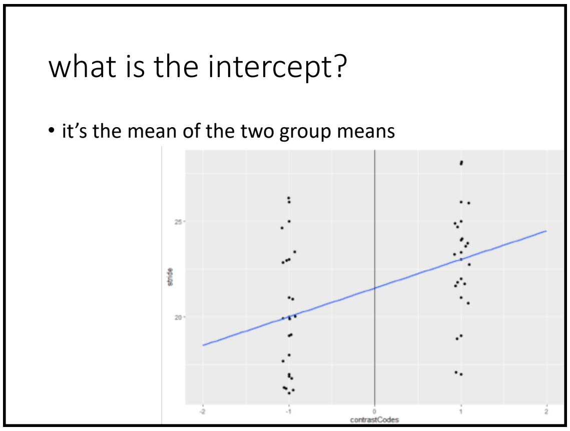
17

what is the slope?

- same as it ever was
- the predicted change as X increase by 1
- in this case, *increases by 1* means changing from X = -1 (the control group) to X = 0 (halfway to the exercise group); this is why the slope is half of the difference between the group means
- in general (with a contrast-coded predictor)

$$b_1 = {\sum \lambda_k \bar{Y}} \Big/ {\sum \lambda_k^2}$$

18



what is the intercept?

- it’s the mean of the two group means

19