Hypothesis testing with linear models

DClin Research Methods 1

Dr Christopher Wilson

Teesside University

Overview

- Hypothesis testing
- Linear models
- Linear models in R

Different types of research hypothesis

- Hypothesis testing is a method for making inferences about a population based on a sample
- In clinical psychology research, for example, we might be interested in the role of attentional bias in anxiety

Different types of research hypothesis

- However, we need to phrase this in terms of a hypothesis that we can test:
  - There is a difference in attentional bias between anxious and non-anxious individuals
  - There is a difference in attentional bias between anxious and non-anxious individuals, but only for threat-related stimuli
  - Level of anxiety predicts level of attentional bias to threat-related stimuli
  - Level of anxiety moderates the relationship between attentional bias and depression

Hypothesis influences research design

- The nature of these hypotheses will determine the design of the study, the variables that are measured, and the statistical analysis that is used
- In many cases people come to analyse their data and find it difficult to know which statistical test to use
- This is not necessarily because of a lack of statistical knowledge

Different analyses for different designs?

- Psych students are often taught to use different statistical tests for different types of designs. For example:
  - t-test for comparing two groups
  - ANOVA for comparing more than two groups
  - Correlation for testing the relationship between two continuous variables
  - Regression for testing the relationship between a continuous and a categorical variable
  - ANCOVA for testing the relationship between a continuous and a categorical variable, controlling for a third continuous variable

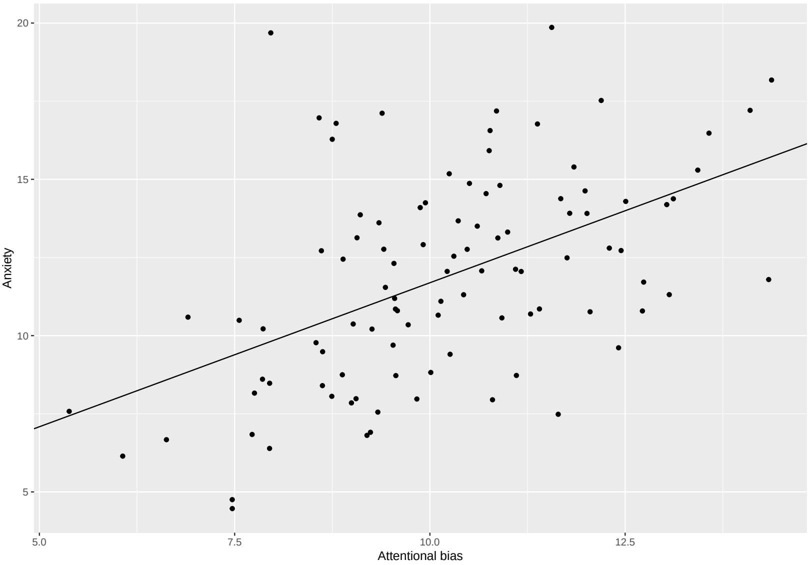
Using linear models to test hypotheses

- However, all of these designs can be analysed using linear models (i.e., regression models)
- This is because all of the above tests can be thought of as special cases of linear models

Teesside
University

Linear models using continuous predictor variables

Linear regression

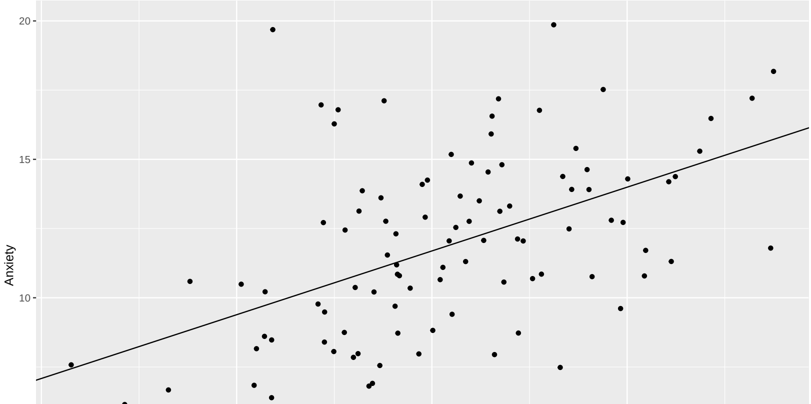


- The regression line is the line of best fit through the data

But what is the null hypothesis in regression?

The null hypothesis in regression

- In regression, the null hypothesis is that the line of best fit is no better at predicting the y variable than the mean of the y variable when the x variable = 0
- In other words, the null hypothesis is that the slope of the line of best fit is 0



Looking at regression output

```r
model1 <- lm(y ~ x, data = df) #<1>

summary(model1) #<2>
```

1. The lm() function is used for linear regression. The model is specified using the formula y ~ x, where y is the outcome variable and x is the predictor variable.
2. The summary() function is used to get a summary of the model. This includes the intercept and slope values, the standard errors, the t-values, and the p-values.

Looking at regression output

```
model1 <- lm(y ~ x, data = df) #<1>

summary(model1) #<2>
```

```
Call:
lm(formula = y ~ x, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7220 -2.0505 -0.2625  1.7419  9.8712

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.4787     1.6579   1.495    0.138
x             0.9213     0.1603   5.747 1.03e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

What if we have a categorical predictor variable?

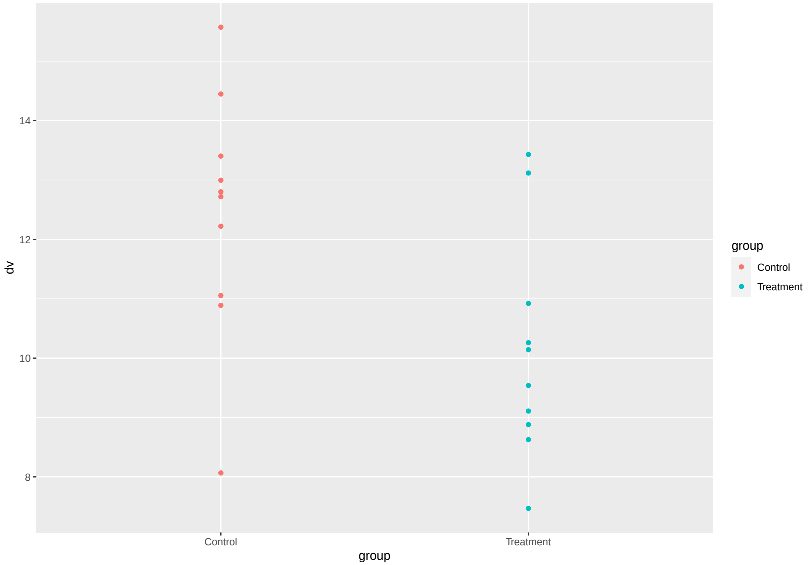
Example: t-test

- A t-test is a special case of a linear model where there is one predictor variable (group) and one outcome variable (DV)

```r
# run a t-test

t.test(dv ~ group, data = df)
```

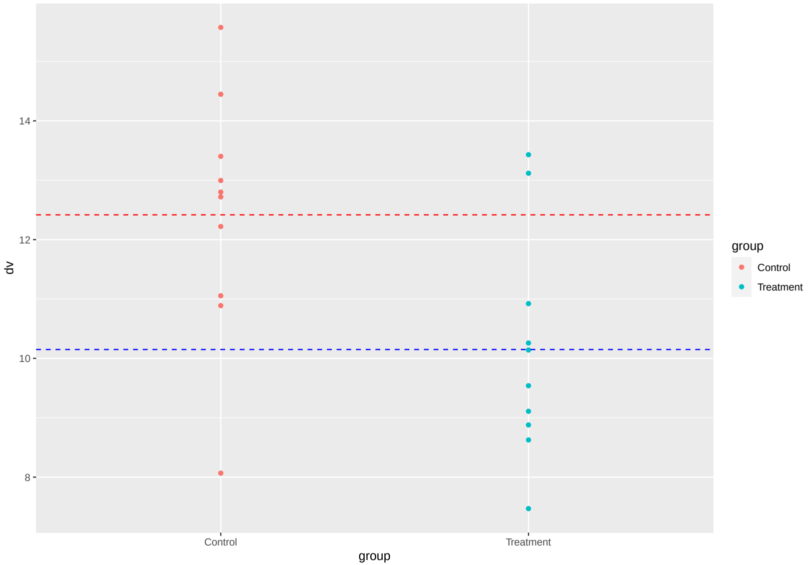
```

    Welch Two Sample t-test

data:  dv by group
t = 2.5438, df = 17.872, p-value = 0.02044
alternative hypothesis: true difference in means between gr
95 percent confidence interval:
 0.3938876 4.1420977
sample estimates:
  mean in group Control mean in group Treatment
               12.41724                10.14925
```

Plotting the data #1



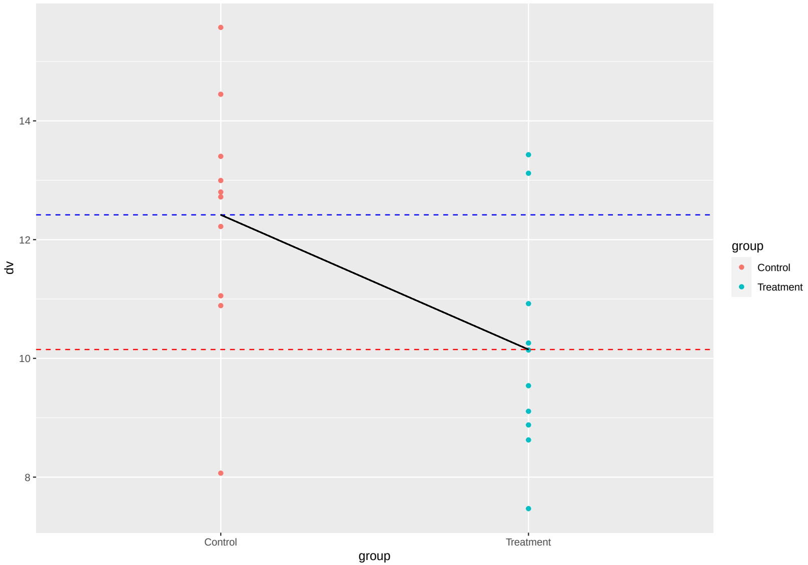
Here we can see the data for each group

Plotting the data #2



Here we can see the mean for each group, represented by the

Plotting the data #3



In a regression model, the intercept is the mean of one of the

Example: t-test as a linear model

```
# run a linear model

lm(dv ~ group, data = df) |> summary()
```

```
Call:
lm(formula = dv ~ group, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3505 -1.2934  0.0505  0.8258  3.2809

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     12.4172     0.6304  19.696 1.25e-13 ***
groupTreatment  -2.2680     0.8916  -2.544   0.0204 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

We can check the confidence intervals from the regression model

```
# show the confidence confidence intervals of the coeffici

lm(dv ~ group, data = df) |> confint()
```

```
                   2.5 %     97.5 %
(Intercept)    11.092729 13.7417583
groupTreatment -4.141139 -0.3948464
```

We can see that the confidence interval of the regression coefficient is the same as the confidence interval of the difference between means in the t-test

What's the value of this approach?

Advantages of using linear models

- Using linear models allows us to test a wide range of hypotheses using the same approach
- This means that we can use the same approach to test hypotheses about:
  - the relationship between two continuous variables
  - the relationship between a categorical predictor and continuous outcome
  - Continuous and categorical predictors in the same model
- We can use this approach regardless of the number of predictor variables or levels in a categorical predictor

ANOVA as regression

One-way ANOVA

```r
# running an ANOVA

aov(dv ~ group, data = df) |> summary()
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)   
group        2  48.97  24.486   6.435 0.00518 **
Residuals   27 102.73   3.805                   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

In this example, we can see that there is a significant effect of group on the outcome variable. However, we do not know which groups are significantly different from each other.

ANOVA as regression

```
lm(dv ~ group, data = df) |> summary()
```

```

Call:
lm(formula = dv ~ group, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3505 -1.2824 -0.1021  0.9322  3.3567

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.1493     0.6168  16.454 1.34e-15 ***
groupGroup 2   2.2680     0.8723   2.600   0.0149 *
groupGroup 3   3.0016     0.8723   3.441   0.0019 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Important points

Important points

- Not all relationships between variables are linear
- There are other approaches (e.g., logistic regression) for testing non-linear relationships
- You need to check the assumptions of linear models before reporting them

Teesside University

Summary

- Many different types of hypothesis can be tested using linear models
- This can allow us to ask questions that are more complex because we can include multiple predictor variables in the same model (next week)
- We can get more information from a regression output than from a t-test or ANOVA (for example)
- However, we need to check the assumptions of linear models before reporting them