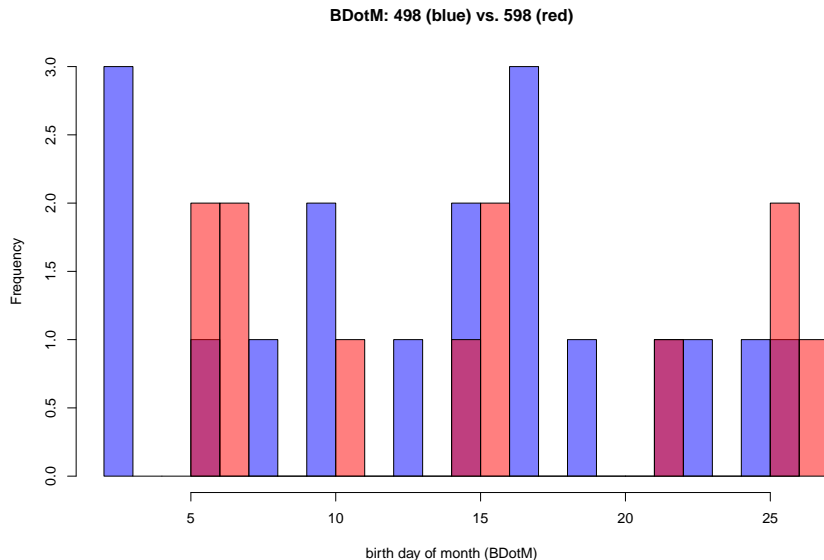


Linear Models: Main Effects

BIOE 498/598 PJ

Spring 2021

Does BDotM differ for the BIOE 498 and BIOE 598 students?



Shortcut method: the t -test

```
t.test(days498, days598, alternative="two.sided")
```

```
##
```

```
##  Welch Two Sample t-test
```

```
##
```

```
## data:  days498 and days598
```

```
## t = -0.4488, df = 22.194, p-value = 0.6579
```

```
## alternative hypothesis: true difference in means is not equal
```

```
## 95 percent confidence interval:
```

```
##  -7.647553  4.925331
```

```
## sample estimates:
```

```
## mean of x mean of y
```

```
##  13.88889  15.25000
```

New technique: Linear Models

Previous approach: Split data into two groups; test for a difference in BDotM

New approach: Build a model that predicts BDotM; ask if 498/598 knowledge helps

Linear Models

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

- ▶ y is the **response**
- ▶ x_i is a **predictor** or **factor**
- ▶ β_0 is the **intercept**
- ▶ $\beta_i, i > 0$ is a **coefficient, effect size, or parameter**
- ▶ ϵ is a **residual**

Re-organizing our data into a **data frame**

```
head(days_data)
```

```
##    day undergrad
## 1  11      FALSE
## 2  22      FALSE
## 3  17       TRUE
## 4  16      FALSE
## 5  27      FALSE
## 6  15       TRUE
```

Working with data frames

Checking the size (# rows and # columns)

```
dim(days_data)
```

```
## [1] 30  2
```

Working with data frames

Checking the size (# rows and # columns)

```
dim(days_data)
```

```
## [1] 30  2
```

Extracting a single entry

```
days_data[12, ]
```

```
##      day undergrad
```

```
## 12  26      FALSE
```


Working with data frames

Checking the size (# rows and # columns)

```
dim(days_data)
```

```
## [1] 30  2
```

Extracting a single entry

```
days_data[12, ]
```

```
##      day undergrad
```

```
## 12  26      FALSE
```

Or just the day

```
days_data[12, "day"]
```

```
## [1] 26
```

Our data frame contains two vectors

```
days_data$day
```

```
## [1] 11 22 17 16 27 15 19 8 10 5 15 26 25 5 7 3 10 2 16  
## [24] 23 7 26 13 15 6 22
```

```
days_data$undergrad
```

```
## [1] FALSE FALSE TRUE FALSE FALSE TRUE TRUE TRUE TRUE FA  
## [12] FALSE TRUE FALSE FALSE TRUE TRUE TRUE FALSE TRUE T  
## [23] FALSE TRUE FALSE TRUE TRUE FALSE TRUE TRUE
```

Building a linear model with `lm`: $y = \beta_0 + \epsilon$

```
lm( days_data$day ~ 1 )
```

```
##
```

```
## Call:
```

```
## lm(formula = days_data$day ~ 1)
```

```
##
```

```
## Coefficients:
```

```
## (Intercept)
```

```
##      14.43
```

Building a linear model with `lm`: $y = \beta_0 + \epsilon$

```
lm( days_data$day ~ 1 )
```

```
##  
## Call:  
## lm(formula = days_data$day ~ 1)  
##  
## Coefficients:  
## (Intercept)  
##      14.43
```

```
mean(days_data$day)
```

```
## [1] 14.43333
```

Modeling with a predictor: $y = \beta_0 + \beta_1 x + \epsilon$

```
lm( days_data$day ~ 1 + days_data$undergrad )
```

```
##
```

```
## Call:
```

```
## lm(formula = days_data$day ~ 1 + days_data$undergrad)
```

```
##
```

```
## Coefficients:
```

```
##           (Intercept)  days_data$undergradTRUE
```

```
##           15.250                -1.361
```

Modeling with a predictor: $y = \beta_0 + \beta_1 x + \epsilon$

```
lm( days_data$day ~ 1 + days_data$undergrad )
```

```
##
```

```
## Call:
```

```
## lm(formula = days_data$day ~ 1 + days_data$undergrad)
```

```
##
```

```
## Coefficients:
```

```
##           (Intercept)  days_data$undergradTRUE
```

```
##           15.250                -1.361
```

```
mean(days498) - mean(days598)
```

```
## [1] -1.361111
```

What does this model mean?

We fit the parameters β_0 and β_1 in the model

$$\text{day} = \beta_0 + \beta_1 \times \text{undergrad} + \epsilon$$

What does this model mean?

We fit the parameters β_0 and β_1 in the model

$$\text{day} = \beta_0 + \beta_1 \times \text{undergrad} + \epsilon$$

Substituting the fitted values:

$$\text{day} = 15.3 - 1.4 \times \text{undergrad} + \epsilon$$

What does this model mean?

We fit the parameters β_0 and β_1 in the model

$$\text{day} = \beta_0 + \beta_1 \times \text{undergrad} + \epsilon$$

Substituting the fitted values:

$$\text{day} = 15.3 - 1.4 \times \text{undergrad} + \epsilon$$

For undergrads: $\text{day} = 15.3 - 1.4 \times 1 = 13.9$

What does this model mean?

We fit the parameters β_0 and β_1 in the model

$$\text{day} = \beta_0 + \beta_1 \times \text{undergrad} + \epsilon$$

Substituting the fitted values:

$$\text{day} = 15.3 - 1.4 \times \text{undergrad} + \epsilon$$

For undergrads: $\text{day} = 15.3 - 1.4 \times 1 = 13.9$

For grad students: $\text{day} = 15.3 - 1.4 \times 0 = 15.3$

What does this model mean?

We fit the parameters β_0 and β_1 in the model

$$\text{day} = \beta_0 + \beta_1 \times \text{undergrad} + \epsilon$$

Substituting the fitted values:

$$\text{day} = 15.3 - 1.4 \times \text{undergrad} + \epsilon$$

For undergrads: $\text{day} = 15.3 - 1.4 \times 1 = 13.9$

For grad students: $\text{day} = 15.3 - 1.4 \times 0 = 15.3$

These are the means for each group.

Let's clean up our calls to `lm`: Intercepts

An intercept is always assumed.

```
lm( y ~ 1 + x )
```

is equivalent to

```
lm( y ~ x )
```

If you don't want an intercept, use a 0

```
lm( y ~ 0 + x )
```

Let's clean up our calls to `lm`: Naming a data frame

We can give `lm` a data frame where it can find our response and predictor variables.

```
lm( days_data$day ~ days_data$undergrad )
```

is equivalent to

```
lm( day ~ undergrad, data=days_data )
```

Storing our model for further analysis

We can assign the output of a model to a variable.

```
model <- lm( day ~ undergrad, data=days_data )  
summary(model)
```

```
##  
## Call:  
## lm(formula = day ~ undergrad, data = days_data)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -11.889  -7.389   0.750   6.340  12.111   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)    15.250     2.306   6.614 3.56e-07 ***  
## undergradTRUE  -1.361     2.976  -0.457   0.651      
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 7.987 on 28 degrees of freedom  
## Multiple R-squared:  0.0074128    Adjusted R-squared:  0.0000000
```

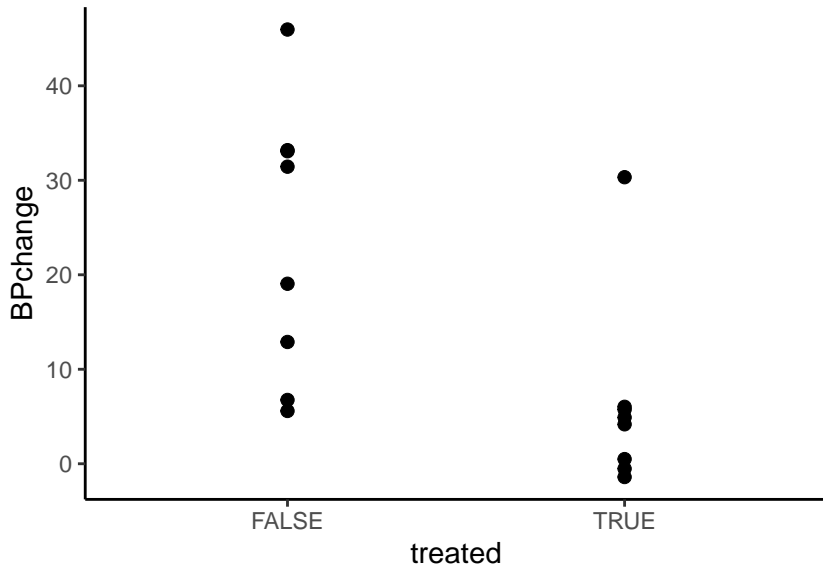
Testing a new blood pressure medication

```
head(bp_data)
```

```
## # A tibble: 6 x 3
##   BPchange treated male
##   <dbl> <lgl>   <lgl>
## 1  -0.525 TRUE    FALSE
## 2   4.17  TRUE    FALSE
## 3   6.03  TRUE     TRUE
## 4  -1.40  TRUE    FALSE
## 5   0.493 TRUE    FALSE
## 6  12.9   FALSE    TRUE
```

Does our BP treatment work?

```
qplot(data=bp_data, x=treated, y=BPchange, size=I(5))
```



Hypothesis testing the treatment effect

```
t.test(BPchange ~ treated, data=bp_data)
```

```
##
```

```
## Welch Two Sample t-test
```

```
##
```

```
## data: BPchange by treated
```

```
## t = 2.7499, df = 12.51, p-value = 0.01704
```

```
## alternative hypothesis: true difference in means is not equal
```

```
## 95 percent confidence interval:
```

```
##      3.649731 30.903566
```

```
## sample estimates:
```

```
## mean in group FALSE mean in group TRUE
```

```
##           23.492472           6.215823
```

A linear model with effect of treatment

```
model <- lm(BPchange ~ treated, bp_data)
summary(model)
```

```
##
## Call:
## lm(formula = BPchange ~ treated, data = bp_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.905   -6.960   -1.678    8.357   24.110
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    23.492     4.442   5.288 0.000115 ***
## treatedTRUE   -17.277     6.283  -2.750 0.015647 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.57 on 14 degrees of freedom
## Multiple R-squared:  0.3507, Adjusted R-squared:  0.3043
## F-statistic: 7.562 on 1 and 14 DF, p-value: 0.01565
```

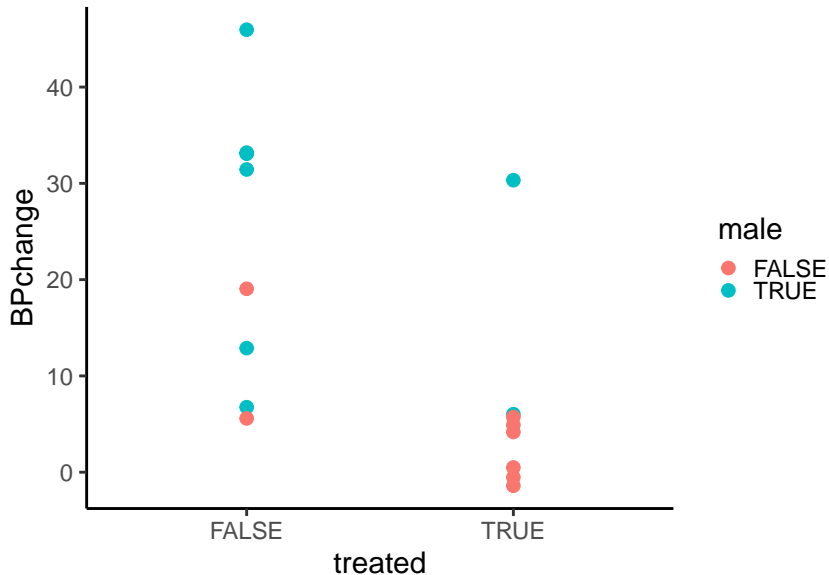
What else could explain the effect?

```
summary(lm(BPchange ~ treated + male, bp_data))
```

```
##
## Call:
## lm(formula = BPchange ~ treated + male, data = bp_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.596  -4.407   2.178   5.756  18.607
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    11.926     5.998   1.988  0.0683 .
## treatedTRUE     -9.566     6.195  -1.544  0.1466
## maleTRUE        15.422     6.195   2.489  0.0271 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.73 on 13 degrees of freedom
## Multiple R-squared:  0.5603, Adjusted R-squared:  0.4927
## F-statistic: 8.283 on 2 and 13 DF, p-value: 0.004791
```

Does our BP treatment work?

```
qplot(data=bp_data, x=treated, y=BPchange, color=male, size=I(5))
```



Summary

- ▶ Linear models can be used for hypothesis testing.

Summary

- ▶ Linear models can be used for hypothesis testing.
- ▶ Multivariate linear models consider how **all** factors affect the response. This is a form of conditioning.

Summary

- ▶ Linear models can be used for hypothesis testing.
- ▶ Multivariate linear models consider how **all** factors affect the response. This is a form of conditioning.
- ▶ Next time: What if the factors interact?