Common Statistical Tests in R

t-tests, linear regression, and ANOVA

Stephanie Zimmer, RTI International

2021-05-26

Topics

- formula notation
- linear regression
- ANOVA
- t-tests
  - one-sample t-tests
  - two-sample t-tests
  - paired t-test

Formula notation

To do any modeling, need to understand how to specify formulas in R

Most basic formula:

```
Y ~ X
Y ~ X + Z
```

- left side of formula is response/dependent variable
- right side of formula is predictor/independent variable(s)

If these are linear models, this is written mathematically as:

$$Y_i = \beta_0 + \beta X_i + \epsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i$$

Sources: https://faculty.chicagobooth.edu/richard.hahn/teaching/formulanotation.pdf,

Formula notation - no intercept

$$Y_i = \beta_1 X_i + \epsilon_i$$

```
Y ~ X - 1
Y ~ X + 0
```

Formula notation - interactions

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i + \epsilon_i$$

```
Y ~ X + Z + X:Z  #X:Z makes interaction between X and Z
Y ~ X * Z  #X*Z includes the variables and the interaction between them
```

Formula notation - most common uses

Symbol	Example	Meaning
+	`+X`	include this variable
-	`-X`	delete this variable
:	`X:Z`	include the interaction between these variables
*	`X*Z`	include these variables and the interactions between them
^n	`(X+Z+Y)^3`	include these variables and all interactions up to n way
I	`I(X-Z)`	as-as: include a new variable which is the difference of these variables

Formula notation - knowledge check

I want to model the following:

$$mpg_i = \beta_0 + \beta_1 cyl_i + \beta_2 disp_i + \beta_3 hp_i + \beta_4 cyl_i disp_i + \beta_5 cyl_i hp_i + \beta_6 disp_i hp_i + \epsilon_i$$

How can you write this formula? Select all that apply:

1. `mpg~cyl:disp:hp`
2. `mpg~(cyl+disp+hp)^2`
3. `mpg~cyl+disp+hp+cyl:disp+cyl:hp+disp:hp`
4. `mpg~cyl*disp*hp`
5. `mpg~cyl*disp+cyl*hp+disp*hp`

Formula notation - knowledge check (solution)

I want to model the following:

$$mpg_i = \beta_0 + \beta_1 cyl_i + \beta_2 disp_i + \beta_3 hp_i + \beta_4 cyl_i disp_i + \beta_5 cyl_i hp_i + \beta_6 disp_i hp_i + \epsilon_i$$

How can you write this formula? Select all that apply:

1. `mpg~cyl:disp:hp` - no, this only has the interactions
2. `mpg~(cyl+disp+hp)^2` - yes
3. `mpg~cyl+disp+hp+cyl:disp+cyl:hp+disp:hp` - yes
4. `mpg~cyl*disp*hp` - no, this also has the 3-way interaction
5. `mpg~cyl*disp+cyl*hp+disp*hp` - yes

There may be other ways as well!!!

Data for exercises

- Using `palmerpenguins` data for examples
- Data were collected and made available by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, a member of the Long Term Ecological Research Network.
- Access data through `palmerpenguins` package https://github.com/allisonhorst/palmerpenguins/

```
library(palmerpenguins)
library(tidyverse)
glimpse(penguins)
```

```
## Rows: 344
## Columns: 8
## $ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel~
## $ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgerse~
## $ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, ~
## $ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, ~
## $ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186~
## $ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, ~
## $ sex               <fct> male, female, female, NA, female, male, female, male~
## $ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007~
```

Linear regression - simple linear regression

Model penguin body mass as function of flipper length

```
o <- lm(body_mass_g ~ flipper_length_mm, data = penguins)
o
```

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm, data = penguins)
##
## Coefficients:
##       (Intercept)  flipper_length_mm
##          -5780.83              49.69
```

Linear regression - simple linear regression

```
summary(o)
```

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm, data = penguins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1058.80  -259.27   -26.88   247.33  1288.69
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -5780.831    305.815  -18.90   <2e-16 ***
## flipper_length_mm    49.686      1.518   32.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 394.3 on 340 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.759,    Adjusted R-squared:  0.7583
## F-statistic:  1071 on 1 and 340 DF,  p-value: < 2.2e-16
```

Linear regression - multiple linear regression

Model penguin body mass as function of flipper length and bill length

```
lm(body_mass_g ~ flipper_length_mm + bill_length_mm, data = penguins) %>%
    summary()
```

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm + bill_length_mm,
##     data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1090.5  -285.7   -32.1   244.2  1287.5
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -5736.897    307.959 -18.629   <2e-16 ***
## flipper_length_mm    48.145      2.011  23.939   <2e-16 ***
## bill_length_mm        6.047      5.180   1.168    0.244
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Linear regression - multiple linear regression

How would I model penguin mody mass as a function of flipper length and bill length and also include an interaction?

```
lm(body_mass_g~######,
     data=penguins)
```

Linear regression - multiple linear regression

How would I model penguin mody mass as a function of flipper length and bill length and also include an interaction?

```
lm(body_mass_g~######,
     data=penguins)
```

```
lm(body_mass_g ~ flipper_length_mm * bill_length_mm, data = penguins) %>%
    summary()
```

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm * bill_length_mm,
##     data = penguins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1040.18  -283.07   -23.94   241.93  1241.40
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    5090.5088  2925.3007   1.740 0.082740 .
```

ANOVA - one-way using lm

Does average penguin mass vary by species?

```
lm(body_mass_g ~ species, data = penguins) %>%
    summary()
```

```
##
## Call:
## lm(formula = body_mass_g ~ species, data = penguins)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1126.02  -333.09   -33.09   316.91  1223.98
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3700.66      37.62   98.37   <2e-16 ***
## speciesChinstrap    32.43      67.51    0.48    0.631
## speciesGentoo     1375.35      56.15   24.50   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 462.3 on 339 degrees of freedom
```

ANOVA - one-way using aov

Does average penguin mass vary by species?

```
aov(body_mass_g ~ species, data = penguins)
```

```
## Call:
##    aov(formula = body_mass_g ~ species, data = penguins)
##
## Terms:
##                   species Residuals
## Sum of Squares  146864214  72443483
## Deg. of Freedom         2       339
##
## Residual standard error: 462.2744
## Estimated effects may be unbalanced
## 2 observations deleted due to missingness
```

```
coefficients(aov(body_mass_g ~ species, data = penguins))
```

```
##      (Intercept) speciesChinstrap    speciesGentoo
##       3700.66225         32.42598       1375.35401
```

ANOVA - two-way

Does average penguin mass vary by species and sex?

```
penguins %>%
    filter(!is.na(sex)) %>%
    ggplot(aes(x = species, y = body_mass_g)) + geom_boxplot() + facet_wrap(~sex)
```

ANOVA - two-way

Does average penguin mass vary by species and sex?

```
lm(body_mass_g ~ species * sex, data = penguins) %>%
    summary()
```

```
##
## Call:
## lm(formula = body_mass_g ~ species * sex, data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -827.21 -213.97   11.03  206.51  861.03
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               3368.84      36.21  93.030  < 2e-16 ***
## speciesChinstrap           158.37      64.24   2.465  0.01420 *
## speciesGentoo             1310.91      54.42  24.088  < 2e-16 ***
## sexmale                    674.66      51.21  13.174  < 2e-16 ***
## speciesChinstrap:sexmale  -262.89      90.85  -2.894  0.00406 **
## speciesGentoo:sexmale      130.44      76.44   1.706  0.08886 .
## ---
```

t-test: one-sample with lm

On average, are penguin body masses significantly different from half a kg (5000 g)?

```
lm(I(body_mass_g - 5000) ~ 1, data = penguins) %>%
    summary()
```

```
##
## Call:
## lm(formula = I(body_mass_g - 5000) ~ 1, data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1501.8  -651.8  -151.8   548.2  2098.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -798.25      43.36  -18.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 802 on 341 degrees of freedom
##   (2 observations deleted due to missingness)
```

t-test: one-sample with t.test (A)

On average, are penguin body masses significantly different from half a kg (5000 g)?

```
t.test(I(body_mass_g - 5000) ~ 1, data = penguins)
```

```
##
##     One Sample t-test
##
## data:  I(body_mass_g - 5000)
## t = -18.408, df = 341, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -883.5417 -712.9496
## sample estimates:
## mean of x
## -798.2456
```

t-test: one-sample with t.test (B)

On average, are penguin body masses significantly different from half a kg (5000 g)?

```
t.test(penguins$body_mass_g - 5000)
```

```
##
##     One Sample t-test
##
## data:  penguins$body_mass_g - 5000
## t = -18.408, df = 341, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -883.5417 -712.9496
## sample estimates:
## mean of x
## -798.2456
```

t-test: two-sample with lm

Is penguin weight significantly different between males and females?

```
lm(body_mass_g ~ sex, data = penguins) %>%
    summary()
```

```
##
## Call:
## lm(formula = body_mass_g ~ sex, data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1295.7  -595.7  -237.3   737.7  1754.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3862.27      56.83  67.963  < 2e-16 ***
## sexmale       683.41      80.01   8.542  4.9e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 730 on 331 degrees of freedom
##   (11 observations deleted due to missingness)
```

t-test: two-sample with t.test

Is penguin weight significantly different between males and females?

```
t.test(body_mass_g ~ sex, data = penguins, var.equal = TRUE)
```

```
##
##     Two Sample t-test
##
## data:  body_mass_g by sex
## t = -8.5417, df = 331, p-value = 4.897e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -840.8014 -526.0222
## sample estimates:
## mean in group female   mean in group male
##             3862.273             4545.685
```

paired t-test: introduce data

Is the cost of books on Amazon cheaper than the bookstore?

```
library(openintro)
glimpse(textbooks)
```

```
## Rows: 73
## Columns: 7
## $ dept_abbr <fct> Am Ind, Anthro, Anthro, Anthro, Art His, Art His, Asia Am, A~
## $ course    <fct>  C170, 9, 135T, 191HB, M102K, 118E, 187B, 191E, C125, M145B,~
## $ isbn      <fct> 978-0803272620, 978-0030119194, 978-0300080643, 978-02262068~
## $ ucla_new  <dbl> 27.67, 40.59, 31.68, 16.00, 18.95, 14.95, 24.70, 19.50, 123.~
## $ amaz_new  <dbl> 27.95, 31.14, 32.00, 11.52, 14.21, 10.17, 20.06, 16.66, 106.~
## $ more      <fct> Y, Y, Y, Y, Y, Y, Y, N, N, Y, Y, N, Y, Y, N, N, N, N, N, N, ~
## $ diff      <dbl> -0.28, 9.45, -0.32, 4.48, 4.74, 4.78, 4.64, 2.84, 17.59, 3.7~
```

```
textbooks %>%
    ggplot(aes(x = ucla_new, y = amaz_new)) + geom_point() + geom_abline(slope = 1,
    intercept = 0)
```

paired t-test: lm

Is the cost of books on Amazon different than the bookstore?

```
lm(ucla_new - amaz_new ~ 1, data = textbooks) %>%
    summary()
```

```
##
## Call:
## lm(formula = ucla_new - amaz_new ~ 1, data = textbooks)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.292  -8.962  -4.532   4.828  53.238
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   12.762      1.668   7.649 6.93e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.26 on 72 degrees of freedom
```

paired t-test: t.test

Is the cost of books on Amazon different than the bookstore?

```
t.test(textbooks$ucla_new, textbooks$amaz_new, paired = TRUE)
```

```
##
##     Paired t-test
##
## data:  textbooks$ucla_new and textbooks$amaz_new
## t = 7.6488, df = 72, p-value = 6.928e-11
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   9.435636 16.087652
## sample estimates:
## mean of the differences
##                12.76164
```

Linear models - good resource

Great resource: https://lindeloev.github.io/tests-as-linear/

Sources

- Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0. https://allisonhorst.github.io/palmerpenguins/
- https://faculty.chicagobooth.edu/richard.hahn/teaching/formulanotation.pdf
- https://stat.ethz.ch/R-manual/R-devel/library/stats/html/formula.html
- https://lindeloev.github.io/tests-as-linear/
- Mine Çetinkaya-Rundel, David Diez, Andrew Bray, Albert Y. Kim, Ben Baumer, Chester Ismay, Nick Paterno and Christopher Barr (2021). openintro: Data Sets and Supplemental Functions from 'OpenIntro' Textbooks and Labs. R package version 2.1.0. https://CRAN.R-project.org/package=openintro
- Diez, David M., Christopher D. Barr, and Mine Çetinkaya-Rundel. OpenIntro statistics, Fourth Edition. OpenIntro, 2019.

ANCOVA

After controlling for flipper length, is body mass the same for all species?

```
penguins %>%
    ggplot(aes(x = flipper_length_mm, y = body_mass_g, colour = species, group = species)) +
    geom_point()
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

ANCOVA

After controlling for flipper length, is body mass the same for all species?

```
ancout <- lm(body_mass_g ~ flipper_length_mm + species, data = penguins)
summary(ancout)
```

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm + species, data = penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -927.70 -254.82  -23.92  241.16 1191.68
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -4031.477    584.151  -6.901 2.55e-11 ***
## flipper_length_mm    40.705      3.071  13.255  < 2e-16 ***
## speciesChinstrap   -206.510     57.731  -3.577 0.000398 ***
## speciesGentoo       266.810     95.264   2.801 0.005392 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Plot ANCOVA

After controlling for flipper length, is body mass the same for all species?

```
peng2 <- penguins %>%
    filter(!is.na(flipper_length_mm), !is.na(body_mass_g), !is.na(species)) %>%
    mutate(yhat = predict(ancout))
peng2 %>%
    ggplot(aes(x = flipper_length_mm, y = body_mass_g, colour = species, group = species)) +
    geom_point() + geom_line(aes(y = yhat), size = 1)
```