Chapter 6: Comparing more than two means

Learning objectives of this chapter:

- Implementing a t-test as a regression in R
- Performing an ANOVA in R
- Implementing an ANOVA as a regression in R
- Introducing a covariate in ANCOVA in R

Assignment 6.1: Implementing a t-test as a regression in R

Many of the statistical tests that we have seen are actually equivalent to a specific form of **linear regression** . To understand how a **t-test** can be implemented as a **linear regression** , let's look at an example. Suppose you work in advertising and show four groups of people an advertisement, where the only difference is in the color/position of the eyes of the model (blue eyes, brown eyes, green eyes, or downward-looking eyes), and you ask them how they rate your brand after seeing this advertisement. For this assignment, we will use the `eyeColor.csv` data file that contains 222 participants' ratings for one of the four groups. As a first question, you want to assess whether there is a difference in the ratings if the model shown in the advertisement has blue eyes rather than brown eyes.

6.1 a) Read in the file `eyeColor.csv` and store the data in an object called `dataset8` .

R code 6.1a:

Note that this data set contains all four groups (`Blue` , `Brown` , `Green` , `Down`). Run the following R code to isolate the scores of the groups that were shown advertisements with `Blue` and `Brown` eyes and store them in the new `ttestData` object.

```
ttestData <- subset(dataset8,
                    dataset8$Group == 'Blue'|
                    dataset8$Group == 'Brown')
```

6.1 b) Write down the **null hypothesis** H_0 and **alternative hypothesis** H_1 for testing whether the **mean** score of the group that was shown `Blue` eyes is equal to the **mean** score of the group that was shown `Brown` eyes. Remember that these are independent **samples** .

Answer 6.1b:

H_0 : ________________ H_1 : ________________

6.1 c) Use the `t.test()` function to test the equality of the two means.

Hint 6.1: Make sure to specify `var.equal = TRUE` *to perform a* **two-sample t-test** *instead of the non-parametric* **Welch's t-test** *.*

R code 6.1c:

6.1 d) What is your conclusion on the basis of these results? Include the following elements:

- ☐ Discuss what the **p-value** is for this test.
- ☐ Discuss whether H_0 is rejected or not.
- ☐ Describe what this tells us about μ_{Blue} and μ_{Brown}.
- ☐ Describe what type of error is relevant *(type-I or type-II)*.

Answer 6.1d:

Now, instead of using the `t.test()` function to test the equality of the two **means** , you can test the equality of the two **means** using a **regression model** . Therefore, you need to add a variable to our data that says whether a participant saw one of the two groups (e.g., only the people that were shown `Brown` eyed models).

Run the following code in R:

```
dummyBrown <- as.numeric(ttestData$Group == 'Brown')
ttestData <- cbind(ttestData, dummyBrown)
```

6.1 e) What are the contents of `dummyBrown` ? What do we call this kind of variable?

Answer 6.1e:

6.1 f) Create a **linear model** in R where you predict the (outcome) variable `Score` using only the (predictor) variable `dummyBrown` . Store the fitted model in an object called `ttestreg` .

R code 6.1f:

6.1 g) Use the `summary()` function to inspect the results of the `ttestreg` model. How does this output correspond to the **t-test** that you performed in assignment 6.1c? Where do you find the **p-value** that you calculated using the `t.test()` function?

R code 6.1g:

Answer 6.1g:

Assignment 6.2: Performing an ANOVA in R

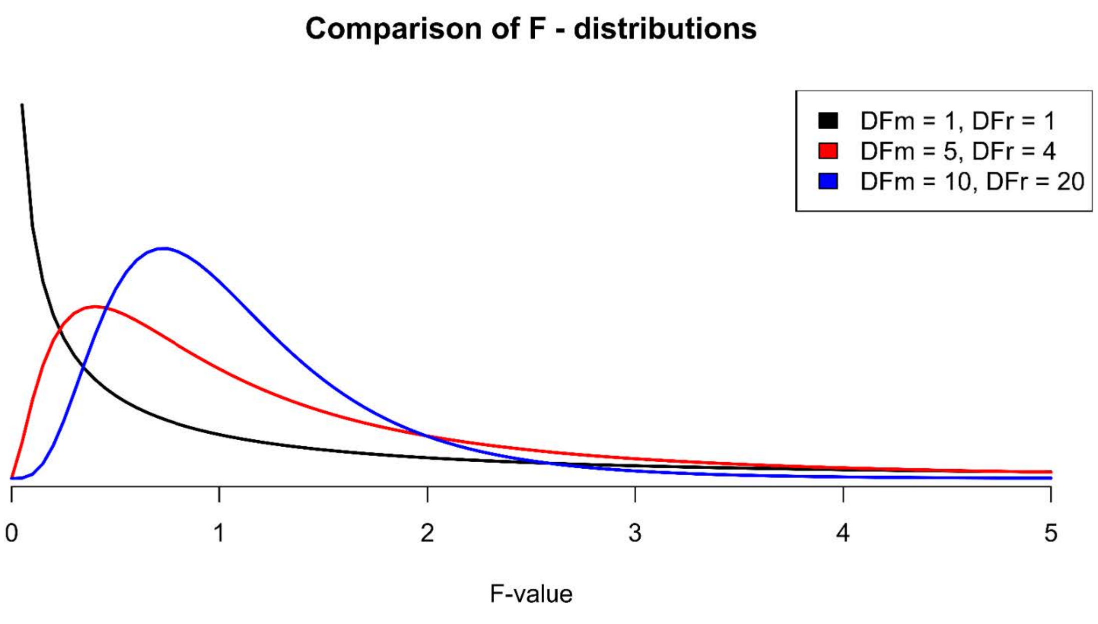
Now let's extend the analysis from assignment 6.1 by comparing all four groups (`Blue` , `Brown` , `Green` , `Down`) instead of only the `Blue` and `Brown` groups. When you are testing more than two **means** , you can use an **ANOVA** test (a specific form of regression). Since you want to compare all four groups, you can leave the `ttestData` from the previous assignment and focus on the data in `dataset8` . Remember that you are interested in testing the effect of the model's eye color on the rating of your brand.

6.2 a) Write down the **null hypothesis** H_0 and the **alternative hypothesis** H_1 for testing whether the **mea** n score of the four groups (`Blue` , `Brown` , `Green` , `Down`) are equal.

Answer 6.2a:

H_0 : ________________ H_1 : ________________

The **ANOVA** uses the **F-distribution** to test for a significant difference between all four **means** . Using the (two types of) **degrees of freedom** of this **F-distribution** , you can calculate the **critical F-value** that is required to reject the **null hypothesis** that the **means** of the four groups are equal. Table 5 on page 99 contains the **critical F-values** for a confidence of 95%.



6.2 b) Calculate the **degrees of freedom** df_M and df_R of the **F-distribution** for the data in `dataset8` .

Hint 6.2: You can find the formulas for df_M and df_R in the formula sheet on page 93.

Answer 6.2b:

df_M : ______________ df_R : ______________

6.2 c) Using the `qf()` function, calculate the **critical F-value** that is required to reject the **null hypothesis** H_0 for these data with 95% confidence.

R code 6.2c:

Answer 6.2c:

`Critical F-value:` ______________

The `aov()` function in R is a wrapper for the `lm()` function. The difference between these two functions is that the `lm()` function can only handle **categorical** predictors with two levels (e.g., a **dummy variable**). The `aov()` function can handle **categorical** predictor variables with more than two levels, since it automatically rewrites the `formula` to include the **dummy variables** .

6.2 d) Use the `aov()` function to perform an **ANOVA** with the dependent (outcome) variable `Score` and the independent (predictor) variable `Group` and store the result in an object named `anovaResult` .

Hint 6.3: You can check more information on the `aov()` function with `?aov` .

R code 6.2d:

6.2 e) Use the `summary()` function to inspect the results of the **ANOVA** in `anovaResult` . What is the **F-value** that is calculated from the **sample** ? What is the **p-value** calculated from the **sample** ?

R code 6.2e:

Answer 6.2e:

F-value: ______________ p-value: ______________

6.2 f) What is your conclusion on the basis of these results? Include the following elements:

- ☐ Discuss what the **p-value** is for this test.
- ☐ Discuss whether H_0 is rejected or not.
- ☐ Describe what this tells us about μ_{Blue}, μ_{Brown}, μ_{Green}, and μ_{Down}.
- ☐ Describe what type of error is relevant *(type-I or type-II)*.

Answer 6.2f:

Assignment 6.3: Implementing an ANOVA as a regression in R

Now that you have seen the results of the **ANOVA**, let's try to replicate these by implementing the same **ANOVA** as a **linear regression**. Remember that this is exactly what you did for the **t-test** in assignment 6.1 by adding one **dummy variable** to your model that isolated the `Brown` group. For the **ANOVA**, you are going to have three **dummy variables** in your model, one that represents `Brown` eyes, one that represents `Blue` eyes, and one that represents `Green` eyes. You first have to add these dummy variables to your data set.

Run the following code in R that adds a **dummy variable** for `Brown` eyes to the data set:

```
dummyBrown <- as.numeric(dataset8$Group == 'Brown')
dataset8 <- cbind(dataset8, dummyBrown)
```

6.3 a) Add two more **dummy variables** to the data in `dataset8`, one for `Blue` eyes and one for `Green` eyes. Name these variables `dummyBlue` and `dummyGreen`.

R code 6.3a:

6.3 b) Create a **linear model** in R where you predict the (outcome) variable `Score` using the (predictor) variables `dummyBrown`, `dummyGreen`, and `dummyBlue`. Store the fitted **linear model** in an object called `anovaReg`.

R code 6.3b:

6.3 c) Use the `summary()` function to inspect the results of the **linear model** stored in `anovaReg`. What is the **F-value** of the model? What is the **p-value** of this model?

R code 6.3c:

Answer 6.3c:

F-value: ______________ p-value: ______________

6.3 d) Do the **`F-value`** and **`p-value`** of this **`linear model`** match those of the **`ANOVA`** in assignment 6.2?

Answer 6.3d:

YES / NO

Assignment 6.4: Introducing a covariate in ANCOVA in R

There might be other determinants that influence people's ratings of your brand that you have not captured by varying the eye color in the advertisements. An example of this might be people's initial rating of your brand. These kinds of variables are called **`covariates`** and you can incorporate them in our **ANOVA** , very smoothly resulting in an **ANCOVA** . In our scenario, we want to incorporate the **`covariate`** for the initial score that our raters gave by adding the `initialScore` variable to our **`linear model`** .

6.4 a) Create a **`linear model`** in R where you predict the (outcome) variable `Score` using the (predictor) variables `dummyBrown` , `dummyGreen` , `dummyBlue` , and the variable `initialScore` . Store the fitted model in an object called `ancovaReg` .

R code 6.4a:

6.4 b) Use the `summary()` function to inspect the results of the **`linear model`** stored in `ancovaReg` . What is the **`F-value`** of the model? What is the **`p-value`** of this model?

R code 6.4b:

Answer 6.4b:

`F-value:` ________________ `p-value:` ________________

6.4 c) What is your conclusion on the basis of these results? Include the following elements:

- ☐ Discuss what the **`p-value`** is for this test.
- ☐ Discuss whether H_0 is rejected or not.
- ☐ Describe what this tells us about μ_{Blue}, μ_{Brown}, μ_{Green}, and μ_{Down}, given the covariate.
- ☐ Describe what type of error is relevant *(type-I or type-II)*.

Answer 6.4c:

____

____

____

6.4 d) Can you tell whether `initialScore` is a good predictor of the `Score` ? On what value can you base your conclusion?

Hint 6.4: First consider which results you would expect if $\beta_3 \neq 0$.

Answer 6.4d:

____

____

To find out whether adding this **`covariate`** is an improvement over the **`linear model`** in assignment 6.3, we can compare the two **`linear models`** `anovaReg` (without `initialScore`) and `ancovaReg` (with `initialScore`) with respect to their proportion of **`explained variance`** (their R^2).

6.4 e) What is the (multiple) R^2 of the `anovaReg` **`model`** ? What is the (multiple) R^2 of the `ancovaReg` **`model`** ? Which **`model`** explains more variation in the outcome variable `Score` ?

Answer 6.4e:

R^2 `anovaReg` : ____________ R^2 `ancovaReg` : ____________

`The` `anovaReg` `/` `ancovaReg` `regression model explains more variation in the outcome variable score.`

6.4 f) Interpret the R^2 for the best model.

Answer 6.4f:

____

____

The R^2 statistic will always increase when you add more (predictor) variables to our **`model`** , since you are adding more information. To reliably compare our two **`models`** , you have to look at a measure that penalizes a **`model`** for including more (predictor) variables. You can use the **`AIC`** value for that. The rule of thumb for the **`AIC`** value is that the **`model`** with the lower **`AIC`** value is the preferred model.

6.4 g) Use the `AIC()` function to calculate the **AIC** value of the **anovaReg** and the **ancovaReg** models.

R code 6.4g:

Answer 6.4g:

AIC **anovaReg** : ________________ AIC **ancovaReg** : ________________

6.4 h) What is the preferred model? How can you use the **AIC** statistic to validate you answer in assignment 6.4d?

Answer 6.4h:

____

____

Assignment 6.5: Using post-hoc tests in R to find differences in means

The state of Iowa in the USA receives many invoices for services that they buy. In turn, these invoices need to be paid in a timely manner. The questions has been raised whether the state of Iowa pays all invoices equally timely. To investigate this you are requested to perform an audit. In this audit you set out to statistically check if you can find differences -between the various services that are bought- in the time that it takes for Iowa to pay an invoice.

For this assignment you will have to download the data file `iowa.RData` from the online resources[2]. `.RData` files are compressed R objects, and are useful when dealing with very large data sets such as this one.

The `iowa.RData` file contains payment transactions recorded in the State of Iowa's central accounting system for the Executive Branch and is real data.

6.5 a) Load the `iowa.RData` data file into the environment using the `load()` function.

R code 6.5a:

The data set is now stored in object called `iowa` .

6.5 b) Give a short description of the data in `iowa` .

Hint 6.5: Search the internet for the source of these data to find out what the columns represent.

Answer 6.5b:

____

____

____

6.5 c) How many rows and columns does the `iowa` data set have?

Answer 6.5c:

Rows: ____________ Columns: ____________

[2] These data are taken from https://data.iowa.gov/State-Government-Finance/State-of-Iowa-Checkbook/cyqb-8ina.

6.5 d) How many unique services are there? Make a **`frequency table`** of these services.

R code 6.5d:

Answer 6.5d:

```
Unique services:     ______________
```

6.5 e) Which service has the most rows? How many rows does this service have?

Answer 6.5e:

```
Service:     ______________          Rows:     ______________
```

6.5 f) How many rows show a difference in invoice date and payment date?

R code 6.5f:

Answer 6.5f:

```
Number of rows that show a difference:     ______________
```

6.5 g) Create a new data set that consists of these differences, and name the new data set **`dataDif`** .

R code 6.5g:

6.5 h) Create an extra column named `dif.days` in `dataDif` that contains the number of days between invoice and payment.

Hint 6.6: Make sure that the column `dif.days` *is numeric.*

R code 6.5h:

6.5 i) Calculate the **minimum**, **maximum**, **mean**, **quartiles**, and **standard deviation** of the column `dif.days`.

R code 6.5i:

Answer 6.5i:

Minimum:	________	Upper quartile:	________
Mean:	________	Lower quartile:	________
Maximum:	________	Standard deviation:	________

6.5 j) Create a histogram of the column `dif.days`. Describe what you see in the histogram.

R code 6.5j:

Answer 6.5j:

____

6.5 k) Again, create a histogram, but now only use the subset of `dif.days` that is in the 5-95% quantile range (so you cut off the bottom and top 5%).

Hint 6.7: Hint: use the `quantile()` *function.*

R code 6.5k:

You don't trust the negative values in `dif.days` as you cannot interpret them, and therefore you will not include them in your investigation. Moreover, you also don't want to include value in `dif.days` that are higher than 365 days.

6.5 l) Create a new data set in which these values are removed and name this data set `dataDif2` .

R code 6.5l:

6.5 m) Create a scatter plot with `dif.days` on the *y-axis* and `Amount` on the *x-axis*.

R code 6.5m:

6.5 n) Compute the **correlation** between the time between invoice and payment, and the amount that is paid.

R code 6.5n:

Answer 6.5n:

```
Correlation: _______________
```

6.5 o) Elaborate on the **correlation coefficient** and it's significance. What does this imply?

Answer 6.5o:

6.5 p) Compute the **mean** `dif.days` per expense category.

Hint 6.8: Use the `aggregate()` *function (for more help on this function see* `?aggregate` *).*

R code 6.5p:

Answer 6.5p:

6.5 q) Use the `aov()` function to test whether the **means** that you computed in assignment 6.5p are statistically different.

Answer 6.5q:

p-value: ____________

Conclusion:

R code 6.5q:

6.5 r) Use **Tukey's Honest Significant Differences** to find out which group **means** are truly different.

Hint 6.9: Use the `TukeyHSD()` *function to find* ***Tukey's Honest Significant Differences*** .

R code 6.5r:

Answer 6.5r:

In assignment 6.5q you have computed an **ANOVA** , but this is statistically not completely sound.

6.5 s) Can you formulate why the **ANOVA** in assignment 6.5q was not statistically sound? What would be an appropriate analysis?

Answer 6.5s:

6.5 t) Why do you think it is a good or bad idea to calculate **p-values** if the number of rows in the data is large?

Answer 6.5t: