STA238 Tutorial 9

Luis Ledesma

2023-03-29

1 Announcements

- You can upload your work on Crowdmark from the end of the tutorial session to 5pm Friday of that week.
- All questions must be solved using RStudio.

2 Recall: Last tutorial

Last tutorial: We reviewed applying linear regression models to a dataset, obtaining fitted values and parameter estimates, along with visualizations of the data. In addition, we set up a hypothesis test to see the significance of an independent variable in the model, and computed confidence intervals for the parameter estimates.

Main takeaways:

1. The function `lm()` will be used to fit linear regression models, and with `summary()` one can extract the parameter estimates, along with their associated standard errors and coefficient of determination R^2.
2. The hypothesis test for linear regression will be of the form $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$. In other words, the null hypothesis will be that the independent variable x_1 has no effect on the dependent variable y.
3. The test statistic $t = \frac{\hat{\beta_1}}{S_{\hat{\beta_1}}}$ will be t_{n-2} under the null hypothesis. The coefficient estimates for the confidence intervals will come from this distribution.

3 Tutorial activity

For this tutorial, we want to:

1. Carry out a one-way ANOVA test, and state the hypothesis test
2. Interpret the results from an ANOVA test
3. Check the conditions and assumptions for an ANOVA test
4. Perform a multiple comparisons test (Tukey)

3.1 Testing different drug effects

3.1.1 Carrying out the ANOVA test

```
library("car")
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
Drug <- c(rep("A", 9), rep("B", 9), rep("C", 9))

DrugA <- c(4, 5, 4, 3, 2, 4, 3, 4, 4)
DrugB <- c(6, 8, 4, 5, 4, 6, 5, 8, 6)
DrugC <- c(6, 7, 6, 6, 7, 5, 6, 5, 5)

Pain <- c(DrugA, DrugB, DrugC)

Drugdf <- data.frame(Drug= Drug, Pain= Pain)
```

The 27 volunteers were assigned to take one out of three drugs, and then they reported their pain levels after the next migraine episode. We would like to know whether the effect of the drugs on migraine pain levels is significantly different across the types of drugs or not. We will carry out **a one-way ANOVA test to compare the mean pain levels for different drugs.**

The hypothesis test will be:

$$H_0 : \mu_A = \mu_B = \mu_C$$

And H_a will be that at least two of the means μ_i are different. One can use the `aov()` function:

```
model1 <- aov(Pain ~ Drug, data = Drugdf)
summary(model1)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Drug         2  28.22  14.111   11.91 0.000256 ***
## Residuals   24  28.44   1.185
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value will be $0.000256 < 0.05$, so we reject H_0 at 0.05 significance. Then, there is enough evidence that at least two of the means are different; or that there are some differences among population means.

3.2 Checking the assumptions for the ANOVA test

To carry out the above test, we assumed the following:

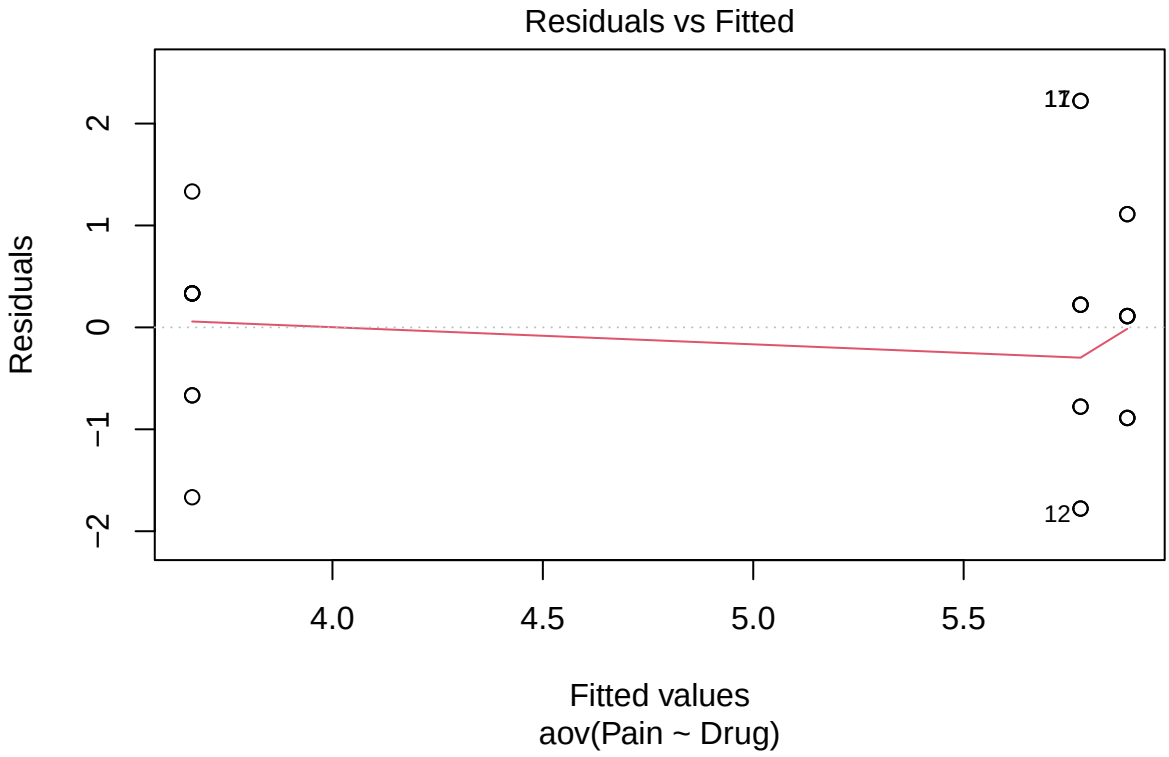
1. The variances are equal, or **homogeneity of variances**.

2. The response variable data comes from a normal distribution.

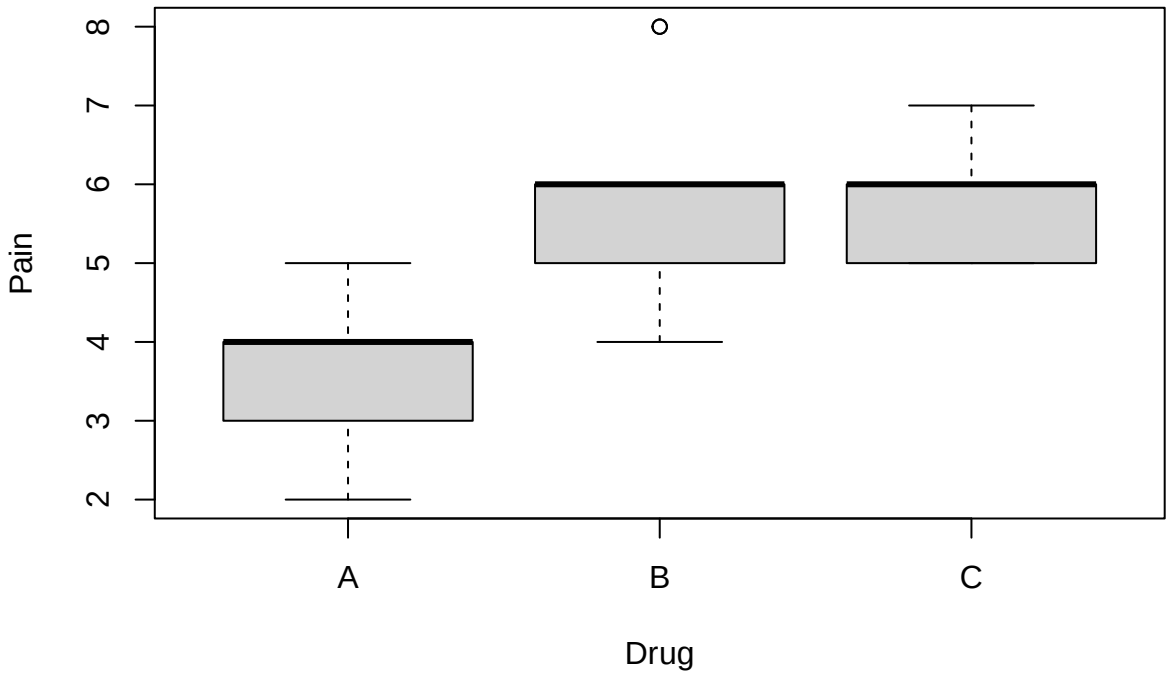
3.2.1 Homogeneity of variances

We can extract the plot of residuals against fitted values, and the boxplots of the data to assess the equal variances assumption:

```
##Homogeneity of variances
plot(model1, 1)
```




```
boxplot(Pain ~ Drug, data = Drugdf)
```

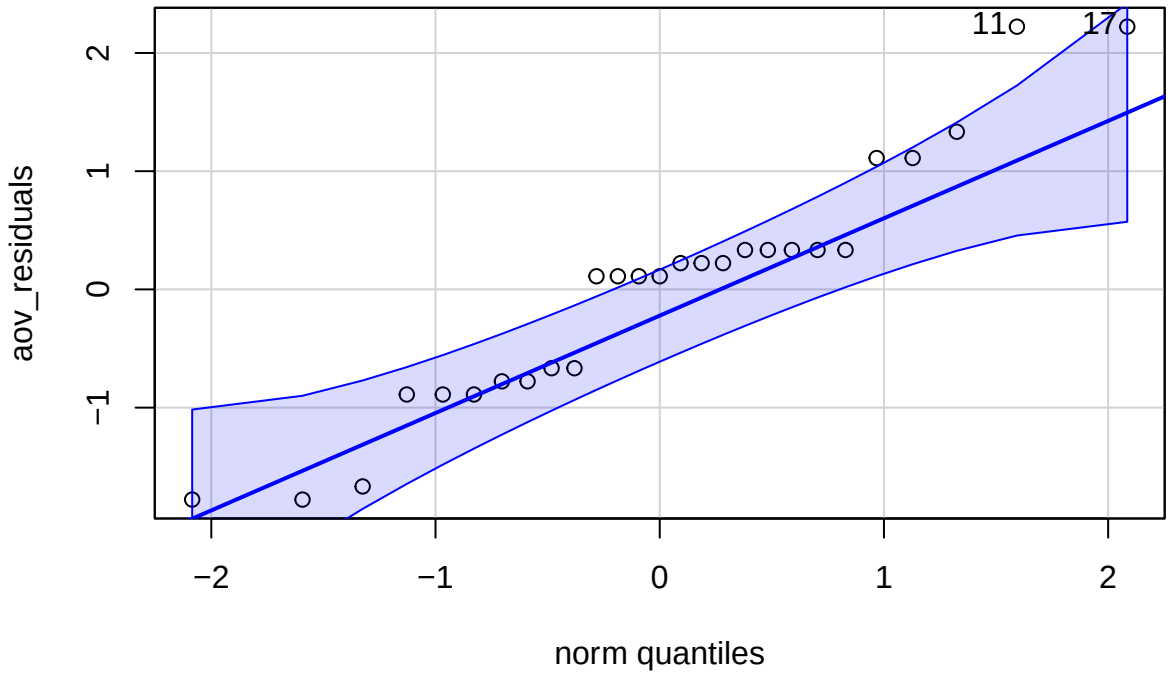


For the plot of residuals against fitted values, we have a straight line that does not vary much at the end, and the residuals are spread out over the x-axis. Likewise, for the boxplots, these also appear to be quite similar. Thus, we may assume that the variances are equally distributed.

3.2.2 Normality assumptions

To test for this assumption, we can look at the QQ-plots and conduct a Shapiro-Wilk test of normality:

```
##Normality assumptions
# Extract the residuals
aov_residuals <- residuals(object = model1)
qqPlot(aov_residuals)
```




```
## [1] 11 17
```

```
# Run Shapiro-Wilk test
shapiro.test(x = aov_residuals )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals
## W = 0.93675, p-value = 0.1013
```

The QQ-plot appears to roughly follow a straight line, and the p-value of the Shapiro-Wilk normality test is $0.1013 > 0.05$, so the assumption of normality is also sensible.

Hence, both assumptions for the one-way ANOVA test appear to be satisfied.

3.3 Comparisons of means

We now want to compare means using a multiple comparisons test, to determine which drugs differ in mean levels of pain reported. To do this, one uses the `TukeyHSD()` function:

```
TukeyHSD(model1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Pain ~ Drug, data = Drugdf)
##
## $Drug
##          diff        lwr      upr     p adj
## B-A 2.1111111  0.8295028 3.392719 0.0011107
## C-A 2.2222222  0.9406139 3.503831 0.0006453
## C-B 0.1111111 -1.1704972 1.392719 0.9745173
```

For the above, the p-values comparing B-A and C-A are significant. Thus, $\mu_B > \mu_A$ and $\mu_C > \mu_A$ (pay close attention to the value of `diff` to determine the sign in the inequality). As the p-value copmaring C-B is not significant (close to 1), then $\mu_B = \mu_C$. To conclude, we have $\mu_B = \mu_C > \mu_A$, or that drug A appears to have a better effect at mitigating pain compared to drugs B and C.