FORE224/STAT202 Assignment 7

Two-way analysis of variance (ANOVA) and modeling with both continuous and categorical predictors

Due at 12 noon, 6 October 2021

As usual, create a new project in its own folder for this assignment.

Two-way ANOVA

[1] Load the `tidyverse` library.

Download the file job_satisfaction2.csv from Learn and save it in your project folder. The file contains data for 106 cases. Three variables are recorded:

- the categorical variable education_level, for the case's highest education level, with values "school", "college", "university" (it is a US dataset; college is roughly between school and university)
- the categorical variable gender, for the case's gender, with values "another", "female", "male"
- the continuous variable score, which is the case's job satisfaction score.

Read the dataset from job_satisfaction2.csv and make sure that the education_level variable is treated as a factor with levels "school", "college", "university" where "school" is the baseline group, and the gender variable is treated as a factor (alphabetical factor levels). You can do this by using `read_csv` to read in the data and then using `mutate` as you did in Assignment 6 but adding `gender = factor(gender)` inside the `mutate` function.

Use `set.seed` to set the random seed to your student ID number and make a subset called my_js of 104 cases from this job satisfaction data using `sample_n` as in previous labs.

[2] Make a jittered scatter plot of score by education level and gender using `ggplot`. The following code will do this assuming that your sample data is called `my_js`:

```
my_js %>% ggplot(aes(
  x = gender, y = score,
  colour = gender
)) +
  geom_jitter(width = 0.1) +
  facet_wrap(~education_level) # does a nice layout :)
```

Optional: use `labs` to add an appropriate title and axis labels to the scatter plot.

Include the scatter plot in your assignment report.

[3] You are going to investigate whether there is evidence of an interaction between education_level and gender in the sampled populations. Explain briefly (in about 70 words) what it would mean for the relationship between population mean job satisfaction scores and education level and gender if there is such an interaction. You may find it helpful to use examples rather than a theoretical explanation.

[4] Use `lm` to create a model for job satisfaction score using education level and gender as the predictors, including the interaction between education level and gender. Show the code to create the model.

Create the usual 4 regression diagnostic plots (eg, as obtained using `plot(...)`). Include the plots in your report. Comment in about 60 words on whether there is anything that causes you concern about the least-squares linear model assumptions.

[5] Show the analysis of variance for the model given by the R function `anova`. State the p-value for the F-test for the interaction between education level and gender shown on the `anova’ output. Explain in about 80 words what probability this p-value represents.

Would you prefer to keep the interaction in the model or drop it? Briefly explain your choice in about 60 words.

Modeling using continuous and categorical predictors

[6] Load the `palmerpenguins` library. You will also need `tidyverse` if it is not already loaded. Use `drop_na` to drop the cases with missing values (usually written as “n/a”, “N/A” or “NA”) from the penguins data in the `palmerpenguins` library (as shown in the code used in lectures), and filter to get the male sample only. The following code will do both steps for you and create a sample called `male_penguins` of the male penguins with no “n/a” values.

```
male_penguins <- penguins %>%
  drop_na() %>%
  filter(sex == "male")
```

Use `set.seed` to set the random seed to your student ID number and make a subset called my_pen_m of 150 cases from this male penguin data using `sample_n` as in previous labs. Use this subset of the male penguin data for the rest of the assignment.

[7] Use `ggplot` to create a scatter plot of bill_depth_mm (y-axis) against bill_length_mm (x-axis) but ignoring species (ignoring species is not a good idea, as we will see very shortly!). Use `geom_smooth` to plot the single linear regression line for response bill_depth_mm, predictor bill_length_mm. The following code will create the required scatter plot and regression line:

```
my_pen_m %>% ggplot(aes(
  x = bill_length_mm,
  y = bill_depth_mm
)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = lm, se = FALSE)
```

Include the scatter plot in your report.

Now also create a second scatter plot with the points coloured by species. You can do this by adding `colour = species` inside the `aes` function in the code above. R will colour the points for each species and `geom_smooth` will make a regression line for each species. Include this second scatter plot in your report as well.

Optional: Use `labs` to add a title and appropriate axis labels to your plots.

Compare the relationships between bill depth and bill length suggested by these two plots, one without species, one with colours by species (about 80 words in total).

[8] You are going to create models for bill depth (response) using bill length and species as predictors with and without an interaction between bill length and species. Describe in about 40 words how the slope(s) of the regression line(s) using the model with interactions could differ from the slope(s) of the regression line(s) using the model without interactions.

Create a model to predict male penguin bill depth using both bill length and species, including the interaction between bill length and species. Show the code to create the model in your report.

Also create a model to predict male penguin bill depth using both bill length and species, but with no interaction between bill length and species. Show the code to create the model in your report.

[9] Perform a nested model ANOVA F-test using R's `anova` function to compare the two models you created in part [8].

State in 60-80 words what null and alternative hypotheses this F-test is testing.

Give your conclusion for this test and explain how the nested model ANOVA output justifies this conclusion (up to 50 words in total).

[10] Give the equations of the regression lines for each of the 3 species using the model including the interaction between bill length and species from part [8] (you will find it helpful to get R's summary output for the model to be able to work out the values of the intercepts and slopes for each species). Two decimal places for each coefficient is sufficient!

[11] Create the usual 4 regression diagnostic plots for the model including the interaction between bill length and species from part [8]. Include the plots in your report. Comment in 40-60 words on anything that causes you concern about the least-squares linear model assumptions.

Include all the answers and code used to obtain them in a Word document and submit it through Learn. Remember to use your own words when answering the questions.