Comparing and Contrasting ANOVA and Regression

Comparing 3+ groups – Stat 230

Scientists are interested in whether organic methods can be used to control harmful insects and limit their effect on sweet corn.[1] In a study of this question, researchers compared the weights of ears of corn under five conditions in an experiment in which sweet corn was grown using organic methods. A total of 60 plots were used in the study. In 12 plots of corn a beneficial soil nematode was introduced. In another 12 plots a parasitic wasp was used. Another 12 plots were treated with both the nematode and the wasp. In a fourth set of 12 plots a bacterium was used. Finally, a fifth set of 12 plots of corn acted as a control in which no special treatment was applied. The plots were all randomly assigned to the treatment conditions. Twenty-five ears of corn from each plot were randomly sampled and each was weighed (in ounces). Those weights were then used to determine an average weight for an ear of corn for each plot.

The data set can be loaded using the code

```
corn <- read.csv("https://aloy.rbind.io/data/cornplot.csv")
```

Notice that the treatment groups were recorded as numeric labels where

treatment label	condition
1	soil nematode
2	parasitic wasp
3	nematode and parasitic wasp
4	bacterium
5	control

Two statistics students analyze these data: one uses a regression model, the other uses a one-way ANOVA model.

- One-way ANOVA model: $y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j}$ for $i = 1, 2, 3, 4, 5$ and $j = 1, \dots, 12$
- Linear regression model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ for $i = 1, 2, \dots, 60$

EDA

To begin, explore the data set by completing the following tasks:

Task 1. Create an individual values plot of the data. Comment on the distribution within each treatment group. Are there outliers? Are the distributions skewed? Are there unequal variances?

Task 2. Reflecting on your answers to task 1, is a transformation of weight needed prior to modeling?

[1] The problem and data set were obtained from *Intermediate Statistical Investigations* by Tintle et al.

ANOVA-based analysis

Now, let's use the one-way ANOVA model to conduct a first analysis.

Task 3. Use R to fit the one-way ANOVA model for this experiment. Use the `summary()` command to print the full ANOVA table.

> **Note**
>
> Since `Treatment` is coded numerically, you need to convert it to a categorical variable so that R knows what to do. Thus, you should use `factor(Treatment)` as the explanatory variable for your ANOVA model.

Task 4. The ANOVA model predicts that every point in a treatment group will be at the group mean. Use the below code to create a plot of the fitted ANOVA model. Comment on the fit of this model to your data.

```
group_means <- corn |>
  group_by(Treatment) |>
  summarize(avg = mean(Weight))

gf_jitter(Weight ~ Treatment, data = corn, width= 0.1) |>
  gf_summary(fun.y = "mean", color = "darkorange", size = 3, geom = "point")
```

Task 5. Construct a plot of the residuals vs. the explanatory variable (Treatment) and a normal Q-Q plot of the residuals for the ANOVA model. Comment on the fit of the ANOVA model.

Post-hoc analysis

In Daily Prep 9.21 you saw that the ANOVA F-test indicated that there is a statistically discernible difference in the treatment means. This is great information to have, but the scientists ultimately want to know *which* treatment groups are different and *how* they are different. To answer this question we need to assess how the means of the treatment groups relate to one another. This is sometimes called **post-hoc analysis**. For five treatment groups, this might involve looking at all ten pairwise comparisons.

You may recall from your introductory statistics course that running a large number of hypothesis tests (or constructing a large number of confidence intervals) quickly increases the chance of making at least one error (i.e., getting a small p-value or a CI that doesn't contain 0 when there's no effect). This is the problem of multiple comparisons. One way to guard against this "runaway" error rate is to only carry out comparisons *after* finding a statistically discernible results from ANOVA. This is called F-protected inference. (If you get interested in multiple comparisons, you can read more about them in Chapter 1.)

Task 6. Calculate 95% confidence intervals for the 10 pairwise comparisons of each of the treatments. To calculate an interval we use the same formula as our two-sample t-based intervals, but set $s_p = \sqrt{MSE}$. Thus, you are calculating $\bar{y}_{i\cdot} - \bar{y}_{j\cdot} \pm t^* \cdot \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$. (Recall that you can obtain the group means and sample sizes from `favstats()`.)

> **Finding $t*$**
>
> Recall that you can calculate the $t*$ value for a confidence interval using `qt(area, df)`. Where `area` is $1-\alpha/2$ (notice how this differs from the confidence level of $1-\alpha$) and `df` are the degrees of freedom.

Task 7. Summarize, broadly, what these intervals tell you about how the treatments compare to each other.

Regression-based analysis

Next, let's use the linear regression model used by the other student to conduct an analysis.

Task 8. Use R to fit the simple linear regression model. If you transformed the data for your ANOVA model, use the same transformation for this question. Use the `summary()` command print the coefficient table containing the results of the t-test for the slope parameter.

Task 9. The regression model uses a linear equation to predict the response for each treatment group. Use the below code to create a plot of the fitted regression model (transform the response variable if necessary). Comment on the fit of this model to your data.

```
gf_jitter(Weight ~ Treatment, data = corn, width= 0.1) |>
  gf_lm(color = "darkorange")
```

Task 10. Construct a plot of the residuals vs. the explanatory variable (Treatment) and a normal Q-Q plot of the residuals for the regression model. Comment on the fit of the regression model.

Task 11. Is the p-value for the ANOVA F-test that $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5$ the same or different from the p-value for the t-test that $H_0 : \beta_1 = 0$? Does this surprise you?

Synthesis

Task 12. Write a brief statement advising the scientists on how they should analyze these data (i.e., which student's analysis to go with). You should clearly justify your choice, linking to specific evidence from earlier tasks.