Exercise 3

PLSCI 7201

9/6/2020

Summary

This exercise is split into two parts. The first focuses on multiple regression and model comparisons. The second involves fitting a one-way ANOVA and making comparisons between treatment groups. You should already by familiar with many of the functions used in this exercise, so I've only included examples where new functions are introduced.

Multiple regression

You are interested in investigating the effects of vegetative traits on grain yield in rice. Seventy-five plants were selected at random from a farmer's field. At harvest grain yield was recorded (g plant), as well as straw biomass (g plant) and straw N content (% straw dry weight). In addition to the primary variables of interest (straw biomass and straw N), flowering time (days from sowing) and panicle length (cm) was recorded. These data can be found in the file `Problem1Data.RDS`. You can read RDS files in using `readRDS()`.

Exploratory analysis

- **Q1.** Calculate the correlation between all variables using Pearson's method. Do any of the predictor variables appear to be correlated? Based on these results do you think any of the predictors are related to yield?

Regression

- **Q2.** Fit a linear model that includes only the vegetative traits (straw biomass and straw N). Generate a set of diagnostic plots (QQ plot, scale location, residual vs fitted). Do you think we are violating any assumptions of the OLS? If so, why? Are there any problematic data points?
- **Q3.** The first hypothesis to test in linear regression is one regarding the overall fit of the model. Here the null hypothesis is $H_0 : \beta_1 = \beta_2 = ...\beta_i = 0$ and the alternative is at least one $\beta_i \neq 0$. Based on the data below, do you believe that we should reject the null hypothesis? Why?
- **Q4.** How much variation in grain yield is explained by these predictors? (hint: this is the ratio of the regression sum of squares and total sum of squares.)

Model comparisons

Although our primary interest is in studying the effects of vegetative traits on grain yield, we have reason to believe that other traits such as flowering time or panicle length may also have an effect on grain yield. Fit a full model that includes all four predictors and compare it to the model above.

- **Q5.** Did we improve the fit of the model by including these additional predictors? If so, by how much? Do you believe this model provides an adequate fit to the data?

We can also test the effect of any given predictor using a t-test. The hypothesis we're testing is $H_0 : \beta_i = 0; H_1 : \beta_i \neq 0$.

- **Q6.** Interpret the cofficient for any significant predictor ($p < 0.05$).

One-way ANOVA

We're interested in evaluating six different oat varieties for yields. The six varieties (Belinda, Corral, Horsepower, Badger, Steve, Tinner) were assigned at random to 30 plots in the field. Each variety was planted in an equal number of plots (five reps for each line). These data can be found in the file Problem2Data.RDS.

OLS with dummy variables

Remember we can express these data using a additive model given by $y_{ij} = \mu + \alpha_i + e_{ij}$, where α_i is the devition of the ith group from the mean. The data is ordered by variety, so the first ten lines of the incidence matrix (matrix of dummy coded variables) would look something like

$$\begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

We can create such a matrix in R using the `model.matrix()` function.

- **Q6.** Create a design matrix for the additive model and fit OLS using these dummy variables using the `lm()` function. The example below can be used to generate the degisn matrix and fit the linear model. Report the means for each group. Explain how you calculated the mean for the line Tinner.

Example

```
oatData <- readRDS("Problem2Data.RDS")
desMat <- model.matrix(~0 + Line, oatData)[, 2:6]

mod1 <- lm(oatData$yield ~ 1 + desMat)
summary(mod1)
```

- **Q7.** You should see an F statistic and a p-value at the bottom of the output. What hypothesis is being tested here? What can you conclude based on this test?

Post-hoc testing

Finally, let's fit an ANOVA the ''easy" way in R (i.e. no dummy coding) using `aov()`, and compare the lines using Tukey's honest significant differnence test. Refer to slides 28 and 29 in the lecture to see how to fit an ANOVA the easy way. `TukeyHSD` takes an `aov()` model as input.

- **Q8.** Based on the output from Tukey's HSD, what can you conclude about the experiment? Do you seeany significant differences between lines ($p < 0.05$)? If so which pairs differ?