STAT 420: Methods of Applied Statistics

Categorical Predictors and ANOVA

Ruoqing Zhu, Ph.D. <rqzhu@illinois.edu>

Course website: https://sites.google.com/site/teazrq/teaching/STAT420

Department of Statistics
University of Illinois at Urbana-Champaign
April 20, 2017

Categorical Predictors

- Categorical Predictors appears in many real applications:
  - A variable indicates college year: freshman, sophomore, junior and senior
  - Genotype for petal color in a pea plant: AA, Aa, aa
- We are usually interested in
  - If a categorical predictor is important
  - If there is an interaction between two categorical predictors
  - If there is an interaction between a categorical predictor and a continuous predictor

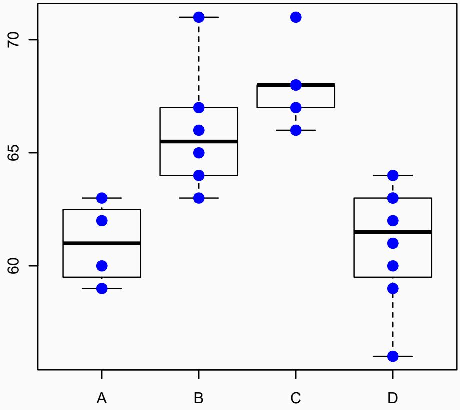
One factor model

- The coagulation data in faraway package
- We want to analyze whether diet (4 categories: A, B, C, D) is associated with the blood coagulation times (coag).
- This is the only variable in the model, and we can write down the model as

$$
\begin{aligned}
Y_i &= \beta_1 \mathbb{1}\{\text{diet} = \text{A}\} + \beta_2 \mathbb{1}\{\text{diet} = \text{B}\} + \\
&= \beta_3 \mathbb{1}\{\text{diet} = \text{C}\} + \beta_4 \mathbb{1}\{\text{diet} = \text{D}\} + \epsilon_i
\end{aligned}
$$

- Can you write down the design matrix?

Coagulation dataset



One way ANOVA

```
> summary(aov(coag ~ diet, data = coagulation))
            Df Sum Sq Mean Sq F value   Pr(>F)
diet         3    228    76.0   13.57 4.66e-05 ***
Residuals   20    112     5.6
>
> summary(lm(coag ~ diet, data = coagulation ))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.100e+01  1.183e+00  51.554  < 2e-16 ***
dietB       5.000e+00  1.528e+00   3.273 0.003803 **
dietC       7.000e+00  1.528e+00   4.583 0.000181 ***
dietD       2.991e-15  1.449e+00   0.000 1.000000
---
Signif. codes:  0 ?**?0.001 ?*?0.01 ??0.05 ??0.1 ??1

Residual standard error: 2.366 on 20 degrees of freedom
Multiple R-squared:  0.6706, Adjusted R-squared:  0.6212
F-statistic: 13.57 on 3 and 20 DF,  p-value: 4.658e-05
```

One factor model

- How to judge whether `diet` is an important predictor?
- This is done using ANOVA (analysis of variance)
- The idea of ANOVA is looking at the variance of the outcome before and after the categorical variable is added.
- Suppose there are J categories, the sample sizes are $n_1, \ldots, n_J$, hence, $n = \sum_j n_j$, then

$$\text{SST} = \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - \bar{\mathbf{y}})^2$$

$$\text{SSE} = \sum_{j=1}^{J} \sum_{i=1}^{n_j} (y_{ij} - \bar{\mathbf{y}}_{\bullet j})^2 = \sum_{j=1}^{J} (n_j - 1) s_j^2$$

$$\text{SSR} = \sum_{j=1}^{J} \sum_{i=1}^{n_j} (\bar{\mathbf{y}}_{\bullet j} - \bar{\mathbf{y}})^2 = \sum_{j=1}^{J} n_j (\bar{\mathbf{y}}_{\bullet j} - \bar{\mathbf{y}})^2$$

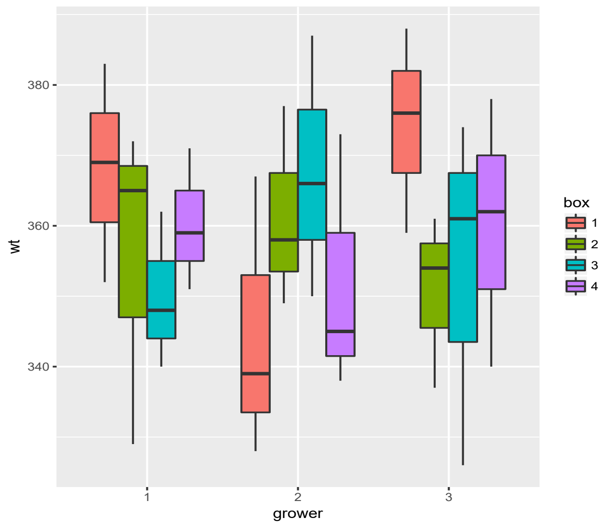
One factor model

- This is equivalent to running an regression with J number of parameters and look at the F test statistics
  - First, write down the design matrix (recall our exam 1, last question)
  - What is the β_j for each category?
  - How the SSE is calculated?
- Is the intercept term needed?
- What's the degrees of freedom for testing this categorical variable?
- Interpretation of the results?

Two-way ANOVA

- What we had previously was a One-Way ANOVA.
- Suppose we have two categorical variables. Example: the broccoli dataset in faraway.
- Model weight using grower (3 levels) and box (4 levels).
- Consider two models:
  (1) weight $=$ {grower type} $+$ {box type} (totally 6 levels)
  (2) weight $=$ {grower type} $\times$ {box type} (totally 12 parameters)
- Two-way ANOVA can be used to test: is that 6 extra parameters needed?
- It can also simply test if any of the single effect exist. But that is trivial.

Broccoli dataset



Is interaction needed?

Broccoli: no interaction

```
> fit = lm(wt ~ factor(grower) + factor(box), data = broccoli)
> summary(aov(fit))
               Df Sum Sq Mean Sq F value Pr(>F)
factor(grower)  2     64    32.2   0.102  0.904
factor(box)     3    222    74.1   0.234  0.872
Residuals      30   9506   316.9
```

Broccoli: with interaction

```
> fit = lm(wt ~ factor(grower)*factor(box), data = broccoli)
> summary(aov(fit))
                            Df Sum Sq Mean Sq F value Pr(>F)
factor(grower)               2     64    32.2   0.106  0.900
factor(box)                  3    222    74.1   0.244  0.865
factor(grower):factor(box)   6   2227   371.1   1.224  0.329
Residuals                   24   7279   303.3
```

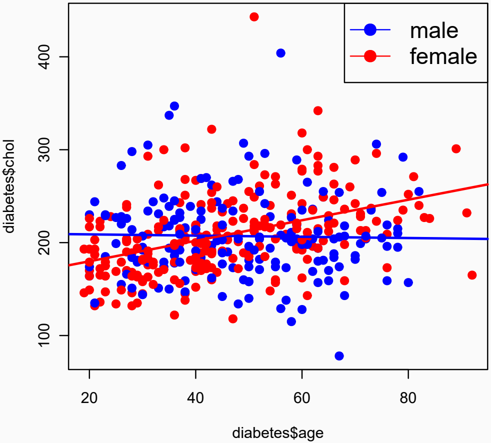
Example

- Load the butterfat data from faraway. Which model the most appropriate based on 0.05 significance level?
  A). Null model
  B). Breed only
  C). Age only
  D). Breed + Age
  E). Breed × Age

ANCOVA

- Load the diabetes dataset
- Model the total cholesterol (chol) using (age).
- We can test:
    - Is there a gender effect at all? (controlling for age)
    - Is this relationship different across different gender?
- ANCOVA can be used

Broccoli dataset



Is interaction needed?

Testing for gender effect

```
> fit = lm(chol ~ age + gender, diabetes)
# why I dont need to specify "factor(gender)"?
> summary(aov(fit))
             Df Sum Sq Mean Sq F value   Pr(>F)
age           1  43048   43048  22.956 2.34e-06 ***
gender        1    857     857   0.457    0.499
Residuals   399 748233    1875
```

Testing for difference of age effect across gender

```
> fit = lm(chol ~ age*gender, diabetes)
> summary(aov(fit))
             Df Sum Sq Mean Sq F value    Pr(>F)
age           1  43048   43048  24.016 1.39e-06 ***
gender        1    857     857   0.478      0.49
age:gender    1  34832   34832  19.433 1.34e-05 ***
Residuals   398 713401    1792
```