Lecture 10: Introduction to Analysis of Variance

Ailin Zhang

2023-06-02

Recap: Qualititative/ Categorical Predictors

```
salary = read.table("salary.txt", header=TRUE)
```

$$
\begin{aligned}
S &= \text{Salary} \\
X &= \text{Experience, in years} \\
E &= \text{Education} \\
& \quad \text{1 if H.S. only,} \\
& \quad \text{2 if Bachelor's only,} \\
& \quad \text{3 if Advanced degree} \\
M &= \text{Management Status} \\
& \quad \text{1 if manager, 0 if non-manager}
\end{aligned}
$$

Indicator Varibales (aka. Dummy Variables)

- For Education (E), it contains 3 categories, so we can start from 3 indicator variables:

$$E_{i1} = \begin{cases} 1 & \text{if sample i has a high school diploma only} \\ 0 & \text{otherwise} \end{cases}$$

$$E_{i2} = \begin{cases} 1 & \text{if sample i has a B.S. only} \\ 0 & \text{otherwise} \end{cases}$$

$$E_{i3} = \begin{cases} 1 & \text{if sample i has an advanced degress only} \\ 0 & \text{otherwise} \end{cases}$$

- Identity holds: $E_1 + E_2 + E_3 = 1$. Must drop one of them.
- In general, a categorical predictor with c categories needs only c-1 indicator variables

Recap: Example

- The command `as.factor()` or `factor()` tells R that E is categorical and the indicator variables E1, E2, E3 are created automatically. By default, R drops the indicator E1 for the lowest level

```
salary$Edu = factor(salary$E,
                    labels=c("High School","College","Advanced"))
lmsalary = lm(S ~ X+Edu, salary)
```

$$Y = \beta_0 + \beta_1 X + \gamma_2 E_2 + \gamma_3 E_3 + \epsilon$$

And we can look at the E(Y) for different education levels:

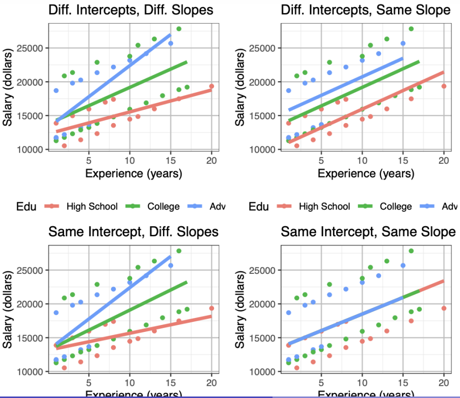
Education(E)	Indicator	E(Y)
1 (HS)	$E_2 = E_3 = 0$	$\beta_0 + \beta_1 X$
2 (B.S.)	$E_2 = 1, E_3 = 0$	$\beta_0 + \beta_1 X + \gamma_2$
3 (Advanced)	$E_2 = 0, E_3 = 1$	$\beta_0 + \beta_1 X + \gamma_3$

Based on the model above, for people w/ the same years of experience (X), the difference in their mean salary are

- γ_2 : B.S.-HS
- γ_3 : Advanced-HS
- $\gamma_3 - \gamma_2$: Advanced-B.S.

Interaction and Additive Models

- If the effect of a predictor on response changes with the level of another predictor, we say there exists interaction(s) between the 2 predictors
- Otherwise, we say their effects are additive.
- The previous model assumes the effects of education (E) and experience (X) on salary are additive



Model with Different Intercepts & Different Slopes (Interaction)

Consider the model:

$$Y = \beta_0 + \beta_1 X + \gamma_2 E_2 + \gamma_3 E_3 + \alpha_2(E_2 \cdot X) + \alpha_3(E_3 \cdot X) + \epsilon$$

$$Y = \begin{cases} \beta_0 + (\beta_1)X + \epsilon & \text{if high school diploma only} \\ \beta_0 + (\beta_1 + \alpha_2)X + \gamma_2 + \epsilon & \text{if BS only} \\ \beta_0 + (\beta_1 + \alpha_3)X + \gamma_3 + \epsilon & \text{if advanced} \end{cases}$$

- This model has different intercepts and different slopes!

```
lm_int =lm(S~X+Edu+X*Edu, salary)
```

Model Summary

```r
summary(lm_int)$coef
```

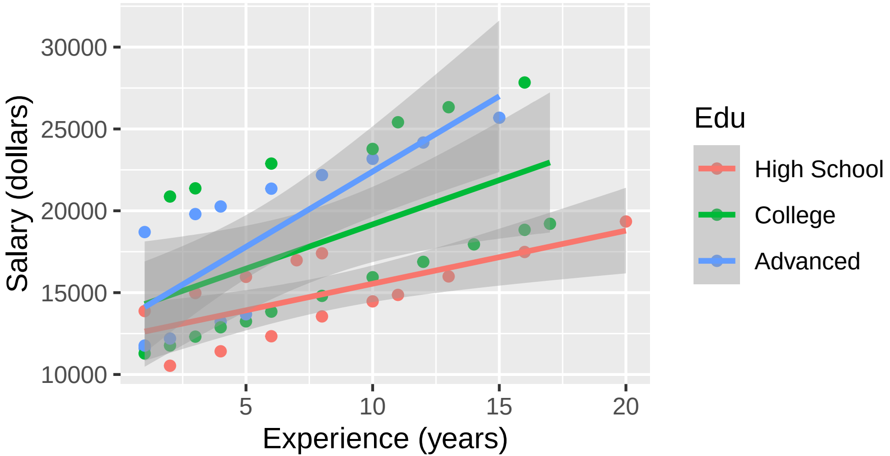
```
##                   Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   12299.0222  1740.3665 7.0669151 1.513652e-08
## X               324.5148   179.6417 1.8064564 7.837470e-02
## EduCollege     1461.1812  2326.3969 0.6280877 5.335164e-01
## EduAdvanced     898.2396  2357.1494 0.3810703 7.051676e-01
## X:EduCollege    216.3477   238.6374 0.9065959 3.700497e-01
## X:EduAdvanced   595.5212   288.8711 2.0615464 4.579092e-02
```

$$\hat{Y} = \begin{cases} 12299 + 324.5X & \text{if high school diploma only} \\ 12299 + (324.5 + 216.3)X + 1461.2 & \text{if BS only} \\ 12299 + (324.5 + 595.5)X + 898.2 & \text{if advanced} \end{cases}$$

On average, every extra year of experience worth

- \$324.5 if HS only
- \$324.5+\$216.3 if BA or BS only
- \$324.5+\$595.5 if Adv. deg.

```
library("ggplot2")
ggplot(salary, aes(x = X, y = S, color=Edu)) +
  geom_point() + geom_smooth(method="lm", formula='y~x') + xla
ylab("Salary (dollars)")
```



Test Whether the Slopes Are Different

```
summary(lm_int)$coef
```

```
##                  Estimate Std. Error  t value     Pr(>|t|)
## (Intercept)   12299.0222  1740.3665 7.0669151 1.513652e-08
## X               324.5148   179.6417 1.8064564 7.837470e-02
## EduCollege     1461.1812  2326.3969 0.6280877 5.335164e-01
## EduAdvanced     898.2396  2357.1494 0.3810703 7.051676e-01
## X:EduCollege    216.3477   238.6374 0.9065959 3.700497e-01
## X:EduAdvanced   595.5212   288.8711 2.0615464 4.579092e-02
```

- X:EduCollege (α_2) is not significant (P-value 0.37)

  No significant difference between the slopes of the lines for HS & College

- X:EduAdvanced (α_3) is slightly significant (P-value 0.045).

  slightly significant difference between the slopes of the lines for HS v.s. advanced degree.

Test of Interactions

To know whether the effect of experience X on salary S changes with education level, one can test

$$H_0 : \alpha_2 = \alpha_3 = 0$$

```
anova(lmsalary, lm_int)
```

```
## Analysis of Variance Table
##
## Model 1: S ~ X + Edu
## Model 2: S ~ X + Edu + X * Edu
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1     42 550853135
## 2     40 497897342  2  52955792 2.1272 0.1325
```

Model with Same Intercept but Different Slopes (Less Common)

Consider the model:

$$Y = \beta_0 + \beta_1 X + \alpha_2(E_2 \cdot X) + \alpha_3(E_3 \cdot X) + \epsilon$$

$$Y = \begin{cases} \beta_0 + (\beta_1)X + \epsilon & \text{if high school diploma only} \\ \beta_0 + (\beta_1 + \alpha_2)X + \epsilon & \text{if BS only} \\ \beta_0 + (\beta_1 + \alpha_3)X + \epsilon & \text{if advanced} \end{cases}$$

- R will automatically include E and X if E*X is included in the model. (Three identical methods as following)

```
lm(S~X+Edu+X*Edu, salary)
lm(S~X+X*Edu, salary)
lm(S~X*Edu, salary)
```

- How to exclude categorical variables?

Model with Same Intercept but Different Slopes (Less Common)

- Unlike `E*X`, `E:X` would not automatically include E and X.

```r
summary(lm(S~X+X:Edu, salary))$coef
```

```
##                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   13144.5736   916.3481 14.344520 8.698772e-18
## X               251.1565   124.2237  2.021808 4.959733e-02
## X:EduCollege    343.0494   125.0421  2.743472 8.901257e-03
## X:EduAdvanced   674.7885   168.2282  4.011150 2.431440e-04
```

Various Interactions

There are various interaction terms we can include in the model.

- X + E + M
- X + E + M + E:X
- X + E + M + M:X
- X + E + M + E:M
- X + E + M + E:X + M:X
- X + E + M + E:X + M:X + E:M
- X + E + M + E:X + M:X + E:M + E:M:X

Factor Analysis : ANOVA

- Partition of the response-variable sum of squares into "explained" and "unexplained"
- Procedures for fitting and testing linear models in which the explanatory variables are categorical.
- Suppose that there are no quantitative explanatory variables—only a single factor in your model:

$$Y_i = \beta_0 + \gamma_2 E_{i2} + \gamma_3 E_{i3} + \epsilon_i$$

- $E(\hat{Y}_i)$ in each category (group, level of factor) is the population group mean, can be denoted by μ_j

One-way ANOVA

Education(E)	Indicator	E(Y)
1 (HS)	$E_2 = E_3 = 0$	$\mu_1 = \beta_0$
2 (B.S.)	$E_2 = 1, E_3 = 0$	$\mu_2 = \beta_0 + \gamma_2$
3 (Advanced)	$E_2 = 0, E_3 = 1$	$\mu_3 = \beta_0 + \gamma_3$

- One-way ANOVA focuses on testing for differences among group means.
- One-way ANOVA examines the relationship between a quantitative response variable and a factor.
- H_0: $\gamma_2 = \gamma_3 = 0$, which implies $\mu_1 = \mu_2 = \mu_3$
- F-statistic for the regression of the response variable on 0/1 dummy regressors constructed from the factor tests for differences in the response means across levels of the factor.

Example: One-way ANOVA

- A treatment (or factor): characteristic, that allows us to distinguish the different populations from one another.

```
anova(lm(S~Edu,salary))
```

```
## Analysis of Variance Table
##
## Response: S
##           Df    Sum Sq  Mean Sq F value Pr(>F)
## Edu        2 109134646 54567323  2.6306 0.0836 .
## Residuals 43 891962932 20743324
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Decision rule:

- p-value Approach: Reject H_0 if p-value$< \alpha$
- Critical Value Approach: Reject H_0 if $F > F_\alpha$

With $\alpha = 0.05$, we don't have sufficient evidence to conclude that the mean salary is not the same at all 3 education levels

One-way ANOVA Table

Source	df	Sum of Squares	Mean Squares	F
Treatment	c-1	SSR	MSR	$F = \frac{MSR}{MSE}$
Error	n-c	SSE	MSE	
Total	n-1	SST		

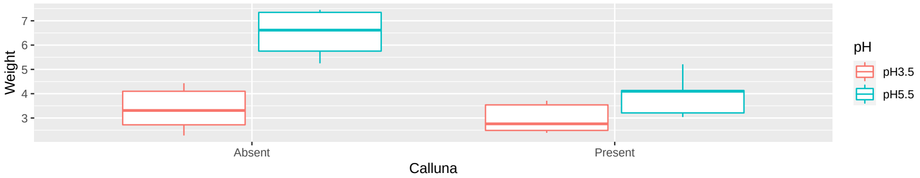
- $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$
- $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{j=1}^{c} n_j(\hat{\mu}_j - \bar{y})^2$
- $SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{j=1}^{c} (n_j - 1)s_j^2$

Where sample variance: $s_j^2 = \frac{\sum_{i \in \text{group}_j} (y_i - \mu_j)^2}{n_j - 1}$

Two-way ANOVA

- Two-way ANOVA allows us to determine whether there are significant differences between the effects of two categorical variables.
- Change of response variable may depend on
  - the level of the categorical variable (additive model)
  - the level of the interaction model.

```
festuca <- read.table(file = "festuca.txt",header = T)
ggplot(data = festuca, aes(x = Calluna, y = Weight,
                           colour = pH)) +
  geom_boxplot()
```



Fitting the ANOVA model

```
festuca_model <- lm(Weight ~ pH + Calluna + pH*Calluna,
                    data = festuca)
summary(festuca_model)$coef
```

```
##                        Estimate Std. Error   t value     Pr
## (Intercept)              3.3680  0.3869111  8.704842 2.9985
## pHpH5.5                  3.1145  0.5803667  5.366435 7.8510
## CallunaPresent          -0.3900  0.5471749 -0.712752 4.8693
## pHpH5.5:CallunaPresent  -2.1545  0.7976377 -2.701101 1.6422
```

$Y = \beta_0 + \beta_1 \text{pH} + \alpha_1 \text{Calluna} + \gamma_1 \text{pH} \cdot \text{Calluna}$

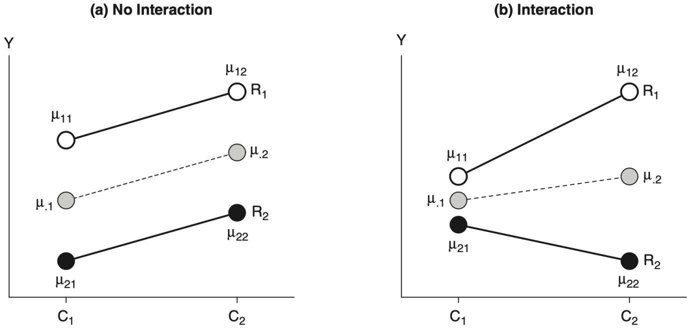
Fitting the ANOVA model

```r
anova(festuca_model)
```

```
## Analysis of Variance Table
##
## Response: Weight
##             Df Sum Sq Mean Sq F value    Pr(>F)
## pH           1 17.026 17.0261 22.7469 0.0002484 ***
## Calluna      1  9.307  9.3070 12.4341 0.0030548 **
## pH:Calluna   1  5.461  5.4610  7.2959 0.0164225 *
## Residuals   15 11.227  0.7485
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Line 1: the main effect of pH: pH has a significant effect on the weight.
- Line 2: the main effect of Calluna: Calluna has a significant effect on the weight.
- Line 3: the third for the interaction between pH and Calluna: The interaction has a significant effect on the weight.
- The presence of an interaction between treatments indicates that the impact of one factor depends on the levels of the other factor.

Understanding the model graphically



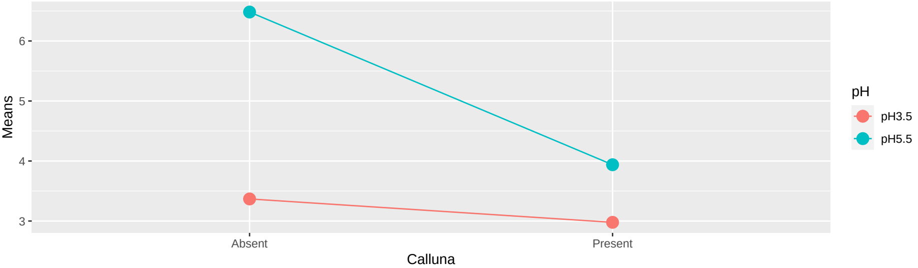
- Interaction Diagram: the lines linking the treatments are parallel when there is no interaction, but become non-parallel when an interaction is present.

Interaction Diagram

```r
# step 1. calculate means for each treatment combination
festuca_means <- festuca %>%
  group_by(Calluna, pH) %>% # <- remember to group by *both* factors
  summarise(Means = mean(Weight))
```

```
## `summarise()` has grouped output by 'Calluna'. You can override using the
## `.groups` argument.
```

```r
# step 2. plot these as an interaction diagram
ggplot(festuca_means,
       aes(x = Calluna, y = Means, colour = pH, group = pH)) +
  geom_point(size = 4) + geom_line()
```



ANOVA and Linear Regression

Model	Terms	Regression Sum of Squares
1	pH, Calluna, pH*Calluna	RSS_1
2	pH, Calluna	RSS_2
3	pH	RSS_3
4	Calluna	RSS_4

The four models will produce the two-way ANOVA table

Source	Model Contrasted	df	Sum of Squares	Mean Squares	F
pH	2-4	1	$SSR_2 - SSR_4$	MSR	$F = \frac{MSR}{MSE}$
Calluna	2-3	1	$SSR_2 - SSR_3$	MSR	$F = \frac{MSR}{MSE}$
Interaction	1-2	1	$SSR_1 - SSR_2$	MSR	$F = \frac{MSR}{MSE}$
Error	from Model 1	n-4	SSE	MSE	
Total		n-1	SST		

Extra: Multiple Comparisons of Means

- If we do reject the null hypothesis, we know that at least two means are significantly different from one another.
- To identify which pairs of means differ from one another, we use a multiple comparison procedure.
- There are multiple methods for multiple comparison of means, but it is not the primary focus for this course.
- For example: Tukey's method, Bonferroni's method, etc..

Example: Tukey's method

```
festuca_aov <- aov(festuca_model)
TukeyHSD(festuca_aov, which = 'pH:Calluna')
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = festuca_model)
##
## $`pH:Calluna`
##                                diff        lwr       upr     p adj
## pH5.5:Absent-pH3.5:Absent    3.1145  1.4417970  4.787203 0.0004089
## pH3.5:Present-pH3.5:Absent  -0.3900 -1.9670395  1.187039 0.8904902
## pH5.5:Present-pH3.5:Absent   0.5700 -1.0070395  2.147039 0.7283665
## pH3.5:Present-pH5.5:Absent  -3.5045 -5.1772030 -1.831797 0.0001201
## pH5.5:Present-pH5.5:Absent  -2.5445 -4.2172030 -0.871797 0.0026762
## pH5.5:Present-pH3.5:Present  0.9600 -0.6170395  2.537039 0.3319010
```

Extra: Two-way Ancova

- Determine whether there is an interaction effect between two independent variables in terms of a continuous dependent variable, after adjusting/controlling for one or more continuous covariates

```
lm_int =lm(S~X+Edu+X*Edu, salary)
anova(lm_int)
```

```
## Analysis of Variance Table
##
## Response: S
##           Df    Sum Sq   Mean Sq F value    Pr(>F)
## X          1 290716721 290716721 23.3556 2.014e-05 ***
## Edu        2 159527722  79763861  6.4081  0.003854 **
## X:Edu      2  52955792  26477896  2.1272  0.132458
## Residuals 40 497897342  12447434
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- There is significant effect of years of experience on salary while controlling for education level
- There is significant effect of education on salary while controlling for years of experience
- There is no significant effect of interaction between education and years of experience on salary