

Eco311 Optional Reading: ANOVA and ANCOVA

(Jing Li, Miami University)

1. Social science such as economics differs from natural science in many ways. From econometrics perspective, one key difference is that experiments and randomized controlled trials (RCT) are routinely used in natural science but not in social science. Accordingly, different terminology is developed for a statistical analysis of experimental data. This note introduces analysis of variance (ANOVA) and analysis of covariance (ANCOVA), and explains how the two methods are closely related to regression analysis.
2. Suppose we run a randomized controlled trial that randomly assigns NBA players to different positions. For instance, players with last digit of SSN being 7 will play center position. The categorical variable “position”, denoted by x , is called treatment, and in this case x takes three levels—center, forward, guard.
3. We are interested in estimating treatment effect—how the treatment affects outcome variable y such as wage. When experimental data are used, estimating treatment effect is easy since the difference in outcome can only be attributed to treatment—other confounding factors are either constant in an experiment or statistically similar in RCT.
4. The null hypothesis of ANOVA states that there is no treatment effect, i.e., the average wages across positions (conditional means) are the same:

$$H_0 : \mu_{center} = \mu_{forward} = \mu_{guard}, \quad (\text{ANOVA Hypothesis}) \quad (1)$$

This hypothesis involves multiple restrictions, so F test is needed.

5. Notice that we can use a two-sample t test if there is no center position, i.e., when there are only two positions (levels). In light of this, ANOVA or F test generalizes the two sample t test to cases where there are more than two levels. Another example of ANOVA is testing the hypothesis of equal average SAT scores across students who attend public schools, private schools, and are home schooled.
6. In order to test hypothesis (1), consider a decomposition of variance for y :

$$var(y) = E(var(y|x)) + var(E(y|x)), \quad (\text{Variance Decomposition}) \quad (2)$$

Proof

$$\text{var}(y) = E(y - E(y))^2 = E(y - E(y|x) + E(y|x) - E(y))^2 \quad (3)$$

$$= E(y - E(y|x))^2 + E(E(y|x) - E(y))^2 + 2E[(y - E(y|x))(E(y|x) - E(y))] \quad (4)$$

$$= E(\text{var}(y|x)) + \text{var}(E(y|x)) \quad (5)$$

where y is wage, x is position. The last equality follows because law of iterated expectation implies that $E[(y - E(y|x))(E(y|x) - E(y))] = 0$, and $E(y - E(y|x))^2 = E(E(y - E(y|x))^2|x) = E(\text{var}(y|x))$

7. The meanings for math terms are as follows: $E(\text{var}(y|x))$ is the average conditional variance, or it entails comparing y to level- i mean. We call this **within** variation. By contrast, $\text{var}(E(y|x))$ is the variance of conditional means, and it involves comparing level- i mean to level- j mean. We call that **between** variation
8. Under the null hypothesis of equal conditional means we have $\text{var}(E(y|x)) = \text{var}(\mu) = 0$. Under the alternative hypothesis $\text{var}(E(y|x)) \neq 0$. This contrast suggests comparing $\text{var}(E(y|x))$ to $E(\text{var}(y|x))$ with a ratio given by

$$\frac{\text{var}(E(y|x))}{E(\text{var}(y|x))} = \frac{\text{between variation}}{\text{within variation}} \quad (6)$$

That is the intuition for the F test.

9. Those population conditional moments in (6) are of course unknown, and need to be estimated by its sample counterparts. Toward that end, let's derive a sample version of the variance decomposition (2). Let i be the index for level ($i = 1$ for center, $i = 2$ for forward, $i = 3$ for guard); and j be the index for observation for the given i -the level (given $i = 1$, $j = 1$ for the first center player, $j = 2$ for the second center player, and so on). Let n_i be the sample size for level- i subgroup. We can define the overall unconditional mean (e.g., mean for all players) as

$$\bar{y} = \frac{\sum_i \sum_j y_{ij}}{\sum_i n_i}$$

and the level- i conditional mean (e.g., mean for all center players) as

$$\bar{y}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$$

The sum square (SS) decomposition is

$$\sum_i \sum_j (y_{ij} - \bar{y})^2 = \sum_i n_i (\bar{y}_i - \bar{y})^2 + \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 \quad (7)$$

Proof

$$\sum_i \sum_j (y_{ij} - \bar{y})^2 = \sum_i \sum_j (y_{ij} - \bar{y}_i + \bar{y}_i - \bar{y})^2 \quad (8)$$

$$= \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 + \sum_i \sum_j (\bar{y}_i - \bar{y})^2 + 2 \sum_i \sum_j (y_{ij} - \bar{y}_i)(\bar{y}_i - \bar{y}) \quad (9)$$

$$= \sum_i \sum_j (y_{ij} - \bar{y}_i)^2 + \sum_i n_i (\bar{y}_i - \bar{y})^2 \quad (10)$$

We may describe the sum square decomposition as

Total Sum Square (TSS) = Within Sum Square (WSS) + Between Sum Square (BSS)

where $TSS \equiv \sum_i \sum_j (y_{ij} - \bar{y})^2$, $WSS \equiv \sum_i \sum_j (y_{ij} - \bar{y}_i)^2$, $BSS \equiv \sum_i n_i (\bar{y}_i - \bar{y})^2$

10. Let df denote degree of freedom. The F test for the null hypothesis (1) is computed as

$$F = \frac{\sum_i n_i (\bar{y}_i - \bar{y})^2 / (k - 1)}{\sum_i \sum_j (y_{ij} - \bar{y}_i)^2 / (n - k)} = \frac{BSS / df.\text{between}}{WSS / df.\text{within}} \quad (11)$$

where $n = \sum_i n_i$ is total number of observations, and k is the number of levels (for position, $k = 3$).

11. When the null hypothesis is true, $\bar{y}_i - \bar{y} \approx 0$, $BSS \approx 0$, so the numerator of F test is close to zero. Thus, a big F test or small p value rejects the null hypothesis, where the p value is from an F distribution.
12. Using NBA data, the F test for the null hypothesis (1) or one way ANOVA is provided by **aov** function in R

```
> library("readxl")
> setwd("C:/Users/lij14/Dropbox/311r")
> data = read_excel("311_nba.xls")
> data = subset(data, is.na(wage)==F)
```

```

> attach(data)
> summary(aov(wage~factor(position)))
              Df Sum Sq Mean Sq F value Pr(>F)
factor(position)  2   5.68   2.8395    2.873 0.0582 .
Residuals        277 273.79   0.9884

```

Given the p value 0.0582 we can reject the null hypothesis (1) at the 10% level.

13. To see how to compute the F test 2.873, use the codes below

```

> mean(wage)
[1] 1.42027
> tapply(wage,position,mean)
      center forward  guard 
1.661021 1.477314 1.273954
> tapply(wage,position,length)
      center forward  guard 
        47      112      121
> ss.between = (1.661021-1.42027)^2*47+(1.477314-1.42027)^2*112+(1.273954-1.42027)^2*121
> df.between = 2
> ss.within = sum((wage[position=="center"]-1.661021)^2)+sum((wage[position=="forward"]-1.477314)^2)+sum((wage[position=="guard"]-1.273954)^2)
> df.within = length(wage)-3
> f = (ss.between/df.between)/(ss.within/df.within)
> f
[1] 2.872805

```

where the unconditional mean is $\bar{y} = 1.42027$. We get conditional mean \bar{y}_i and n_i using **tapply**. Then we apply formula (11). Note that R displays the following sum squares in **aov** output

```

> ss.between
[1] 5.679032
> ss.within
[1] 273.7902

```

Mean squares are computed as sum squares divided by degree of freedom. The F test is the ratio of two mean squares

$$2.873 = \frac{5.68/2}{273.79/277} = \frac{2.8395}{0.9884}$$

14. Next I will explain how ANOVA is related to a dummy variable regression. First, rewrite the null hypothesis (1) as

$$H_0 : \mu_{forward} - \mu_{center} = 0, \mu_{guard} - \mu_{center} = 0$$

The two differences of conditional means are β_1 and β_2 in the following regression

$$wage = \beta_0 + \beta_1 forward + \beta_2 guard + u \quad (12)$$

where the dummy variable *forward* equals one if a player is forward, and dummy variable *guard* equals one if a player is guard. To avoid dummy variable trap we exclude the dummy variable for center. Thus, center is based group, and is captured by β_0

```
> summary(lm(wage~factor(position)))$coef
              Estimate Std. Error  t value    Pr(>|t|)
(Intercept)    1.6610213   0.1450174  11.453944 4.000835e-25
factor(position)forward -0.1837070   0.1727864  -1.063203 2.886161e-01
factor(position)guard  -0.3870676   0.1708764  -2.265190 2.427344e-02
> summary(lm(wage~factor(position)))$fstat[1]
2.872816
```

It is reassuring to see that the F value 2.8728 reported by this dummy variable regression is the same as the F value based on formula (11). This finding is expected because the regression-based F test is about

$$H_0 : \beta_1 = 0, \beta_2 = 0$$

which is the same as testing $H_0 : \mu_{forward} - \mu_{center} = 0, \mu_{guard} - \mu_{center} = 0$

15. In practice the regression-based Wald test is preferred because formula (11) is valid only

when homoskedasticity holds. It is easy to account for heteroskedasticity by reporting a heteroskedasticity-robust Wald test.

16. It is straightforward to conduct a two-way ANOVA where two categorical variables are treatments. For instance, the categorical variable “marr” is the second treatment, and it equals one if a player is married.

```
> summary(aov(wage~factor(position)+marr))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
factor(position)	2	5.68	2.840	2.945	0.05427 .
marr	1	7.66	7.658	7.942	0.00518 **
Residuals	276	266.13	0.964		

The F value for marr is computed as

$$7.942 = \frac{7.658/1}{266.13/276}$$

17. The F value for marr 7.942 can be equivalently obtained from a regression:

```
> m = lm(wage~factor(position)+marr)
> library("car")
> linearHypothesis(m, c("marr"))
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	277	273.79				
2	276	266.13	1	7.6581	7.9421	0.005179 **

The regression-based F value is computed as

$$7.942 = \frac{(273.79 - 266.13)/1}{266.13/276}$$

where 273.79 is the RSS of restricted regression that drops marr

```
> m.r = lm(wage~factor(position))
> sum(resid(m.r)^2)
[1] 273.7902
```

and 266.13 is the RSS of unrestricted regression

```
> m.u = lm(wage~factor(position)+marr)
> sum(resid(m.u)^2)
[1] 266.1321
```

18. Finally, analysis of covariance (ANCOVA) is about using a categorical treatment and quantitative control variable to explain outcome. We call the corresponding multiple regression DVR II. For instance, DVR II below relates wage to position and points (the quantitative control)

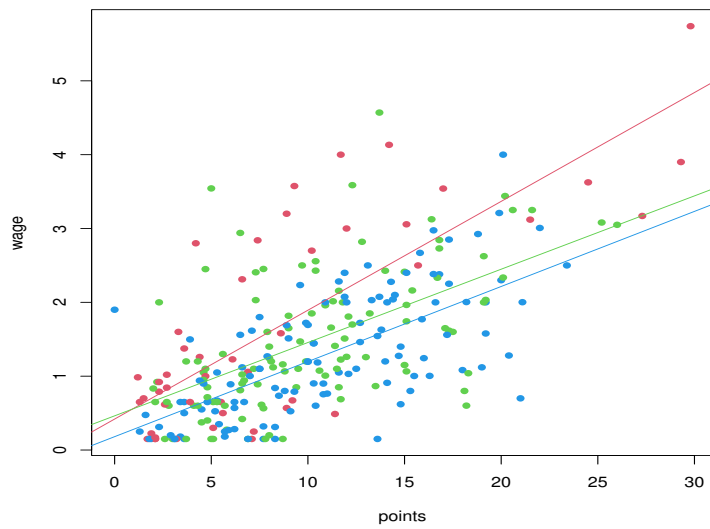
```
> summary(lm(wage~factor(position)+points+factor(position):points))$coef
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.41944639	0.15970290	2.6264169	9.123537e-03
factor(position)forward	0.04360719	0.21783127	0.2001879	8.414849e-01
factor(position)guard	-0.23845826	0.21824598	-1.0926124	2.755415e-01
points	0.14739586	0.01416269	10.4073314	1.545759e-21
factor(position)forward:points	-0.04810815	0.01913995	-2.5134938	1.253889e-02
factor(position)guard:points	-0.04570421	0.01877657	-2.4341090	1.557956e-02

we see that relative to center, there are no differences in intercepts for forward $t = 0.2001879$ and guard $t = -1.0926124$. However, the differences in slopes are significant— $t = -2.5134938, -2.4341090$, which implies that as points rises wages of forward and guard grow more slowly than center.

19. We can visualize the treatment effect of position on wage (after controlling for points) as

```
plot(points, wage, pch=16, col=ifelse(position=="center", 2, ifelse(position=="forward", 3, 4))
abline(lm(wage[position=="center"]~points[position=="center"]), col=2)
abline(lm(wage[position=="forward"]~points[position=="forward"]), col=3)
abline(lm(wage[position=="guard"]~points[position=="guard"]), col=4)
```



20. In short, ANOVA and ANCOVA can be conducted by running F tests based on regressions that involve dummy variables.