ANCOVA

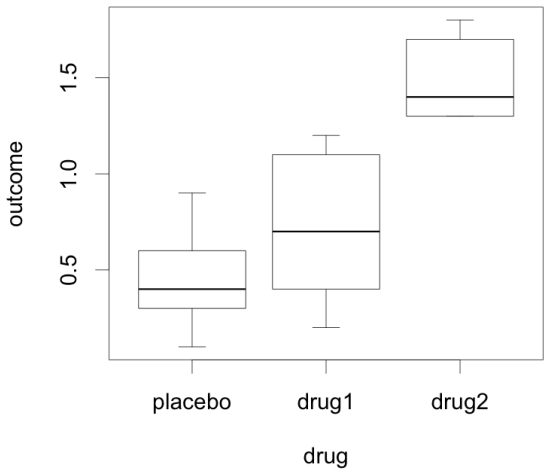
2018-02-08
Andres Kroker

Let's assume we have the following data:

- Outcome score of drug/therapy
- Drug given to patient
- Therapy given to patient

Score	Drug	Therapy
0.5	placebo	no
0.3	Placebo	no
0.1	placebo	no
0.6	drug 1	no
0.4	drug 1	no
0.2	drug 1	no
1.4	drug 2	no
1.7	drug 2	no
1.3	drug 2	no
0.6	placebo	yes
0.9	Placebo	yes
0.3	placebo	yes
1.1	drug 1	yes
0.8	drug 1	yes
1.2	drug 1	yes
1.8	drug 2	yes
1.3	drug 2	yes
1.4	drug 2	yes

For now let's just look at the relationship between **drug** and **score.** With the predictor being categorical we can look at the data as a box plot.



Naturally, we want to run an ANOVA. Actually, this is the very example we used in the ANOVA lecture, so here is a quick recap.

ANOVA

H_o: $Y_{ik} = \mu + \varepsilon_{ik}$
H_1: $Y_{ik} = \mu_k + \varepsilon_{ik}$

where *i* denotes the *i*th observation, and *k* the *k*th drug group.

$$SS_T = SS_M + SS_R$$

we know that:

$$SS_T = \sum_{k=1}^{G} \sum_{i=1}^{n_k} (Y_{ik} - \bar{Y})^2$$

where G is the total number of drug groups. Also

$$SS_R = \sum_{k=1}^{G} \sum_{i=1}^{n_k} (Y_{ik} - \bar{Y}_k)^2$$

$$SS_M = \sum_{k=1}^{G} n_k (\bar{Y}_k - \bar{Y})^2$$

Let's calculate F:

$$MS_M = \frac{SS_M}{df_m} \quad \text{, with } df_m = G\text{-}1$$

$$MS_R = \frac{SS_R}{df_R} \quad \text{, with } df_R = n - G$$

which in our case results in:
$MS_M = \frac{SS_M}{df_m} = \frac{3.453}{3-1} = 1.7267$ and $MS_R = \frac{SS_R}{df_R} = \frac{1.392}{18-3} = 0.0928$ with $F = \frac{MS_M}{MS_R} = \frac{1.7267}{0.0928} = 18.61$

To summarize: $F(2, 15)=18.61$, $p<0.001$

This math looks identical to last weeks MLR lecture. So let's try to represent this ANOVA model as an MLR

ANOVA model as MLR

First we need to find a way to represent a categorical variable numerically.

Drug	X.1	X.2
placebo	0	0
drug 1	1	0
drug 2	0	1

We need G-1 dummy variables, as the reference 'level' placebo is encoded as (0, 0). Let's updated the data

Score	Drug	Therapy	X.1	X.2
0.5	placebo	no	0	0
0.3	Placebo	No	0	0
0.1	placebo	no	0	0
0.6	drug 1	no	1	0
0.4	drug 1	no	1	0
0.2	drug 1	no	1	0
1.4	drug 2	no	0	1
1.7	drug 2	no	0	1
1.3	drug 2	no	0	1
0.6	placebo	yes	0	0
0.9	Placebo	yes	0	0
0.3	placebo	yes	0	0
1.1	drug 1	yes	1	0
0.8	drug 1	yes	1	0
1.2	drug 1	yes	1	0
1.8	drug 2	yes	0	1
1.3	drug 2	yes	0	1
1.4	drug 2	yes	0	1

Now we can create the linear model

$$Y_i = b_2 X_{i.2} + b_1 X_{i.1} + b_0 + \varepsilon_i$$

But what does this encoding mean in regards to our linear model? Let's take a look at our estimated model and the cases of our dummy variable combinations.

$$\hat{Y}_i = \hat{b}_2 X_{i.2} + \hat{b}_1 X_{i.1} + \hat{b}_0$$

Case 1: Xi.1=0; Xi.2=0 $\rightarrow \hat{b}_0$ is the mean of the reference group

$$\begin{aligned} \hat{Y}_i &= \hat{b}_2 X_{i.2} + \hat{b}_1 X_{i.1} + \hat{b}_0 \\ &= \hat{b}_2 0 + \hat{b}_1 0 + \hat{b}_0 \\ &= \hat{b}_0 \end{aligned}$$

Case 2: Xi.1=0; Xi.2=0 $\rightarrow \hat{b}_1$ is the difference in means of the reference group and group 1

$$\begin{aligned} \hat{Y}_i &= \hat{b}_2 X_{i.2} + \hat{b}_1 X_{i.1} + \hat{b}_0 \\ &= \hat{b}_2 0 + \hat{b}_1 1 + \hat{b}_0 \\ &= \hat{b}_1 + \hat{b}_0 \end{aligned}$$

Case 3: Xi.1=0; Xi.2=0 $\rightarrow \hat{b}_2$ is the difference in means of the reference group and group 2

$$\begin{aligned}\hat{Y}_i &= \hat{b}_2 X_{i.2} + \hat{b}_1 X_{i.1} + \hat{b}_0 \\ &= \hat{b}_2 1 + \hat{b}_1 0 + \hat{b}_0 \\ &= \hat{b}_2 + \hat{b}_0\end{aligned}$$

Now let's perform the whole model test on our model.

H_o: $Y_i = b_0 + \varepsilon_i$
H_1: $Y_i = b_2 X_{i.2} + b_1 X_{i.1} + b_0 + \varepsilon_i$

Comparing 2 nested models → F-test

$$SS_T = SS_{M0} + SS_{R0}$$
$$SS_T = SS_{M1} + SS_{R1}$$

Thus, the model sums of squares are

$$SS_{M0} = SS_T - SS_{R0}$$
$$SS_{M1} = SS_T - SS_{R1}$$

We want to know if there is a difference between the two models

$$\Delta SS_M = SS_{M1} - SS_{M0} = SS_{R0} - SS_{R1} = \sum_{i=1}^{n} (\hat{Y}_i^{(1)} - \hat{Y}_1^{(0)})^2$$

$$MS_{\Delta M} = \frac{\Delta SS_M}{df_{\Delta M}} = \frac{3.4533}{2-0} = 1.72667$$

with

$$df_{\Delta M} = p_1 - p_0$$

were p_0 is the number of predictors (or dummy variables) model H_0 and p_1 is the number of predictors (or dummy variables) of model H_1

$$MS_R = \frac{SS_R}{df_R} = \frac{1.3917}{18-2-1} = 0.09278$$

with

$$df_R = n - p_1 - 1$$

Thus, our F-statistic is

$$F = \frac{MS_{\Delta M}}{MS_R} = 18.611$$

To summarize: $F(2, 15)=18.61$, $p<0.001$

Here is the whole thing in R

```
> model1 = aov(outcome~drug, data=data)
summary(model1)

            Df Sum Sq Mean Sq F value   Pr(>F)
drug         2  3.453  1.7267   18.61 8.65e-05 ***
Residuals   15  1.392  0.0928
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


data$dummy1=as.numeric(data$drug=='drug1')
data$dummy2=as.numeric(data$drug=='drug2')
model2 = lm(outcome ~ dummy1 + dummy2, data=data)
summary(model2)

Call:
lm(formula = outcome ~ dummy1 + dummy2, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.51667 -0.17500 -0.08333  0.20000  0.48333

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.4500     0.1244   3.619  0.00253 **
dummy1        0.2667     0.1759   1.516  0.15021
dummy2        1.0333     0.1759   5.876 3.05e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3046 on 15 degrees of freedom
Multiple R-squared:  0.7128,     Adjusted R-squared:  0.6745
F-statistic: 18.61 on 2 and 15 DF,  p-value: 8.646e-05
```

If you want a full ANOVA table, you need to create null hypothesis model and then create an ANOVA table comparing it with the alternative model

```
model3 = lm(outcome ~ 1, data=data)
anova(model3, model2)

Analysis of Variance Table

Model 1: outcome ~ 1
Model 2: outcome ~ dummy1 + dummy2
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     17 4.8450
2     15 1.3917  2    3.4533 18.611 8.646e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

If you define the ‘drug’ variable as a factor, R know which model to define as null hypothesis and will run the tests for you automatically. So you can just type the following three commands, and get the same data tables. Just do not forget to define ‘drug’ as a factor with a reference level ‘placebo’.

```
model4 = lm(outcome~drug, data=data)
summary(model4)
anova(model4)
```

ANCOVA (Analysis of Covariance)

Now that we have seen that ANOVA is just a special case of MLR, why should we limit ourselves to analyzing either the relationship of continuous or categorical predictor to a continuous outcome variable. Adding both to the model is what we call ANCOVA.

Let's take a look at the following data

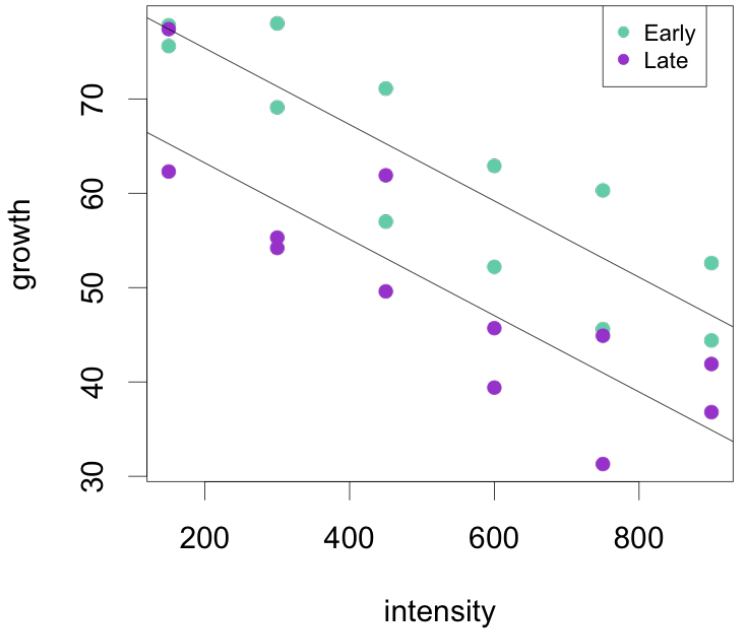
Plant growth	Light intensity	Light timing	X.E
77.8	150	Early	1
75.6	150	Early	1
69.1	300	Early	1
78.0	300	Early	1
57.0	450	Early	1
71.1	450	Early	1
62.9	600	Early	1
52.2	600	Early	1
60.3	750	Early	1
45.6	750	Early	1
52.6	900	Early	1
44.4	900	Early	1
62.3	150	Late	0
77.4	150	Late	0
55.3	300	Late	0
54.2	300	Late	0
49.6	450	Late	0
61.9	450	Late	0
39.4	600	Late	0
45.7	600	Late	0
31.3	750	Late	0
44.9	750	Late	0
36.8	900	Late	0
41.9	900	Late	0

Let's run our ANCOVA model

```
> model5 = lm(growth ~ light + intensity, data=data)
summary(model5)

Call:
lm(formula = growth ~ light + intensity, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-9.652 -4.139 -1.558  5.632 12.165

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  83.464167   3.273772  25.495  < 2e-16 ***
lightLate    12.158333   2.629557   4.624 0.000146 ***
intensity    -0.040471   0.005132  -7.886 1.04e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 6.441 on 21 degrees of freedom
Multiple R-squared:  0.7992,     Adjusted R-squared:   0.78
F-statistic: 41.78 on 2 and 21 DF,  p-value: 4.786e-08
```

....so what does this model actually mean? Let's take a look at the data and plot the model over it.



Where are the two lines coming from? Turns out, this particular ANCOVA model is a parallel lines model with 2 lines. Here is why.

$$Y_i = b_2 X_{i.I} + b_1 X_{i.E} + b_0 + \varepsilon_i$$

Case 1: Xi.E = 0

$$\begin{aligned} \hat{Y}_i &= \hat{b}_2 X_{i.I} + \hat{b}_1 X_{i.E} + \hat{b}_0 \\ &= \hat{b}_2 X_{i.I} + \hat{b}_1 0 + \hat{b}_0 \\ &= \hat{b}_2 X_{i.I} + \hat{b}_0 \end{aligned}$$

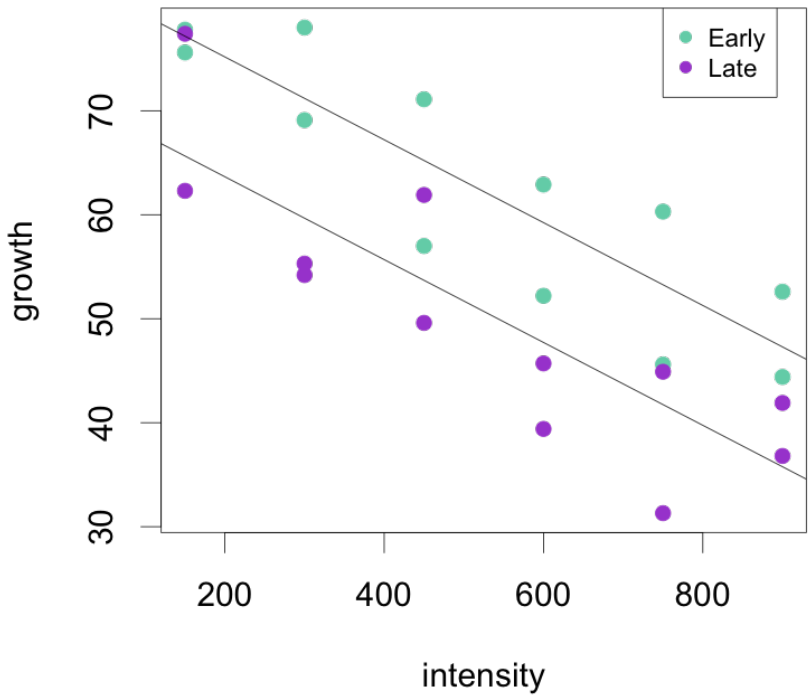
Case 1: Xi.E = 1

$$\begin{aligned} \hat{Y}_i &= \hat{b}_2 X_{i.I} + \hat{b}_1 X_{i.E} + \hat{b}_0 \\ &= \hat{b}_2 X_{i.I} + \hat{b}_1 1 + \hat{b}_0 \\ &= \hat{b}_2 X_{i.I} + (\hat{b}_1 + \hat{b}_0) \end{aligned}$$

→ 2 lines with the same slope $\hat{b}_2$ but two different intercepts. 2 parallel lines.

But now that we think about it, could light intensity depend on the time of the day? And thus differ early mornings to late afternoons? Let's add an interaction term.

```
> model6 = lm(growth ~ light + intensity + light:intensity, data=data)
summary(model6)

Call:
lm(formula = growth ~ light + intensity + light:intensity, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-9.516 -4.276 -1.422  5.473 11.938

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          83.146667   4.343305  19.144 2.49e-14 ***
lightLate            11.523333   6.142360   1.876   0.0753 .
intensity            -0.039867   0.007435  -5.362 3.01e-05 ***
lightLate:intensity  -0.001210   0.010515  -0.115   0.9096
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.598 on 20 degrees of freedom
Multiple R-squared:  0.7993,     Adjusted R-squared:  0.7692
F-statistic: 26.55 on 3 and 20 DF,  p-value: 3.549e-07
```



while in this case the interaction was not significant, we can see that the slopes are now different, but what does the coefficient of the interaction term mean?

$$Y_i = b_3 X_{i.I} X_{i.E} + b_2 X_{i.I} + b_1 X_{i.E} + b_0 + \varepsilon_i$$

Case 1: Xi.E = 0

$$\begin{aligned}\hat{Y}_i &= \hat{b}_3 X_{i.I} X_{i.E} + \hat{b}_2 X_{i.I} + \hat{b}_1 X_{i.E} + \hat{b}_0 \\ &= \hat{b}_3 X_{i.I} 0 + \hat{b}_2 X_{i.I} + \hat{b}_1 0 + \hat{b}_0 \\ &= \hat{b}_2 X_{i.I} + \hat{b}_0\end{aligned}$$

Case 1: Xi.E = 1

$$\begin{aligned}\hat{Y}_i &= \hat{b}_3 X_{i.I} X_{i.E} + \hat{b}_2 X_{i.I} + \hat{b}_1 X_{i.E} + \hat{b}_0 \\ &= \hat{b}_3 X_{i.I} 1 + \hat{b}_2 X_{i.I} + \hat{b}_1 1 + \hat{b}_0 \\ &= (\hat{b}_3 + \hat{b}_2) X_{i.I} + (\hat{b}_1 + \hat{b}_0)\end{aligned}$$

The coefficient $\hat{b}_3$ is the difference in slope between the line in case 1 and the line in case 2. Now you can apply standard MLR techniques to e.g. test if the interaction adds any valuable information by comparing the two models with one another

```
> anova(model6, model5)

Analysis of Variance Table

Model 1: growth ~ light + intensity + light:intensity
Model 2: growth ~ light + intensity
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     20 870.66
2     21 871.24 -1  -0.57604 0.0132 0.9096
```