STAT 425

ANCOVA Models

ANCOVA Models

- ANCOVA stands for ANalysis of COVAriance: These are regression problems where some predictors are quantitative (i.e. numerical) and some are qualitative (i.e. categorical).
- For simplicity we will focus on examples with just two predictors: X (numerical) and D (categorical).

A two-level example

- Suppose we model the response Y by two predictors X and D, where X is a numerical variable and D is categorical with two levels (such as male, female)
- You can code D as 0 or 1,e.g.,1 for male and 0 for female. Note: you can code the two levels using any two different values, which will not change $\hat{y}$, but the interpretation of the estimated coefficients.
- In general, a factor with k levels corresponds to $k-1$ variables, when there is an additional intercept.

Consider the general model:

$$y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 (x \cdot d) + e$$

The cats example revisited

We want to build a model to predict *Hwt* based on *Bwt*. For simplicity, assume we have $n = 4$ observations and the first two are female. What are the possible regression models?

1. Coincident regression lines (the simplest model): the same regression line for both groups, i.e., the categorical variable D has no effect on Y:

$$y = \beta_0 + \beta_1 x + e$$

1' Two-mean model (another simplest model): the numerical variable X has no effect on Y:

$$y = \beta_0 + \beta_2 d + e = \begin{cases} \beta_0 + e, & d = 0 \\ (\beta_0 + \beta_2) + e, & d = 1 \end{cases}$$

2. Parallel regression lines: the categorical variable D only changes the intercept, i.e., it produces an additive effect only:

$$y = \beta_0 + \beta_2 d + \beta_1 x + e = \begin{cases} \beta_0 + \beta_1 x + e, & d = 0 \\ (\beta_0 + \beta_2) + \beta_1 x + e, & d = 1 \end{cases}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & x_1 \\ 1 & 0 & x_2 \\ 1 & 1 & x_3 \\ 1 & 1 & x_4 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_2 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

β_2: measures the change of the additive effect (i.e., difference of the intercept).

Alternative choices for the design matrix (they should give us the same $\hat{y}$)

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & x_1 \\ 1 & 0 & x_2 \\ 1 & 1 & x_3 \\ 1 & 1 & x_4 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_2 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & 1 & x_1 \\ 1 & 1 & x_2 \\ 1 & 2 & x_3 \\ 1 & 2 & x_4 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_2 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

3. Regression lines with equal intercepts but different slopes: the categorical variable D only changes the effect of X on Y:

$$\begin{aligned} y =& \beta_0 + \beta_1 x + \beta_3 (x \cdot d) + e \\ =& \begin{cases} \beta_0 + \beta_1 x + e, & d = 0 \\ \beta_0 + (\beta_1 + \beta_3) x + e, & d = 1 \end{cases} \end{aligned}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & x_1 & 0 \\ 1 & x_2 & 0 \\ 1 & x_3 & x_3 \\ 1 & x_4 & x_4 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

β_3: measures the change of the slope.

4. Unrelated regression lines (the most general model): the categorical variable D produces an additive change in Y and also changes the effect of X on Y. Then, should we just divide the data into two sets and run *lm* separately on them?

$$
\begin{aligned}
y =& \beta_0 + \beta_1 x + \beta_2 d + \beta_3 (x \cdot d) + e \\
=& \begin{cases} \beta_0 + \beta_1 x + e, & d = 0 \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x + e, & d = 1 \end{cases}
\end{aligned}
$$

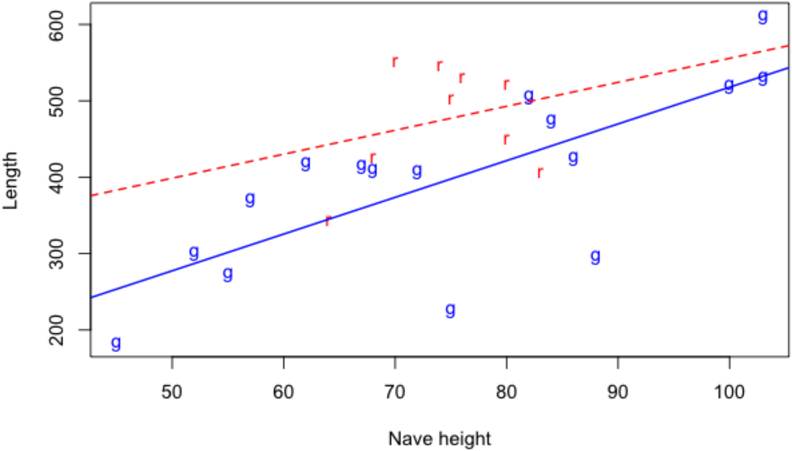
$$
\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & x_1 & 0 \\ 1 & 0 & x_2 & 0 \\ 1 & 1 & x_3 & x_3 \\ 1 & 1 & x_4 & x_4 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_2 \\ \beta_1 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}
$$

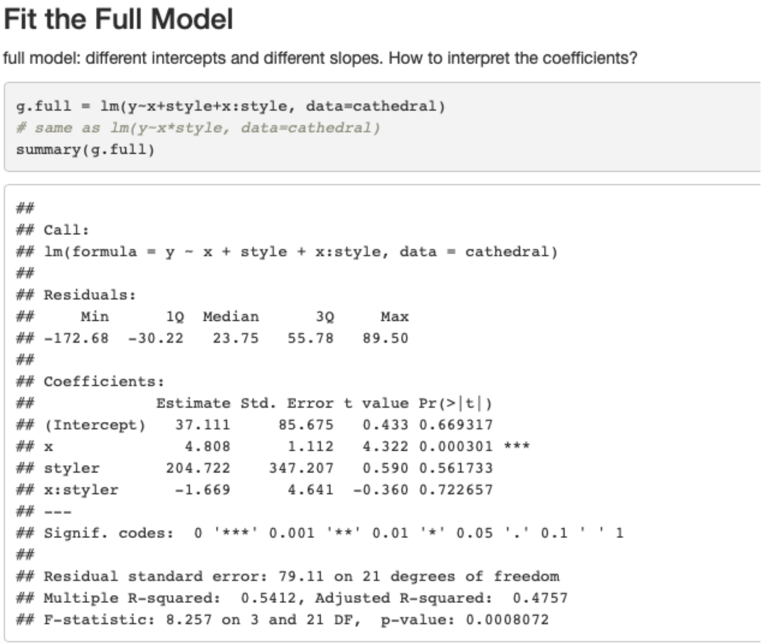
How to interpret the LS coefficients from model 4?

- The usual: β_1 *measures the effect of* X_1 *on* Y *when other predictors are held unchanged*, does not make much sense for models with interactions. We cannot change x while holding d and $(x \cdot d)$ unchanged.
- Check the Cathedral **R** example.

Example: cathedral data set

$x=$ nave height; $y=$ total length in feet for English medieval cathedrals. r represents Romanesque style and g represents Gothic style.





Fit the Full Model

full model: different intercepts and different slopes. How to interpret the coefficients?

```
g.full = lm(y~x+style+x:style, data=cathedral)
# same as lm(y~x*style, data=cathedral)
summary(g.full)
```

```
##
## Call:
## lm(formula = y ~ x + style + x:style, data = cathedral)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -172.68  -30.22   23.75   55.78   89.50
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   37.111     85.675   0.433 0.669317
## x              4.808      1.112   4.322 0.000301 ***
## styler       204.722    347.207   0.590 0.561733
## x:styler      -1.669      4.641  -0.360 0.722657
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 79.11 on 21 degrees of freedom
## Multiple R-squared:  0.5412, Adjusted R-squared:  0.4757
## F-statistic: 8.257 on 3 and 21 DF,  p-value: 0.0008072
```

Which model to pick?

You can use F-test to select the appropriate model.

- First test whether the interaction term is significant:

  $$H_0 : \text{model } 2 \quad H_a : \text{model } 4$$

  If reject the null, stop and take model 4.
  Otherwise, decide whether you can further reduce model 2 to model 1 or model $1'$.

- What if β_3 (the interaction) is significant, but, β_1 or β_2, is not significant? What about model 3?

- The Hierarchical Rule for interactions: an interaction term will be included in a model only if all its main effects have been included. Due to this rule, we would include both β_1 and β_2, once β_3 is significant.
- In practice we could test $\beta_1 = 0$ or $\beta_2 = 0$. We just need to understand what the model looks like when β_1 or β_2 equals zero.

- When $\beta_1 = 0$ (it doesn't mean that X is not significant):

$$y = \begin{cases} \beta_0 + e, & d = 0 \\ (\beta_0 + \beta_2) + \beta_3 x + e, & d = 1 \end{cases}$$

- When $\beta_2 = 0$ (gives us model 3; it does not mean D is not significant):

$$y = \begin{cases} \beta_0 + \beta_1 x + e, & d = 0 \\ \beta_0 + (\beta_1 + \beta_3) x + e, & d = 1 \end{cases}$$

A Multi-level example

- Model the response Y by two predictors X and D, where X is a numerical variable and D is categorical with k levels.
- We need to generate $k-1$ dummy variables: $D_2, \ldots, D_k$ where:

$$D_i = \begin{cases} 0 & \text{if not level i} \\ 1 & \text{if level i} \end{cases}$$

Level 1 is the reference level.

The main purpose of the analysis is to decide which of the following models fits the data:

- Model 0: $Y \sim 1$
- Model 1a: $Y \sim X$
- Model 1b: $Y \sim D$
- Model 2: $Y \sim D + X$
- Model 4: $Y \sim D + X + D : X$

The major tool is the F-test. Note that when D has more than two levels, the difference between model parameter number may not be one, so t-test is no longer appropriate.

1) Compare models:

$$H_0 : Y \sim X + D \quad vs. \quad H_a : Y \sim D + X + D : X$$

If the interaction $D : X$ is significant, stop.

2a) If X is significant, keep X.

2b) If D is significant, keep D.

3) If neither X nor D are significant, report the intercept model $Y \sim 1$.

2a) and 2b) are a little bit tricky:

2a) Is X is significant?
Test the marginal contribution of X:

$$H_0 : Y \sim 1 \quad vs. \quad H_a : Y \sim X$$

Test the contribution of X in addition to D:

$$H_0 : Y \sim D \quad vs. \quad H_a : Y \sim X + D$$

2b) Is D is significant?

$$H_0 : Y \sim 1 \quad vs. \quad H_a : Y \sim D$$

$$H_0 : Y \sim X \quad vs. \quad H_a : Y \sim D + X$$

Sequential ANOVA

We can use the **anova** function to get sequential F-tests. The sequence of F -tests given by $anova(lm(Y \sim X + D + X : D))$

H_0	H_a
$Y \sim 1$	$Y \sim X$
$Y \sim X$	$Y \sim X + D$
$Y \sim X + D$	$Y \sim X + D + X : D$

The sequence of F -tests given by $anova(lm(Y \sim D + X + X : D))$ is given by:

H_0	H_a
$Y \sim 1$	$Y \sim D$
$Y \sim D$	$Y \sim D + X$
$Y \sim D + X$	$Y \sim D + X + X : D$

Be aware that: Some of the F-stats and p-values from the sequential ANOVA table are different from the ones we calculated based on usual F-test (we learned) for comparing two nested models.

Suppose we want to compare:

$$H_0 : Y \sim X \quad vs \quad H_a : Y \sim X + D$$

- The usual F-stat is given by:

$$\frac{(RSS_0 - RSS_a)/(k-1)}{RSS_a/(n-p_a)} = \frac{(RSS_0 - RSS_a)/(k-1)}{\hat{\sigma}_a^2}$$

  which follows $F_{k-1,n-p_a}$ under the null hypothesis. k is the total number of categories of variable D

- The F-stat from the sequential ANOVA table:

$$\frac{(RSS_0 - RSS_a)/(k-1)}{RSS_A/(n-p_A)} = \frac{(RSS_0 - RSS_a)/(k-1)}{\hat{\sigma}_A^2}$$

which follows $F_{k-1,n-p_A}$ under the null hypothesis, where RSS_A denotes the RSS from the biggest model $Y \sim X + D + X : D$ and $p_A = 2k$

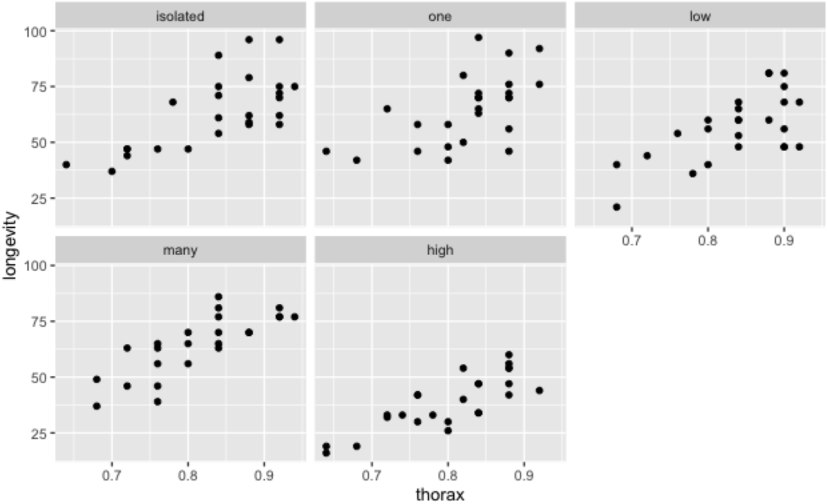
Example: fruitfly data set

The *fruitfly* data frame has 9 rows and 3 columns. 125 fruit flies were divided randomly into 5 groups of 25 each. The response is the longevity of the fruit fly in days. The following groups or categories describe the sexual activity:

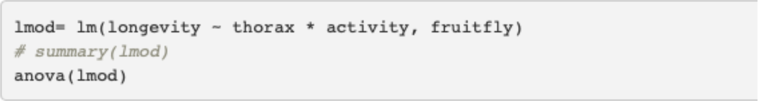
- One group was kept solitary (isolated)
- One group was kept with a virgin female each day (low)
- One group was kept with 8 virgin females per day (high)
- One group was kept with one pregnant female per day (one)
- One group was kept with eight pregnant female per day (many)

Pregnant fruit flies will not mate. The thorax length of each male was measured as this was known to affect longevity. One observation in the many group has been lost. So the total sample size is 124.

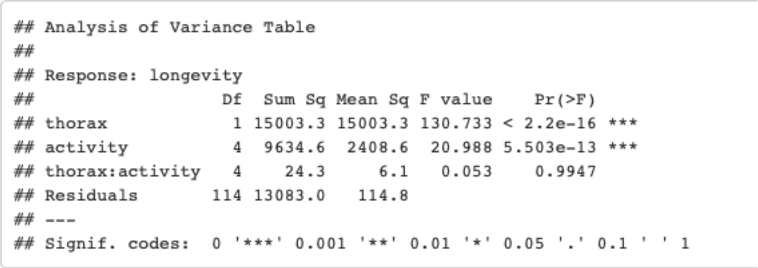
Response: Longevity (days); Predictors: Thorax length (numerical) and activity (categorical)



Sequential ANOVA



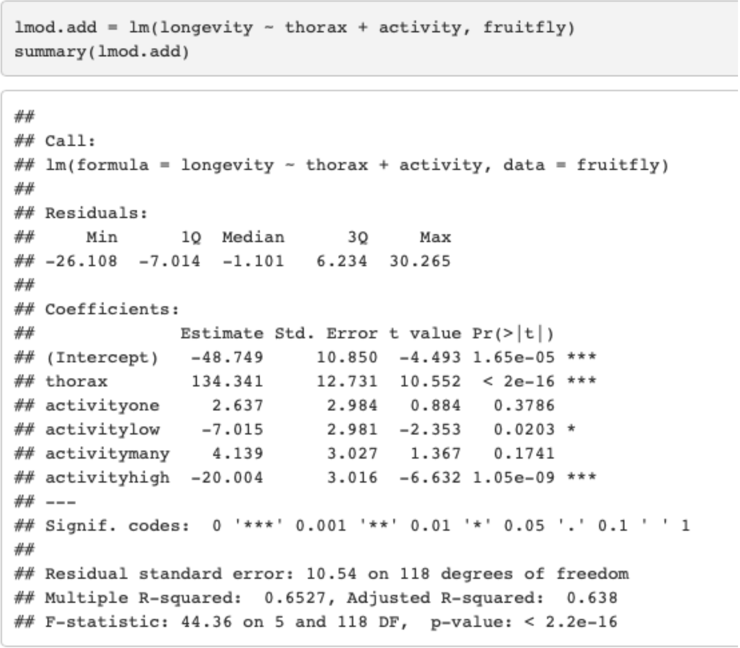
```
lmod= lm(longevity ~ thorax * activity, fruitfly)
# summary(lmod)
anova(lmod)
```



```
## Analysis of Variance Table
##
## Response: longevity
##                   Df  Sum Sq Mean Sq F value    Pr(>F)
## thorax             1 15003.3 15003.3 130.733 < 2.2e-16 ***
## activity           4  9634.6  2408.6  20.988 5.503e-13 ***
## thorax:activity    4    24.3     6.1   0.053    0.9947
## Residuals        114 13083.0   114.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: F-stat for the activity variable: $F = \frac{9634.6/4}{13083.0/114} = 20.988$. Under the null hypothesis that $\beta_{activity}$ is non-significant, $F \sim F_{4,114}$. From these results we conclude that the interaction term is not significant.

Additive Model



```
lmod.add = lm(longevity ~ thorax + activity, fruitfly)
summary(lmod.add)
```

```
##
## Call:
## lm(formula = longevity ~ thorax + activity, data = fruitfly)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.108  -7.014  -1.101   6.234  30.265
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -48.749     10.850  -4.493 1.65e-05 ***
## thorax        134.341     12.731  10.552  < 2e-16 ***
## activityone     2.637      2.984   0.884   0.3786
## activitylow    -7.015      2.981  -2.353   0.0203 *
## activitymany    4.139      3.027   1.367   0.1741
## activityhigh  -20.004      3.016  -6.632 1.05e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.54 on 118 degrees of freedom
## Multiple R-squared:  0.6527, Adjusted R-squared:  0.638
## F-statistic: 44.36 on 5 and 118 DF,  p-value: < 2.2e-16
```