ANCOVA

- ANCOVA $=$ ANalysis of COVAriance: regression problems where some predictors are quantitative (i.e., numerical) and some are qualitative (i.e., categorical).

- For simplicity, focus on examples where we have just two predictors: X (numerical) and D (categorical).

A Two-Level Example

- 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 or female).

- 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.

Recall the `cats` data, where we want to build a model to predict `Hwt` based on `Bwt`. For simplicity, assume $n = 4$ and first two are female.

What are the possible regression models?

1. Coincident regression line (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 only an additive effect.

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

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & x_1 \\ 1 & 0 & x_2 \\ 1 & 1 & x_3 \\ 1 & 1 & x_4 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_2 \\ \beta_1 \end{pmatrix} + \mathbf{e}$$

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 $\hat{y}$)

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 & x_1 \\ 1 & 0 & x_2 \\ 0 & 1 & x_3 \\ 0 & 1 & x_4 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_2 \\ \beta_1 \end{pmatrix} + \mathbf{e}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & 1 & x_1 \\ 1 & 1 & x_2 \\ 1 & 2 & x_3 \\ 1 & 2 & x_4 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_2 \\ \beta_1 \end{pmatrix} + \mathbf{e}$$

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

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

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & x_1 & 0 \\ 1 & x_2 & 0 \\ 1 & x_3 & x_3 \\ 1 & x_4 & x_4 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_3 \end{pmatrix} + \mathbf{e}$$

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?

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

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

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.
- Let's look at the Cathedral Example.

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 $\beta_1 = 0$ (doesn't mean 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; doesn't mean D is not significant)

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

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 1: $Y \sim X$
- Model 1': $Y \sim D$
- Model 2: $Y \sim D + X$
- Model 4: $Y \sim D + X + D : X$

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

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

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

2) If X is significant, keep X.

2') If D is significant, keep D.

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

2) and 2') are a little tricky.

2) Is X is significant?

Test the marginal contribution of X

$$H_0 : Y \sim 1, \quad H_a : Y \sim X$$

Test the contribution of X in addition to D

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

2') Is D is significant?

$$H_0 : Y \sim 1, \quad H_a : Y \sim D$$

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

The Sequential ANOVA

The sequence of F-tests given by `anova(lm(Y ~ 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 ~ D + X + X:D))`

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

Here is the catch: 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

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

- The usual F-stat

$$\frac{(\mathsf{RSS}_0 - \mathsf{RSS}_a)/(k-1)}{\mathsf{RSS}_a/(n-p_a)} = \frac{(\mathsf{RSS}_0 - \mathsf{RSS}_a)/(k-1)}{\hat{\sigma}_a^2}$$

which follows $F_{k-1,n-1-p}$ under the null.

- The F-stat from the sequential ANOVA table

$$\frac{(\mathsf{RSS}_0 - \mathsf{RSS}_a)/(k-1)}{\mathsf{RSS}_A/(n-p_A)} = \frac{(\mathsf{RSS}_0 - \mathsf{RSS}_a)/(k-1)}{\hat{\sigma}_A^2}$$

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