13.13 The Linear Model for ANOVA

- In the regression models we have studied so far, the independent variable(s) have been numerical variables that we treat as either continuous in nature, or if discrete, having a large number of levels.
- We could also conceive of a linear model in which the independent variable is a <u>factor</u> with only a few <u>levels</u> of interest.
- Such models often arise in experimental studies in which the experimenter sets some factor at a few different levels and observes how a related response variable varies.

Example: An agricultural scientist applies four different types of fertilizer to various plots in a field and then measures the resulting yields of corn crops.

Example: We compare the mean amounts of active ingredients across five different brands of aspirin.

- A linear model with only factors (i.e., having a few categories of interest) as its independent variables is called an Analysis of Variance (ANOVA) model.
- A linear model with only a single factor is called a One-way ANOVA model.
- The primary goal of the one-way ANOVA is to compare $E(Y)$ across the k levels of the factor.

- The one-way ANOVA model can be written:

where n_i is the number of observations at the i-th level of the factor;

μ_i is the population mean response for an observation at the i-th level

and

- To fit this as a linear model, we define (k−1) <u>dummy variables</u>.
- One level is designated as the <u>baseline level</u> (R uses the first level as the baseline, SAS uses the last level as the baseline).

- The dummy variables take on values 0 or 1:
- Letting the k-th level be the baseline, define:

- If an observation comes from level k, then

for that observation.

Example: 15 subjects were categorized into 4 groups: Aerobics Class Exercisers (A), Walkers (W), Joggers (J), and Sedentary (S). After 5 weeks, depression scores were measured for each subject:

A	43.7, 54.2, 51.5, 47.7
W	46.7, 58.2, 50.5, 45.6
J	58.6, 51.1, 49.1, 54.9
S	68.4, 57.7, 56.4

- The linear model is

Let "S" be the baseline level. Then

- Of interest is whether the mean depression score varies across the four groups.
- We can answer this by testing

with a Complete vs. Reduced F-test.

- If we reject this H_0, we conclude there is a significant difference in mean depression across groups.
- We have

- Matrix algebra will show that

and

- We see that $\hat{\beta}_0 =$

- Also, as we might expect:

  $\hat{\beta}_1 =$

  $\hat{\beta}_2 =$

  $\hat{\beta}_3 =$

- Our test of $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ has test statistic and P-value =
- If $\alpha = 0.10$, we could ______________ and conclude the mean depression score

- To compare each pair of level means, we could use the Bonferroni simultaneous comparison method discussed in Sec 13.12 (see R code on course web page).
- STAT 515 and especially STAT 516 discuss ANOVA models in much greater detail.