Notes for Tuesday January 20, 1997

1 ANOVA/ANCOVA and Designed Experiments

In this part of the course we will learn about ANOVA models, and their application to designed experiments. "ANOVA" is one of the most confusing terms in applied statistics.

1.1 What ANOVA is not, Part I

You already know about the ANOVA table used to describe a multiple regression.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

we might build the table

Source	*SS*	*df*	*MS*
X_1	$SSR(X_1)$=17.16	1	17.16
$X_2 \vert X_1$	$SSR(X_2 \vert X_1)$=33.67	1	33.67
$X_3 \vert X_1, X_2$	$SSR(X_3 \vert X_1, X_2)$= 3.98	1	3.98
Regression	$SSR(X_1, X_2, X_3)$=54.80	3	18.27
Error	$SSE(X_1, X_2, X_3)$=27.67	$n-4=107$	0.26
Total	SST=82.47	$n-1=110$	

which we can read off almost directly from the SPLUS output

```
402 > anova(mlogoz.lm)
Analysis of Variance Table

Response: logoz

Terms added sequentially (first to last)
           Df Sum of Sq  Mean Sq  F Value        Pr(F)
      rad   1  17.15661 17.15661  66.3508 0.0000000000
     temp   1  33.66815 33.66815 130.2069 0.0000000000
     wind   1   3.97777  3.97777  15.3835 0.0001552777
Residuals 107  27.66744  0.25857
```

This sort of ANOVA table will be useful when we build and assess ANOVA models, but ANOVA is not merely this table.

1.2 What ANOVA is not, Part II

Someone who understands regression models (like me) will often say, *"oh well, ANOVA, that's just multiple regression where all the predictor variables are dummy variables!"*. This is mathematically correct (and we will exploit this idea often!) but ANOVA modeling is much more than just running `lm()` with a bunch of dummy variables for X's.

1.3 What ANOVA is

Analysis of variance models (ANOVA models) are used to study the relationship between predictor variables that *are not* necessarily quantitative and response variables that *are* quantitative. They are structured like linear regression models, and have the same basic assumptions, but they do not assume a linear relationship between the predictor and response variables.

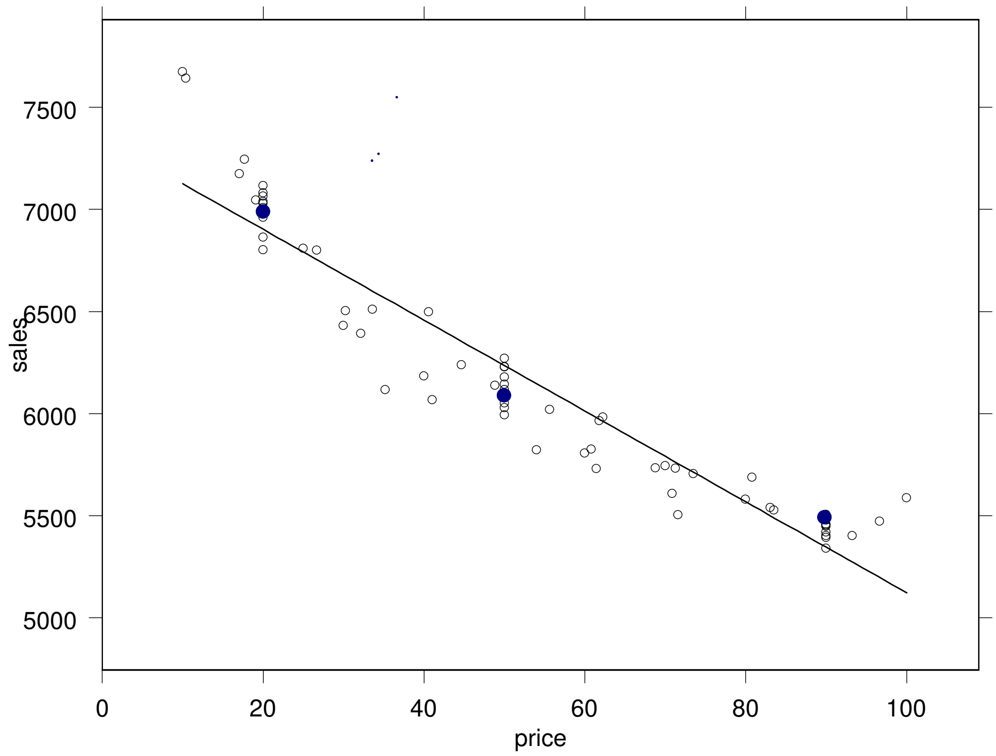
Example.

Suppose we want to study the relationship between X = *the price of a product*, and Y = *the sales volume in one month.*

As a linear regression problem, this is simple. We want to fit the model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

But this explicitly assumes that the relationship between price and sales is linear, and this need not be the case, as shown in the figure (for about 70 months' worth of data).



An alternative to fitting the complex nonlinear models we talked about at the end of class last time is to choose a few price points, such as \$20.00, \$50.00, and \$90.00, and define dummy variables

$$X_1 = \begin{cases} 1 & \text{if price=\$20.00} \\ 0 & \text{else} \end{cases} \qquad X_2 = \begin{cases} 1 & \text{if price=\$50.00} \\ 0 & \text{else} \end{cases} \qquad X_3 = \begin{cases} 1 & \text{if price=\$90.00} \\ 0 & \text{else} \end{cases}$$

and then fit the regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \tag{1}$$

This is an example of a *"one-way ANOVA" model.*

There are three things to note about this approach:

- Even though the model is still the familiar linear regression model, it is possible to detect and explore nonlinearity in the relationship between X and Y, by looking at just these three price levels, or perhaps by adding two or three more price levels to get a better look at the curve. This can be much more efficient than fitting a polynomial or nonlinear regression model.
- It may also be more efficient to collect data for the one-way ANOVA model than for the general regression model; maybe we can get 5 months' worth of data at each of three price points, and have the data collected in a year or so. It is much harder to say how to collect data for this problem if all possible prices are allowed.
- In matrix form this (naive form of the) ANOVA model is

$$\underline{y} = \begin{bmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n_1 1} \\ y_{12} \\ y_{22} \\ \vdots \\ y_{n_2 2} \\ y_{13} \\ y_{23} \\ \vdots \\ y_{n_3 3} \end{bmatrix} = X\underline{\beta} + \underline{\epsilon} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ & \vdots & & \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ & & \vdots & \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ & & & \vdots \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{n_1 1} \\ \epsilon_{12} \\ \epsilon_{22} \\ \vdots \\ \epsilon_{n_2 2} \\ \epsilon_{13} \\ \epsilon_{23} \\ \vdots \\ \epsilon_{n_3 3} \end{bmatrix}$$

  showing that the model is overparameterized (linearly dependent columns).

  All ANOVA models are overparametrized in their "natural" form, and we often must make a choice about what a sensible "reduced" parametrization is. □

2 One-way ANOVA analysis in SPLUS

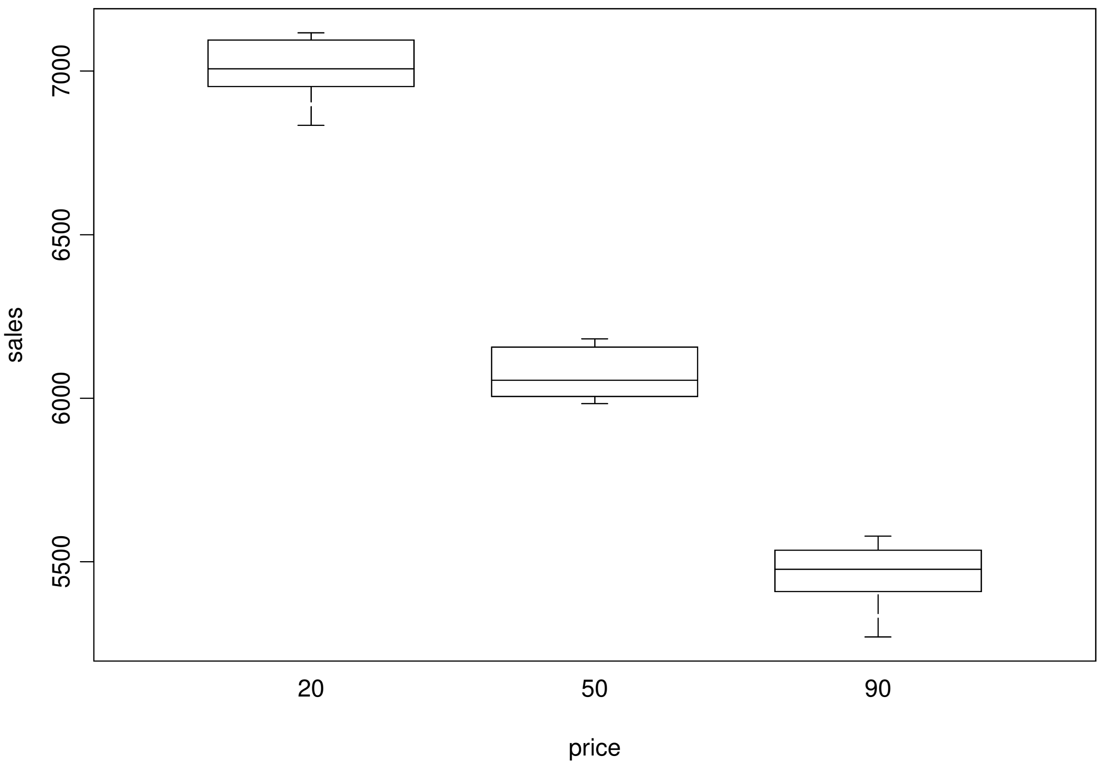
We can illustrate the one-way analysis of variance using the sales data:

```
402 > attach(salesdata)
402 > sales205090 _ salesdata[((price==20)|(price==50)|(price==90)),]
402 > sales205090$price _ factor(sales205090$price)
402 > table(sales205090$price)  # There are 11 months of sales at each price
 20 50 90
 11 11 11

402 > view(sales205090$price)

      structure(.Data = c(1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3,
      1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3), .Label =
      c("20", "50", "90"), class = "factor")

402 > plot.factor(sales205090,style.bxp="old")
```




```
402 > sales.aov _ aov(sales ~ price,data=sales205090)
402 > anova(sales.aov)
Analysis of Variance Table

Response: sales

Terms added sequentially (first to last)
```

```
          Df Sum of Sq Mean Sq  F Value Pr(F)
    price  2  13332276 6666138 895.2989     0
Residuals 30    223371    7446
402 >  13332276/( 13332276 +  223371 )  # calculate R^2
[1] 0.9835219
```

This tells us that the "price" variable explains much of the variability in the "sales" data, but isn't very informative about details, like what the mean sales was at each price level, etc.

To obtain this, we need to think about the model in a different way. In ANOVA notation, we would write the model (1) as

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

where $i = 1, 2$ or 3, depending on whether the price was 20, 50 or 90, and j indexes the 11 observations at each price level.

Nomenclature:

- μ is called the *overall mean* or *grand mean*
- α_1, α_2 and α_3 are called "effects". Just to be confusing, they are also called "contrasts."

Note that there are four parameters but only three groups of data—cells—from which to estimate them.

We can add one linear constraint to take away the overparametrization. The usual one for the "grand means plus effects" model is to make the effects sum to zero:

$$\sum_{i=1}^{\#cells} \alpha_i = 0$$

The $\alpha's$ can then be estimated with the `model.tables()` function:

```
402 > model.tables(sales.aov,type="effects")
Refitting model to allow projection
Tables of effects

 price
    20     50     90
 826.3 -106.6 -719.7
402 > sum(scan())        # verify that the alpha's sum to zero
1:  826.3 -106.6 -719.7
4:
[1] -1.136868e-013
```

An alternative parameterization with just three parameters is to take out the intercept and just estimate the mean sales for each price level, this is called the *cell means* model:

$$y_{ij} = \mu_i + \epsilon_{ij}$$

Again, we can estimate the cell means with `model.tables()`:

```
02 > model.tables(sales.aov,type="means")
Refitting model to allow projection
Tables of means
Grand mean

 6185

 price
   20   50   90
 7011 6078 5465
```

Let's compare these results with fitting the model "by hand" using `lm()`.

```
402 > lmsales _ sales205090$sales
402 > lmprice _ sales205090$price
402 > p20 _ ifelse(lmprice==20,1,0)
402 > p50 _ ifelse(lmprice==50,1,0)
402 > p90 _ ifelse(lmprice==90,1,0)
402 > sales.lm _ lm(lmsales ~ p20 + p50 + p90)
Error in lm.fit.qr(x, y): computed fit is singular, rank 3
Dumped
402 > # oops, forgot to delete a column
402 > sales.lm _ lm(lmsales ~ p50 + p90)
402 > summary(sales.lm)

Call: lm(formula = lmsales ~ p50 + p90)
Residuals:
    Min     1Q Median    3Q   Max
 -194.8 -55.67 -4.224 78.53 113.8

Coefficients:
                 Value Std. Error   t value  Pr(>|t|)
(Intercept)  7010.8662    26.0170  269.4728    0.0000
        p50  -932.8187    36.7936  -25.3528    0.0000
        p90 -1545.9577    36.7936  -42.0171    0.0000

Residual standard error: 86.29 on 30 degrees of freedom
Multiple R-Squared: 0.9835
F-statistic: 895.3 on 2 and 30 degrees of freedom, the p-value is 0

Correlation of Coefficients:
    (Intercept)     p50
p50 -0.7071
p90 -0.7071      0.5000
```

Actually, when `lm()` sees a `factor` variable in a regression, it will construct dummy variables automatically, but unfortunately it then transforms the X matrix to make the columns orthogonal; we will learn more about SPLUS's coding of dummy variables later.

3 Uses of ANOVA models

ANOVA models are used to study the effects of the independent (predictor) variable(s) on the dependent (response variable), without making assumptions about the "shape" of the relationship. They are used in randomized experiments, where the assumption of random uncorrelated (and perhaps normal) errors in the regression model is carefully built into the data by the experimenter, and in observational studies, where the assumption of random uncorrelated (perhaps normal) errors must be argued into the problem.

Many of the ideas in this section were taken from Chapter 14 of

Neter, J., Wasserman, W., and Kutner, M. H. (1990). *Applied linear statistical models.* Homewood IL: Irwin.

3.1 Experimental and Observational Studies

Example 1.

The effect of four cooking temperatures on the fluffiness of omelets prepared from a mix was studied by randomly assigning five packages of mix to each of four cooking temperatures. This is an *experiment* because the independent variable of interest (cooking temperature) was controlled by the researcher. The experimental design was a *completely randomized design*, since the 20 packages of mix were randomly distributed among the cooking temperatures; this (together with some care in performing each of the 20 "runs") guarantees that the errors in the ANOVA/regression model are random and independent; it may *not* guarantee that they are normal, or have the same variability in each of the four cells (cooking temperatures) of the design.

In a real industrial setting (say, to decide what instructions to write on boxes of omelet mix), probably all four cooking temperatures produce somewhat fluffy omelets. What is of interest is which one is better, not what the overall average fluffiness is. Since this is a common situation in ANOVA applications, ANOVA software often doesn't bother to report the grand mean, even in the "grand mean plus effects" parameterization.

If the packages were also chosen randomly from the company's output, it may be reasonable to assume that these results will be generalizable to the hypothetical population of all boxes of omelet mix of this type produced by this company. □

Example 2.

A study of the effects of education and type of past employment experience of salespeople on their sales volumes was made by selecting a random sample of salespeople currently employed at the company and obtaining information on highest degree obtained, type of experience, and sales volume for each selected employee. This was an *observational study* because the independent variables (education, type of experience) was not controlled by the experimenter. Random sampling does guarantee that the errors will be random and independent (but again normality and equal variances across cells is harder).

Because of the random sampling the results of this survey may generalize to a hypothetical population of employees like those working in sales for this company right now. However, it seems less likely that this is a population of interest—and even if it is, that it will remain stable over time so that the results can be used with future employees—than in **Example 1**.

Example 3.

An appliance manufacturer operates three training centers in the United States for training mechanics to service the company's products. At each center, two different training programs were studied, with trainees randomly assigned to each program. The outcome is a score on a standardized repair exam. □

3.2 Factors

The word *factor* is just another word for independent variable. Usually factors are qualitative variables; the different values that a factor can have (such as sales prices of \$20, \$50 or \$90) are called *levels* of the factor.

- In **Example 1** there is just one factor: the cooking temperature. It has four, ordered levels.
- In **Example 2** there are two factors: Education, and type of experience. Education could be a *quantitative factor* whose values are just the number of years of education the salesperson has had. More likely it is an *ordered qualitative factor* with levels like *High School*, *College Degree*, *Master's Degree*, and *PhD*. Type of past employment experience is probably an *unordered qualitative factor*, with levels like *laborer*, *office work*, *sales*, *consulting*, etc.
- In **Example 3** again there are two qualitative factors (both unordered): which center the employee is at, and which training program the employee was assigned to.

ANOVA applied to single-factor studies is called *one-way ANOVA*. ANOVA applied to multi-factor studies is called *multi-way ANOVA*.

Many studies have a mix of "experimental" and "observational" components. In a study, the factors that were controlled and randomly assigned by the researcher are called *experimental factors*. Those that merely classify subjects, but are not under the control of the researcher, are called *classification factors*. **Example 3** is an example with both an **experimental factor** (which of the two programs the examinee was assigned to) and a **classification factor** (which training center the employee was at).

If we tried to analyze **Example 3** as a one-way ANOVA using only the training program, we might mask some important differences among the centers (suppose at Center A, method 2 is much better than method 1, but at centers B and C, method 1 is moderately better than method 2; if the data is "pooled across centers" then we might conclude that the two methods are equally effective.

If we analyze **Example 3** as a two-way ANOVA, we still must be careful: if method 1 is better than method 2 at all three sites, the evidence in favor of method 1 is strong. On the other hand, if method 2 does better at some centers and worse at others, these differences are hard to interpret without getting more information: perhaps one center has better teachers for method 2, or some centers are in geographic locations where better education is provided. At least the differences among levels of the classification factor (training site) suggest what sort of information is needed to understand the results.

3.3 Treatments

In a one-factor study, the treatments are the same as the levels of the factor. In a multifactor study, each unique combination of factors is a treatment. For example, a two-factor study of three kinds of deodorant and four levels of exercise has twelve treatments.

Choosing which treatments (which factors, how many levels, etc.) to study is not easy. This is part of experimental design, which we will discuss more in a couple of weeks. For now, here are some main points to keep in mind:

- There are generally two goals of a study or series of studies: one is to understand a process, and the other is to make a prediction or comparison.

  Studies that are intended to help in understanding a process may start out small: the initial investigation may involve only a few factors, or a few levels of the factors; and then after the main factors of interest have been identified a more complete study may be undertaken, to provide a more detailed picture of the process.

  Studies that are intended to help answer a specific practical question tend to have fewer factors, but even here including some extra levels to help provide some understanding of the findings is useful. For example if your company is considering changing to a new manufacturing process that is expected to allow you to increase the number of units per day that you produce, and you want to study defects in the manufactured products, you may want to try include not only the "old" method under the current production schedule, the "new" method under the current production schedule, and the "new" method under the accellerated production schedule. This way you can disentangle the effect of the change in process from the effect of speeding up the production schedule, on the number of defects you see.

- The definition of "What is a treatment" is not always obvious. This is especially clear in educational studies. Suppose two teaching methods were being compared, and different teachers were using the different methods. Since the teacher has an effect on how well a method is implemented, it's may not be clear whether the treatment is just the teaching method, or a combination of teacher and method.

- A *control treatment* consists of withholding all active treatments but still applying the same experiemental procedures as would be needed for the active treatments. Controls are needed, for example,

  - when the general effectiveness of all the active treatments is not known or not consistent; or
  - in studies where "halo", "placebo", or "experimenter bias" effects could make an ineffective active treatment look effective.

  Some studies, such as comparisons of a new treatment for a disease vs. a current, already effective treatment for the disease, do not need a control since there is an effective treatment standard, the current treatment, to which we may compare. In such studies it may also be unethical to include a control treatment (withholding treatment from those in need).

3.4 Units of Study

Finally, in addition to ambiguities in what constitutes a "treatment" and which treatments should be administered, there may be substantial ambiguity about what constitues a "unit of study" in an experiment or observational study.

The "units of study" are the rows in the Y vector and X matrix in the matrix form of the regression equations.

Example.

> In a study of a new method of teaching mathematics, school district officials were able to get 60 students in each of nine schools to participate. In each school, 30 children were randomly assigned to the "current" method, and 30 were assigned to the "new" method. Three teachers taught three classes each of the "current" method classes , and three more taught three classes each of the "new" method classes. □

In the example, $60 \times 9 = 540$ children participated in the study, with 270 children experiencing each teaching method. We might think of these children as the experimental units. However,

- Because the children interact with one another in the classroom (in both methods), their responses to the end-of-semester test will be dependent on one another within each classroom (even if no one cheats on the exam!). Thus perhaps we should think of the classrooms as the experimental units; now there are only 9 units in the "current" method and 9 units in the "new" method.
- Finally, perhaps the teacher also has an influence on the outcome of the students; then all the children who have the same teacher will have dependent responses, so perhaps classes within teacher should be the units of study; now there are only 3 units in the "current" method, and 3 units in the "new" method.

This example is frustrating because each of the possible units: children, classrooms, and teachers, contributes something to the outcome. Mixed effects linear models, which we will discuss later, provide one way to "hedge our bets" across the three possible definitions of units in this problem.

Nevertheless, identifying and clearly stating what the units of study are is an important point in both communication and in understanding what the data are telling us.