Indicator Variables & the Regression Approach to ANOVA

Dr. Austin Brown

Kennesaw State University

Introduction

- Some of the materials in today's lecture were adapted from materials created by Dr. Taasoobshirazi and my former professor, Dr Khalil Shafie (thanks Drs S & T!).

Introduction

- When we think about linear regression in a traditional sense, we often are discussing the relationship between some continuous response and some set of continuous predictors.

- However, there are lots of instances one can imagine where a categorical predictor would also being useful in improving the explanatory power of the regression model.
  - For example, in the Boston dataset, when we were looking at the relationship between median home price and number of rooms, we can reasonably infer that the neighborhood a home is in is likely also a contributing factor to the variability of home price.

- We can probably think of lots of instances where we'd want to include categorical predictors. How do we go about doing that? Through the use of *indicator* or *dummy* variables.

Indicator Variables

- Suppose we want to predict MLB player salary using the number of years they've been in the league and also whether they are a pitcher or a fielder (player position).

- Player position is obviously categorical. To sort of "brute force" it to be quantitative, we say, alright, if you're a pitcher, that will be denoted with a value of "1" and if you're a fielder, that will be denoted with a value of "0"
  - This simple concept is how an indicator or dummy variable is created.

$$\text{Player Position}_i = \begin{cases} 1 & \text{if Pitcher} \\ 0 & \text{if Fielder} \end{cases}$$

Indicator Variables

- So our regression model, assuming years in the league is x_1 and position is x_2, can be written as:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

- What we've done is, in effect, create two different albeit parallel, regression lines.

Indicator Variables

- Here's our model for pitchers (dropping ε_i for convenience's sake):

$$y_i = \beta_0 + \beta_2 + \beta_1 x_{1i}$$

- And here's our model for fielders:

$$y_i = \beta_0 + \beta_1 x_{1i}$$

Indicator Variables

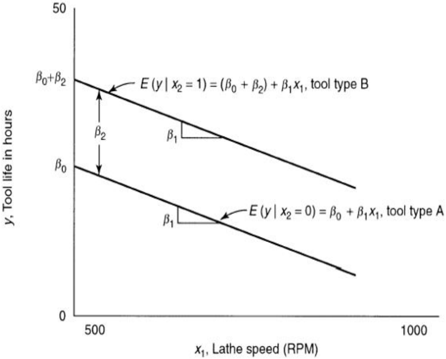

Figure 8.1 Response functions for the tool life example.

Figure 1: Fig 8.1 in your text

Indicator Variables

- Okay, well that's not too bad for a categorical variable with two levels. What happens if we have a categorical variable with say, three levels? What do we do then?

- For example, suppose we want to parse out fielders in our MLB salary example, into infielders and outfielders.

- We would create two indicator variables in this case: say one is for pitchers, one is for infielders, and if they're both 0, that indicates outfielders.

Indicator Variables

Table 1: Dummy Variable Assignment

x_2	x_3	Player Position
1	0	Pitcher
0	1	Infielder
0	0	Outfielder

- So now our regression model is:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

Indicator Variables

- Similar to the previous example, we are creating effectively three different regression lines all with the same slope (β_1), but differing intercepts.

- In general, this concept can be extended to a categorical variable with any finite number of levels. If a categorical predictor has k levels, then we'll create $k - 1$ indicator variables for it.

- Now, you may be asking yourself, "well, why do we do $k - 1$? Why not create an indicator variable for all k?"
  - Great question!

Indicator Variables

- There is a practical and a technical reason why we do it this way (sort of like with everything!):
- Practically, as you saw, $k-1$ indicator variables fully specifies our observations into their respective levels. So there's not really a need for k variables.
- More technically, k indicator variables would give us perfect multicollinearity, and thus, our X^TX matrix would no longer be invertible (remember, we need that property to uniquely solve the normal equations).
  - Let's see why with a quick illustration using a categorical predictor with two levels.

Indicator Variables

$$X = \begin{bmatrix} 1 & x_{11} & 1 & 0 \\ 1 & x_{12} & 1 & 0 \\ 1 & x_{13} & 0 & 1 \\ 1 & x_{14} & 0 & 1 \end{bmatrix}$$

- If we add together columns 3 and 4, we get column 1. So column 1 is a perfect linear combination of two other columns meaning we have perfect multicollinearity so $(X^TX)^{-1}$ doesn't exist.

- Let's look at a couple of examples in R.

Indicator Variables

- In the prior examples, we assumed that each level of the categorical predictor had the same slope, but just different intercepts. Is it always reasonable to assume that?
  - Maybe sometimes, but maybe not in others.
- How can we use a regression model in order to control for not only a difference in intercepts, but also a possible difference in slopes?
- Fortunately, we can do this through the inclusion of an interaction effect.

Indicator Variables

- Let's go back to first baseball example. Suppose we wanted to see if pitchers and fielders have not only different intercepts, but also different slopes. To check on this, we would include an additional variable in the model which is the product of years in the league and our indicator variable.
  - This is called an "interaction effect."

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i}$$

- How does this determine if the slopes differed (which practically means, the nature of the linear relationship between salary and years in the league)?

Indicator Variables

- Model for pitchers:

$$y_i = \beta_0 + \beta_2 + \beta_1 x_{1i} + \beta_3 x_{1i}$$

$$y_i = \beta_0 + \beta_2 + (\beta_1 + \beta_3) x_{1i}$$

Indicator Variables

- Model for fielders:

$$y_i = \beta_0 + \beta_1 x_{1i}$$

- From this, we can see that a significant β_2 implies a difference in intercepts (practically meaning that one position makes more, on average, than the other) and that a significant β_3 would imply a difference in slopes (again, which practically means that the nature of the linear relationship between salary and years in the league is different between the position groups).

- Let's see how we can do this in R.

Regression Approach to ANOVA

- You may recall from prior courses a commonly used statistical method called "ANOVA" (an acronym for analysis of variance) where we are interested in comparing groups.

- For example, in a one-way ANOVA model, we are comparing group means across the levels of a single categorical variable.
  - From the baseball example, we may want to know if pitchers, infielders, or outfielders have differing mean salaries.

Regression Approach to ANOVA

- The typical fixed, treatment-effect ANOVA model is:

$$y_i = \mu + \alpha_j + \varepsilon_i$$

- where μ is the overall mean of our y_i's and α_j is the difference between the jth group's mean and the overall mean, μ (e.g., $\bar{y}_j = \mu + \alpha_j \implies \alpha_j = \bar{y}_j - \mu$).

Regression Approach to ANOVA

- Interestingly, we have the ability to convert the standard fixed effect ANOVA model into a regression model using indicator variables. Considering the baseball example where we want to compare salaries across three groups:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

- where x_{1i} is an indicator variable for pitchers and x_{2i} is an indicator variable for infielders.

Regression Approach to ANOVA

- With this approach:

Table 2: Comparison of ANOVA & Regression

Mean of Group	ANOVA	Regression
Outfielders	$\mu + \alpha_1$	β_0
Pitchers	$\mu + \alpha_2$	$\beta_0 + \beta_1$
Infielders	$\mu + \alpha_3$	$\beta_0 + \beta_2$

- Let's see how this works using an example in R.

Regression Approach to ANOVA

- This can be extended to higher order ANOVA models (e.g., two-way ANOVA, three-way ANOVA, etc.).

- So what are these two approaches doing differently in theory? Remember, in regression, the overall (or omnibus) F-test is testing the hypotheses:

$$H_0 : \beta_j = 0 \quad \forall j$$

$$H_1 : \beta_j \neq 0 \quad \text{for at least one } j$$

- More practically, it's asking the question: is at least one level of my categorical variable significantly different than my reference level? Then the *t* test statistics and their associated p-values are like pairwise independent *t*-tests between each level of the categorical variable and the reference group/level.

Regression Approach to ANOVA

- Whereas with one-way ANOVA, our hypotheses are slightly different:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_j$$

$$H_1 : \mu_i \neq \mu_j \quad \text{for at least one pair of } i \text{ and } j$$

- So why would I prefer one over the other? The answer is that it really depends on what your goals are. Regression is more geared toward prediction and less toward group comparison whereas ANOVA is just the opposite.

Regression Approach to ANOVA

- Now, when considering model building in regression, where our goal is to identify the best subset of variables which predict or fit our outcome, we may run into instances where some of the β coefficients for a categorical predictor aren't significant whereas others are. What do we do then?

- Consider omitting, say, outfielders from our initial analysis using years in the league and player position to predict player salary. Is is appropriate to omit outfielder salaries all together? No! You'd want to recode outfielders into one of the other levels of player position in order to retain all of our observations.

- What happens if there's not good rationale behind grouping two (or more) levels of a categorical variable together? Then don't. At least that's my opinion.