Project 2: Equity in Athletics Data Analysis

McCourt School of Public Policy, Georgetown University

Overview

This project will focus on categorical variables. In quant class, you've likely discussed categorical variables such as race and highest level of education completed.

We will be investigating data collected from college athletic programs. Look for Download Data files on the page linked below. http://ope.ed.gov/athletics (http://ope.ed.gov/athletics) Download the "Data for academic year 2013-2014"

Alternatively, the direct link is: http://ope.ed.gov/athletics/dataFiles/EADA%202013-2014.zip (http://ope.ed.gov/athletics/dataFiles/EADA%202013-2014.zip)

Background information on this data can be found here: http://www2.ed.gov/finaid/prof/resources/athletics/eada.html (http://www2.ed.gov/finaid/prof/resources/athletics/eada.html)

Week 2:

Key Ideas:

- Interaction terms

Key Commands / Concepts:

- factor variables (help fvvarlist)
- encode
- recode
- separate

Questions

2.12 Sport profitability

- Last week we looked at two continuous outcome variables: total revenue and total expenditures.
- This week, we will look at total profit.
- Create a new variable for total profit for each sports team, equal to (total revenue - total

expenditure).

- Create a bar graph of the average profit for each type of sports team.
- Make sure you choose the appropriate orientation, vertical or horizontal.
- Are most college sports teams profitable?
- Which type of sports team is the most profitable, on average?

2.13 Sport categories

- We are interested in the profitability of different types of sports, but there are currently too many different sports here.
- We need to group some sports together into larger categories.
- First, use the `encode` command to create a labeled numeric variable named `sportid`.
- Then use the `recode` command to generate a new variable named `sportcat` containing the categories below.
- See the final example given on the `recode` help page for a useful tip on organizing a long recode command.
- You may want to use `label list` to list the value labels of the `sportid` variable before starting the recode.
- Sport Categories (Some sports will be in their own category)
    - Baseball/Softball
    - Basketball
    - Football
    - Soccer
    - Other
- Verify your variable creation with the following command: `bysort sportcat: tab sportid, m`
- Create a bar graph of total revenue and total expenses over the new sport categories.

2.14 Female Head Coach

- Investigate the following two variables:
    - `SUM_FTHDCOACH_FEM` Number of full-time female head coaches for sports team
    - `SUM_PTHDCOACH_FEM` Number of part-time female head coaches for sports team
- Create a new indicator variable, `femhdcoach`, that is equal to one if a team has either a full-time or part-time female head coach, and zero otherwise.
- Create a bar graph showing average profit for teams with a female head coach and for those without a female head coach.
- Are teams with a female head coach profitable, on average?

2.15 Categorical-Categorical Interactions

- As we saw last week, the idea behind dummy variables is to allow the average value of a variable to change over categories.

- Categorical-categorical interaction terms allow the average value of a variable to change for every unique combination of the categories.
- Suppose we want to look at the average profit of teams with and withoug a female head coach, separately for each sport category.
- You can visualize these relationships in a bar graph with multiple over() categories.
- Create a bar graph of average profitability over `femhdcoach`, then over `sportcat`.
- Is the relationship between having a female head coach and team profit the same for each sport category?

2.16 Factor Variable Interactions

- Interaction terms, like dummy variables, can be included in a regression using factor variable notation.
- To include interaction terms, just include both variable names, jointed by the symbol `#`.
- You also need to include the dummy variables using `i.` notation, so each variable name will appear twice.
- Run a regression of profit on sport category dummies, the female head coach dummy, and interactions for all sports.
- One interaction term is omitted, because there are no football teams with female head coaches.

2.17 Testing Interactions

- Testing factor variable interaction terms works just like testing factor variable dummy variables.
- You can use the `coeflegend` option to get the names of individual interaction coeffiecients.
- Or you can test all the interactions as a group using the `testparm` command.
- Test the null hypothesis that the interaction term of female and baseball is equal to the coefficient for the female and soccer interaction.
- Test the null hypothesis that all the interaction terms are jointly equal to zero.

2.18 Continuous-Categorical Interactions

- Consider a relationship between two continous variables
- Continuous-categorical interactions allow this relationship to change for the different categories.
- As an example, let's look at the relationship between team profit and number of participants.
- Visualize this relationship with a scatter plot combined with a linear fit plot.
- Run a simple bivariate regression of profit on number of participants.
- These results suggest that, as the number of participants increases, profit will also increase.
- We've seen that different sports have very different average profit levels, so it is likely

that this relationship is different for different sports.

- A convenient tool for graphing this relationship over different sports is the `separate` command.

```
separate profit , by(sportcat) gen(prof_) shortlabel
browse sportcat profit prof_*
twoway scatter prof_* numparticipants
```

2.19 Continuous-Categorical Factor Variables

- Use the factor variables to regress profit on number of participants, sport category dummies, and interaction terms for number of participants and each sports category.
- Your regression should include nine independent variables plus a constant term, for a total of 10 coefficients.
- When you are using factor variable notation with continuous variables, you must put `c.` in front of the variable name.

2.20 Graphing Predicted Values

- You may want to create fitted lines to visualize the relationship between profit and number of participants for different categories.
- It is easiest to do this using predicted values after the interacted regression.
- Follow these steps to generate a graph with fitted lines for each category.
- Use the `predict` command to generate predicted values from the previous regression.
- Use the `separate` command to separate the predicted values into new variables, by `sportcat`.
- Create a twoway line graph of the new variables by number of participants.