EWMBA 296 (Fall 2015)

Section 7

GSI: Fenella Carpena

December 8, 2015

Agenda for Today

- Dummy Variables: Basic Concepts
- 3 Forms of Regressions With Dummy Variables
  - Case 1: SRM where the only X variable is a dummy
  - Case 2: MRM where one of the X variables is a dummy
  - Case 3: Interactions involving dummy variables
- Practice Problems/Final Exam Review
  - Practice Problem: Effect of iPad on College GPA
  - Practice Final Exam 1, Q7
  - Practice Problem: Lightbulbs
- **Note:** Slides posted in bCourses; I may ***not*** be able to write section notes on dummy variables due to time constraints.

Dummy Variables: Basic Concepts

- So far, we've been working with variables that have **quantitative** meaning.
  - Sales in thousands of $; Size in square feet of a retail store; Weight of diamond.
- But in some cases, we may want to incorporate **qualitative** variables in our regression.
  - Gender or race of a person, industry of a firm (manufacturing, retail, etc.), region in the US (south, north, west, etc.)
- These qualitative variables cannot be measured on a numerical scale.
- **How can we include such qualitative info in our regression?** We use a **dummy variable** (also called binary variable). It is a variable that takes on only two values: 0 or 1.

Dummy Variables: Basic Concepts

- Example: Suppose we are interested in comparing the election outcomes between Democratic and Republican candidates.
  - We can define *Dem* to be a dummy variable $= 1$ if the candidate is Democrat, 0 if Republican.
  - The same information can be captured by defining *Rep* to be 1 if the candidate is Republican, and 0 otherwise.
  - The choice of which category to assign to 1 is arbitrary.
- In practice, the name of the dummy variable is not important for the regression (we can use any name we want). But it is often useful to choose names that signify what category "turns on" when the dummy $= 1$.
  - Example: A variable name like *party* won't make it clear whether *party* $= 1$ means a Democrat or a Republican.

Dummy Variables: Basic Concepts

We often encounter 3 forms of regressions that include dummy variables.

In the examples below, let *salary* be an employee's annual salary in dollars, $MBA = 1$ if the government has an MBA and 0 otherwise, and *exper* is years of work experience.

- **Case 1: SRM where the only X variable is a dummy.**

$$\widehat{salary} = b_0 + b_1 \cdot MBA$$

- **Case 2: MRM where one of the X variables is a dummy.**

$$\widehat{salary} = b_0 + b_1 \cdot MBA + b_2 \cdot exper$$

- **Case 3: Interactions involving dummy variables.**

$$\widehat{salary} = b_0 + b_1 \cdot MBA + b_2 \cdot exper + b_3 \cdot MBA \cdot exper$$

Important: The interpretation of the dummy variables varies in each of the above 3 cases.

Case 1: SRM where the only X variable is a dummy

$$\widehat{salary} = b_0 + b_1 \cdot MBA$$

(1) Regression if $MBA = 0$:

(2) Regression if $MBA = 1$:

(3) Interpretation of b_0:

(4) Interpretation of $b_0 + b_1$:

(5) Interpretation of b_1:

Case 2: MRM where one of the X variables is a dummy

$$\widehat{salary} = b_0 + b_1 \cdot MBA + b_2 \cdot exper$$

(1) Regression if $MBA = 0$:

(2) Regression if $MBA = 1$:

(3) Interpretation of b_0:

(4) Interpretation of $b_0 + b_1$:

(5) Interpretation of b_1:

(6) Interpretation of b_2:

Case 3: Interactions involving dummy variables

$$\widehat{salary} = b_0 + b_1 \cdot MBA + b_2 \cdot exper + b_3 \cdot MBA \cdot exper$$

(1) Regression if $MBA = 0$:

(2) Regression if $MBA = 1$:

(3) Interpretation of b_0:

(4) Interpretation of $b_0 + b_1$:

(5) Interpretation of b_1:

(6) Interpretation of b_2:

(7) Interpretation of $b_2 + b_3$:

(8) Interpretation of b_3:

Practice Problem: Effect of iPad on College GPA

Suppose that the Haas Undergraduate Program Office wishes to examine the potential educational benefits of providing students with an iPad. Using data from 141 students, the table below summarizes the mean and SD of the variable *collGPA* (college GPA) for students with and without an iPad. Assume that all sample size conditions are met.

	Number	Mean	SD
Without iPad	85	2.989	0.321
With iPad	56	3.159	0.421

(a) Let $\overline{X}_1$ be the sample mean collage GPA of students with an iPad, and $\overline{X}_0$ be the sample mean collage GPA of students without an iPad. Calculate:

(i) $\overline{X}_1 - \overline{X}_0$

(ii) $se(\overline{X}_1 - \overline{X}_0)$

Practice Problem: Effect of iPad on College GPA

(b) The program office does not plan to provide iPads to students unless there is evidence that having an iPad increases students' GPAs. What is the one-sided null and alternative hypothesis that the program office would like to test?

(c) Carry out the test from part (b) at the 1% level. Do you reject or fail to reject the null hypothesis?

	Number	Mean	SD
Without iPad	85	2.989	0.321
With iPad	56	3.159	0.421

Practice Problem: Effect of iPad on College GPA

(d) Can the difference in the GPA between students with and without an iPad be attributed to iPad ownership alone? Explain why or why not.

	Number	Mean	SD
Without iPad	85	2.989	0.321
With iPad	56	3.159	0.421

Practice Problem: Effect of iPad on College GPA

(e) Suppose that an analyst at the program office instead estimated an SRM of college GPA on *HasIPAD*, a dummy variable equal to 1 if the student has an iPad, and 0 otherwise, obtaining the following regression results. Interpret each of the coefficients in the regression.

	Coefficients	Standard Error
Intercept	2.989	0.040
HasIPAD	0.170	0.063

Practice Problem: Effect of iPad on College GPA

(f) Consider again the null and alternative hypothesis from part (b).

(i) How can use the regression coefficients in the SRM from part (e) to test the same null and alternative hypothesis as in part (b)?

(ii) Carry out the test using the regression coefficients at the 1% level (Recall that the sample size is 141).

(iii) What does question (f)(i) and (f)(ii) tell us about how we can use regressions to carry out a test for the differences of means across two groups?

	Coefficients	Standard Error
Intercept	2.989	0.040
HasIPAD	0.170	0.063

Practice Problem: Effect of iPad on College GPA

(g) Now, suppose the following two explanatory variables were added to the SRM from part (f): *skipped* is the average number of lectures the student misses per week, and *hsGPA* is the student's high school GPA. The MRM regression results are shown below. Interpret each of the coefficients in this regression.

	Coefficients	Standard Error
Intercept	1.527	0.300
HasIPAD	0.129	0.057
skipped	-0.065	0.026
hsGPA	0.456	0.086

Practice Problem: Effect of iPad on College GPA

(h) Why is the estimated GPA differential between students with and without an iPad obtained in the MRM in part (g) smaller than what is obtained in the difference in means from part (a)(i)?

Regression from part (g)

	Coefficients	Standard Error
Intercept	1.527	0.300
HasIPAD	0.129	0.057
skipped	-0.065	0.026
hsGPA	0.456	0.086

Sample means from part (a)

	Number	Mean	SD
Without iPad	85	2.989	0.321
With iPad	56	3.159	0.421

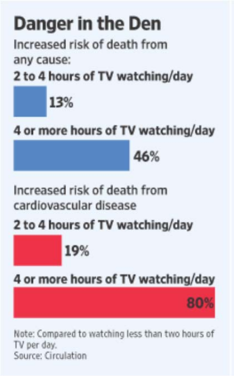
Practice Problem: Effect of iPad on College GPA

(i) How would you augment the model estimated in part (g) to allow the effect of having an iPad on college GPA to vary by gender?

Final Review: Practice Final Exam 1, Q7

Television and Risk of Death. A recent NYTimes article discussed a study about television and the risk of death.

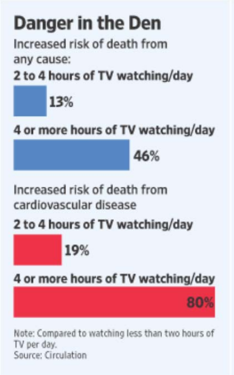
(A) Based on the info reported in this figure, write out the regression equation(s) that you think the authors estimated to produce these results. Define the response variable and all the explanatory variables as completely as possible.



Final Review: Practice Final Exam 1, Q7

Television and Risk of Death. A recent NYTimes article discussed a study about television and the risk of death.

(B) The authors of the study concluded from these results that watching television increases the risk of death. Do you agree with the authors' interpretation. Explain.



Practice Problem: Lightbulbs

Suppose that a lightbulb manufacturing plant produces bulbs with a mean life of 2000 hours and a standard deviation of 200 hours. An inventor claims to have developed an improved process that produces bulbs with a longer mean life and the same standard deviation.

(a) Let μ denote the mean of the new process. What is the null and alternative hypothesis that the plant manager wishes to test?

Practice Problem: Lightbulbs

(b) The plant manager randomly selects a sample of 100 bulbs produced by the new process to test the inventor's claim. Now, suppose that the plant manager uses the following procedure: she will believe the inventor's claim if the sample mean life of the bulbs is > 2100. Assume all SSCs hold.

(i) What is the probability that the plant manager makes a Type I error?

Practice Problem: Lightbulbs

(b) The plant manager randomly selects a sample of 100 bulbs produced by the new process to test the inventor's claim. Now, suppose that the plant manager uses the following procedure: she will believe the inventor's claim if the sample mean life of the bulbs is > 2100. Assume all SSCs hold.

(ii) Suppose that the new process is in fact better and has a mean bulb life of 2150 hours. What is the probably that the plant manager's procedure correctly rejects H_0?

Practice Problem: Lightbulbs

(b) The plant manager randomly selects a sample of 100 bulbs produced by the new process to test the inventor's claim. Now, suppose that the plant manager uses the following procedure: she will believe the inventor's claim if the sample mean life of the bulbs is > 2100. Assume all SSCs hold.

(iii) What action threshold should the plant manager use if she wants the probability of falsely rejecting the null hypothesis to be 5%?


KEEP
CALM
AND
SURVIVE
FINALS WEEK