Fox Module 13 Dummy variables

(The attached PDF file has better formatting.)

REGRESSION ANALYSIS DUMMY VARIABLES PRACTICE PROBLEMS

Much actuarial work uses dummy variables, such as male = 1 and female = 0 or urban = 1 and rural = 0 or normal blood pressure = 0 and high blood pressure = 1. John Fox applies regression to social issues and medical research, which has similar dummy variables, such as vote = 1 and did not vote = 0 or ill = 1 and healthy = 0.

*Question 13.1: Dummy Variables

To forecast auto insurance rates, an actuary uses a multiple regression model with dummy variables for territory and driver age. The state has 15 territories and 6 driver age groups. Each car is in one and only one territory and each driver is in one and only one age group.

The premium rate is the base rate plus the territorial relativity plus the age group relativity. For example, if the base rate is $1,200, territory 01 has a relativity of +$500, and age group 21-25 has a relativity of +$600, the premium rate for drivers age 21-25 in territory 01 is $2,300.

There are no interaction terms. For example, if the age group relativity for 21-25 year old drivers is +$750 in territory 01, it is +$750 in all territories.

How many dummy variables are needed for this regression?

A. 19
B. 20
C. 21
D. 88
E. 89

Answer 13.1: A

For mutually exclusive qualitative values with no N/A or abstain choice, we need one less dummy variable than the number of choices: (15 – 1) + (6 – 1) = 19

Illustration: For male vs female, we need one dummy variable, since if the driver is male, he is not female, and if she is female, she is not male.

If the state has 12 territories, and every car is in one and only one territory, we need 11 dummy variables. If the car is in one of the 11 territories represented by these 11 dummy variables, it has a 1 for that variable and a 0 for the other variables. If it has a 0 for all 11 variables, it must be in the twelfth territory.

Jacob: What does an interaction term mean?

Rachel: Suppose we have two qualitative dimensions, sex and age, with two values in each dimension: male vs female and youthful vs adult.

- For males, annual premium rates are $2,000 for adult and $5,000 for youthful.
- For females, annual premium rates are $1,000 for adult and $2,000 for youthful.

The youthful vs adult difference is higher for males than for females. We need three dummy variables, or 2 × 2 – 1, not two dummy variables, or (2 – 1) × (2 – 1).

Jacob: Interaction terms can be important. Do we maximize the number of interaction terms?

Rachel: Interactions terms make the regression analysis less efficient. We want orthogonal class dimensions, so each variable is measuring something different and no variables overlap. Orthogonal class dimensions means the variables are proxies for the same item.

Illustration: More aggressive drivers have higher accident frequencies. Young males have more testosterone than older males, and young males have more testosterone than young females. A two-class system is more efficient than a one class system. We want two classes that are not proxies for the same thing.

Actuaries related sex and age to character traits like risk-taking and aggression over sixty years ago, but they could not use hormone tests to quantify these relations. Insurers can’t test their policyholders for testosterone. Over the past fifteen years, scientists have re-examined these relations with hormone levels, studying high- vs low-testosterone males and females. Many sex differences affecting behavior, such as competitiveness and risk-taking, seem to be a function hormone levels.

*Question 13.2: Dummy Variables

To forecast auto insurance rates, John uses a multiple regression model with dummy variables for five class dimensions, each of which has two values:

- males vs female
- young vs adult
- married vs unmarried
- urban resident vs suburban
- good credit score vs poor credit score

Nancy uses dummy variables for a single class dimension with $2^5 = 32$ values, such as adult, married woman living in the suburbs with a good credit score.

Let Y be the number of dummy variables used by Nancy, and Y be the number of dummy variables used by John. What is Y – Z?

A. –26
B. –22
C. 0
D. 22
E. 26

Answer 13.2: E

- Z: John needs 5 dummy variables: one for each rating dimension: 5 × (2 – 1) = 5.
- Y: Nancy needs 31 dummy variables, or 32 – 1.

Y – Z = 31 – 5 = 26

*Question 13.3: Dummy Variables

An actuary regresses personal auto claim frequency on miles driven, deriving an intercept α and a slope parameter β. The actuary believes the intercept or the slope may differ for male vs female drivers.

The actuary wants to test three possible scenarios:

1. The intercept differs for men vs women, but the slope is assumed to be the same.
2. The slope differs for men vs women, but the intercept is assumed to be the same.
3. Both the intercept and the slope differ for men vs women.

For which of these three scenarios might the actuary use dummy variables?

A. 1 and 2 only
B. 1 and 3 only
C. 2 and 3 only
D. 1, 2, and 3
E. None of A, B, C, or D is correct

Fox shows graphics for each of these.

If the intercept differs but the slope is the same, the two regression lines are parallel. If D is the dummy variable, the two intercepts are α_1 and α_1 + D × α_2.

If the intercept is the same but the slope differs, the two regression lines intersect at the intercept. If D is the dummy variable, the two slopes are β_1 and β_1 + D × β_2

In much social research, the intercept is arbitrary. In much actuarial work, the intercept has meaning.

Answer 13.3: D