Dummy variables: Coding categorical explanatory variables

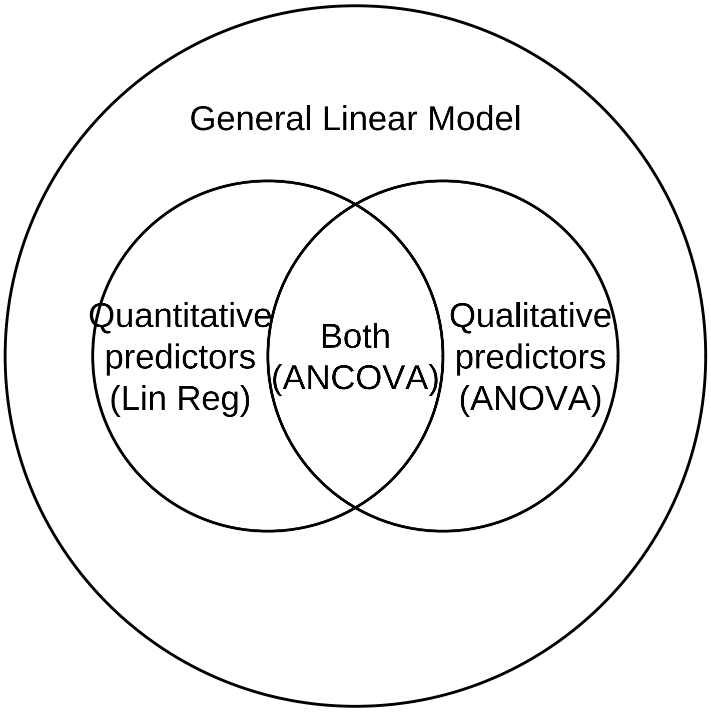
Biometry 755

Spring 2009

Introduction

So far, the predictor variables in our regression analyses have been quantitative, i.e. variables that take on values on a continuous scale. However, many predictors of interest are qualitative (categorical) variables. Common examples include gender (male/female), race (black/white/other), smoking status (ever/never), and treatment arm of a randomized trial (placebo/intervention). In this lecture we will investigate some of the particular issues that arise when qualitative predictors are incorporated into a regression model.

Terminology



General Linear Model

Quantitative predictors (Lin Reg)

Both (ANCOVA)

Qualitative predictors (ANOVA)

SLR on a qualitative predictor

Consider the variable MEDSCHL coded 1 if a hospital in the study is affiliated with a medical school and 2 if it is not. This variable's values are not meaningful in and of themselves, other than to distinguish between two classes of hospitals. We could have used any coding scheme so long as the values we selected were unique. In regression analysis, it is convenient to code categorical variables using a coding scheme of zeros and ones called *reference cell coding*. We call such variables *indicator* variables or *dummy* variables, since their values lack any intrinsic meaning.

Defining MEDSCHL using dummy variables

$$\text{MEDSCHL} = \begin{cases} 1 & \text{if YES} \\ 2 & \text{if NO.} \end{cases}$$

The following describes how to create a dummy variable for MEDSCHL using a reference cell coding scheme.

$$\text{MSIND} = \begin{cases} 1 & \text{if Med school affiliation} \\ 0 & \text{if No med school affiliation.} \quad \leftarrow \text{ reference group} \end{cases}$$

Alternatively, we could have defined

$$\text{MSIND2} = \begin{cases} 1 & \text{if No med school affiliation} \\ 0 & \text{if Med school affiliation.} \qquad \leftarrow \text{ reference group} \end{cases}$$

Constructing these variables in SAS

```
data one;
    set betsy.senicfull;
    if medschl = . then msind = .;
    else
    if medschl = 1 then msind = 1;
    else msind = 0;

    label msind = 'MedSchl: yes vs no';
run;
```

SLR of INFRISK on INDMS

```
proc reg data = one;
    model infrisk = msind;
run;
```

```
                       Parameter Estimates

                                   Parameter  Standard
Variable   Label               DF  Estimate   Error     t Value   Pr > |t|

Intercept  Intercept           1   4.22396    0.13369     31.60    <.0001
msind      MedSchl: yes vs no  1   0.87016    0.34467      2.52    0.0130
```

How do we interpret the parameter estimates from this model?

Interpreting the output

The model we fit is

$$\text{INFRISK} = \beta_0 + \beta_1 \text{MSIND} + \varepsilon.$$

Then

No med school affiliation (MSIND = 0)

$$\text{INFRISK} = \beta_0 + \varepsilon \quad \Rightarrow \quad \widehat{\text{INFRISK}} = \hat{\beta}_0.$$

Med school affiliation (MSIND = 1)

$$\text{INFRISK} = \beta_0 + \beta_1 + \varepsilon \quad \Rightarrow \quad \widehat{\text{INFRISK}} = \hat{\beta}_0 + \hat{\beta}_1.$$

Interpreting the output (cont.)

The fitted model is

$$\widehat{\text{INFRISK}} = 4.22 + 0.87 \times \text{MSIND}.$$

Then

<u>No med school affiliation (MSIND = 0)</u>

$$\widehat{\text{INFRISK}} = \hat{\beta}_0 = 4.22$$

<u>Med school affiliation (MSIND = 1)</u>

$$\widehat{\text{INFRISK}} = \hat{\beta}_0 + \hat{\beta}_1 = 4.22 + 0.87 = 5.09$$

Summary: SLR on an indicator variable

In a SLR of Y on X where X is a dummy variable with reference cell coding:

- The intercept is the average response for the reference group ($X = 0$)
- The slope is the relative increase (or decrease) in average response comparing the '$X = 1$' level to the reference group
- The sum of the slope and intercept is the average response for the non-reference group ($X = 1$).

This is why a SLR on a dummy variable is equivalent to a t-test with equal variances.

Indicator variables for REGION

Consider the variable REGION, defined as follows:

$$\text{REGION} = \begin{cases} 1 & \text{if NE} \\ 2 & \text{if NC} \\ 3 & \text{if S} \\ 4 & \text{if W.} \end{cases}$$

Indicator variables for REGION (cont.)

Even though REGION has four categories, it will only take three indicator variables to represent all its levels.

$$\text{REGIND1} = \begin{cases} 1 & \text{if NC (REGION = 2)} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{REGIND2} = \begin{cases} 1 & \text{if S (REGION = 3)} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{REGIND3} = \begin{cases} 1 & \text{if W (REGION = 4)} \\ 0 & \text{otherwise} \end{cases}$$

Which region is the reference group?

Constructing the indicators in SAS

```
data one;
    set betsy.senicfull;
    if region = . then regind1 = .;
    else
    if region = 2 then regind1 = 1;
    else regind1 = 0;

    if region = . then regind2 = .;
    else
    if region = 3 then regind2 = 1;
    else regind2 = 0;

    if region = . then regind3 = .;
    else
    if region = 4 then regind3 = 1;
    else regind3 = 0;
run;
```

Regression of INFRISK on REGION

```
proc reg data = one;
    model infrisk = regind1 regind2 regind3;
run;
```

```
                            Parameter Estimates

                               Parameter Standard
Variable  Label             DF  Estimate    Error  t Value Pr > |t|

Intercept Intercept         1   4.86071   0.24778    19.62   <.0001
regind1   Region: NC vs NE  1  -0.46696   0.33929    -1.38   0.1716
regind2   Region: S vs NE   1  -0.93369   0.32842    -2.84   0.0053
regind3   Region: W vs NE   1  -0.47946   0.41090    -1.17   0.2458
```

Interpreting the output

The model we fit is:

$$\text{INFRISK} = \beta_0 + \beta_1\text{INDR1} + \beta_2\text{INDR2} + \beta_3\text{INDR3} + \varepsilon.$$

NE (REGIND1 = REGIND2 = REGIND3 = 0)

$$\text{INFRISK} = \beta_0 + \varepsilon \;\;\Rightarrow\;\; \widehat{\text{INFRISK}} = \hat{\beta}_0.$$

NC (REGIND1 = 1; REGIND2 = REGIND3 = 0)

$$\text{INFRISK} = \beta_0 + \beta_1 + \varepsilon \;\;\Rightarrow\;\; \widehat{\text{INFRISK}} = \hat{\beta}_0 + \hat{\beta}_1.$$

S (REGIND2 = 1; REGIND1 = REGIND3 = 0)

$$\text{INFRISK} = \beta_0 + \beta_2 + \varepsilon \;\;\Rightarrow\;\; \widehat{\text{INFRISK}} = \hat{\beta}_0 + \hat{\beta}_2.$$

W (REGIND3 = 1; REGIND1 = REGIND2 = 0)

$$\text{INFRISK} = \beta_0 + \beta_3 + \varepsilon \;\;\Rightarrow\;\; \widehat{\text{INFRISK}} = \hat{\beta}_0 + \hat{\beta}_3.$$

Interpreting the output (cont.)

The fitted model is:

$$\widehat{\text{INFRISK}} = 4.86 - 0.47 \times \text{REGIND1} - 0.93 \times \text{REGIND2} - 0.48 \times \text{REGIND3}.$$

NE (REGIND1 = REGIND2 = REGIND3 = 0)

$$\widehat{\text{INFRISK}} = 4.86$$

NC (REGIND1 = 1; REGIND2 = REGIND3 = 0)

$$\widehat{\text{INFRISK}} = 4.86 - 0.47 = 4.39$$

S (REGIND2 = 1; REGIND1 = REGIND3 = 0)

$$\widehat{\text{INFRISK}} = 4.86 - 0.93 = 3.93$$

W (REGIND3 = 1; REGIND1 = REGIND2 = 0)

$$\widehat{\text{INFRISK}} = 4.86 - 0.48 = 4.38$$

Summary: Reference cell coding

- A categorical variable with k levels requires $k - 1$ indicator variables
- The reference group is identified as that for which *all* indicator variables are equal to ‘0’.
- The intercept is the average response for the reference group.
- For a given indicator variable, the corresponding slope is the relative increase (decrease) in average response comparing the group represented by the indicator to the reference group.
- For a given indicator variable, the sum of the corresponding slope and intercept is the average response for the group represented by the indicator.

Who should be the reference group?

Typically in clinical studies, the goals are to identify risk factors for adverse health outcomes. In that context, the reference group is typically selected to be that which has a more favorable average response so that the estimated slope parameters reflect the increase risk for the outcome for subjects in the non-reference category relative to those in the reference category.

However, there are instances where this rule of thumb does not hold. Ultimately it doesn’t matter who the reference group is, although interpreting the coefficients of the regression can be made more straightforward by a judicious choice. This is particularly true for analyses that use logistic regression or hazard regression, but less so for linear regression.

Inference

How do we interpret the results of the t-tests?

```
                            Parameter Estimates

                               Parameter Standard
Variable  Label             DF  Estimate    Error  t Value Pr > |t|

Intercept Intercept         1   4.86071   0.24778    19.62   <.0001
regind1   Region: NC vs NE  1  -0.46696   0.33929    -1.38   0.1716
regind2   Region: S vs NE   1  -0.93369   0.32842    -2.84   0.0053
regind3   Region: W vs NE   1  -0.47946   0.41090    -1.17   0.2458
```

Inference (cont.)

How do we assess the association between region of the US and risk of nosocomial infection?

```
proc reg data = one ;
    model infrisk = regind1 regind2 regind3;
    Region_test: test regind1, regind2, regind3;
run;
quit;
```

```
 Test Region_test Results for Dependent Variable INFRISK
                                   Mean
Source              DF           Square    F Value    Pr > F

Numerator            3          4.66565       2.71    0.0484
Denominator        109          1.71911
```

Can’t SAS create dummy variables for me?

Yes ... but not in PROC REG. Instead, use PROC GLM, which stands for General Linear Model (see Slide 3).

```
proc sort data = one;
    by descending region;
run;

proc glm data = one order = data;
    class region;
    model infrisk = region/solution ss3;
run;
quit;
```

PROC GLM output

```
Source     DF     Type III SS     Mean Square    F Value    Pr > F

REGION      3     13.99693932      4.66564644       2.71    0.0484

                                             Standard
Parameter             Estimate                  Error    t Value    Pr > |t|

Intercept          4.860714286 B       0.24778368      19.62      <.0001
REGION     W      -0.479464286 B       0.41090274      -1.17      0.2458
REGION     S      -0.933687259 B       0.32841918      -2.84      0.0053
REGION     NC     -0.466964286 B       0.33929177      -1.38      0.1716
REGION     NE      0.000000000 B        .                .          .

NOTE: The X’X matrix has been found to be singular, and a generalized
      inverse was used to solve the normal equations.  Terms whose
      estimates are followed by the letter ’B’ are not uniquely estimable.
```