General Concepts on Categorical Variables as Predictors

Dr. Kiah Wah Ong

Qualitative/Categorical Variables

In many modeling situations, it is necessary to use predictors/regressors such as sex, educational level, employment status and so on.

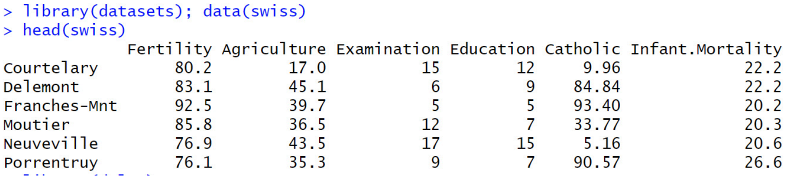
These nonnumerical variables are known as **qualitative** or **categorical** variables.

We need a way to do regression analysis for these models that involve non-quantitative variables.

Qualitative/Categorical Variables

We will investigate this new concept using a dataset from *R* which is called *Swiss Fertility and Socioeconomic Indicators (1888) Data*.

Here is how the first 6 observations look like:



```
> library(datasets); data(swiss)
> head(swiss)
             Fertility Agriculture Examination Education Catholic Infant.Mortality
Courtelary        80.2        17.0          15        12     9.96             22.2
Delemont          83.1        45.1           6         9    84.84             22.2
Franches-Mnt      92.5        39.7           5         5    93.40             20.2
Moutier           85.8        36.5          12         7    33.77             20.3
Neuveville        76.9        43.5          17        15     5.16             20.6
Porrentruy        76.1        35.3           9         7    90.57             26.6
```

Qualitative/Categorical Variables

The "Swiss" data has 47 observations on 6 variables:

(1) **Fertility**, I_g, using common standardized fertility measure
(2) **Agriculture**, % of males involved in agriculture as occupation
(3) **Examination**, % draftees receiving highest mark on army exam
(4) **Education**, % education beyond primary school for draftees.
(5) **Catholic**, % Catholic as opposed to Protestant
(6) **Infant Mortality**, % live births who live less than 1 year

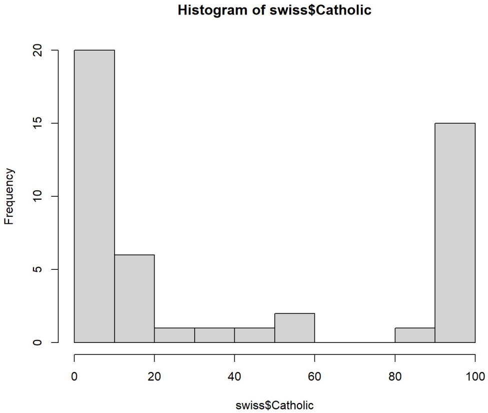
The data collected are for 47 French-speaking provinces at about 1888.

I_g is equal to the total number of children born to married women divided by the maximum conceivable number of children, obtained from data on the Hutterites, an Anabaptist sect that does not practice any form of fertility limitations.

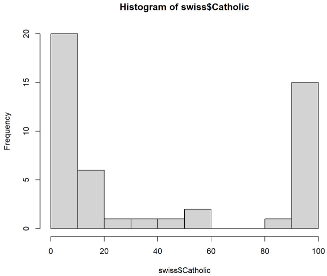
Qualitative/Categorical Variables

Suppose we are interested in "Agriculture", x_1, and "Catholic", x_2, as regressor variables and "Fertility" as an outcome, y.

Let us examine the data on "Catholic". Using $\mathrm{hist(swiss\$Catholic)}$ we obtain



Qualitative/Categorical Variables

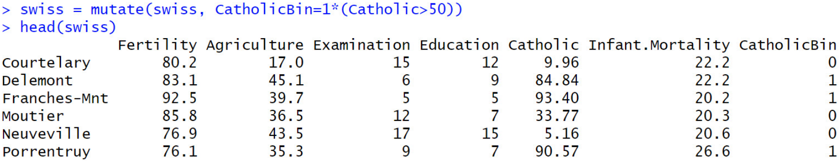


Because of the bimodel nature of “Catholic”, we want to create an **indicator variable** x_2 as follows:

$$x_2 = \begin{cases} 1 & \text{if the province is over 50\% Catholic} \\ 0 & \text{otherwise} \end{cases}$$

Qualitative/Categorical Variables

We can do this in *R* using dplyr as shown below:



```
> swiss = mutate(swiss, CatholicBin=1*(Catholic>50))
> head(swiss)
             Fertility Agriculture Examination Education Catholic Infant.Mortality CatholicBin
Courtelary        80.2        17.0          15        12     9.96             22.2           0
Delemont          83.1        45.1           6         9    84.84             22.2           1
Franches-Mnt      92.5        39.7           5         5    93.40             20.2           1
Moutier           85.8        36.5          12         7    33.77             20.3           0
Neuveville        76.9        43.5          17        15     5.16             20.6           0
Porrentruy        76.1        35.3           9         7    90.57             26.6           1
```

Qualitative/Categorical Variables

Using the indicator variable x_2, we can formulate a linear regression model given below

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Notice that this model can be rewritten as:

$$y = \begin{cases} \beta_0 + \beta_1 x_1 + \epsilon & \text{if } x_2 = 0 \;\; \text{(non-Catholic)} \\ (\beta_0 + \beta_2) + \beta_1 x_1 + \epsilon & \text{if } x_2 = 1 \;\; \text{(Catholic)} \end{cases}$$

Interpretation

Let us consider the meaning of the regression coefficients in the mean response functions below.

$$E(y) = \begin{cases} \beta_0 + \beta_1 x_1 & \text{if } x_2 = 0 \quad \text{(non-Catholic)} \\ (\beta_0 + \beta_2) + \beta_1 x_1 & \text{if } x_2 = 1 \quad \text{(Catholic)} \end{cases}$$

We see that the mean fertility, $E(y)$ is a linear function of “Agriculture”, x_1 (the % of males involved in agriculture as occupation).

Also, the $E(y)$ here has the same slope β_1 whether the province is majority Catholic or not.

Interpretation

Let us examine the function here again:

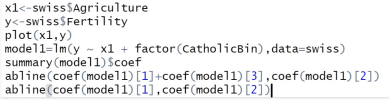
$$E(y) = \begin{cases} \beta_0 + \beta_1 x_1 & \text{if } x_2 = 0 \quad \text{(non-Catholic)} \\ (\beta_0 + \beta_2) + \beta_1 x_1 & \text{if } x_2 = 1 \quad \text{(Catholic)} \end{cases}$$

Notice that β_2 indicates how much higher (or lower) the mean response for a province that is majority Catholic is than the one that is majority non-Catholic, for any given percentage of males involved in agricultural occupation.

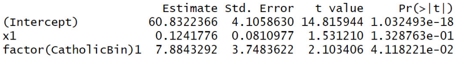
In general, β_2 shows how much higher (or lower) the mean response line is for the class coded 1 than the line for the class coded 0, for any given level of x_1.

Interpretation

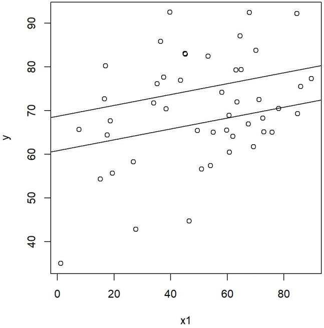
Using *R*, we set up the following regression model:



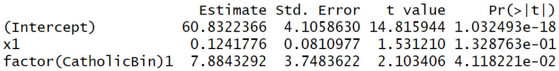
```
x1<-swiss$Agriculture
y<-swiss$Fertility
plot(x1,y)
model1=lm(y ~ x1 + factor(CatholicBin),data=swiss)
summary(model1)$coef
abline(coef(model1)[1]+coef(model1)[3],coef(model1)[2])
abline(coef(model1)[1],coef(model1)[2])
```



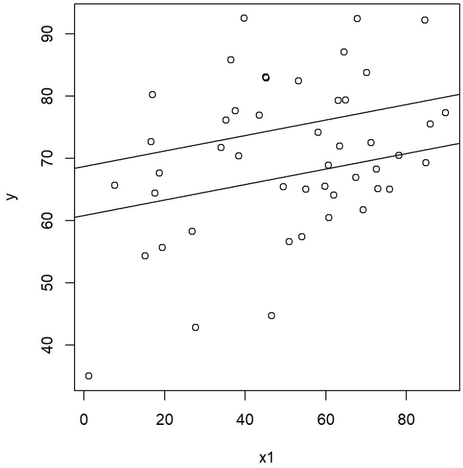
```
                     Estimate Std. Error  t value     Pr(>|t|)
(Intercept)        60.8322366  4.1058630 14.815944 1.032493e-18
x1                  0.1241776  0.0810977  1.531210 1.328763e-01
factor(CatholicBin)1 7.8843292  3.7483622  2.103406 4.118221e-02
```



Interpretation



```
                      Estimate Std. Error   t value     Pr(>|t|)
(Intercept)         60.8322366  4.1058630 14.815944 1.032493e-18
x1                   0.1241776  0.0810977  1.531210 1.328763e-01
factor(CatholicBin)1 7.8843292  3.7483622  2.103406 4.118221e-02
```



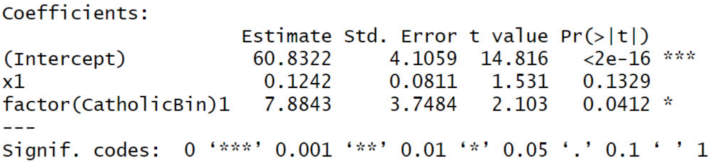
The mean response of “Fertility” is about 7.88% higher in the majority Catholic provinces as compared to the majority Protestant provinces, for any given % of males involved in agriculture occupation.

Interpretation

Looking at the summary from R, for a test

$$H_0 : \beta_2 = 0 \ \text{ vs } \ H_1 : \beta_2 \neq 0$$

we will reject H_0 if the significant is set at $\alpha = 0.05$.

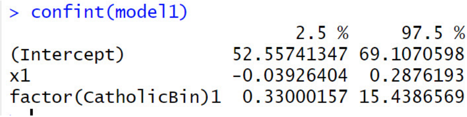


```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)            60.8322     4.1059  14.816   <2e-16 ***
x1                      0.1242     0.0811   1.531   0.1329    
factor(CatholicBin)1    7.8843     3.7484   2.103   0.0412 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation

From the 95% confidence interval, we see that

$$0.33 \leq \beta_2 \leq 15.44$$



```
> confint(model1)
                           2.5 %     97.5 %
(Intercept)          52.55741347 69.1070598
x1                   -0.03926404  0.2876193
factor(CatholicBin)1  0.33000157 15.4386569
```

Hence, we are 95% confident that for all value of x_1, being in a "Catholic" province will increase the mean response by between 0.33 and 15.44 in its corresponding unit.