Qualitative Predictor with Two or More Classes

Dr. Kiah Wah Ong

Qualitative Predictor with Two Classes

Let us look at the Swiss data set we used in the previous lesson. Now, suppose we define the following two indicator variables x_2 and x_3 as follows:

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

$$x_3 = \begin{cases} 1 & \text{if the province is over 50\% Protestant} \\ 0 & \text{otherwise} \end{cases}$$

then we can write our regression model as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

Qualitative Predictor with Two Classes

However, this intuitive approach of setting up an indicator variable for each class of the qualitative predictor variable will leads to computational difficulties.

To see why, suppose we look at 4 observations from our data set, in which the first two were majority Catholic ($x_2 = 1, x_3 = 0$) and the second two being majority non-Catholic ($x_2 = 0, x_3 = 1$). Hence the $\mathbf{X}$ would be

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & 1 & 0 \\ 1 & x_{21} & 1 & 0 \\ 1 & x_{31} & 0 & 1 \\ 1 & x_{41} & 0 & 1 \end{bmatrix}$$

Qualitative Predictor with Two Classes

Observe that the first column of $\mathbf{X}^T\mathbf{X}$ is the sum of the last two columns. Hence the columns of $\mathbf{X}^T\mathbf{X}$ are linearly dependent, and therefore not possessing an inverse. A consequence of this is that no unique estimators of the regression coefficients can be found.

$$
\begin{aligned}
\mathbf{X}^T\mathbf{X} &= \begin{bmatrix} 1 & 1 & 1 & 1 \\ x_{11} & x_{21} & x_{31} & x_{41} \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & x_{11} & 1 & 0 \\ 1 & x_{21} & 1 & 0 \\ 1 & x_{31} & 0 & 1 \\ 1 & x_{41} & 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 4 & \sum_{i=1}^4 x_{i1} & 2 & 2 \\ \sum_{i=1}^4 x_{i1} & \sum_{i=1}^4 x_{i1}^2 & \sum_{i=1}^2 x_{i1} & \sum_{i=3}^4 x_{i1} \\ 2 & \sum_{i=1}^2 x_{i1} & 2 & 0 \\ 2 & \sum_{i=3}^4 x_{i1} & 0 & 2 \end{bmatrix}
\end{aligned}
$$

Qualitative Predictor with Two Classes

What we need to do in this situation is to drop one of the indicator variables. In our example, we drop x_3 and avoided the above mentioned difficulties.

In general, we shall follow the principle:

A qualitative variable with c classes will be represented by $c - 1$ indicator variables, each taking the values 0 and 1.

Qualitative Predictor with More than Two Classes

Tool wear is the gradual failure of cutting tools due to regular operation. Let us consider a regression model of tool wear y, on tool speed, x_1, and tool models A, B, C and D.

Since we have a qualitative variable with four classes, using the above mentioned principle, we therefore require three indicator variables as follow:

$$x_2 = \begin{cases} 1 & \text{for tool model A} \\ 0 & \text{otherwise} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{for tool model B} \\ 0 & \text{otherwise} \end{cases}$$

$$x_4 = \begin{cases} 1 & \text{for tool model C} \\ 0 & \text{otherwise} \end{cases}$$

Qualitative Predictor with More than Two Classes

The regression model will be of the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$$

and the data input for x variables are given as follow:

Model	x_1	x_2	x_3	x_4
A	x_{i1}	1	0	0
B	x_{i1}	0	1	0
C	x_{i1}	0	0	1
D	x_{i1}	0	0	0

Also notice that the mean respond for the tool wear variable y is given by

$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}$$

Qualitative Predictor with More than Two Classes

Example

Using the Catdata,
$\mathrm{Catdata} < -\mathrm{read.csv}(\text{"Cat1.CSV"}, \mathrm{header} = \mathrm{TRUE}, \mathrm{sep} = \text{","})$ we have:

	group	x1	y
1	A	4.902189	59.26074
2	A	5.165739	55.55838
3	A	4.797517	47.44312
4	A	5.064308	47.28059
5	A	5.093231	41.35536
6	A	5.35219	52.35375
21	B	4.950873	51.34033
22	B	5.212557	47.03896
23	B	4.919606	60.66743
24	B	4.560823	55.86374
25	B	5.223331	53.7338
26	B	4.576883	48.84746
41	C	5.144738	46.69197
42	C	5.227464	46.13791
43	C	4.304202	39.90763
44	C	4.862149	47.33207
45	C	4.668385	52.17364
46	C	5.120878	46.14416
61	D	4.922059	45.45312
62	D	4.576648	48.30953
63	D	4.807593	42.94058
64	D	5.033737	51.08771
65	D	5.126781	53.35063
66	D	5.116051	48.5607

Qualitative Predictor with More than Two Classes

Creating indicating function in R

$$x_2 = \begin{cases} 1 & \text{for tool model A} \\ 0 & \text{otherwise} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{for tool model B} \\ 0 & \text{otherwise} \end{cases}$$

$$x_4 = \begin{cases} 1 & \text{for tool model C} \\ 0 & \text{otherwise} \end{cases}$$

```
x2<-ifelse(Catdata$group=='A',1,0)
x3<-ifelse(Catdata$group=='B',1,0)
x4<-ifelse(Catdata$group=='C',1,0)
```

Qualitative Predictor with More than Two Classes

	toolwear	speed	x2	x3	x4
1	59.26074	4.902189	1	0	0
2	55.55838	5.165739	1	0	0
3	47.44312	4.797517	1	0	0
4	47.28059	5.064308	1	0	0
5	41.35536	5.093231	1	0	0
6	52.35375	5.352190	1	0	0
21	51.34033	4.950873	0	1	0
22	47.03896	5.212557	0	1	0
23	60.66743	4.919606	0	1	0
24	55.86374	4.560823	0	1	0
25	53.73380	5.223331	0	1	0
26	48.84746	4.576883	0	1	0
41	46.69197	5.144738	0	0	1
42	46.13791	5.227464	0	0	1
43	39.90763	4.304202	0	0	1
44	47.33207	4.862149	0	0	1
45	52.17364	4.668385	0	0	1
46	46.14416	5.120878	0	0	1
61	45.45312	4.922059	0	0	0
62	48.30953	4.576648	0	0	0
63	42.94058	4.807593	0	0	0
64	51.08771	5.033737	0	0	0
65	53.35063	5.126781	0	0	0
66	48.56070	5.116051	0	0	0

Qualitative Predictor with More than Two Classes

If you run the regression in R as shown below, what do all these output mean?

```
x2<-ifelse(Catdata$group=='A',1,0)
x3<-ifelse(Catdata$group=='B',1,0)
x4<-ifelse(Catdata$group=='C',1,0)
df_new<-data.frame(toolwear=Catdata$y, speed=Catdata$x1, x2, x3, x4)
df_new
model1=lm(toolwear~speed+factor(x2)+factor(x3)+factor(x4),data=df_new)
summary(model1)
summary(model1)$coef
```

```
> summary(model1)$coef
              Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 29.7302164  10.680868 2.7835020 0.006801382
speed        3.8833220   2.133754 1.8199486 0.072756899
factor(x2)1  2.1128304   1.642591 1.2862791 0.202302342
factor(x3)1  1.4536097   1.643256 0.8845913 0.379204223
factor(x4)1  0.2910097   1.645463 0.1768558 0.860098372
```

ILLINOIS TECH

Qualitative Predictor with More than Two Classes

To understand the meaning of these regression coefficients, recall

$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}$$

and consider the case when we have tool model D for which $x_2 = 0, x_3 = 0,$ and $x_4 = 0$. In this case the mean response of y is given by

$$E(y) = \beta_0 + \beta_1 x_1 \qquad \text{Tool Model D}$$

Similarly, we see that

$$E(y) = (\beta_0 + \beta_2) + \beta_1 x_1 \quad \text{Tool Model A}$$

$$E(y) = (\beta_0 + \beta_3) + \beta_1 x_1 \quad \text{Tool Model B}$$

$$E(y) = (\beta_0 + \beta_4) + \beta_1 x_1 \quad \text{Tool Model C}$$

Qualitative Predictor with More than Two Classes

From

$$
\begin{aligned}
E(y) &= \beta_0 + \beta_1 x_1 && \text{Tool Model D} \\
E(y) &= (\beta_0 + \beta_2) + \beta_1 x_1 && \text{Tool Model A} \\
E(y) &= (\beta_0 + \beta_3) + \beta_1 x_1 && \text{Tool Model B} \\
E(y) &= (\beta_0 + \beta_4) + \beta_1 x_1 && \text{Tool Model C}
\end{aligned}
$$

we see that the coefficients β_2, β_3, and β_4 indicate, respectively, how much higher (lower) the mean response for tool model A, B, and C are than the one for tool model D.

Thus, β_2, β_3 and β_4 measure the differential effects of the qualitative variable classes on the height of the mean response function for any given level of x_1.

Qualitative Predictor with More than Two Classes

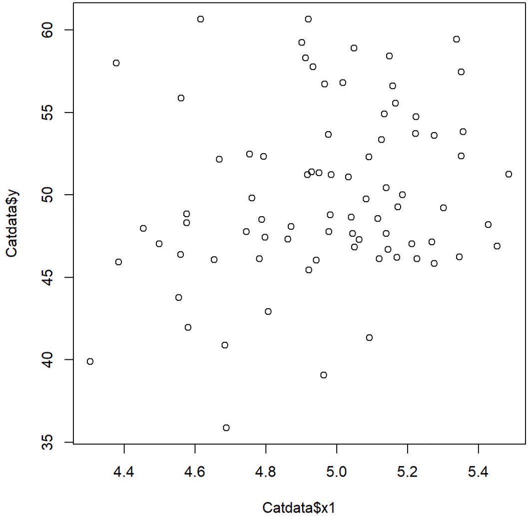
Let us examine these again:

```
x2<-ifelse(Catdata$group=='A',1,0)
x3<-ifelse(Catdata$group=='B',1,0)
x4<-ifelse(Catdata$group=='C',1,0)
df_new<-data.frame(toolwear=Catdata$y, speed=Catdata$x1, x2, x3, x4)
df_new
model1=lm(toolwear~speed+factor(x2)+factor(x3)+factor(x4),data=df_new)
summary(model1)
summary(model1)$coef
```

```
> summary(model1)$coef
               Estimate Std. Error   t value    Pr(>|t|)
(Intercept) 29.7302164  10.680868 2.7835020 0.006801382
speed        3.8833220   2.133754 1.8199486 0.072756899
factor(x2)1  2.1128304   1.642591 1.2862791 0.202302342
factor(x3)1  1.4536097   1.643256 0.8845913 0.379204223
factor(x4)1  0.2910097   1.645463 0.1768558 0.860098372
```

Qualitative Predictor with More than Two Classes

From $\text{plot}(\text{Catdata\$x1}, \text{Catdata\$y})$ we obtain

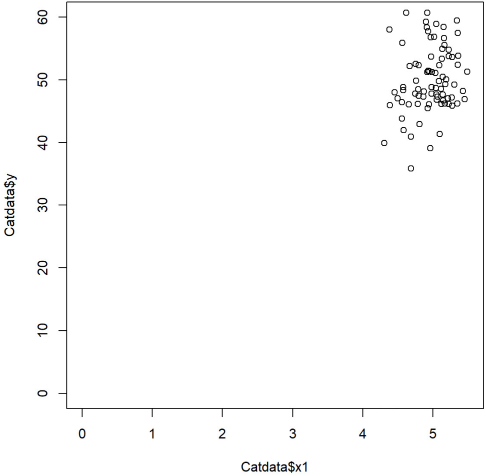


Qualitative Predictor with More than Two Classes

To see the y-intercept, we do

$$\text{plot}(\text{Catdata\$x1}, \text{Catdata\$y}, \text{xlim} = \text{c}(0, 5.5), \text{ylim} = \text{c}(0, 60))$$

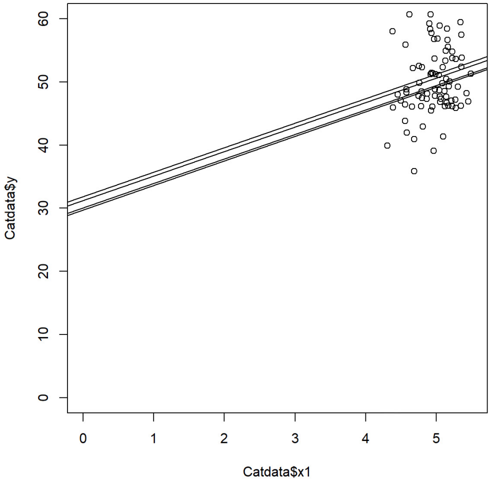
to obtain



Qualitative Predictor with More than Two Classes

```r
model1=lm(toolwear~speed+factor(x2)+factor(x3)+factor(x4),data=df_new)
summary(model1)$coef
abline(coef(model1)[1],coef(model1)[2])
abline(coef(model1)[1]+coef(model1)[3],coef(model1)[2])
abline(coef(model1)[1]+coef(model1)[4],coef(model1)[2])
abline(coef(model1)[1]+coef(model1)[5],coef(model1)[2])
```

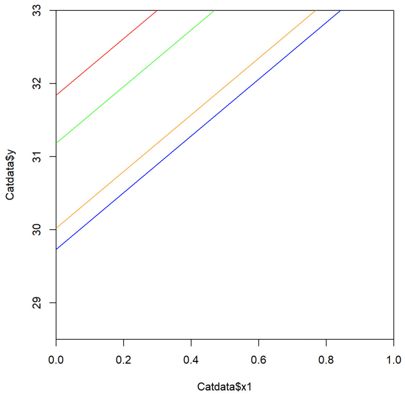
```
> summary(model1)$coef
              Estimate Std. Error   t value    Pr(>|t|)
(Intercept) 29.7302164  10.680868 2.7835020 0.006801382
speed        3.8833220   2.133754 1.8199486 0.072756899
factor(x2)1  2.1128304   1.642591 1.2862791 0.202302342
factor(x3)1  1.4536097   1.643256 0.8845913 0.379204223
factor(x4)1  0.2910097   1.645463 0.1768558 0.860098372
```



Qualitative Predictor with More than Two Classes

```
plot(Catdata$x1,Catdata$y, xlim = c(0, 1), ylim = c(28.5, 33))
abline(coef(model1)[1],coef(model1)[2], col="blue")
abline(coef(model1)[1]+coef(model1)[3],coef(model1)[2], col="red")
abline(coef(model1)[1]+coef(model1)[4],coef(model1)[2], col="green")
abline(coef(model1)[1]+coef(model1)[5],coef(model1)[2], col="orange")
```

```
> summary(model1)$coef
             Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 29.7302164  10.680868 2.7835020 0.006801382
speed        3.8833220   2.133754 1.8199486 0.072756899
factor(x2)1  2.1128304   1.642591 1.2862791 0.202302342
factor(x3)1  1.4536097   1.643256 0.8845913 0.379204223
factor(x4)1  0.2910097   1.645463 0.1768558 0.860098372
```



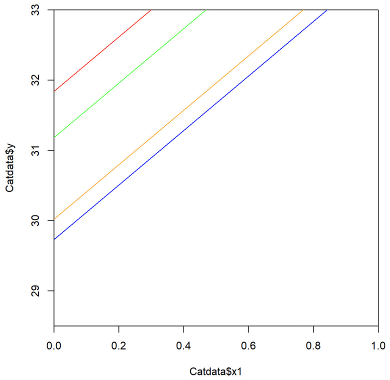
$$E(y) = \hat{\beta}_0 + \hat{\beta}_1 x_1 = 29.73 + 3.88 x_1 \quad \text{Blue D}$$

$$E(y) = (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 x_1 = 31.84 + 3.88 x_1 \quad \text{Red A}$$

$$E(y) = (\hat{\beta}_0 + \hat{\beta}_3) + \hat{\beta}_1 x_1 = 31.18 + 3.88 x_1 \quad \text{Green B}$$

$$E(y) = (\hat{\beta}_0 + \hat{\beta}_4) + \hat{\beta}_1 x_1 = 30.02 + 3.88 x_1 \quad \text{Orange C}$$

Qualitative Predictor with More than Two Classes




```
> summary(model1)$coef
             Estimate Std. Error   t value    Pr(>|t|)
(Intercept) 29.7302164  10.680868 2.7835020 0.006801382
speed        3.8833220   2.133754 1.8199486 0.072756899
factor(x2)1  2.1128304   1.642591 1.2862791 0.202302342
factor(x3)1  1.4536097   1.643256 0.8845913 0.379204223
factor(x4)1  0.2910097   1.645463 0.1768558 0.860098372
```

$E(y) = \hat{\beta}_0 + \hat{\beta}_1 x_1$ Blue D

$E(y) = (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 x_1$ Red A

$E(y) = (\hat{\beta}_0 + \hat{\beta}_3) + \hat{\beta}_1 x_1$ Green B

$E(y) = (\hat{\beta}_0 + \hat{\beta}_4) + \hat{\beta}_1 x_1$ Orange C

Now we can interpret the result as following:

For example, we say that the mean response for "tool wear" y is about 2.11 unit ($\hat{\beta}_2$) higher when using tool model A as compare to when using tool model D for all tool speed x_1.

Qualitative Predictor with More than Two Classes

Suppose you wish to estimate differential effects other than against tool model D. This can be done by estimating the difference between regression coefficients.

For example, $\hat{\beta}_3 - \hat{\beta}_4 \approx 1.45 - 0.29 = 1.16$ unit, hence we can say the following:

The mean response for tool wear y is about 1.16 unit higher when using tool model B as compare to when using tool model C for all tool speed x_1.

	Estimate
(Intercept)	29.7302164
speed	3.8833220
factor(x2)1	2.1128304
factor(x3)1	1.4536097
factor(x4)1	0.2910097

$E(y) = \hat{\beta}_0 + \hat{\beta}_1 x_1$	Blue D
$E(y) = (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 x_1$	Red A
$E(y) = (\hat{\beta}_0 + \hat{\beta}_3) + \hat{\beta}_1 x_1$	Green B
$E(y) = (\hat{\beta}_0 + \hat{\beta}_4) + \hat{\beta}_1 x_1$	Orange C

Qualitative Predictor with More than Two Classes

Notice that because model D is coded 0 for all the indicator variables x_2, x_3 and x_4, model D is then implicitly serves as the baseline category to which other models are compared.

Model	x_1	x_2	x_3	x_4
A	x_{i1}	1	0	0
B	x_{i1}	0	1	0
C	x_{i1}	0	0	1
D	x_{i1}	0	0	0

The choice of a baseline category is essentially arbitrary, however, the indicator coefficients β depend on which category is chosen as the baseline.

In some experiment, it is natural to select a particular category as a baseline, for example, an experiment that includes a "control group".