Math 455 Class 32 April 15

Last time: categorical variables are variables taking non-numerical values. We can put them in a regression with methods of making them numerical.

One such method is "dummy variables".

We studied "dummy codings".

Our example: X = red, yellow, or blue.

Choose reference level: blue

introduce $D_1 = \begin{cases} 1 & \text{if } X = \text{red} \\ 0 & \text{otherwise} \end{cases}$

$D_2 = \begin{cases} 1 & \text{if } X = \text{yellow} \\ 0 & \text{otherwise.} \end{cases}$

Another way to write this:

X	D_1	D_2
B	0	0
R	1	0
Y	0	1

An alternative to dummy codings: Sum codings.

X	D_1	D_2
B	-1	-1
R	1	0
Y	0	1

Sum codings replace the 0 in the "reference level" with -1.

This changes the interpretation of the regression coefficients of D_1 and D_2.

We consider the regression:

$$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \varepsilon.$$

With dummy codings the interpretations are

$$\hat{\beta}_0 = \bar{Y}_{blue} \qquad \beta_0 = E[Y \mid X = blue]$$

$$\hat{\beta}_1 = \bar{Y}_{red} - \bar{Y}_{blue} \qquad \beta_1 = E[Y \mid X = R] - E[Y \mid X = B]$$

$$\hat{\beta}_2 = \bar{Y}_{yellow} - \bar{Y}_{blue} \qquad \beta_2 = E[Y \mid X = yellow] - E[Y \mid X = B]$$

With sum codings the interpretations are

$$\beta_0 = \text{"grand mean"}$$

$$= \frac{E[Y \mid X = B] + E[Y \mid X = R] + E[Y \mid X = \text{yellow}]}{3}$$

$$\hat{\beta}_0 = \frac{\bar{Y}_{\text{blue}} + \bar{Y}_{\text{red}} + \bar{Y}_{\text{yellow}}}{3} \quad \text{"grand mean" (in data)}$$

⚠ Note "grand mean" $\neq \bar{Y}$:

there might be more blues than reds or yellows

$$\beta_1 = E[Y \mid X = R] - (\text{grand mean}).$$

$$\hat{\beta}_1 = \bar{Y}_{\text{red}} - (\text{grand mean})$$

$$\beta_2 = E[Y \mid X = \text{yellow}] - (\text{grand mean})$$

$$\hat{\beta}_2 = \bar{Y}_{\text{yellow}} - (\text{grand mean})$$

Question: What are $E[Y \mid X = \text{blue}] - (\text{grand mean})$

and $\bar{Y}_{\text{blue}} - (\text{grand mean})$?

$E[Y \mid X = \text{blue}]$ - grand mean

↙ from previous page

$= \beta_0 + \beta_1(-1) + \beta_2(-1) - \beta_0$

$= -\beta_1 - \beta_2$

$\bar{Y}_{\text{blue}}$ - grand mean $= 3 \cdot$ grand mean $- \bar{Y}_{\text{red}} - \bar{Y}_{\text{yellow}}$ - grand mean

$= (\text{grand mean} - \bar{Y}_{\text{red}}) + (\text{grand mean} - \bar{Y}_{\text{yellow}})$

$= -\hat{\beta}_1 - \hat{\beta}_2$ (from previous page)

Q: OK, all this references the previous page and the claims there. How are these claims proved?

A: Good news! The proofs are easy.
Bad news! The proofs are so easy that you might be asked to produce them on a test.

Q: Let's see a few of these proofs?

A: Look at the sum-coding case, first.

$$E[Y \mid X = R] = E[\beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot 0 + \varepsilon]$$

$$= \beta_0 + \beta_1$$

$$E[Y \mid X = \text{yellow}] = E[\beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 1 + \varepsilon]$$

$$= \beta_0 + \beta_2$$

$$E[Y \mid X = B] = E[\beta_0 + \beta_1 \cdot (-1) + \beta_2 \cdot (-1) + \varepsilon]$$

$$= \beta_0 - \beta_1 - \beta_2$$

$$\text{Grand Mean} = \frac{E[Y \mid X = R] + E[Y \mid X = \text{yellow}] + E[Y \mid X = B]}{3}$$

$$= \beta_0$$

Now let's consider a calculation in the dummy-coding case. We will show $\hat{\beta}_0 = \bar{Y}_{blue}$

Note that the error vector e is orthogonal to the columns of the design matrix

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & D_1 & D_2 \\ 1 & 1 & 1 \end{bmatrix}$$

$$(D_1)_i = \begin{cases} 1 & x_i = \text{red} \\ 0 & \text{otherwise} \end{cases}$$

$$(D_2)_i = \begin{cases} 1 & x_i = \text{yellow} \\ 0 & \text{otherwise} \end{cases}$$

Thus we have $e \cdot \vec{1} = e \cdot \vec{D_1} = e \cdot \vec{D_2} = 0$.

that is $\sum_{x_i = \text{red}} e_i = \sum_{x_i = \text{yellow}} e_i = \sum_{i} e_i = 0$.

It follows: $\sum_{x_i = \text{blue}} e_i = 0$.

$$\sum_{x_i = \text{blue}} Y_i = \sum_{x_i = \text{blue}} \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \hat{\beta}_2 \cdot 0 + e_i$$

$$= \sum_{x_i = \text{blue}} \hat{\beta}_0 + \cancel{\sum_{x_i = \text{blue}} e_i}^{\,0}$$

$$= n_{blue} \cdot \hat{\beta}_0$$

Thus $\hat{\beta}_0 = \bar{Y}_{blue}$

Problem 12. (3 + 4 + 3 marks) In Problem 3, we expressed history of alcohol use by using the following coding.

Alcohol Use	X_7	X_8
None	0	0
Moderate	1	0
Severe	0	1

(a) Suppose I just build a model of y on X_7 and X_8, then I get the following `lm` output. Interpret the coefficients of X_7 and X_8 in the following output.

```
> lmodInd = lm(y~x7+x8, data = surgicalunit)
> summary(lmodInd)

Call:
lm(formula = y ~ x7 + x8, data = surgicalunit)

Residuals:
    Min      1Q  Median      3Q     Max
-528.30 -229.81  -43.06  109.82 1296.70

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   599.80      94.99   6.315 6.57e-08 ***
x7             36.51     117.00   0.312  0.75628
x8            446.50     150.19   2.973  0.00449 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 367.9 on 51 degrees of freedom
Multiple R-squared:  0.1753,    Adjusted R-squared:  0.143
F-statistic: 5.421 on 2 and 51 DF,  p-value: 0.007336
```

(b) Now, suppose we express history of alcohol use by using the following coding.

Alcohol Use	X_7	X_8
None	-1	-1
Moderate	1	0
Severe	0	1

Now I build a second model of y on X_7 and X_8, then I get the following `lm` output. Interpret the coefficients of X_7 and X_8 in the following output.

```
> summary(lmodSum)

Call:
lm(formula = y ~ x7 + x8, data = SurgicalUnit1)

Residuals:
    Min      1Q  Median      3Q     Max
-528.30 -229.81  -43.06  109.82 1296.70

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   760.80      55.00  13.833  < 2e-16 ***
x7           -124.49      67.68  -1.839  0.07167 .
x8            285.50      86.81   3.289  0.00183 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 367.9 on 51 degrees of freedom
Multiple R-squared:  0.1753,    Adjusted R-squared:  0.143
F-statistic: 5.421 on 2 and 51 DF,  p-value: 0.007336
```

(c) What do you conclude about the effect of alcohol use on survival time? Does your conclusion change with change in codings at 5% level of significance?

Q: How do we interpret the estimated regression coefficient in the output?

A: Let's look at a past test question

(see next page)

$$\hat{Y} = \beta_0 + \beta_7 X_7 + \beta_8 X_8 + \varepsilon$$

So, if patient has "none" history of alcohol use,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_7 \cdot 0 + \hat{\beta}_8 \cdot 0 = \hat{\beta}_0.$$

Thus $\hat{\beta}_0$ is the estimated (fitted) value of Y when the patient has "none" history of alcohol use.

What is the interpretation of $\hat{\beta}_7$?

If the patient has "moderate" history,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_7 \cdot 1 + \hat{\beta}_8 \cdot 0 = \hat{\beta}_0 + \hat{\beta}_7.$$

Interpretation: $\hat{\beta}_7$ is the predicted (fitted) difference in response between a patient with "moderate"

history and a patient with "none" history.

Similarly: $\hat{\beta}_8$ is the predicted diff in Y between a patient with "Severe" and "none".

ICBST: $\hat{\beta}_0$ is the average value of Y for patients with "none" history..

$\hat{\beta}_0 + \hat{\beta}_7$ is the avg value of Y for patients with "moderate" history.

$\hat{\beta}_0 + \hat{\beta}_8$ is the avg value of Y for patients with "Severe" history

---

Alternate Codings "Non-Dummy Codings"

What do we require of the assignment of values to the dummy variables (X_7 and X_8 in the test question)?

Only that the info in X_7 & X_8 is exactly the info in the categorical variable.

In part (b), and alternate coding is given.

Many alternate codings are possible.

In part (b) we see "sum coding" (see ch. 14).

The coding changes the interpretation of the regression coefficients.

The predicted (fitted) values $\hat{Y}$ do NOT change.

The new interpretation: the value of $\hat{\beta}_7$ is the difference in Y between the average for patients with "moderate" and the "overall average" or "grand mean". (for all patients, not weighted by size of patient groups)

The value of $\hat{\beta}_8$ is the difference in average Y between patients with "severe" and this "grand mean".

Q: What about patients with "none"?

A: The difference in average Y between

patients with "none" and the "grand mean" is $-\hat{\beta}_7 * -\hat{\beta}_8$.

Q: Why don't we introduce a dummy var for every level and avoid this ugly $-\hat{\beta}_7 - \hat{\beta}_8$?

A: If we do this, we have that the sum of the dummy variables is the constant 1.

∴ This means that we have collinearity, which breaks our regression.

Problem 12:

Notice that the fitted values are the same with both codings

In (a)

if "None" is the level of alcohol use:

$\hat{Y} = 599.80 + 0 \cdot (36.51) + 0 \cdot (446.50)$

$= 599.80$.

if "Moderate", $X_7 = 1$ and $X_8 = 0$

$\hat{Y} = 599.80 + 1 \cdot (36.51) + 0 \cdot (446.50)$

$= 636.51$

if "Severe", $X_7 = 0$ and $X_8 = 1$

$\hat{Y} = 599.80 + 0 \cdot (36.51) + 1 \cdot 446.50$

$= 1046.3$

In (b):

if "None" the new X_7 and X_8 are both -1

$$\hat{Y} = 760.80 + (-1)(-124.49) + (-1)(285.50)$$

$$= 599.79 \quad \text{(same up to rounding)}$$

if "Moderate", $X_7 = 1$, $X_8 = 0$

$$\hat{Y} = 760.80 + 1 \cdot (-124.49) + 0 \cdot (285.50)$$

$$= 636.31 \quad \text{(same again!)}$$

if "Severe", $X_7 = 0$, $X_8 = 1$

$$\hat{Y} = 760.80 + 0 \cdot (-124.49) + 1 \cdot (285.50)$$

$$= 1046.3 \quad \text{(same!)}$$

Changing the coding of categorical variables does not change the fitted values.

We can observe that $2 \cdot 0 - 1 = -1$

$2 \cdot 1 - 1 = 1$

Interpretation:

In (a) the intercept is the average value of survival time for patients whose alcohol use is "None".

The coefficient of X_7 is the difference in $\hat{}^{avg}$ survival time for patients whose level of alcohol use is "moderate" vs. patients whose level of alcohol use is "None".

The coefficient of X_8 is the difference in avg survival time between patients whose alcohol use is "Severe" and patients whose alcohol use is "None".

Interpretation in (b):

A: The intercept is the overall average survival time.

The coefficient of X_7 is the difference in avg survival time between a patient with "Moderate" alcohol use and the overall average.

Q: Why is this coeff. negative in (b) and positive in (a)?

A: We are comparing the "Moderate" group to a different population. In (a) we compare to the "None" group and in (b) we compare to everyone. "Everyone" includes the "Severe" group who on average survived much longer.

Finally, the coeff of X_8 is the difference in avg. survival time between the "Severe" group and the average of everyone.