STA305/1004 - Class 13

October 24, 2019

Today's Class

- Coding qualitative predictors in regression models
- Estimating treatment effects using least squares
- In-class problem

Coding Qualitative Predictors in Regression Models

- A dummy or indicator variable in a regression takes on a finite number of values so that different categories of a nominal variable can be identified.
- The term dummy reflects the fact that the values taken on by such variables (e.g., 0, 1, -1) do not indicate meaningful measurements but rather categories of interest. (Kleinbaum et al., 1998)

Coding Qualitative Predictors in Regression Models

Consider a regression model: $y = \beta_0 + \beta_1 X_i + \epsilon$

Examples of dummy variables are:

$$X_1 = \begin{cases} 1 & \text{if treatment A} \\ 0 & \text{otherwise} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if subject is male} \\ -1 & \text{if subject is female} \end{cases}$$

The variables X_1, X_2 are nominal variables describing treatment group and sex respectively.

Coding Qualitative Predictors in Regression Models

The following rule should be applied to avoid collinearity in defining a dummy variable for regression analysis:

> *if the nominal independent variable of interest has k categories then exactly $k-1$ dummy variables should be defined to index the categories if the regression model contains an intercept term.*

Dummy Coding

- Dummy coding compares each level to the reference level.
- The intercept is the mean of the reference group.
- Suppose that we would like to compare the mean number of candy colours in each box. The data from 3 smarties boxes are below.

colour	count
Yellow	4
Yellow	3
Yellow	4
Purple	3
Purple	1
Purple	4
Green	2
Green	5
Green	1
Pink	1
Pink	2
Pink	4

Dummy Coding

The average and sd of each colour is:

```
#Get means for each flavour
sapply(split(count,colour),mean)
```

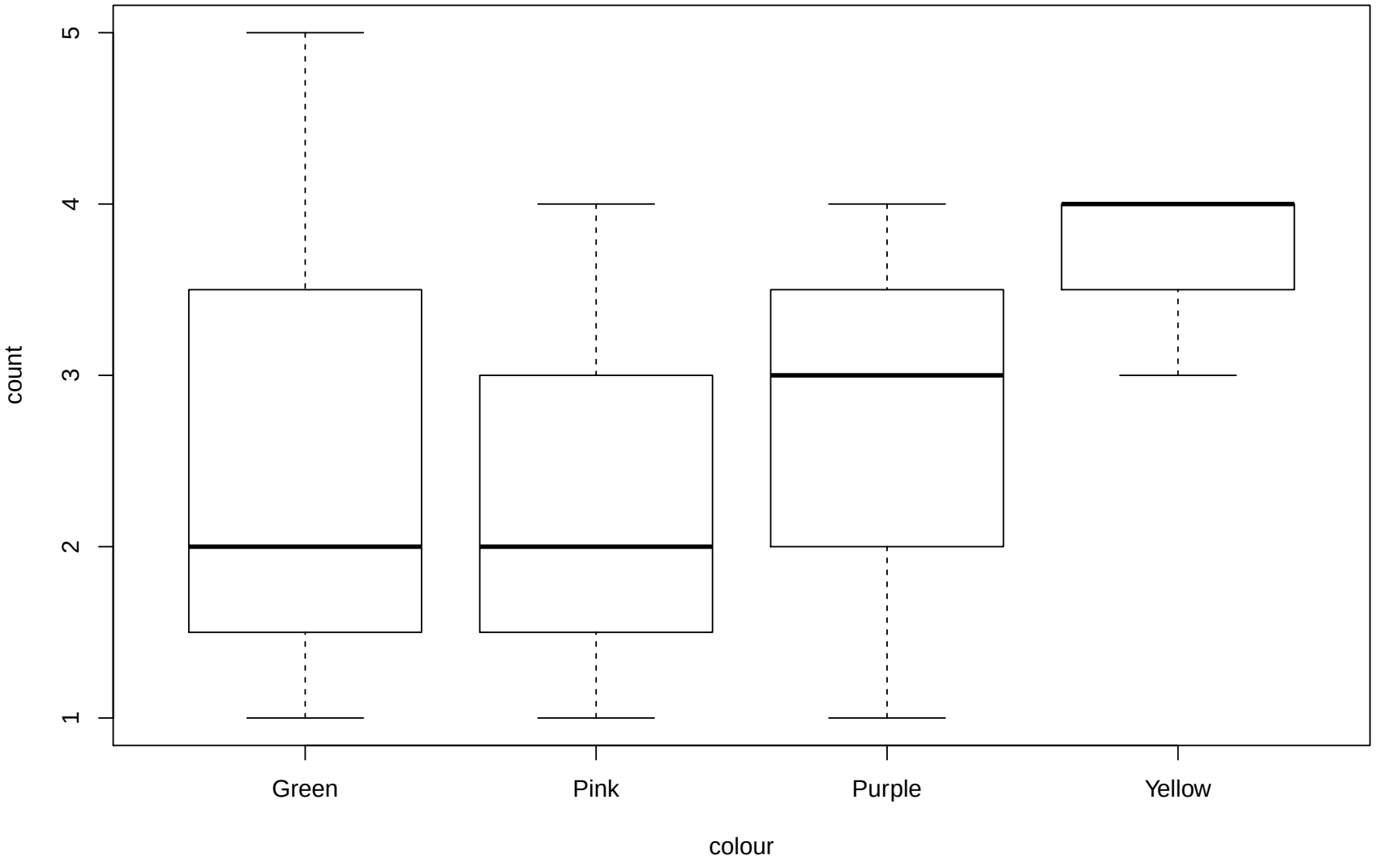
```
##    Green     Pink   Purple   Yellow
## 2.666667 2.333333 2.666667 3.666667
```

```
#Get standard deviations for each flavour
sapply(split(count,colour),sd)
```

```
##     Green      Pink    Purple    Yellow
## 2.0816660 1.5275252 1.5275252 0.5773503
```

Dummy Coding

```
boxplot(count~colour)
```



Dummy Coding

Dummy coding is the default in R and the most common coding scheme. It compares each level of the categorical variable to a fixed reference level.

```
contrasts(colour) <- contr.treatment(4)
contrasts(colour)  # print dummy coding - base is Green
```

```
##        2 3 4
## Green  0 0 0
## Pink   1 0 0
## Purple 0 1 0
## Yellow 0 0 1
```

Green is the reference category. The first column compares Pink to Green, the second column compares Purple to Green, and the third column compares Yellow to Green. The the three columns define three dummy variables:

Dummy Coding

$$X_1 = \begin{cases} 1 & \text{if smartie is pink} \\ 0 & \text{otherwise} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if smartie is purple} \\ 0 & \text{otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if smartie is yellow} \\ 0 & \text{otherwise} \end{cases}$$

If $X_1 = X_2 = X_3 = 0$ then the colour of the smartie is green - the reference category. This shows that we only require 3 dummy variables to define a nominal variable with 4 categories.

Dummy Coding

```
smarties <- data.frame(colour, count)
summary(lm(count ~ colour, data = smarties))$coefficients
```

```
##                  Estimate Std. Error       t value  Pr(>|t|)
## (Intercept)  2.666667e+00  0.8819171  3.023716e+00 0.0164661
## colour2     -3.333333e-01  1.2472191 -2.672612e-01 0.7960287
## colour3      4.710277e-16  1.2472191  3.776624e-16 1.0000000
## colour4      1.000000e+00  1.2472191  8.017837e-01 0.4458383
```

```
#Get means for each flavour
sapply(split(count,colour),mean)
```

```
##    Green     Pink   Purple   Yellow
## 2.666667 2.333333 2.666667 3.666667
```

Dummy Coding

To change the reference level change the value of base in `contr.treatment()`.

```
contrasts(colour) <- contr.treatment(4,base = 2) # Now reference is pink
contrasts(colour)
```

```
##        1 3 4
## Green  1 0 0
## Pink   0 0 0
## Purple 0 1 0
## Yellow 0 0 1
```

```
contrasts(colour) <- contr.treatment(4,base = 4) # Now reference is yellow
contrasts(colour)
```

```
##        1 2 3
## Green  1 0 0
## Pink   0 1 0
## Purple 0 0 1
## Yellow 0 0 0
```

Deviation Coding

- This coding system compares the mean of the dependent variable for a given level to the overall mean of the dependent variable.
-

$$X_1 = \begin{cases} 1 & \text{if smartie is green} \\ -1 & \text{if smartie is yellow} \\ 0 & \text{otherwise} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if smartie is pink} \\ -1 & \text{if smartie is yellow} \\ 0 & \text{otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if smartie is purple} \\ -1 & \text{if smartie is yellow} \\ 0 & \text{otherwise} \end{cases}$$

- 1 is used to compare a level to all other levels and -1 is assigned to yellow because it's the level that will never be compared to the other levels.

Deviation Coding

- In R the variables can be created using the `contr.sum()` function.
- The argument of 4 in `contr.sum(4)` indicates the number of levels of the factor.

```
contrasts(colour) <- contr.sum(4)
contrasts(colour)
```

```
##        [,1] [,2] [,3]
## Green     1    0    0
## Pink      0    1    0
## Purple    0    0    1
## Yellow   -1   -1   -1
```

Deviation Coding

```
summary(lm(count ~ colour))$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  2.8333333  0.4409586  6.4253960 0.0002035567
## colour1     -0.1666667  0.7637626 -0.2182179 0.8327229152
## colour2     -0.5000000  0.7637626 -0.6546537 0.5310577712
## colour3     -0.1666667  0.7637626 -0.2182179 0.8327229152
```

```
#Get means for each flavour
sapply(split(count,colour),mean)
```

```
##    Green     Pink   Purple   Yellow
## 2.666667 2.333333 2.666667 3.666667
```

Example - blood coagualtion study

The table below gives coagulation times for blood samples drawn from 24 animals receiving four different diets A, B, C, and D.

	A	B	C	D
	60	65	71	62
	63	66	66	60
	59	67	68	61
	63	63	68	64
	62	64	67	63
	59	71	68	56
Treatment Average	61	66	68	61
Grand Average	64	64	64	64
Difference	-3	2	4	-3

Estimating treatment effects using least squares

y_{ij} is the j^{th} observation under the i^{th} treatment. Let μ be the overall mean. The model for diet $y_{ij} = \mu + \tau_i + \epsilon_{ij}$, $\epsilon_{ij} \sim N(0, \sigma^2)$ can be written in terms of the dummy variables X_1, X_2, X_3 as:

$$y_{ij} = \mu + \tau_1 X_{1j} + \tau_2 X_{2j} + \tau_3 X_{3j} + \epsilon_{ij},$$

where,

$$X_{1j} = \begin{cases} 1 & \text{if jth unit recieves diet 2} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{2j} = \begin{cases} 1 & \text{if jth unit recieves diet 3} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{3j} = \begin{cases} 1 & \text{if jth unit recieves diet 4} \\ 0 & \text{otherwise} \end{cases}$$

Estimating treatment effects using least squares

It follows that $E(y_{Aj}) = \mu_A = \mu$ is the mean of diet A so

$$
\begin{aligned}
E(y_{Bj}) = \mu_B = \mu_A + \tau_1 \Rightarrow \tau_1 = \mu_B - \mu_A \\
E(y_{Cj}) = \mu_C = \mu_A + \tau_2 \Rightarrow \tau_2 = \mu_C - \mu_A \\
E(y_{Dj}) = \mu_D = \mu_A + \tau_3 \Rightarrow \tau_3 = \mu_D - \mu_A
\end{aligned}
$$

The least squares estimates are:

$$
\begin{aligned}
\hat{\mu} &= \bar{y}_{1\cdot}, \\
\hat{\tau}_1 &= \bar{y}_{2\cdot} - \bar{y}_{1\cdot}, \\
\hat{\tau}_2 &= \bar{y}_{3\cdot} - \bar{y}_{1\cdot}, \\
\hat{\tau}_3 &= \bar{y}_{3\cdot} - \bar{y}_{1\cdot}.
\end{aligned}
$$

Estimating treatment effects using least squares

- This model can also be written in matrix notation

$$y = X\beta + \epsilon$$

,

where $\beta = (\mu, \tau_1, \tau_2, \tau_3)$, $X = (\mathbf{1}, X_{i1}, X_{i2}, X_{i3})$, and $\epsilon = (\epsilon_{ij})$.

- X is an 24×4 design matrix with $\mathbf{1}$ is a 24×1 column vector of 1s, and ϵ is an 24×1 column vector.
- Note that τ_4 corresponding to the 4th treatment is implicitly set to 0. It is used as a constraint so that that $(X'X)^{-1}$ exists.

Example - blood coagualtion study (treatment coding)

```
contrasts(tab0401$diets)
```

```
  B C D
A 0 0 0
B 1 0 0
C 0 1 0
D 0 0 1
```

```
lm.diets <- lm(y~diets,data=tab0401);round(summary(lm.diets)$coefficients,2)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)       61       0.97   63.14        0
dietsB             5       1.37    3.66        0
dietsC             7       1.37    5.12        0
dietsD             0       1.37    0.00        1
```

Example - blood coagualtion study (treatment coding)

The averages for each of the four diets are in the table below.

Diet	A ($j = 1$)	B ($j = 2$)	C ($j = 3$)	D ($j = 4$)
Average ($\bar{y}_{j\cdot}$)	61	66	68	61

$$
\begin{aligned}
\bar{y}_{1\cdot} &= 61, \\
\hat{\tau}_1 &= \bar{y}_{2\cdot} - \bar{y}_{1\cdot} = 5 \\
\hat{\tau}_2 &= \bar{y}_{3\cdot} - \bar{y}_{1\cdot} = 7 \\
\hat{\tau}_3 &= \bar{y}_{3\cdot} - \bar{y}_{1\cdot} = -9.9 \times 10^{-15}.
\end{aligned}
$$

Example - blood coagualtion study (treatment coding)

The design matrix (first 12 observations) is

```
model.matrix(lm.diets)[1:12,]
```

```
##    (Intercept) dietsB dietsC dietsD
## 1            1      0      0      0
## 2            1      0      0      0
## 3            1      0      0      0
## 4            1      0      0      0
## 5            1      0      0      0
## 6            1      0      0      0
## 7            1      1      0      0
## 8            1      1      0      0
## 9            1      1      0      0
## 10           1      1      0      0
## 11           1      1      0      0
## 12           1      1      0      0
```

Example - blood coagualtion study (treatment coding)

The design matrix (first 12 observations) with the observations y and treatment variable `diets` (first 12 observations) is

```
cbind(tab0401$y,tab0401$diets,model.matrix(lm.diets))[1:12,]
```

```
##          (Intercept) dietsB dietsC dietsD
## 1  62 1            1      0      0      0
## 2  60 1            1      0      0      0
## 3  63 1            1      0      0      0
## 4  59 1            1      0      0      0
## 5  63 1            1      0      0      0
## 6  59 1            1      0      0      0
## 7  63 2            1      1      0      0
## 8  67 2            1      1      0      0
## 9  71 2            1      1      0      0
## 10 64 2            1      1      0      0
## 11 65 2            1      1      0      0
## 12 66 2            1      1      0      0
```

Example - blood coagualtion study (deviation coding)

If deviation coding was used then the parameter estimates would represent different treatment effects. In the regression model the dummy variables would be defined as

$$X_1 = \begin{cases} 1 & \text{if diet is A} \\ -1 & \text{if diet is D} \\ 0 & \text{otherwise} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if diet is B} \\ -1 & \text{if diet is D} \\ 0 & \text{otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if diet is C} \\ -1 & \text{if diet is D} \\ 0 & \text{otherwise} \end{cases}$$

Example - blood coagualtion study (deviation coding)

It follows that

$$
\begin{aligned}
E(y_{Aj}) &= \mu_A = \tau_0 + \tau_1 \\
E(y_{Bj}) &= \mu_B = \tau_0 + \tau_2 \\
E(y_{Cj}) &= \mu_C = \tau_0 + \tau_3 \\
E(y_{Dj}) &= \mu_D = \tau_0 - \tau_1 - \tau_2 - \tau_3
\end{aligned}
$$

So,

$$
\begin{aligned}
\tau_0 &= \frac{\mu_A + \mu_B + \mu_C + \mu_D}{4} \\
\tau_1 &= \mu_A - \frac{\mu_A + \mu_B + \mu_C + \mu_D}{4} \\
\tau_2 &= \mu_B - \frac{\mu_A + \mu_B + \mu_C + \mu_D}{4} \\
\tau_3 &= \mu_C - \frac{\mu_A + \mu_B + \mu_C + \mu_D}{4}
\end{aligned}
$$

Example - blood coagualtion study (deviation coding)

```
attach(tab0401)
contrasts(tab0401$diets) <- contr.sum(4)
contrasts(tab0401$diets)
```

```
  [,1] [,2] [,3]
A    1    0    0
B    0    1    0
C    0    0    1
D   -1   -1   -1
```

```
lm.diets <- lm(y~diets,data=tab0401)
round(summary(lm.diets)$coefficients,2)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)       64       0.48  132.49     0.00
diets1            -3       0.84   -3.59     0.00
diets2             2       0.84    2.39     0.03
diets3             4       0.84    4.78     0.00
```

- The estimate of the intercept $\hat{\tau}_0$ is the grand average.
- The slope estimates $\hat{\tau}_1, \hat{\tau}_2, \hat{\tau}_3$ are the differences between the treatment averages and grand average of diets A, B, and C.