R02 - Regression with Categorical Independent Variables

HCI/PSYCH 522
Iowa State University

March 3, 2022

Binary independent variable

Recall the simple linear regression model

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2).$$

If we have a binary independent variable, i.e. the independent variable only has two levels say level A and level B, we can code it as

$$X_i = \mathrm{I}(\text{observation } i \text{ is level A})$$

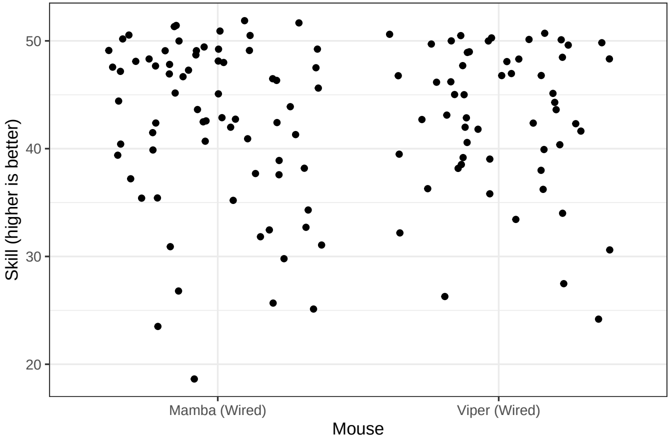
where I(statement) is an indicator function that is 1 when "statement" is true and 0 otherwise. Then

- β_0 is the mean response for level B,
- $\beta_0 + \beta_1$ is the mean response for level A, and
- β_1 is the mean difference in response (level A minus level B).

Player Skill Data

```
mouse <- read_csv("mouse.csv", show_col_types = FALSE) %>% mutate(Mouse = factor(Mouse))
head(mouse)

# A tibble: 6 x 2
  Skill Mouse
  <dbl> <fct>
1  35.5 Dell
2  35.4 Dell
3  34.9 Dell
4  34.8 Dell
5  33.8 Dell
6  33.5 Dell

summary(mouse)

     Skill                        Mouse
 Min.   : 6.4   Basilisk (Wired)    :57
 1st Qu.:31.8   Dell                :49
 Median :39.5   Mamba (Wired)       :71
 Mean   :38.8   Mamba (Wireless)    :56
 3rd Qu.:46.9   Viper (Wired, light):60
 Max.   :54.6   Viper (Wired)       :56
```

Player Skill Plot



Regression model for skill

Let

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

where Y_i is the Skill of the ith individual and

$$X_i = \mathrm{I}(\text{Mouse for observation } i \text{ is Viper (Wired)})$$

then

- mean skill using Mamba (Wired) is β_0,
- mean skill using Viper (Wired) is $\beta_0 + \beta_1$, and
- mean difference in skill [Viper (Wired) minus Mamba (Wired)] is β_1.

R code

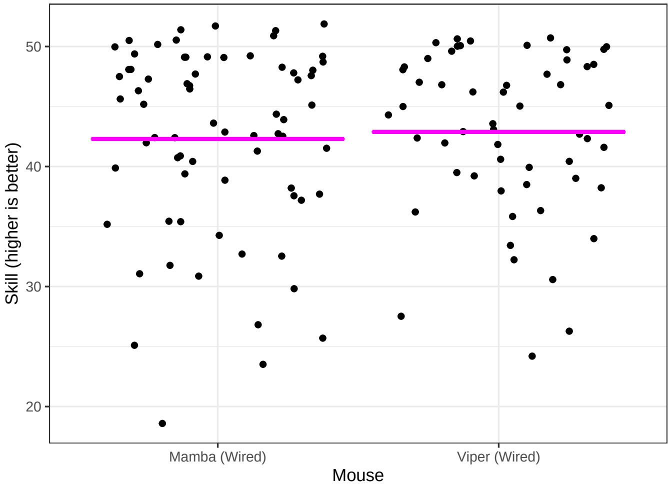
```
two_mice <- mouse %>% filter(Mouse %in% c("Viper (Wired)","Mamba (Wired)"))
two_mice$X <- ifelse(two_mice$Mouse == "Viper (Wired)", 1, 0)
m <- lm(Skill ~ X, data = two_mice)
summary(m)
```

```

Call:
lm(formula = Skill ~ X, data = two_mice)

Residuals:
    Min      1Q  Median      3Q     Max
-23.697  -3.991   1.414   5.803   9.603

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  42.2972     0.8676   48.75   <2e-16 ***
X             0.5885     1.3066    0.45    0.653
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.311 on 125 degrees of freedom
Multiple R-squared:  0.001621,Adjusted R-squared:  -0.006367
F-statistic: 0.2029 on 1 and 125 DF,  p-value: 0.6532
```

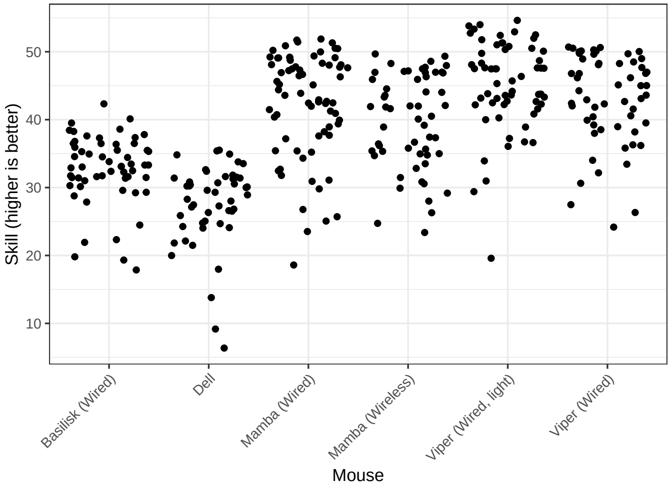
R code

```
m <- lm(Skill ~ Mouse, data = two_mice)
summary(m)

Call:
lm(formula = Skill ~ Mouse, data = two_mice)

Residuals:
    Min      1Q  Median      3Q     Max
-23.697  -3.991   1.414   5.803   9.603

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         42.2972     0.8676   48.75   <2e-16 ***
MouseViper (Wired)   0.5885     1.3066    0.45    0.653
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.311 on 125 degrees of freedom
Multiple R-squared:  0.001621,Adjusted R-squared:  -0.006367
F-statistic: 0.2029 on 1 and 125 DF,  p-value: 0.6532

emmeans(m, "Mouse")

 Mouse         emmean    SE  df lower.CL upper.CL
 Mamba (Wired)   42.3 0.868 125     40.6     44.0
 Viper (Wired)   42.9 0.977 125     41.0     44.8

Confidence level used: 0.95
```

Mice Skills


Skill (higher is better)
50
40
30
20
Mamba (Wired)
Viper (Wired)
Mouse

Using a categorical variable as an independent variable.



Regression with a categorical variable

1. Choose one of the levels as the reference level, e.g. Basilisk (Wired)
2. Construct dummy variables using indicator functions, i.e.

$$\mathrm{I}(A) = \begin{cases} 1 & A \text{ is TRUE} \\ 0 & A \text{ is FALSE} \end{cases}$$

   for the other levels, e.g.

$$\begin{aligned}
X_{i,1} &= \mathrm{I}(\text{Mouse for observation } i \text{ is Dell}) \\
X_{i,2} &= \mathrm{I}(\text{Mouse for observation } i \text{ is Mamba (Wired)}) \\
X_{i,3} &= \mathrm{I}(\text{Mouse for observation } i \text{ is Mamba (Wireless)}) \\
X_{i,4} &= \mathrm{I}(\text{Mouse for observation } i \text{ is Viper (Wired, light)}) \\
X_{i,5} &= \mathrm{I}(\text{Mouse for observation } i \text{ is Viper (Wired)})
\end{aligned}$$

3. Estimate the parameters of a multiple regression model using these dummy variables.

Regression model

Our regression model becomes

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \beta_4 X_{i,4} + \beta_5 X_{i,5}, \sigma^2)$$

where

- β_0 is the mean skill in the Basilisk (Wired) group
- $\beta_0 + \beta_1$ is the mean skill in the Dell group
- $\beta_0 + \beta_2$ is the mean skill in the Mamba (Wired) group
- $\beta_0 + \beta_3$ is the mean skill in the Mamba (Wireless) group
- $\beta_0 + \beta_4$ is the mean skill in the Viper (Wired, light) group
- $\beta_0 + \beta_5$ is the mean skill in the Viper (Wired) group

and thus β_p for $p > 0$ is the difference in mean skills between one group and the reference group.

R code

```
m <- lm(Skill ~ Mouse, data = mouse)
m

Call:
lm(formula = Skill ~ Mouse, data = mouse)

Coefficients:
            (Intercept)                MouseDell        MouseMamba (Wired)     MouseMamba (Wireless)
                 32.691                   -5.289                     9.606                     6.994
MouseViper (Wired, light)        MouseViper (Wired)
                 12.425                   10.194
```

```
confint(m)

                              2.5 %    97.5 %
(Intercept)               30.951394 34.431062
MouseDell                 -7.848142 -2.730232
MouseMamba (Wired)         7.269897 11.942013
MouseMamba (Wireless)      4.523030  9.465943
MouseViper (Wired, light)  9.995893 14.854984
MouseViper (Wired)         7.723030 12.665943
```

R code (cont.)

```
summary(m)

Call:
lm(formula = Skill ~ Mouse, data = mouse)

Residuals:
     Min       1Q   Median       3Q      Max
-25.5167  -3.3857   0.8143   5.1833  10.0143

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                32.6912     0.8846  36.958  < 2e-16 ***
MouseDell                  -5.2892     1.3010  -4.065 5.95e-05 ***
MouseMamba (Wired)          9.6060     1.1877   8.088 1.06e-14 ***
MouseMamba (Wireless)       6.9945     1.2565   5.567 5.25e-08 ***
MouseViper (Wired, light)  12.4254     1.2352  10.059  < 2e-16 ***
MouseViper (Wired)         10.1945     1.2565   8.113 8.88e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.678 on 343 degrees of freedom
Multiple R-squared:  0.4543,Adjusted R-squared:  0.4463
F-statistic:  57.1 on 5 and 343 DF,  p-value: < 2.2e-16
```

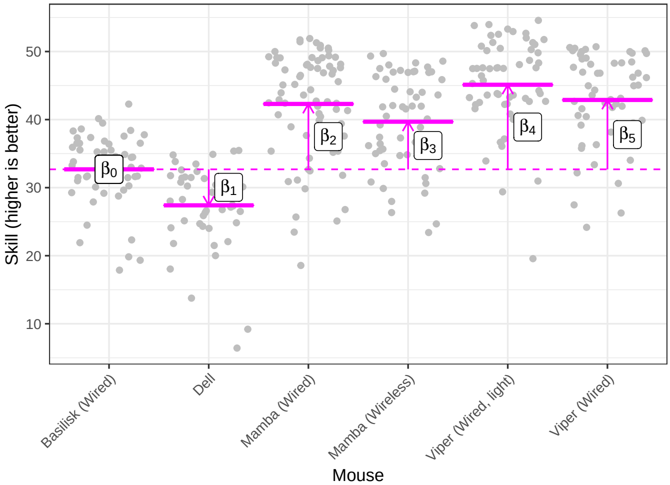
Interpretation

- β_0, i.e. mean of the dependent variable for the reference level
- $\beta_p, p > 0$: mean change in the dependent variable when moving from the reference level to the level associated with the p^{th} dummy variable

For example,

- The mean skill using the Basilisk (Wired) mouse is 32.7 (31,34.4).
- The mean increase in skill when using the Viper (Wired, light) mouse compared to the Basilisk (Wired) mouse is 12.4 (10,14.9).
- This model explains 45% of the variability in skill.

Using a categorical variable as an independent variable.



Group means with 95% credible intervals

```
em <- emmeans(m, pairwise ~ Mouse)
em$emmeans

 Mouse                  emmean    SE  df lower.CL upper.CL
 Basilisk (Wired)         32.7 0.885 343     31.0     34.4
 Dell                     27.4 0.954 343     25.5     29.3
 Mamba (Wired)            42.3 0.793 343     40.7     43.9
 Mamba (Wireless)         39.7 0.892 343     37.9     41.4
 Viper (Wired, light)     45.1 0.862 343     43.4     46.8
 Viper (Wired)            42.9 0.892 343     41.1     44.6

Confidence level used: 0.95
```

Group comparisons with 95% credible intervals

```
confint(em$contrasts)

 contrast                                   estimate   SE  df lower.CL upper.CL
 Basilisk (Wired) - Dell                       5.289 1.30 343    1.561    9.018
 Basilisk (Wired) - Mamba (Wired)             -9.606 1.19 343  -13.010   -6.202
 Basilisk (Wired) - Mamba (Wireless)          -6.994 1.26 343  -10.596   -3.393
 Basilisk (Wired) - Viper (Wired, light)     -12.425 1.24 343  -15.965   -8.885
 Basilisk (Wired) - Viper (Wired)            -10.194 1.26 343  -13.796   -6.593
 Dell - Mamba (Wired)                        -14.895 1.24 343  -18.450  -11.341
 Dell - Mamba (Wireless)                     -12.284 1.31 343  -16.028   -8.540
 Dell - Viper (Wired, light)                 -17.715 1.29 343  -21.400  -14.029
 Dell - Viper (Wired)                        -15.484 1.31 343  -19.228  -11.740
 Mamba (Wired) - Mamba (Wireless)              2.611 1.19 343   -0.809    6.032
 Mamba (Wired) - Viper (Wired, light)         -2.819 1.17 343   -6.176    0.537
 Mamba (Wired) - Viper (Wired)                -0.589 1.19 343   -4.009    2.832
 Mamba (Wireless) - Viper (Wired, light)      -5.431 1.24 343   -8.987   -1.875
 Mamba (Wireless) - Viper (Wired)             -3.200 1.26 343   -6.817    0.417
 Viper (Wired, light) - Viper (Wired)          2.231 1.24 343   -1.325    5.787

Confidence level used: 0.95
Conf-level adjustment: tukey method for comparing a family of 6 estimates
```

Changing the reference level

If you want to change the reference level, you can

```
mouse <- mouse %>%
  mutate(Mouse = relevel(Mouse, ref = "Dell"))

m <- lm(Skill ~ Mouse, data = mouse)
summary(m)


Call:
lm(formula = Skill ~ Mouse, data = mouse)

Residuals:
     Min       1Q   Median       3Q      Max
-25.5167  -3.3857   0.8143   5.1833  10.0143

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                 27.402      0.954  28.722  < 2e-16 ***
MouseBasilisk (Wired)        5.289      1.301   4.065 5.95e-05 ***
MouseMamba (Wired)          14.895      1.240  12.009  < 2e-16 ***
MouseMamba (Wireless)       12.284      1.306   9.403  < 2e-16 ***
MouseViper (Wired, light)   17.715      1.286  13.776  < 2e-16 ***
MouseViper (Wired)          15.484      1.306  11.852  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.678 on 343 degrees of freedom
Multiple R-squared:  0.4543,Adjusted R-squared:  0.4463
```

(Almost) equivalence to our multiple group model

Recall that we had a multiple group model

$$Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma_j^2)$$

for groups $j = 0, 1, 2, \ldots, 5$.

Our regression model is (almost) a reparameterization of the multiple group model:

Basilisk (Wired):	μ_0	$= \beta_0$
Viper (Wired, light):	μ_1	$= \beta_0 + \beta_1$
Mamba (Wired):	μ_2	$= \beta_0 + \beta_2$
Dell:	μ_3	$= \beta_0 + \beta_3$
Viper (Wired):	μ_4	$= \beta_0 + \beta_4$
Mamba (Wireless):	μ_5	$= \beta_0 + \beta_5$

assuming the groups are labeled appropriately.

Summary

When you run a regression with a categorical variable, you are

1. Choosing one of the levels as the reference level.
2. Constructing dummy variables using indicator functions for all other levels, e.g.

$$X_i = \mathrm{I}(\text{observation } i \text{ is } <\text{some non-reference level}>).$$

3. Estimating the parameters of a multiple regression model using these dummy variables.