

## R04 - Regression with Categorical Explanatory Variables

STAT 587 (Engineering)  
Iowa State University

October 26, 2020

## Binary explanatory 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 explanatory variable, i.e. the explanatory variable only has two levels say level A and level B, we can code it as

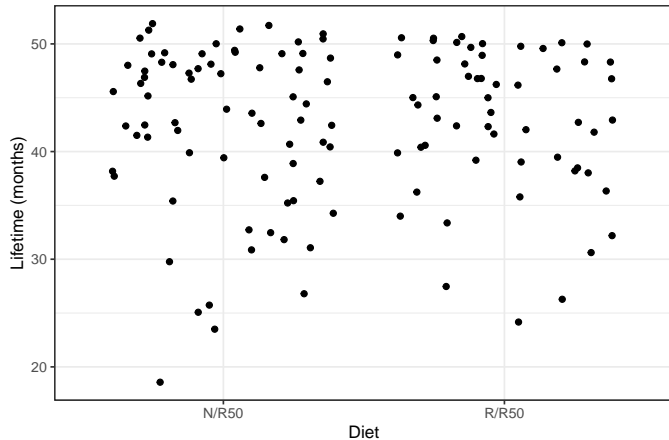
$$X_i = \text{I}(\text{observation } i \text{ is level A})$$

where  $\text{I}(\text{statement})$  is an **indicator function** that is 1 when *statement* is true and 0 otherwise. Then

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

# Mice lifetimes

Sleuth3::case0501



## Regression model for mice lifetimes

Let

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

where  $Y_i$  is the lifetime of the  $i$ th mouse and

$$X_i = I(\text{Diet}_i = \text{N/R50})$$

then

$$\begin{aligned} E[\text{Lifetime} | \text{R/R50}] &= E[Y_i | X_i = 0] = \beta_0 \\ E[\text{Lifetime} | \text{N/R50}] &= E[Y_i | X_i = 1] = \beta_0 + \beta_1 \end{aligned}$$

and

$$\begin{aligned} E[\text{Lifetime difference}] &= E[\text{Lifetime} | \text{N/R50}] - E[\text{Lifetime} | \text{R/R50}] \\ &= (\beta_0 + \beta_1) - \beta_0 = \beta_1. \end{aligned}$$

# R code

```

case0501$X <- ifelse(case0501$Diet == "N/R50", 1, 0)
(m <- lm(Lifetime ~ X, data = case0501, subset = Diet %in% c("R/R50", "N/R50")))

Call:
lm(formula = Lifetime ~ X, data = case0501, subset = Diet %in%
  c("R/R50", "N/R50"))

Coefficients:
(Intercept)          X
   42.8857      -0.5885

confint(m)

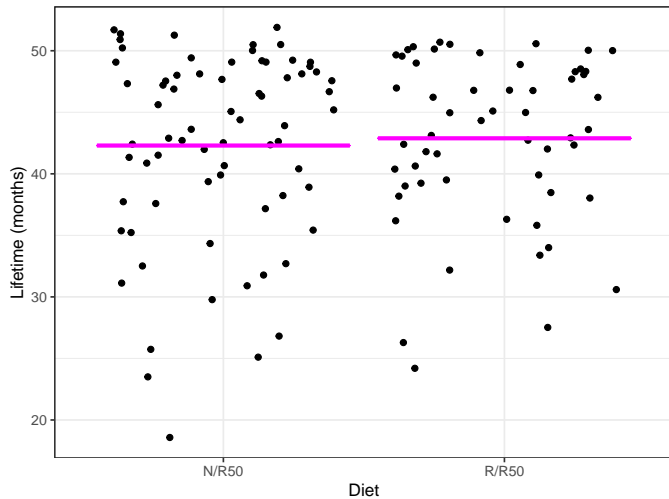
              2.5 %    97.5 %
(Intercept) 40.952257 44.819172
X           -3.174405  1.997342

predict(m, data.frame(X=1), interval = "confidence") # Expected lifetime on N/R50

      fit      lwr      upr
1 42.29718 40.58007 44.0143

```

# Mice lifetimes



## Equivalence to a two-sample t-test

Recall that our two-sample t-test had the model

$$Y_{ij} \overset{ind}{\sim} N(\mu_j, \sigma^2)$$

for groups  $j = 0, 1$ . This is equivalent to our current regression model where

$$\begin{aligned}\mu_0 &= \beta_0 \\ \mu_1 &= \beta_0 + \beta_1\end{aligned}$$

assuming

- $\mu_0$  represents the mean for the R/R50 group and
- $\mu_1$  represents the mean for N/R50 group.

When the models are effectively the same, but have different parameters we the model is **reparameterized**.

# Equivalence

```
summary(m)$coefficients[2,4] # p-value
```

```
[1] 0.6531748
```

```
confint(m)
```

```

          2.5 %    97.5 %
(Intercept) 40.952257 44.819172
X           -3.174405  1.997342
```

```
t.test(Lifetime ~ Diet, data = case0501, subset = Diet %in% c("R/R50", "N/R50"), var.equal=TRUE)
```

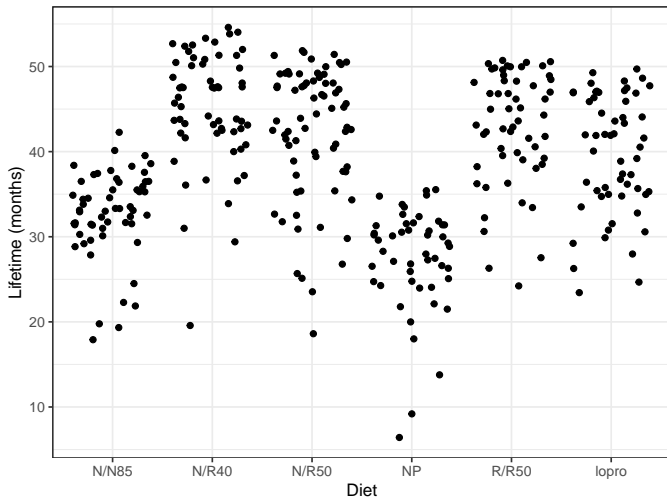
Two Sample t-test

```

data: Lifetime by Diet
t = -0.45044, df = 125, p-value = 0.6532
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.174405  1.997342
sample estimates:
mean in group N/R50 mean in group R/R50
    42.29718         42.88571
```



# Using a categorical variable as an explanatory variable.



## Regression with a categorical variable

1. Choose one of the levels as the **reference** level, e.g. N/N85
2. Construct dummy variables using indicator functions, i.e.

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

for the other levels, e.g.

$$X_{i,1} = I(\text{diet for observation } i \text{ is N/R40})$$

$$X_{i,2} = I(\text{diet for observation } i \text{ is N/R50})$$

$$X_{i,3} = I(\text{diet for observation } i \text{ is NP})$$

$$X_{i,4} = I(\text{diet for observation } i \text{ is R/R50})$$

$$X_{i,5} = I(\text{diet for observation } i \text{ is loopro})$$

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

- $\beta_0$  is the expected lifetime for the N/N85 group
- $\beta_0 + \beta_1$  is the expected lifetime for the N/R40 group
- $\beta_0 + \beta_2$  is the expected lifetime for the N/R50 group
- $\beta_0 + \beta_3$  is the expected lifetime for the NP group
- $\beta_0 + \beta_4$  is the expected lifetime for the R/R50 group
- $\beta_0 + \beta_5$  is the expected lifetime for the lopro group

and thus  $\beta_p$  for  $p > 0$  is the difference in expected lifetimes between one group and a **reference** group.

# R code

```
case0501 <- case0501 %>%
  mutate(X1 = Diet == "N/R40",
         X2 = Diet == "N/R50",
         X3 = Diet == "NP",
         X4 = Diet == "R/R50",
         X5 = Diet == "lopro")

m <- lm(Lifetime ~ X1 + X2 + X3 + X4 + X5, data = case0501)
m
```

Call:

```
lm(formula = Lifetime ~ X1 + X2 + X3 + X4 + X5, data = case0501)
```

Coefficients:

(Intercept)	X1TRUE	X2TRUE	X3TRUE	X4TRUE	X5TRUE
32.691	12.425	9.606	-5.289	10.194	6.994

```
confint(m)
```

	2.5 %	97.5 %
(Intercept)	30.951394	34.431062
X1TRUE	9.995893	14.854984
X2TRUE	7.269897	11.942013
X3TRUE	-7.848142	-2.730232
X4TRUE	7.723030	12.665943
X5TRUE	4.523030	9.465943

# R code (cont.)

```
summary(m)
```

Call:

```
lm(formula = Lifetime ~ X1 + X2 + X3 + X4 + X5, data = case0501)
```

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 ***
X1TRUE	12.4254	1.2352	10.059	< 2e-16 ***
X2TRUE	9.6060	1.1877	8.088	1.06e-14 ***
X3TRUE	-5.2892	1.3010	-4.065	5.95e-05 ***
X4TRUE	10.1945	1.2565	8.113	8.88e-15 ***
X5TRUE	6.9945	1.2565	5.567	5.25e-08 ***

---  
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

- $\beta_0 = E[Y_i | \text{reference level}]$ , i.e. expected response for the reference level

Note: the only way  $X_{i,1} = \dots = X_{i,p} = 0$  is if all indicators are zero, i.e. at the reference level.

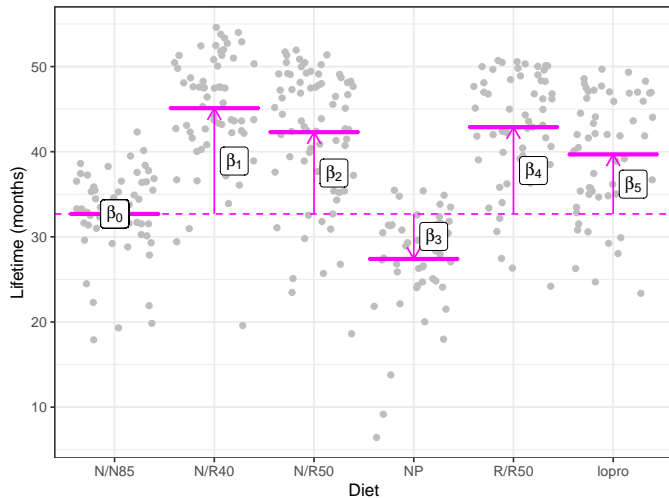
- $\beta_p, p > 0$ : expected change in the response moving from the reference level to the level associated with the  $p^{\text{th}}$  dummy variable

Note: the only way for  $X_{i,p}$  to increase by one is if initially  $X_{i,1} = \dots = X_{i,p} = 0$  and now  $X_{i,p} = 1$

For example,

- The expected lifetime for mice on the N/N85 diet is 32.7 (31.0, 34.4) months.
- The expected increase in lifetime for mice on the N/R40 diet compared to the N/N85 diet is 12.4 (10.0, 14.9) months.
- The model explains 45% of the variability in mice lifetimes.

# Using a categorical variable as an explanatory variable.



## Equivalence to multiple group model

Recall that we had a multiple group model

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

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

Our regression model is a **reparameterization** of the multiple group model:

$$\begin{array}{lll} N/N85 : & \mu_0 & = \beta_0 \\ N/R40 : & \mu_1 & = \beta_0 + \beta_1 \\ N/R50 : & \mu_2 & = \beta_0 + \beta_2 \\ NP : & \mu_3 & = \beta_0 + \beta_3 \\ R/R50 : & \mu_4 & = \beta_0 + \beta_4 \\ lopro : & \mu_5 & = \beta_0 + \beta_5 \end{array}$$

assuming the groups are labeled appropriately.



# Summary

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

$$X_i = I(\text{observation } i \text{ is } \langle \text{some non-reference level} \rangle).$$

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