

R04 - Categorical explanatory variables

STAT 401 (Engineering) - Iowa State University

March 25, 2018

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 0 and level 1, we can code it as

$$X_i = \begin{cases} 1 & \text{observation } i \text{ is level 1} \\ 0 & \text{observation } i \text{ is level 0} \end{cases}$$

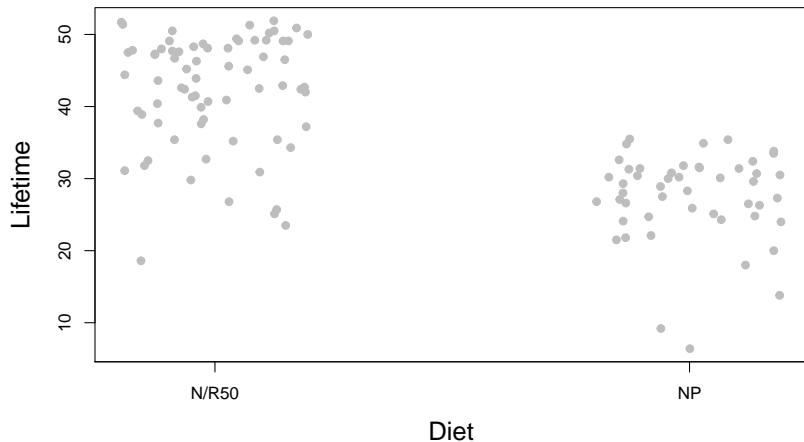
$$= I(\text{observation } i \text{ is level 1})$$

where $I(A)$ is an **indicator function** that is 1 when A is true and 0 otherwise. Then

- β_0 is the expected response for level 0,
- $\beta_0 + \beta_1$ is the expected response for level 1, and therefore
- β_1 is the expected difference in response (level 1 minus level 0).

Mice lifetimes

Reconsider the mice lifetime data set but only consider the diets NP and N/R50:



Regression model for mice lifetimes

Considering only the NP and N/R50 diets. 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

$$X_i = I(\text{Diet}_i = \text{N/R50}) = \begin{cases} 1 & \text{ith mouse diet is N/R50} \\ 0 & \text{ith mouse diet is NP} \end{cases}$$

then

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

and

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

R code

```

case0501$X <- ifelse(case0501$Diet == "N/R50", 1, 0)
unique(case0501$X)

[1] 0 1

(m <- lm(Lifetime ~ X, data = case0501, subset = Diet %in% c("NP", "N/R50")))

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

Coefficients:
(Intercept)          X
      27.4         14.9

confint(m)

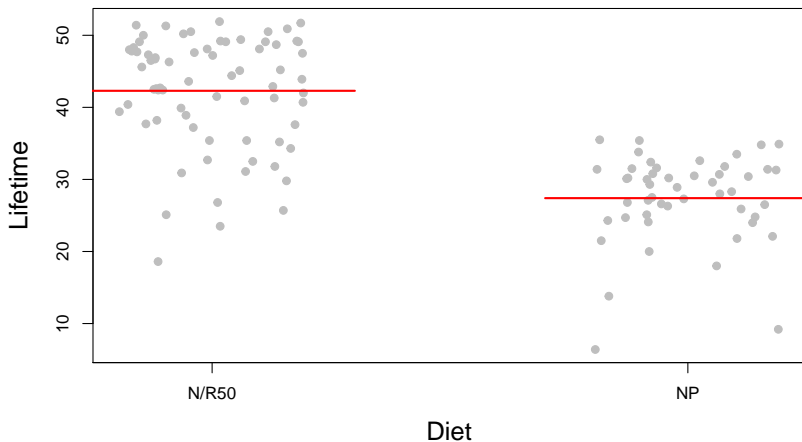
              2.5 %    97.5 %
(Intercept) 25.37974 29.42434
X           12.26605 17.52424

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

      fit      lwr      upr
1 42.29718 40.61717 43.9772

```

Mice lifetimes



Equivalence to model for two-sample t-test

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

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

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

- μ_0 represents the mean for the NP group and
- μ_1 represents the mean for N/R50 group.

When the models are effectively the same, but have different parameters we call it a **reparameterization**.

Equivalence

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

```
[1] 2.534716e-20
```

```
confint(m)
```

```

                2.5 %   97.5 %
(Intercept) 25.37974 29.42434
X           12.26605 17.52424
```

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

Two Sample t-test

data: Lifetime by Diet

t = 11.219, df = 118, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

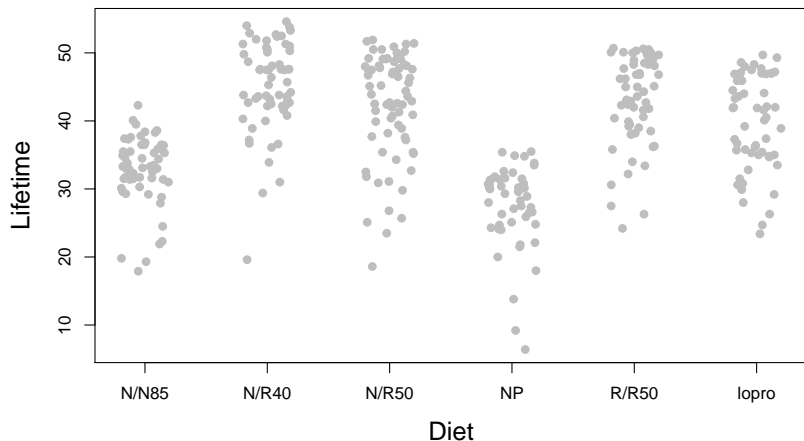
95 percent confidence interval:

12.26605 17.52424

sample estimates:

mean in group N/R50	mean in group NP
42.29718	27.40204

Using a categorical variable as an explanatory variable.



Regression with a categorical variable

- Choose one of the levels as the **reference** level, e.g. N/N85
- 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.

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

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

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

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

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

- 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 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 β_p for $p > 0$ is the difference in expected lifetimes.

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

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.

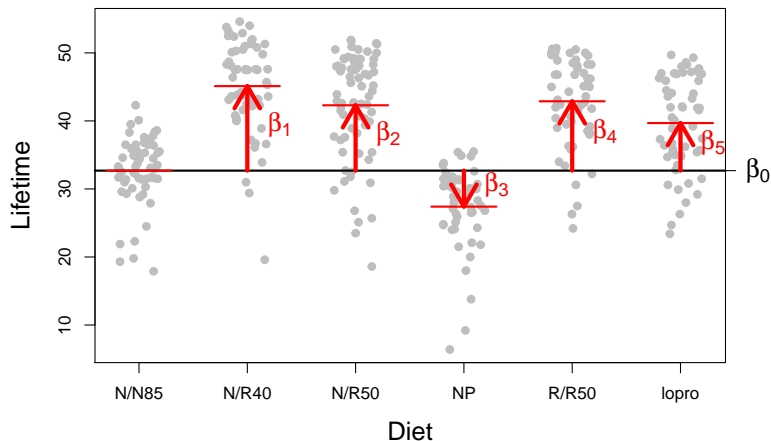
- $\beta_p, p > 0$: expected change in the response moving from the reference level to the level associated with the p^{th} dummy variable

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) weeks.
- 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) weeks.
- 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 $j = 0, 1, 2, \dots, 5$. This is equivalent to our current regression model where

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

assuming the groups are labeled appropriately.