Interactions Involving Categorical Predictors

Tutorial 12

Ravleen Bajaj

Categorical Predictors and Interactions

- Interactions allow the effect of one variable to depend on the level or value of another.
- Today:
  - Interactions between **two categorical predictors**
  - Interactions between a **categorical** and a **numeric** predictor

Case 1: 1 Categorical, 1 Numerical Variable

Suppose we have a dataset where treatment group is coded as:

$$\text{treatment} = \begin{cases} 1 & \text{Placebo} \\ 2 & \text{Low Dose} \\ 3 & \text{High Dose} \end{cases}$$

- These numeric labels **must be converted to a factor in R**.
- Otherwise, R interprets them as numeric and fits a line through 1–3.

We create indicator variables:

$$x_{1i} = \begin{cases} 1 & \text{if patient i gets low dose} \\ 0 & \text{otherwise} \end{cases} \qquad x_{2i} = \begin{cases} 1 & \text{if patient i gets high dose} \\ 0 & \text{otherwise} \end{cases}$$

Model With Numeric Predictor Interaction

We want to model the tumor size Y_i of the i^{th} individuals using their Age (x_{3i}), and the treatment administered to them.

- Categorical predictor (factor): treatment group (Placebo, Low, High)
- Numeric predictor: age (x_{3i})

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 (x_{1i} x_{3i}) + \beta_5 (x_{2i} x_{3i}) + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$ and $\varepsilon_i's$ are independent.

Interpretation

Placebo:

$$Y_i = \beta_0 + \beta_3 x_{3i} + \varepsilon_i$$

Low Dose:

$$Y_i = \beta_0 + \beta_1 + \beta_3 x_{3i} + \beta_4 x_{3i} + \varepsilon_i$$

High Dose:

$$Y_i = \beta_0 + \beta_2 + \beta_3 x_{3i} + \beta_5 x_{3i} + \varepsilon_i$$

- β_0 is the mean blood pressure for those who were administered the placebo and were of age 0.
- β_4 is the difference in slopes, i.e., the effect of age between the Low Dose and Placebo groups.
- β_5 is the difference in slopes, i.e., the effect of age between the high Dose and Placebo groups.

Case 2: Interaction between Categorical Variables

Another variable represents disease severity:

$$\text{severity} \in \{\text{"malignant"}, \text{"benign"}\}$$

- R orders character levels alphabetically: benign, malignant.
- This may not be scientifically meaningful.
- We can reorder them manually:

```
severity <- factor(severity, levels =
c("malignant","malign"))
```

$$x_{1i} = \begin{cases} 1 & \text{if patient i gets low dose} \\ 0 & \text{otherwise} \end{cases} \qquad x_{2i} = \begin{cases} 1 & \text{if patient i gets high dose} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{3i} = \begin{cases} 1 & \text{if tumor of patient i is malignant} \\ 0 & \text{otherwise} \end{cases}$$

Two Categorical Predictors With Interaction

Now, we model tumor size Y_i using:

- Treatment group (Placebo, Low, High)
- Severity (malignant, benign)

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 (x_{1i} x_{3i}) + \beta_5 (x_{2i} x_{3i}) + \varepsilon_i$$

This allows a separate mean response for each of the 3×2 combinations.

Also, the mean structure is characterized by 6 parameters!

Interpretation of Categorical Interaction

- There are 3 treatment groups and 2 severity groups.
- Without interaction:
  - Treatment effect is the same across severity groups.
  - Severity effect is the same across treatment groups.
- With interaction:
  - Treatment effect may differ by severity level.
  - Severity effect may differ by treatment group.
- Interaction terms quantify how the effect of one factor changes across the other factor.

Interpretation

- **Placebo + Benign**: $Y_i = \beta_0 + \varepsilon_i$
- **Low Dose + Benign**: $Y_i = \beta_0 + \beta_1 + \varepsilon_i$
- **High Dose + Benign**: $Y_i = \beta_0 + \beta_2 + \varepsilon_i$
- **Placebo + Malignant**: $Y_i = \beta_0 + \beta_3 + \varepsilon_i$
- **Low Dose + Malignant**: $Y_i = \beta_0 + \beta_1 + \beta_3 + \beta_4 + \varepsilon_i$
- **High Dose + Malignant**: $Y_i = \beta_0 + \beta_2 + \beta_3 + \beta_5 + \varepsilon_i$
- β_0: Mean tumor size for Placebo dosage for Benign tumors
- β_1: Effect of Low Dose vs Placebo for Benign tumors
- β_2: Effect of High Dose vs Placebo for Benign tumors
- β_3: Effect of Malignant vs Benign for Placebo group
- β_4: Interaction between Low Dose and Malignant (how the Low Dose effect differs between Malignant and Benign)
- β_5: Interaction between High Dose and Malignant (how the High Dose effect differs between Malignant and Benign)

Formally, all of β_1 through β_5 represent changes in the mean response across groups.

Key Takeaways

- Categorical predictors must be encoded properly:
  - Convert numeric labels (1,2,3) into factors.
  - Manually reorder character levels when needed.
- Interactions between categorical variables allow the effect of one variable to depend on another variable.
- Categorical-numeric interactions allow the coefficient of the numeric variable to differ across levels of a factor.

Common Mistakes from Midterm

- True Coefficients: β_i
- Estimated Coefficients: $\hat{\beta}_i$
- Coefficients Estimators: $\tilde{\beta}_i$
- Y: Random variable
- y: realization of a random variable
- $\beta_0, \beta_1, \sigma^2$: True parameters (fixed, unknown constants)
- ε_i: Random error term
- x_i: Predictor value (known, fixed)

Common Mistakes from Midterm (contd.)

True Model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

Fitted Model:

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- $\hat{\mu}_i$: Estimated mean response of i^{th} individual (a number).
- $\hat{\beta}_0, \hat{\beta}_1$: Estimated coefficients (numbers from our data)
- **No error term!** This is our estimate of the mean

What we actually see: $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\epsilon}_i$
or equivalently: $\hat{\epsilon}_i = y_i - \hat{\mu}_i$

- y_i: Observed response of the i^{th} individual (actual data point)
- $\hat{\epsilon}_i$: i^{th} Residual (calculated)

Common Mistakes from Midterm (contd.)

Mistake: true model followed by the same model with the estimated parameters substituted for the true parameters

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{True model}$$

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i \quad \text{Wrong!}$$

Common Mistakes from Midterm (contd.)

Output:

```
model <- lm(PIQ ~ Brain + Height + Weight)
summary(model)
# Coefficients:
#               Estimate Std. Error t value Pr(>|t|)
# (Intercept)  1.114e+02  6.297e+01   1.768 0.085979 .
# Brain        2.060e+00  5.634e-01   3.657 0.000856 ***
# Height      -2.732e+00  1.229e+00  -2.222 0.033034 *
# Weight       5.599e-04  1.971e-01   0.003 0.997750
```

Correct:

$$\hat{y}_i \text{ or } \hat{\mu}_i = 111.4 + 2.060 \cdot \text{Brain}_i - 2.732 \cdot \text{Height}_i + 0.0005599 \cdot \text{Weight}_i$$

Thank You!

Questions?