STAT 302 - Chapter 4: Additional Topics in Regression - Part 3

Harsha Perera

Coding Categorical Variables

- In this section we will be looking categorical predictors with more than two levels

- In chapter 3, we discussed how to incorporate a categorical predictors with two levels

- In this section we extend that idea for more than two levels

Categorical Predictor with Two Levels

Variable: Sex
Levels: Male / Female

Sex	IMale	IFemale
Female	0	1
Male	1	0
Female	0	1
Female	0	1
Male	1	0
Male	1	0

Categorical Predictor with Three Levels

Variable: Car type
Levels: Accord, Maxima, Mazda6

Car type	IMazda6	IMaxima	IAccord
Accord	0	0	1
Maxima	0	1	0
Mazda6	1	0	0
Accord	0	0	1
Accord	0	0	1
Maxima	0	1	0
Mazada6	1	0	0

Example 4.9: Car Prices - Multiple Linear Regression

- Objective is to predict prices of used car based on Mileage and Car Type ($n = 90$)

- Response: Price

- Predictors
  - Mileage
  - Car type: IAccord, IMazda6, IMaxima

$$Price = \beta_0 + \beta_1 \cdot IAccord + \beta_2 \cdot IMazda6 + \beta_3 \cdot IMaxima + \in$$

Example 4.9: Car prices - Multiple Linear Regression

- Any one of the indicator variables is exactly a linear function of the other two

- If we know the values of two indicator variables we can predict the other one

  e.g. if $IAccord = 0$ and $IMaxima = 0$, then $IMazda6 = 1$

- Therefore We don't need to include all three indicator variables

Example 4.9: Car Prices - Multiple Linear Regression

- Fitted model:

$$\widehat{Price} = 14.28 - 2.77 \cdot IMazda6 + 1.19 \cdot IMaxima$$

- When $IMaxima = 0$ and $IMazda = 0$, $IAccord = 1$

  $\widehat{Price} = 14.28 - 2.77 \cdot (0) + 1.19 \cdot (0) = 14.28$
  Average price of a used Accord is \$14,280

- When $IMaxima = 0$ and $IMazda = 1$, $IAccord = 0$
  $\widehat{Price} = 14.28 - 2.77 \cdot (1) + 1.19 \cdot (0) = 11.51$

  Average price of a used Mazda6 is \$11,510

- When $IMaxima = 1$ and $IMazda = 0$, $IAccord = 0$
  $\widehat{Price} = 14.28 - 2.77 \cdot (0) + 1.19 \cdot (1) = 15.47$

  Average price of a used Maxima is \$15,470

Coding Categorical Variables

- When a categorical predictor has K levels,

  We need $K - 1$ indicator variables to represent each level of that variables

- The Indicator variable to be excluded from the model is called: **Reference category / level**

Price vs. Mileage for Three Car types

- Objective is to find the mean price of one of the three car types at a given mileage
- Model :

$$Price = \beta_0 + \beta_1 \cdot Mileage + \beta_2 \cdot IMazda6 + \beta_3 \cdot IMaxima + \in$$

- From R :

$$\widehat{Price} = 21.09 - 0.1249 \cdot Mileage - 1.2616 \cdot IMazda6 + 1.5397 \cdot IMaxima$$

Price vs. Mileage for Three Car Types...

- Intercept: $\hat{\beta}_0 = 21.09$

  When $Mileage = 0$, $IMaxima = 0$ and $IMazda6 = 0$, model provides the base price of an *Accord*

  Base price of an *Accord* $= \$21,087$

- When $Mileage = 0$, $IMaxima = 0$ and $IMazda6 = 1$, model provides the base price of a $Mazda6$

  Base price of a $Mazda6 = \$21,087 - \$1261.6 = \$19,827$

  $\hat{\beta}_2$: indicates that predicted base price of a Mazda6 is \$1261.6 less than an Accord

Price vs. Mileage for Three Car Types...

- When *Mileage* $= 0$, *IMaxima* $= 1$ and *IMazda6* $= 0$, model provides the base price of a *Maxima*

  Base price of a *Maxima* $= \$21,087 + \$1539.7 = \$22,626.7$

  $\hat{\beta}_3$: indicates that predicted base price of a Maxima is \$1539.7 more than an Accord

- $\hat{\beta}_1 = -0.12$

  The price of any of the three cars will decrease by \$0.124 for every mile driven

  The model assumes rate of depreciation is same all three car models. Three different lines (different intercepts) with same slope

  If not we need to introduce an interaction term between Car type and mileage.

Price vs. Mileage for Three Car types: With the Interaction

- Is the assumption of a common slope reasonable?
- To test we need to fit a model with the interaction between Mileage and Car type

$$Price = \beta_0 + \beta_1 \cdot Mileage + \beta_2 \cdot IMazda + \beta_3 \cdot IMaxima +$$

$$\beta_4 \cdot Mileage \cdot IMazda + \beta_5 \cdot Mileage \cdot IMaxima + \in$$

Price vs. Mileage for Three Car types: With the Interaction

- For an Accord: $IMaxima = 0$ and $IMazda6 = 0$

$$Price = \beta_0 + \beta_1 \cdot Mileage$$

- For Mazda6: $IMaxima = 0$ and $IMazda6 = 1$
  vspace12pt

$$Price = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) \cdot Mileage$$

β_2: change in the intercept of Mazda6 in comparison to an Accord

β_4: change in the slope of Mazda6 in comparison to an Accord

Price vs. Mileage for Three Car types: With the Interaction

- For Maxima: $IMaxima = 1$ and $IMazda6 = 0$

$$Price = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) \cdot Mileage$$

β_3: change in the intercept of Maxima in comparison to an Accord

β_5: change in the slope of Maxima in comparison to an Accord

Hypothesis test for Interaction terms - Nested F test

- Hypotheses: $H_0 : \beta_4 = \beta_5 = 0$ vs. $H_1 : \beta_4 \neq 0$ or $\beta_5 \neq 0$

  p-value > 0.05: fail to reject the null hypothesis

  Interaction term is insignificant. Assumption of common slope is reasonable.