Lesson 19. Coding Categorical Predictors – Part 1

Note. In Part 2 of this lesson, you can run the R code that generates the outputs here in Part 1.

1 Overview

- We have seen how to include categorical predictor variables when there are only two categories
- Now we'll see what to do when there are more than two categories

Example 1. Let's look at the data in `ThreeCars2017` from the `Stat2Data` library, which contains information on 90 randomly selected used cars. In particular, we will consider *CarType* (Accord, Maxima, or Mazda6), *Price* (in $1000s), and *Mileage* (in 1000s of miles).

We want to predict a car's price based on its mileage and type. In particular, can we answer the following questions:

- Are car type and mileage useful predictors of price?
- What is the predicted price of a car with given characteristics?
- For a fixed mileage, does the price of a car differ by car type, on average?
- After accounting for car type, how is a car's mileage related to its price, on average?

- We can run the following R code:

```
library(Stat2Data)
data(ThreeCars2017)
head(ThreeCars2017)
```

Here is the output:

A data.frame: 6 × 7

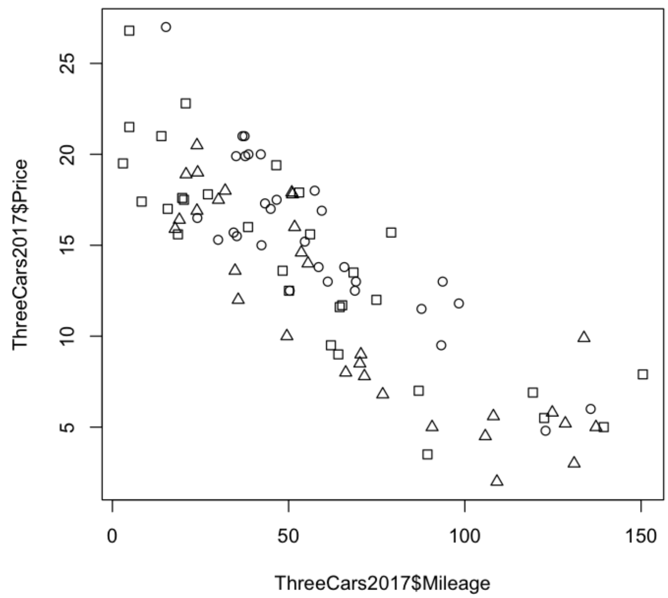
	CarType	Age	Price	Mileage	Mazda6	Accord	Maxima
	<fct>	<int>	<dbl>	<dbl>	<int>	<int>	<int>
1	Mazda6	3	15.9	17.8	1	0	0
2	Mazda6	2	16.4	19.0	1	0	0
3	Mazda6	1	18.9	20.9	1	0	0
4	Mazda6	2	16.9	24.0	1	0	0
5	Mazda6	2	20.5	24.0	1	0	0
6	Mazda6	1	19.0	24.2	1	0	0

- Let's visualize the data by creating a scatterplot, with different point shapes (the `pch` parameter) for each *CarType*

Note the use of nested `ifelse()` to assign 3 shapes for 3 categories

```
plot(ThreeCars2017$Mileage, ThreeCars2017$Price,
     pch=ifelse(ThreeCars2017$CarType == "Accord", 0,
                ifelse(ThreeCars2017$CarType == "Maxima", 1, 2)))
```

Here is the output:



2 Including categorical predictors into a regression model

- To include a categorical variable with more than 2 categories:
  - Select one group to be the **reference category**
  - Include an indicator variable for each other category
- So, if we have ℓ categories, we will have $\ell - 1$ indicator variables
- Note! Do not code a categorical variable as one predictor with groups labeled by numerical values (e.g., $X \in \{0, 1, 2\}$)
  - This forces the group intercepts to be equally spaced – not what we're going for
  - This also yields different intercepts if we assign the group labels differently – also not what we're going for
  - See STAT2 Section 4.5 for a cautionary demonstration of this incorrect approach

3 Allowing different intercepts for each group

Example 2. Continuing with Example 1...

- For *CarType*, let Accord be the reference category
- Then, we define indicator variables for Maxima and Mazda6:

- Our model:

- For Accords, our model reduces to:

- For Maximas, our model reduces to:

- For Mazda6s, our model reduces to:

- Coefficients:
  - β_0:
  - β_1:
  - β_2:
  - β_3:

- We can fit this model with the following R code:

```
fit <- lm(Price ~ Mileage + as.factor(CarType), data = ThreeCars2017)
summary(fit)
```

⚠ See Part 2 for other ways to do this in R – in particular, if you want to change the reference category

The output is as follows:

```
Call:
lm(formula = Price ~ Mileage + as.factor(CarType), data = ThreeCars2017)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4208 -2.1225 -0.2257  1.6904  6.7866

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              21.087383   0.682805  30.883   <2e-16 ***
Mileage                  -0.124906   0.008252 -15.136   <2e-16 ***
as.factor(CarType)Maxima  1.539735   0.726685   2.119   0.0370 *
as.factor(CarType)Mazda6 -1.261552   0.733145  -1.721   0.0889 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.813 on 86 degrees of freedom
Multiple R-squared:  0.7518,	Adjusted R-squared:  0.7431
F-statistic: 86.81 on 3 and 86 DF,  p-value: < 2.2e-16
```

Example 3. Continuing with Example 2...

a. What is the fitted model?

b. Predict the price of a Maxima with 30,000 miles.

c. Carefully interpret the coefficient of the *Mazda6* indicator variable.

d. For a fixed car type, describe the estimated relationship between mileage and price.

e. Is the relationship you described in part d statistically significant?

4 Allowing different intercepts and slopes for each group

Example 4. Continuing Example 3...

- The model that would allow for different intercepts and slopes is:

- We can fit this model with the following R code:

```
fit <- lm(Price ~ Mileage + as.factor(CarType) + Mileage:as.factor(CarType),
          data = ThreeCars2017)
summary(fit)
```

The output is as follows:

```
Call:
lm(formula = Price ~ Mileage + as.factor(CarType) + Mileage:as.factor(CarType),
    data = ThreeCars2017)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5984 -2.0047 -0.1778  1.8321  6.7536

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      20.809613   0.876372  23.745  < 2e-16 ***
Mileage                          -0.119812   0.012964  -9.242 1.93e-14 ***
as.factor(CarType)Maxima          2.461613   1.467904   1.677   0.0973 .
as.factor(CarType)Mazda6         -1.016487   1.355525  -0.750   0.4554
Mileage:as.factor(CarType)Maxima -0.016325   0.022540  -0.724   0.4709
Mileage:as.factor(CarType)Mazda6 -0.004603   0.018668  -0.247   0.8058
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.837 on 84 degrees of freedom
Multiple R-squared:  0.7533,  Adjusted R-squared:  0.7386
F-statistic:  51.3 on 5 and 84 DF,  p-value: < 2.2e-16
```

Example 5. Continuing with Example 4...

a. What is the fitted model?

b. How does the car type affect the relationship between *Mileage* and *Price*?

- In a future lesson, we will learn how to formally test if there is a significant difference among the slopes by testing for a significant difference between the coefficients of subsets of predictors