Multiple Linear Regression with Interaction Terms

Andrew Levine

November 2021

```
UsedCars <- read.csv(file = "/Users/andrewlevine/Downloads/Statistics II/UsedCars.csv")
```

Response variable: Car price; Quantitative variable
Explanatoey variables: Car age, quantitative variable; Whether the car is domestic or foreign, categorical variable with two levels

Turning origin of car into a factor:

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
str(UsedCars)
```

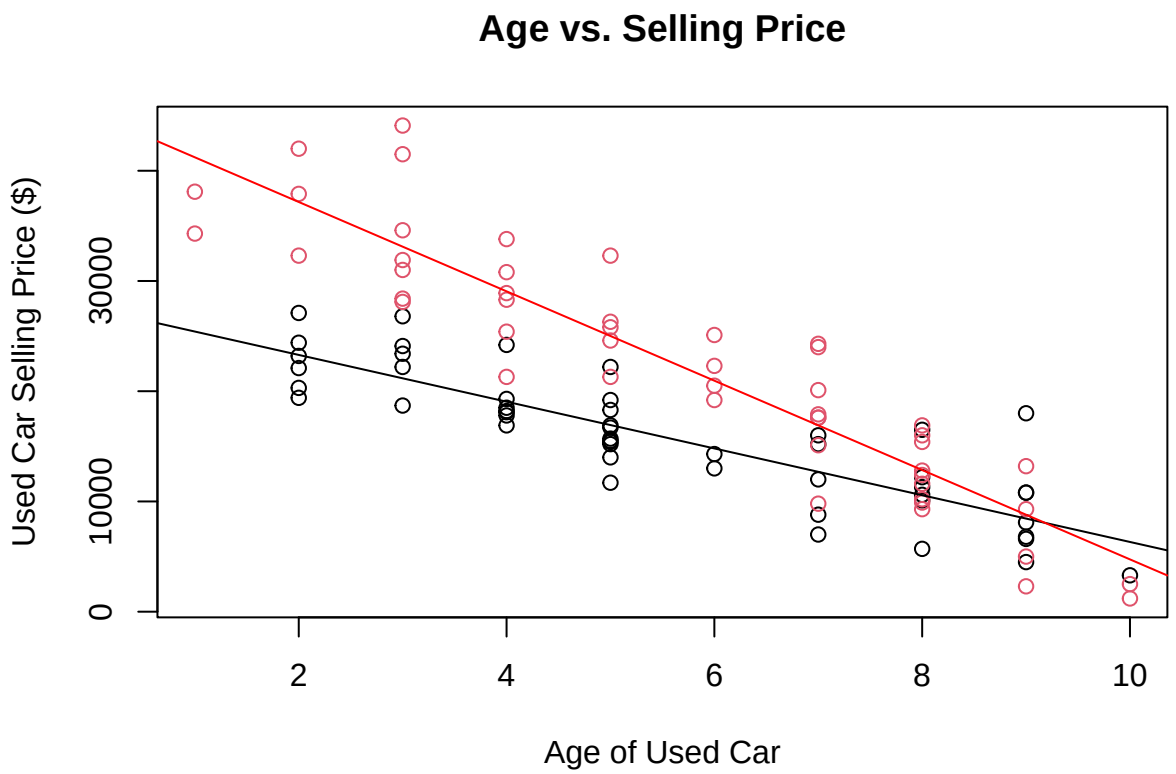
```
## 'data.frame':    100 obs. of  3 variables:
##  $ Age  : int  2 2 2 2 2 2 3 3 3 3 ...
##  $ Price: int  22100 24400 19400 23200 20300 27100 23400 24100 22200 26800 ...
##  $ Type : chr  "Domestic" "Domestic" "Domestic" "Domestic" ...
```

```
UsedCars$Type <- factor(UsedCars$Type)
contrasts(UsedCars$Type)
```

```
##          Foreign
## Domestic       0
## Foreign        1
```

Scatter plot for car age vs. selling price based on car's origin:

```
domestic <- filter(UsedCars, Type == "Domestic")
foreign <- filter(UsedCars, Type == "Foreign")
lm_info_domestic <- lm(Price ~ Age, data = domestic)
lm_info_foreign <- lm(Price ~ Age, data = foreign)
plot(x = UsedCars$Age, y = UsedCars$Price,
  col = UsedCars$Type,
  xlab = "Age of Used Car", ylab = "Used Car Selling Price ($)", main = "Age vs. Selling Price"
)
abline(a = lm_info_domestic$coefficients[1], lm_info_domestic$coefficients[2])
abline(a = lm_info_foreign$coefficients[1], b = lm_info_foreign$coefficients[2], col = "red")
```



The slopes of the regression lines of selling prices of used cars differ based on whether the car's origin is foreign or domestic. More specifically, the slope of the regression line for foreign cars is stronger than the one for domestic cars, which tells us that as car age increases, the value of the foreign car falls at a greater rate than the value of a domestic car. The fact that these two slopes are not parallel to each other suggests that interaction between age and type may be occurring here, which means we will investigate this observation further using significance tests.

Building the regression model:

```
lm_info_UsedCars <- lm(Price ~ Age + Type + Age*Type, data = UsedCars)
summary(lm_info_UsedCars)
```

```
##
## Call:
## lm(formula = Price ~ Age + Type + Age * Type, data = UsedCars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -7760.7 -2301.1  -428.9  2239.0 10986.6
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      27533.9     1354.3  20.331  < 2e-16 ***
## Age              -2120.9      224.2  -9.459 2.16e-15 ***
## TypeForeign      17737.7     1910.4   9.285 5.12e-15 ***
## Age:TypeForeign  -1931.9      311.8  -6.196 1.44e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3741 on 96 degrees of freedom
## Multiple R-squared:  0.8419, Adjusted R-squared:  0.837
## F-statistic: 170.5 on 3 and 96 DF,  p-value: < 2.2e-16
```

$\hat{y}_{Price|Age,Type} = b_0 + b_1(Age_i) + b_2(Type_i) + b_3(Age_i)(Type_i)$
$\mu_{Price|Age,Type} = 27533.947 - 2120.854(Age_i) + 17737.712(Type_i) - 1931.884(Age_i)(Type_i)$

Testing for structural multicollinearity:

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
vif(lm_info_UsedCars)
```

```
##      Age     Type Age:Type
## 2.073297 6.520899 7.754012
```

Structural multicollinearity might be present due to the fact that we are creating a new explanatory variable using the values of two current ones. The interaction variable we created came from both age and type, which means we must be careful here, since we are already taking these variables into account when creating the model. Since not every VIF for the predictors are below the threshold of 5, structural multicollinearity is present. We will center the model to adjust for this.

Centering the model to ease structural multicollinearity:

```
UsedCars$AgeCentered <- UsedCars$Age - mean(UsedCars$Age)
lm_info_UsedCars2 <- lm(Price ~ AgeCentered + Type + AgeCentered*Type, data = UsedCars)
vif(lm_info_UsedCars2)
```

```
##      AgeCentered             Type AgeCentered:Type
##         2.073297         1.001112         2.072109
```

Once we center the data, we can clearly see the VIFs of each explanatory variable decrease, which means we have eased the issue of structural multicollinearity.

Building the new linear regression model, post-centering and hypothesis testing:

```
summary(lm_info_UsedCars2)
```

```
##
## Call:
## lm(formula = Price ~ AgeCentered + Type + AgeCentered * Type,
##     data = UsedCars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7760.7 -2301.1  -428.9  2239.0 10986.6
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               15572.3      529.3  29.420  < 2e-16 ***
## AgeCentered               -2120.9      224.2  -9.459 2.16e-15 ***
## TypeForeign                6841.9      748.5   9.140 1.05e-14 ***
## AgeCentered:TypeForeign   -1931.9      311.8  -6.196 1.44e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3741 on 96 degrees of freedom
## Multiple R-squared:  0.8419, Adjusted R-squared:  0.837
## F-statistic: 170.5 on 3 and 96 DF,  p-value: < 2.2e-16
```

$\hat{y}_{Price|Age,Type}$ = 15572.3 - 2120.9(Age_i) + 6841.9($Type_i$) - 1931.9(Age_i)($Type_i$)

Hypotheses and Significance Level:

H_0: $b_1 = b_2 = b_3 = 0$
H_a: At least one $b_j \neq 0$ (j=1,2,3)
$\alpha = 0.05$
Our overall p-value of $< 2.2 * 10^{-16}$ is less than our α of 0.05, and so we reject H_0. The model that contains used car age, whether the car is domestic or foreign, and their interaction is useful in explaining used car selling price.

Hypothesis test for the interaction term:

Hypotheses and Significance Level:
H_0: $b_3 = 0$
H_a: $b_3 \neq 0$
$\alpha = 0.05$
The interaction term's p-value of $1.44 * 10^{-8}$ is less than our α of 0.05, and so we reject H_0. The interaction between used car age and whether the car is domestic or foreign is significant, and so we should keep this interaction in the model.

How does the average selling price of used foreign cars change as you add one additional year in age?
$\hat{y}_i$ = 15572.332 - 2120.854(Age_i) + 6841.887(1) - 1931.884(Age_i)(1)

$\hat{y}_i$ = 22414.22 - 4052.738(Age_i)
For each additional year in age, the average selling price of foreign used cars sold on the website decreases by $4,052.74.

How does the average selling price of a 10 year old used foreign car differ from the average selling price of a 10 year old used domestic car?
$\hat{y}_i$ = 27533.947 - 2120.854(10) + 6841.887($Type_i$) - 1931.884(10)($Type_i$)
$\hat{y}_i$ = 6325.407 - **12476.95**($Type_i$)
The average selling price of a 10-year-old used foreign car is $12,476.95 **less** than the average selling price of a 10-year-old used domestic car.