Hw4

Hongkai Lou

2024-10-19

```
library(car)
```

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

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.3.3
```

```
##
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
auto <- read.table('auto.txt', header = T)
auto$origin = factor(auto$origin, 1:3, c("US", "Europe", "Japan"))
fit = lm(log(mpg) ~ origin*log(displacement) + year, data=auto)
```

Let $O_E = 1$ if the origin is Europe and 0 otherwise. $O_J = 1$ If the origin is Japan and 0 otherwise. Our model is $y = B_0 + B_1 * year + B_2 * O_E + B_3 * O_J + B_4 * log(displacement) + B_5 * O_E * log(displacement) + B_6 * O_J * log(displacement) + \epsilon$.

3(b)

```
drop1(fit, test = "F")
```

```
## Single term deletions
##
## Model:
## log(mpg) ~ origin * log(displacement) + year
##                           Df Sum of Sq    RSS     AIC  F value    Pr(>F)
## <none>                                  7.207 -1577.5
## year                       1    4.1374 11.344 -1399.4 223.8920 < 2.2e-16 ***
## origin:log(displacement)   2    0.2270  7.434 -1569.2   6.1422  0.002364 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The null hypothesis is that $B_5 = B_6 = 0$. The alternative hypothesis is that at least one of the B_5 or B_6 is not equal to 0. The F value is 6.1422, with our p-value is 0.002364. Since $0.002364 < 0.05$, we reject the null hypothesis, meaning that the effect of log(mpg) and the effect of log(displacement) does depend on origin.

3(c): The DF column refers to the degrees of freedom. It is 1 for numerical variable and the number of category minus one for categorical variable. We have three origins, so DF is 2. The sum of square means how many sum of square errors is being added to the model if this variable is removed from the model. If we remove year, the sum of squared errors will increase by 4.1374. RSS refers to the Residual sum of squares of the model if we drop that specific independent variable. F value is calculated as (change in SSE/df) / SSE/(df of residual), which calculate the importance of the dropped variable. So for the origin:log(displacement) column, F value is (0.2270/2)/(7.207/390). The p value is the associated p value of the F value, with df of the dropped variable and df of the residual. We can conclude that both year and the interaction variable are important.

```
summary(fit)
```

```
##
## Call:
## lm(formula = log(mpg) ~ origin * log(displacement) + year, data = auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.50957 -0.07570  0.00506  0.07388  0.56393
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      3.241432   0.222751  14.552  < 2e-16 ***
## originEurope                     1.276599   0.409746   3.116  0.00197 **
## originJapan                     -0.416641   0.353779  -1.178  0.23964
## log(displacement)               -0.480425   0.020793 -23.105  < 2e-16 ***
## year                             0.030578   0.002044  14.963  < 2e-16 ***
## originEurope:log(displacement) -0.276179   0.086682  -3.186  0.00156 **
## originJapan:log(displacement)   0.090394   0.075651   1.195  0.23286
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1359 on 390 degrees of freedom
## Multiple R-squared:  0.8426, Adjusted R-squared:  0.8402
## F-statistic: 348.1 on 6 and 390 DF,  p-value: < 2.2e-16
```

3(d): $B_6 = 0.0904$. It means that for the autos where their origin is Japan, for every unit increase in log(displacement), log(mpg) will increase by 0.0904. It is worth noting that the p value of this interaction, 0.23286, is larger than 0.05 and will not pass the hypothesis test, meaning it is not a significant variable.

3(e): For US: $mpg = exp(3.2414 - 0.4804 * log(displacement))$.

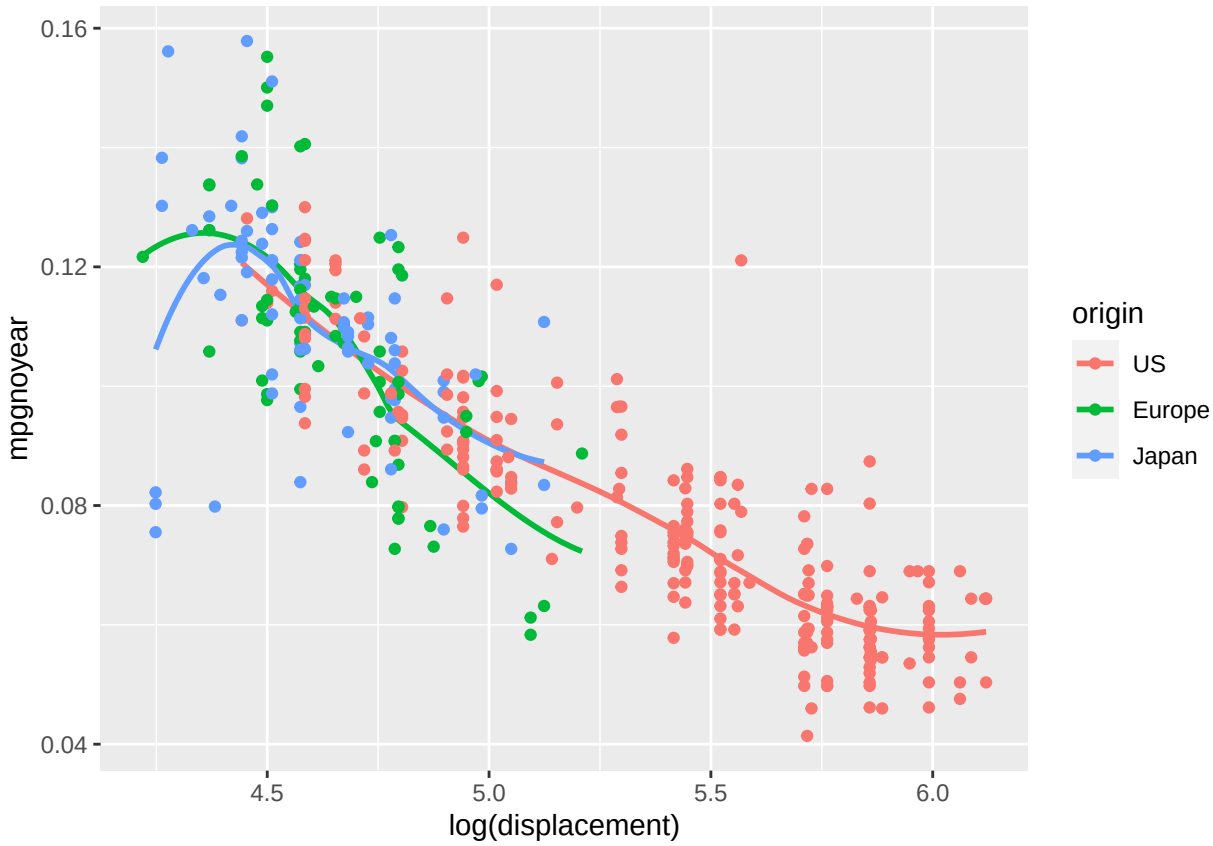
For Europe: $mpg = exp(3.2414 - 0.4804 * log(displacement) + 1.2766 - 0.2762 * log(displacement)$. Simplify, we get $mpg = 4.518 - 0.7566 * log(displacement)$.

For Japan: $mpg = exp(3.2414 - 0.4804 * log(displacement) - 0.4166 + 0.0904 * log(displacement)$. Simplify, we get $mpg = 2.8248 - 0.39 * log(displacement)$.

3(f)

```
auto$mpgnoyear <- exp(log(auto$mpg)-fit$coefficients[1]-fit$coefficients[5]*auto$year)
ggplot(data = auto, aes(x = log(displacement), y = mpgnoyear, group = origin,col=origin))+geom_smooth(s
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```


0.16
0.12
0.08
0.04
mpgnoyear
4.5
5.0
5.5
6.0
log(displacement)
origin
US
Europe
Japan