Lesson 10: Cars - Weight vs. Acceleration

CDT I.M. Smart

```
### Loading packages
library(tidyverse)
library(ggResidpanel)
## Load data
car <- read.table('https://raw.githubusercontent.com/jkstarling/MA376/main/carmpg.txt', header = TRUE,

# setwd('C:/Users/james.starling/OneDrive - West Point/Teaching/MA376/JimsLessons_AY23-1/Block II/LSN10
# car <- read.table('carmpg.txt', header = TRUE, stringsAsFactors = TRUE)
```

(0): Recall that we discussed *confounding* earlier in our course. We can define confounding as...

(1): Today we will investigate the question, "Do heavier cars accelerate faster?"

Here is some background on the data we will use:

One commonly cited metric in car commercials is the time it takes the car to accelerate from 0 to 60 mph. In this exploration, we will consider what car features might help explain faster acceleration (shorter times) using data on 406 cars (Quinlan, "Combining Instance-based and Model-based Learning," Proceedings on the Tenth International Conference of Machine Learning, 236–243, 1993; used in the 1983 ASA Data Graphics Exposition). The carmpg file contains data on the weight of the car (whether or not the car weighs over 2800 kg) and the time it takes to accelerate to 60 mph (measured in seconds). (This dataset has been around the internet for a while but details on the data collection are limited. We use the version with 392 complete records.)

Was this an observational study or a randomized experiment? How do you know?

(2): What are the two main variables of interest to the study authors? Which of these is the response variable and which is the explanatory variable? Write down a good statistical model.

(3): Plot the distributions of acceleration times between light and heavy cars, and calculate the mean acceleration times for heavy and light cars.

Comment on the size and direction of the difference in means. Explain why the direction of this association could be viewed by some as surprising. (Hint: Remember that smaller acceleration times are faster cars)

```
# car %>% ggplot(aes(x = acceleration, fill = weight))+geom_histogram()+
#   facet_wrap(~weight, ncol = 1)+
#   theme_classic()
#
# car %>% group_by(weight) %>% xxx
```

(4): Conduct a one-variable analysis (multiple-mean model) evaluating the association between acceleration and car weight.

Is there a statistically significant relationship between acceleration time and car weight? (Note: use effect coding when creating your model)

```
# contrasts(car$weight) = contr.sum
# contrasts(car$weight)
#
# weight.mod <- lm(xxx ~ xx, data=car)
# summary(weight.mod)
# wt.anova <- anova(weight.mod)
# wt.anova
```

(5): What proportion of variation in acceleration times is explained by whether or not the car is heavy?

From your anova table in question 4, calculate and save the SSTotal, SSE, and Rsquared values.

```
# SST_w <-
# SSE_w <-
# c(SST_w, SSE_w)
# Rsquared_w <-
# Rsquared_w
```

(6): Horsepower may also impact acceleration time. Even if this is not the main research question of interest, we might want to explain some of the unexplained variation using horsepower as a factor.

Calculate the means and number of observations, grouped by weight and horsepower. What do you notice about the sample sizes? What about the means compared to the overall mean?

```
# car %>% group_by(xxx, xxx) %>%
#   summarise(mn=mean(acceleration), n=n())
```

(7): Nevertheless, let us ignore this fact and use the same approach we used in previous sections. First conduct a one-variable analysis evaluating the association between acceleration and horsepower.

```
# contrasts(car$horsepower) = contr.sum
# contrasts(car$horsepower)
#
# h.mod <- lm(xxx ~ xxx, data=car)
# summary(h.mod)
# h.anova <- anova(h.mod)
# h.anova
# SST_h <- xxx
# SSE_h <- xxx
```

(8): If we were to conduct a two-variable analysis and there was no association between weight and horsepower we would expect the SSError to be decreased by the SSModel from the `hp.mod`. Is this the case?

```
# w.hp.mod <- lm(xxx ~ xxx, data=car)
# summary(w.hp.mod)
# w.hp.anova <- anova(w.hp.mod)
# w.hp.anova
# SST_whp <- xxx
# SSE_whp <- xxx
```

(9): Up to this point we have been calculating what are called Type I Sums of Squares. These are done sequentially. We first find the Sums of Squares due to Factor A and then find the Sums of Squares due to Factor B *given that* factor A is in the model.

Create a new model `hp.w.mod` which reverses the order of the two explanatory variables. However, notice the difference between model ‘w.hp.mod’ and model ‘hp.w.mod’. What do we notice?

```
# hp.w.mod <- lm(xxx ~ xxx, data=car)
# hp.w.anova <- anova(hp.w.mod)
# hp.w.anova
# SST_hpw <- xxx
```

(10): We should calculate Type III Sums of Squares. Note that our statistical model does not change, but to get the appropriate ANOVA table we need `library(car)` installed (it provides the `Anova()` function).

```
#install.packages('car')
# library(car)
# w.hp.mod = lm(xxx ~ xxx, data=car)
# summary(w.hp.mod)
# Anova3.wt.hp <- Anova(xxx)
# Anova3.wt.hp
```

(11): An interesting note here is that the sums of squares no longer equal the total sums of squares. The extra sums of squares can be thought of as variation that cannot be disentangled from weight or horsepower. Our book calls this SSCovariation. It is variability that still exists but we cannot attribute it to either factor so we basically shrug our shoulders. Calculate the SSCovariation:

```
# SST_A3 <- xxx
# SSCovariation = xxx - xxx
# SSCovariation
```

(12): What proportion of variation in acceleration times is explained by this model?

```
# xxx
```

Type I Sums of Squares vs Type III Sums of Squares

- Type I: preferable when order matters. The factor we are more concerned with goes first. Used for 1 variable analysis and experiments with a balanced design.
- Type III: all effects are conditional on *everything else in the model*. Used for observational studies.

References

Nathan Tintle et al.(2019). Intermediate Statistical Investigations.