BACS HW12 - 109006234

Credit: 109006278

May 7th 2023

Loading the data

```
cars <- read.table("auto-data.txt", header=FALSE, na.strings = "?")
names(cars) <- c("mpg", "cylinders", "displacement",
                 "horsepower", "weight", "acceleration",
                 "model_year", "origin", "car_name")
vars <- c("mpg", "weight", "acceleration",
          "model_year", "origin", "cylinders")
cars <- cars[vars]
cars_log <- with(cars, data.frame(log(mpg), log(weight),
              log(acceleration), model_year, origin, log(cylinders)))
```

Problem 1

(a) Let's visualize how weight might moderate the relationship between acceleration and mpg:

```
weight_mean_log <- log(mean(cars$weight))
```

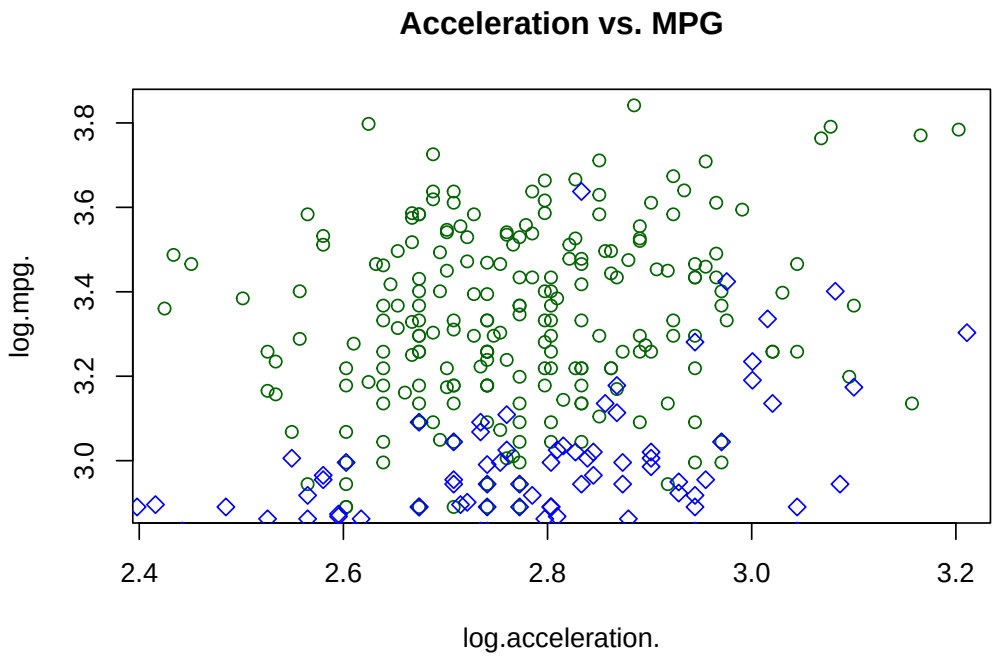
(i) Create two subsets of your data, one for light-weight cars (less than mean weight) and one for heavy cars (higher than the mean weight)

```
#Light Cars Regression
light_log <- subset(cars_log, log.weight. < weight_mean_log)
light_reg <- with(light_log, lm(log.mpg. ~ log.acceleration.))

#Heavy Cars Regression
heavy_log <- subset(cars_log, log.weight. >= weight_mean_log)
heavy_reg <- with(heavy_log, lm(log.mpg. ~ log.acceleration.))
```

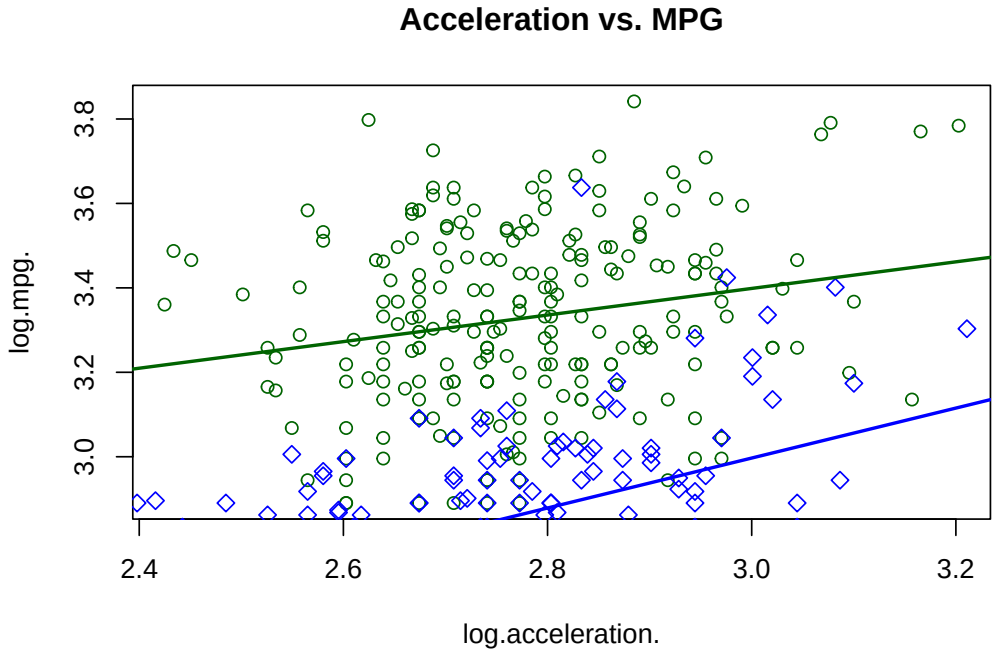
(ii) Create a single scatter plot of acceleration vs. mpg, with different colors and/or shapes for light versus heavy cars

```
with(light_log, plot(log.acceleration., log.mpg., pch=1,
          col="darkgreen", main = "Acceleration vs. MPG"))
with(heavy_log, points(log.acceleration., log.mpg., pch=5, col="blue"))
```



(iii) Draw two slopes of acceleration-vs-mpg over the scatter plot: one slope for light cars and one slope for heavy cars

```
with(light_log, plot(log.acceleration., log.mpg., pch=1,
        col="darkgreen", main = "Acceleration vs. MPG"))
with(heavy_log, points(log.acceleration., log.mpg., pch=5, col="blue"))
abline(light_reg, col="darkgreen", lwd=2)
abline(heavy_reg, col="blue", lwd=2)
```



(b) Report the full summaries of two separate regressions for light and heavy cars where log.mpg. is dependent on log.weight., log.acceleration., model_year and origin

Light Cars

```
regr_light_full <- lm(log.mpg.~ log.weight. + log.acceleration. +
                      model_year + factor(origin), data=light_log)
summary(regr_light_full)
```

```
##
## Call:
## lm(formula = log.mpg. ~ log.weight. + log.acceleration. + model_year +
##     factor(origin), data = light_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36464 -0.07181  0.00349  0.06273  0.31339
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         6.86661    0.52767  13.013   <2e-16 ***
## log.weight.        -0.83437    0.05662 -14.737   <2e-16 ***
## log.acceleration.   0.10956    0.05630   1.946   0.0529 .
## model_year          0.03383    0.00198  17.079   <2e-16 ***
## factor(origin)2     0.05129    0.01980   2.590   0.0102 *
## factor(origin)3     0.02621    0.01846   1.420   0.1571
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1112 on 221 degrees of freedom
## Multiple R-squared:  0.7292, Adjusted R-squared:  0.7231
## F-statistic:   119 on 5 and 221 DF,  p-value: < 2.2e-16
```

Heavy Cars

```
regr_heavy_full <- lm(log.mpg.~ log.weight. + log.acceleration. +
                      model_year + factor(origin), data=heavy_log)
summary(regr_heavy_full)
```

```
##
## Call:
## lm(formula = log.mpg. ~ log.weight. + log.acceleration. + model_year +
##     factor(origin), data = heavy_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36811 -0.06937  0.00607  0.06969  0.43736
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.188679   0.759983   9.459  < 2e-16 ***
## log.weight.        -0.822352   0.077206 -10.651  < 2e-16 ***
## log.acceleration.   0.040140   0.057380   0.700   0.4852
## model_year          0.030317   0.003573   8.486 1.14e-14 ***
## factor(origin)2     0.091641   0.040392   2.269   0.0246 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1212 on 166 degrees of freedom
## Multiple R-squared:  0.7179, Adjusted R-squared:  0.7111
## F-statistic: 105.6 on 4 and 166 DF,  p-value: < 2.2e-16
```

(c) Using your intuition only: What do you observe about light versus heavy cars so far?

From the plot above, we can confirm that there are more light cars compared to heavy cars, as there are more dark-green data points. I think that the slope of heavy cats are more fitted compared to the other data available.

Problem 2

(a) Considering weight and acceleration, use your intuition and experience to state which of the two variables might be a moderating versus independent variable, in affecting mileage.

In my opinion, acceleration might be a moderating vs. independent variable, while at the same time it will affect mpg.

(b) Use various regression models to model the possible moderation on log.mpg.

(i) Report a regression without any interaction terms

```r
no_inter <- lm(log.mpg. ~ log.weight. +
               log.acceleration. + model_year +
               factor(origin), data=cars_log)
summary(no_inter)
```

```
##
## Call:
## lm(formula = log.mpg. ~ log.weight. + log.acceleration. + model_year +
##     factor(origin), data = cars_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38275 -0.07032  0.00491  0.06470  0.39913
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.431155   0.312248  23.799  < 2e-16 ***
## log.weight.       -0.876608   0.028697 -30.547  < 2e-16 ***
## log.acceleration.  0.051508   0.036652   1.405  0.16072
## model_year         0.032734   0.001696  19.306  < 2e-16 ***
## factor(origin)2    0.057991   0.017885   3.242  0.00129 **
## factor(origin)3    0.032333   0.018279   1.769  0.07770 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1156 on 392 degrees of freedom
```

```
## Multiple R-squared:  0.8856, Adjusted R-squared:  0.8841
## F-statistic: 606.8 on 5 and 392 DF,  p-value: < 2.2e-16
```

(ii) Report a regression with an interaction between weight and acceleration

```
weight_acc_inter <- lm(log.mpg. ~ log.weight. + log.acceleration. +
                     log.weight. * log.acceleration., data=cars_log)
summary(weight_acc_inter)
```

```
##
## Call:
## lm(formula = log.mpg. ~ log.weight. + log.acceleration. + log.weight. *
##     log.acceleration., data = cars_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49728 -0.10145 -0.01102  0.09665  0.56416
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     16.0249     3.6950   4.337 1.84e-05 ***
## log.weight.                     -1.6878     0.4578  -3.687 0.000259 ***
## log.acceleration.               -1.8252     1.3537  -1.348 0.178351
## log.weight.:log.acceleration.    0.2529     0.1681   1.505 0.133123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1613 on 394 degrees of freedom
## Multiple R-squared:  0.7763, Adjusted R-squared:  0.7746
## F-statistic: 455.7 on 3 and 394 DF,  p-value: < 2.2e-16
```

(iii) Report a regression with a mean-centered interaction term

```
mean_weight <- scale(cars_log$log.weight., center = TRUE, scale = FALSE)
mean_acc <- scale(cars_log$log.acceleration., center = TRUE, scale = FALSE)
mean_mpg <- scale(cars_log$log.mpg., center = TRUE, scale = FALSE)

mean_reg <- lm(mean_mpg ~ mean_acc + mean_weight + mean_acc*mean_weight)
summary(mean_reg)
```

```
##
## Call:
## lm(formula = mean_mpg ~ mean_acc + mean_weight + mean_acc * mean_weight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49728 -0.10145 -0.01102  0.09665  0.56416
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.005447   0.008857   0.615 0.538884
## mean_acc              0.187500   0.051862   3.615 0.000339 ***
## mean_weight          -0.997466   0.031930 -31.239  < 2e-16 ***
## mean_acc:mean_weight  0.252948   0.168071   1.505 0.133123
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1613 on 394 degrees of freedom
## Multiple R-squared:  0.7763, Adjusted R-squared:  0.7746
## F-statistic: 455.7 on 3 and 394 DF,  p-value: < 2.2e-16
```

(iv) Report a regression with an orthogonalized interaction term

```
dot_product <- cars_log$log.weight. * cars_log$log.acceleration.
inter_reg <- lm(dot_product ~
                cars_log$log.weight. + cars_log$log.acceleration.)
inter_ortho <- inter_reg$residuals
ortho_inter <- lm(log.mpg. ~ log.weight. + log.acceleration. +
                inter_ortho, data=cars_log)
summary(ortho_inter)
```

```
##
## Call:
## lm(formula = log.mpg. ~ log.weight. + log.acceleration. + inter_ortho,
##     data = cars_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49728 -0.10145 -0.01102  0.09665  0.56416
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       10.48669    0.33430  31.369  < 2e-16 ***
## log.weight.       -1.00048    0.03187 -31.395  < 2e-16 ***
## log.acceleration.  0.21084    0.04949   4.260 2.56e-05 ***
## inter_ortho        0.25295    0.16807   1.505    0.133
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1613 on 394 degrees of freedom
## Multiple R-squared:  0.7763, Adjusted R-squared:  0.7746
## F-statistic: 455.7 on 3 and 394 DF,  p-value: < 2.2e-16
```

(c) For each of the interaction term strategies above (raw, mean-centered, orthogonalized) what is the correlation between that interaction term and the two variables that you multiplied together?

```
#Raw
a <- cor(cars_log$log.weight.*cars_log$log.acceleration., cars_log$log.weight.)
b <- cor(cars_log$log.weight.*cars_log$log.acceleration., cars_log$log.acceleration.)

#Mean-centered
c <- as.vector(cor(mean_acc*mean_weight, mean_weight))
d <- as.vector(cor(mean_acc*mean_weight, mean_acc))

#Orthogonalized
e <- cor(inter_reg$residuals, cars_log$log.weight.)
f <- cor(inter_reg$residuals, cars_log$log.acceleration.)
```

```r
cor_mat <- matrix(c(a,b,c,d,e,f), ncol=2,byrow=TRUE)
rownames(cor_mat) <- c("raw", "mean-centered", "orthogonalized")
colnames(cor_mat) <- c("log.weight.","log.acceleration")
round(cor_mat,2)
```

```
##                log.weight. log.acceleration
## raw                   0.11             0.85
## mean-centered        -0.20             0.35
## orthogonalized        0.00             0.00
```

Problem 3

Let's check whether weight mediates the relationship between cylinders and mpg, even when other factors are controlled for. Use log.mpg., log.weight., and log.cylinders as your main variables, and keep log.acceleration., model_year, and origin as control variables

(a) Let's try computing the direct effects first:

(i) Model 1: Regress log.weight. over log.cylinders. only

```r
reg_weight_cyl <- lm(log.weight.~ log.cylinders., data=cars_log)
summary(reg_weight_cyl)
```

```
##
## Call:
## lm(formula = log.weight. ~ log.cylinders., data = cars_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35473 -0.09076 -0.00147  0.09316  0.40374
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     6.60365    0.03712  177.92   <2e-16 ***
## log.cylinders.  0.82012    0.02213   37.06   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1329 on 396 degrees of freedom
## Multiple R-squared:  0.7762, Adjusted R-squared:  0.7757
## F-statistic:  1374 on 1 and 396 DF,  p-value: < 2.2e-16
```

(ii) Model 2: Regress log.mpg. over log.weight. and all control variables

```r
regr_mpg_weight <- lm(log.mpg.~ log.weight., data=cars_log)
summary(regr_mpg_weight)
```

```
##
## Call:
## lm(formula = log.mpg. ~ log.weight., data = cars_log)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52408 -0.10441 -0.00805  0.10165  0.59384
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.5219     0.2349   49.06   <2e-16 ***
## log.weight.  -1.0583     0.0295  -35.87   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.165 on 396 degrees of freedom
## Multiple R-squared:  0.7647, Adjusted R-squared:  0.7641
## F-statistic:  1287 on 1 and 396 DF,  p-value: < 2.2e-16
```

(b) What is the indirect effect of cylinders on mpg?

```
reg_weight_cyl$coefficients[2] * regr_mpg_weight$coefficients[2]
```

```
## log.cylinders.
##     -0.8679111
```

(c) Let's bootstrap for the confidence interval of the indirect effect of cylinders on mpg

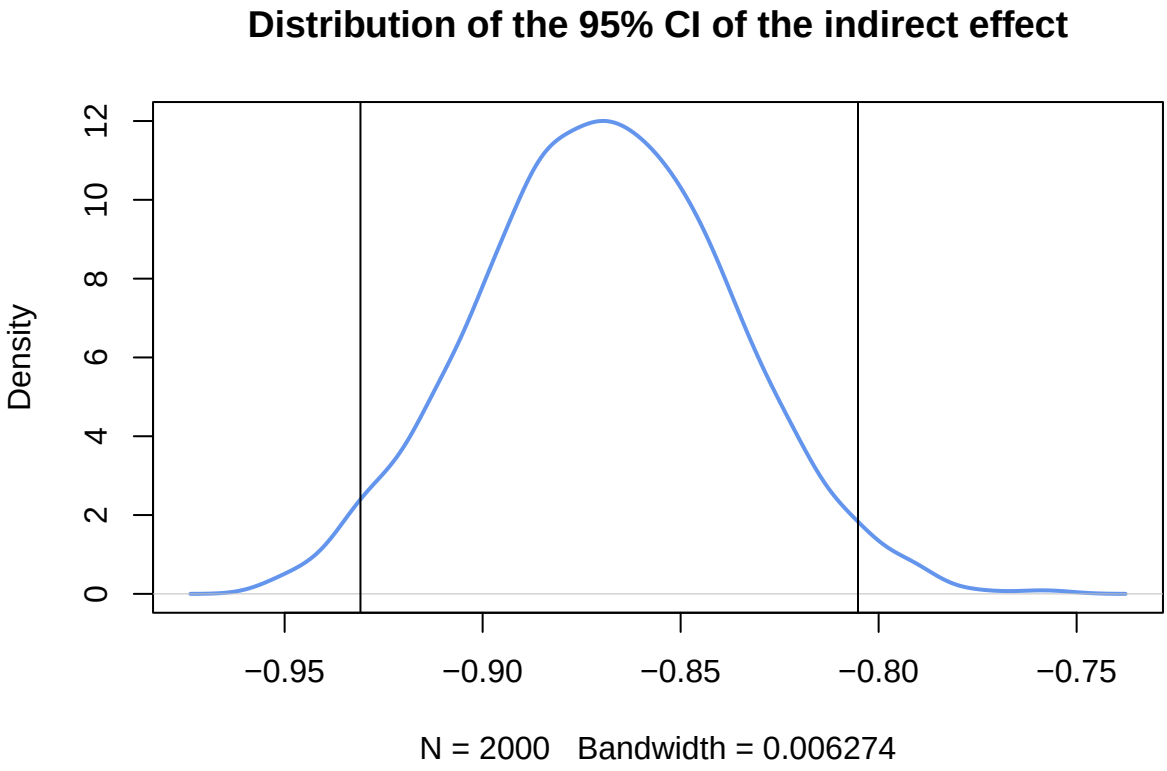
```
boot_mediation <- function(model1, model2, dataset) {
  boot_index <- sample(1:nrow(dataset), replace=TRUE)
  data_boot <- dataset[boot_index, ]
  regr1 <- lm(model1, data_boot)
  regr2 <- lm(model2, data_boot)
  return(regr1$coefficients[2] * regr2$coefficients[2])
}
```

(i) Bootstrap regression models 1 & 2, and compute the indirect effect each time: What is its 95% CI of the indirect effect of log.cylinders. on log.mpg.?

```
set.seed(42)
indirect <- replicate(2000,
        boot_mediation(reg_weight_cyl, regr_mpg_weight, cars_log))
boot_ci <- quantile(indirect, probs=c(0.025, 0.975))
```

(ii) Show a density plot of the distribution of the 95% CI of the indirect effect

```
plot(density(indirect), lwd=2, col="cornflowerblue",
        main= "Distribution of the 95% CI of the indirect effect")
abline(v=quantile(indirect, probs=c(0.025, 0.975)))
```


Distribution of the 95% CI of the indirect effect
Density
0
2
4
6
8
10
12
−0.95
−0.90
−0.85
−0.80
−0.75
N = 2000 Bandwidth = 0.006274