BACS - HW 12 106073401

Question 1) Let's visualize how weight and acceleration are related to mpg.

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

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

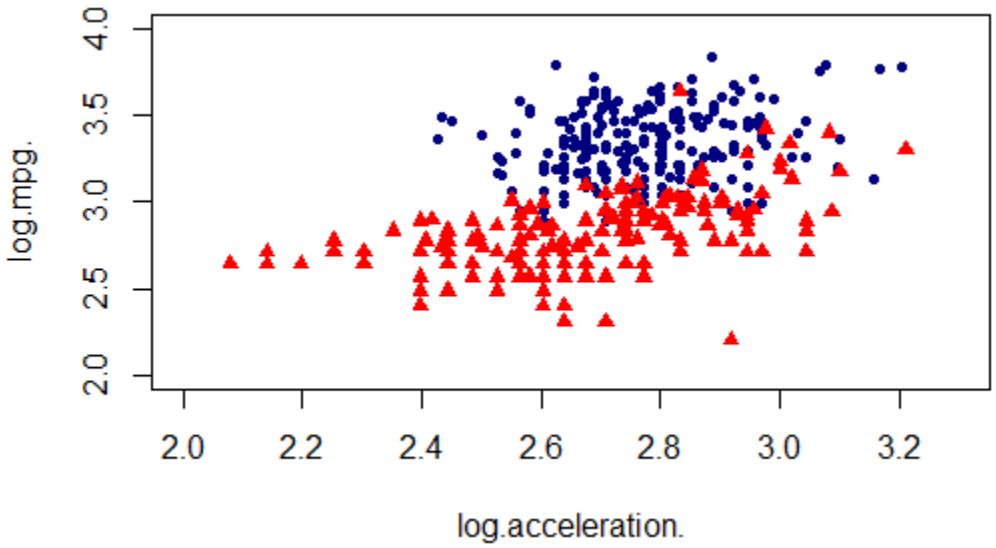
i. Create two *subsets* of your data, one for light-weight cars (less than mean weight) and one for heavy cars (higher than the mean weight)

```
cars_light <- cars[cars$weight<mean(cars$weight),]
cars_log_light <- with(cars_light,data.frame(log(mpg),log(weight),log(acceleration),
 model_year, origin))

cars_heavy <- cars[cars$weight>mean(cars$weight),]
cars_log_heavy <- with(cars_heavy, data.frame(log(mpg), log(weight),log(acceleratio
n), model_year, origin))
```

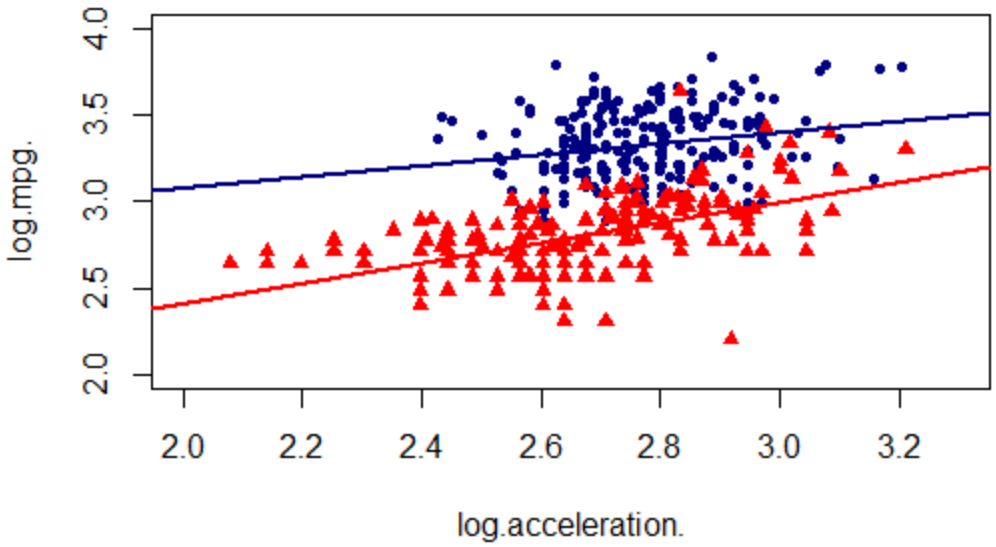
ii. Create a *single* scatter plot of acceleration vs. mpg, with different colors and/or shapes for light versus heavy cars

```
plot(cars_log_light$log.acceleration.,cars_log_light$log.mpg.,pch = 20, col="navy",
ylim = c(2,4), xlim = c(2,3.3),ylab = "log.mpg.", xlab = "log.acceleration.")
points(cars_log_heavy$log.acceleration.,cars_log_heavy$log.mpg.,pch = 17, col="red")
```



iii. Draw <u>two slopes</u> of acceleration-vs-mpg over the scatter plot: one slope for light cars and one slope for heavy cars (distinguish them by appearance)

```
abline(lm(log.mpg.~log.acceleration. , data = cars_log_light), col ="navy", lwd = 2)
abline(lm(log.mpg.~log.acceleration. , data = cars_log_heavy), col = "red", 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`

```
summary(lm(log.mpg.~log.acceleration.+ log.weight. + model_year + origin ,data = car
s_log_light))

Residuals:
     Min       1Q   Median       3Q      Max
-0.37941 -0.07219 -0.00307  0.06759  0.34454

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      7.059570   0.526938  13.397   <2e-16 ***
log.acceleration.  0.108295   0.056775   1.907   0.0578 .
log.weight.       -0.849942   0.056655 -15.002   <2e-16 ***
model_year        0.032895   0.001951  16.858   <2e-16 ***
origin            0.012824   0.009310   1.377   0.1698
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1121 on 222 degrees of freedom
Multiple R-squared:  0.7233,   Adjusted R-squared:  0.7183
F-statistic: 145.1 on 4 and 222 DF,  p-value: < 2.2e-16

summary(lm(log.mpg.~log.acceleration.+ log.weight. + model_year + origin ,data = car
s_log_heavy))

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)      7.097038   0.762942   9.302  < 2e-16 ***
log.acceleration.  0.040140   0.057380   0.700   0.4852
log.weight.       -0.822352   0.077206 -10.651  < 2e-16 ***
model_year        0.030317   0.003573   8.486 1.14e-14 ***
origin            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. *(not graded)* Using your intuition only: What do you observe about light versus heavy cars so far?
Heavy cars can run less miles than light cars, and acceleration plays a more important role in heavy cars than light cars.

Question 2) Using the fully transformed dataset from above (`cars_log`), to test whether we have moderation.

a. *(not graded)* Between weight and acceleration ability (in seconds), use your intuition and experience to state which variable might be a moderating versus independent variable, in affecting mileage.
Weight is independent variable and acceleration is moderator.

b. Use various regression models to model the possible moderation on `log.mpg.`:
(use `log.weight.`, `log.acceleration.`, `model_year` and `origin` as independent variables)

i. Report a regression *without any interaction terms*

```
summary(lm(log.mpg.~log.acceleration.+log.weight.+model_year+origin, data=cars_log))
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     7.539281   0.314707  23.956   <2e-16 ***
log.acceleration.  0.062145   0.036679   1.694   0.0910 .
log.weight.       -0.889384   0.028466 -31.243   <2e-16 ***
model_year         0.032106   0.001690  18.999   <2e-16 ***
origin             0.018352   0.009165   2.002   0.0459 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1164 on 393 degrees of freedom
Multiple R-squared:  0.8836,   Adjusted R-squared:  0.8825
F-statistic: 746.1 on 4 and 393 DF,  p-value: < 2.2e-16
```

ii. Report a regression *with an interaction between weight and acceleration*

```
summary(lm(log.mpg.~log.acceleration.+log.weight.+log.acceleration.*log.weight.+model
_year+origin, data = cars_log))
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                   1.773573   2.763699   0.642   0.5214
log.acceleration.             2.162941   1.001155   2.160   0.0313 *
log.weight.                  -0.179842   0.339101  -0.530   0.5962
model_year                    0.032933   0.001728  19.057   <2e-16 ***
origin                        0.016595   0.009164   1.811   0.0709 .
log.acceleration.:log.weight. -0.261526   0.124550  -2.100   0.0364 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1159 on 392 degrees of freedom
Multiple R-squared:  0.8849,   Adjusted R-squared:  0.8835
F-statistic:   603 on 5 and 392 DF,  p-value: < 2.2e-16
```

iii. Report a regression *with a mean-centered interaction term*

```
log_weight_mc <- scale(cars_log$log.weight., center = TRUE, scale = FALSE)
log_acc_mc <- scale(cars_log$log.acceleration.,center = TRUE, scale =FALSE)
summary(lm(log.mpg.~log_weight_mc+log_acc_mc+log_weight_mc*log_acc_mc+model_year+orig
in,data = cars_log))
```

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.566397   0.132258   4.283 2.33e-05 ***
log_weight_mc            -0.893616   0.028415 -31.448  < 2e-16 ***
log_acc_mc                0.082003   0.037725   2.174   0.0303 *
model_year                0.032933   0.001728  19.057  < 2e-16 ***
origin                    0.016595   0.009164   1.811   0.0709 .
log_weight_mc:log_acc_mc -0.261526   0.124550  -2.100   0.0364 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1159 on 392 degrees of freedom
Multiple R-squared:  0.8849,   Adjusted R-squared:  0.8835
F-statistic:   603 on 5 and 392 DF,  p-value: < 2.2e-16
```

iv. Report a regression *with an orthogonalized interaction term*

```
ortho <- lm(log.weight.*log.acceleration.~log.weight.+log.acceleration.,data = cars_log)$residuals
summary(lm(log.mpg.~log.acceleration.+log.weight.+ortho+model_year+origin, data = cars_log))

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        7.499651   0.313919  23.890   <2e-16 ***
log.acceleration.  0.057873   0.036577   1.582   0.1144
log.weight.       -0.890495   0.028349 -31.412   <2e-16 ***
ortho             -0.261526   0.124550  -2.100   0.0364 *
model_year         0.032933   0.001728  19.057   <2e-16 ***
origin             0.016595   0.009164   1.811   0.0709 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1159 on 392 degrees of freedom
Multiple R-squared:  0.8849,   Adjusted R-squared:  0.8835
F-statistic:   603 on 5 and 392 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?

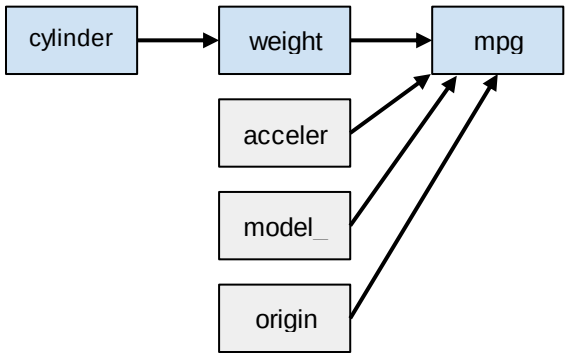
```
#raw
cor(cars_log$log.weight.,cars_log$log.acceleration.*cars_log$log.weight.)
[1] 0.1083055
cor(cars_log$log.acceleration.,cars_log$log.acceleration.*cars_log$log.weight.)
[1] 0.852881

#mean-centered
cor(log_weight_mc,log_weight_mc*log_acc_mc)
           [,1]
[1,] -0.2026948
cor(log_acc_mc,log_weight_mc*log_acc_mc)
          [,1]
[1,] 0.3512271

#orthogonalized
cor(cars_log$log.weight.,ortho)
[1] 2.468461e-17
cor(cars_log$log.acceleration.,ortho)
[1] -6.804111e-17
```

Question 3) We saw earlier that the number of cylinders does not seem to *directly* influence mpg when car weight is also considered. But might cylinders have an *indirect* relationship with mpg through its weight?
(see blue variables in diagram)
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* (see gray variables in diagram).

Conceptual Path Diagram of Mediated Model



a. Let's try computing the direct effects first:
   i. Model 1: Regress **`log.weight.`** over **`log.cylinders.`** only
   (check whether number of cylinders has a significant direct effect on weight)

```
regr_md1 <- lm(log.weight.~log.cylinders.,data=cars_log)
summary(regr_md1)

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
```

Cylinders has a significant direct effect on weight.

   ii. Model 2: Regress **`log.mpg.`** over **`log.weight.`** and all control variables
   (check whether weight has a significant direct effect on mpg with other variables statistically controlled?)

```
regr_md2 <- lm(log.mpg. ~ log.weight.+log.acceleration.+model_year+origin,data = ca
rs_log)
summary(regr_md2)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        7.539281   0.314707  23.956   <2e-16 ***
log.weight.       -0.889384   0.028466 -31.243   <2e-16 ***
log.acceleration.  0.062145   0.036679   1.694   0.0910 .
model_year         0.032106   0.001690  18.999   <2e-16 ***
origin             0.018352   0.009165   2.002   0.0459 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1164 on 393 degrees of freedom
Multiple R-squared:  0.8836,   Adjusted R-squared:  0.8825
F-statistic: 746.1 on 4 and 393 DF,  p-value: < 2.2e-16
```

b. What is the *indirect effect* of cylinders on mpg? *(use the product of slopes between model 1 & 2)*

```
regr_md1$coefficients["log.cylinders."] * regr_md2$coefficients["log.weight."]
log.cylinders.
    -0.7294051
```

c. Let's bootstrap for the confidence interval of the *indirect effect* of cylinders on mpg

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.`?

```
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])
 }
set.seed(67)
indirect <- replicate(800, boot_mediation(regr_md1, regr_md2, cars_log))
quantile(indirect, probs = c(0.025, 0.975))
      2.5%      97.5%
-0.7881702 -0.6683161
```

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

```
plot(density(indirect),main = "Desity plot of indirect effect", col = "navy", lwd
= 2)
abline(v = quantile(indirect,c(0.025,0.975)),lty = "dashed", col = "RED" )
```