HW12

109006206

Set Working Directories & Reading Files

```
library(ggplot2)
setwd("/Users/olivia/Documents/Documents/Study/Semester 6/BACS/HW12")
cars<-read.table("auto-data.txt", header=FALSE, na.strings = "?")
names(cars) <- c("mpg", "cylinders", "displacement", "horsepower", "weight",
                 "acceleration", "model_year", "origin", "car_name")
cars_log <- with(cars, data.frame(log(mpg), log(cylinders), log(displacement), log(horsepower), log(wei
```

QUESTION 1

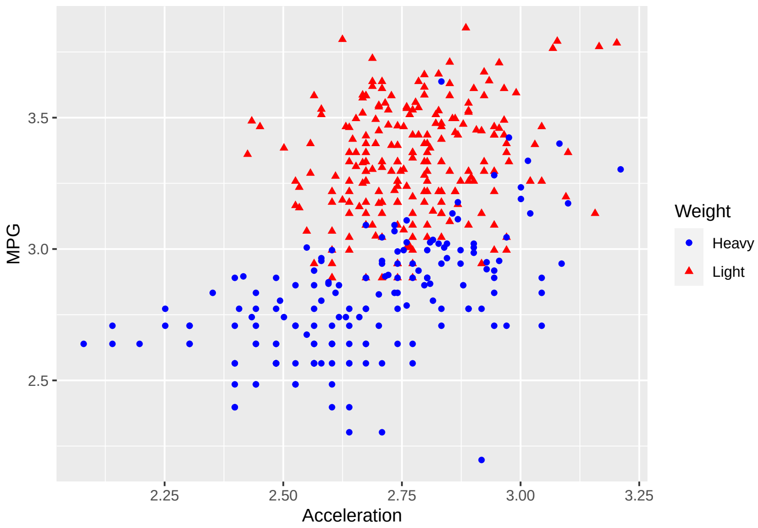
A) Let's visualize how weight might moderate the relationship between acceleration and mpg:

1) 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<-cars_log[cars_log$log.weight. < log(mean(cars$weight)), ]
heavy_cars<-cars_log[cars_log$log.weight. >= log(mean(cars$weight)), ]
```

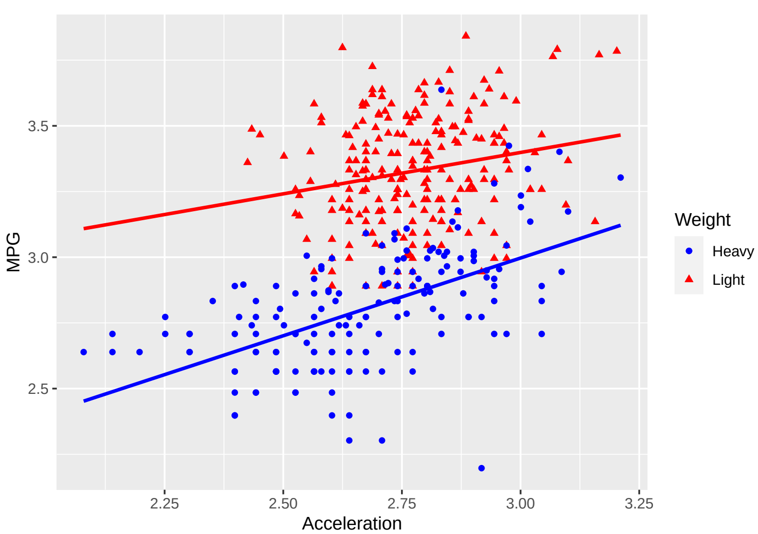
2) Create a single scatter plot of acceleration vs. mpg, with different colors and/or shapes for light versus heavy cars

```
ggplot() +
  geom_point(data = light_cars, aes(x = log.acceleration., y = log.mpg., color = "Light", shape = "Ligh
  geom_point(data = heavy_cars, aes(x = log.acceleration., y = log.mpg., color = "Heavy", shape = "Heav
  scale_color_manual(values = c(Light = "red", Heavy = "blue")) +
  scale_shape_manual(values = c(16, 17)) +
  labs(x = "Acceleration", y = "MPG", color = "Weight", shape = "Weight")
```



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

```
ggplot() +
  geom_point(data = light_cars, aes(x = log.acceleration., y = log.mpg., color = "Light", shape = "Ligh
  geom_point(data = heavy_cars, aes(x = log.acceleration., y = log.mpg., color = "Heavy", shape = "Heav
  geom_smooth(data = light_cars, aes(x = log.acceleration., y = log.mpg.), method = "lm", se = FALSE,ful
  geom_smooth(data = heavy_cars, aes(x = log.acceleration., y =log.mpg.), method = "lm", se = FALSE,ful
  scale_color_manual(values = c(Light = "red", Heavy = "blue")) +
  scale_shape_manual(values = c(16, 17)) +
  labs(x = "Acceleration", y = "MPG", color = "Weight", shape = "Weight")
```



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_regr <- lm(log.mpg. ~ log.weight. + log.acceleration. + model_year + origin, data = light_cars)
summary(light_regr)
```

```
##
## Call:
## lm(formula = log.mpg. ~ log.weight. + log.acceleration. + model_year +
##     origin, data = light_cars)
##
## 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.weight.       -0.849942   0.056655 -15.002   <2e-16 ***
## log.acceleration.  0.108295   0.056775   1.907   0.0578 .
## 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
```

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

```
##
## Call:
## lm(formula = log.mpg. ~ log.weight. + log.acceleration. + model_year +
##     origin, data = heavy_cars)
##
## 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.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 ***
## 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) Using your intuition only: What do you observe about light versus heavy cars so far?
Answer : Based on the scatter plot, both light and heavy cars have the same characteristic as the acceleration increases, the mpg also increases.

QUESTION 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.
Answer : Acceleration is likely to be a moderating variable

B) Use various regression models to model the possible moderation on log.mpg.:

1) Report a regression without any interaction terms

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

```
##
## 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
```

2) Report a regression with an interaction between weight and acceleration

```
regr_int <- lm(log.mpg. ~ log.weight. + log.acceleration. + model_year + factor(origin) + log.weight.*l
summary(regr_int)
```

```
##
## Call:
## lm(formula = log.mpg. ~ log.weight. + log.acceleration. + model_year +
##     factor(origin) + log.weight. * log.acceleration., data = cars_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37807 -0.06868  0.00463  0.06891  0.39857
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)                     1.089642   2.752872   0.396  0.69245    
## log.weight.                    -0.096632   0.337637  -0.286  0.77488    
## log.acceleration.               2.357574   0.995349   2.369  0.01834 *  
## model_year                      0.033685   0.001735  19.411  < 2e-16 ***
## factor(origin)2                 0.058737   0.017789   3.302  0.00105 ** 
## factor(origin)3                 0.028179   0.018266   1.543  0.12370    
## log.weight.:log.acceleration. -0.287170   0.123866  -2.318  0.02094 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.115 on 391 degrees of freedom
## Multiple R-squared:  0.8871, Adjusted R-squared:  0.8854 
## F-statistic: 512.2 on 6 and 391 DF,  p-value: < 2.2e-16
```

3) Report a regression with a mean-centered interaction term

```
weight_mc <- scale(cars_log$log.weight., center=TRUE, scale=FALSE)
acceleration_mc <- scale(cars_log$log.acceleration., center=T, scale=F)
mpg_mc <- scale(cars_log$log.mpg., center=T, scale=F)
regr_interaction <- summary(lm(mpg_mc ~ weight_mc + acceleration_mc + weight_mc*acceleration_mc))
regr_interaction
```

```
## 
## Call:
## lm(formula = mpg_mc ~ weight_mc + acceleration_mc + weight_mc * 
##     acceleration_mc)
## 
## 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    
## weight_mc                 -0.997466   0.031930 -31.239  < 2e-16 ***
## acceleration_mc            0.187500   0.051862   3.615 0.000339 ***
## weight_mc:acceleration_mc  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
```

4) Report a regression with an orthogonalized interaction term

```
weight_acc <- cars_log$log.weight.*cars_log$log.acceleration.
interaction_regr <- lm(weight_acc ~ cars_log$log.weight. + cars_log$log.acceleration.)
interaction_ortho <- interaction_regr$residuals
summary(lm(log.mpg.~log.weight.+log.acceleration.+interaction_ortho,data=cars_log))
```

```
## 
## Call:
```

```
## lm(formula = log.mpg. ~ log.weight. + log.acceleration. + interaction_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 ***
## interaction_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?

1. Raw

```
cor(cars_log$log.weight., cars_log$log.weight. * cars_log$log.acceleration.)
```

```
## [1] 0.1083055
```

```
cor(cars_log$log.acceleration., cars_log$log.weight. * cars_log$log.acceleration.)
```

```
## [1] 0.852881
```

2. Mean Centered

```
cor(weight_mc, weight_mc*acceleration_mc)
```

```
##            [,1]
## [1,] -0.2026948
```

```
cor(acceleration_mc, weight_mc*acceleration_mc)
```

```
##           [,1]
## [1,] 0.3512271
```

3. Orthogonalized

```r
cor(interaction_ortho, cars_log$log.weight.)
```

```
## [1] 2.468461e-17
```

```r
cor(interaction_ortho, cars_log$log.acceleration.)
```

```
## [1] -6.804111e-17
```

QUESTION 3

A) Let's try computing the direct effects first:

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

```
model1 <- lm(log.weight. ~ log.cylinders., data = cars_log)
summary(model1)
```

```
##
## 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
```

Answer : Yes, based on the p-value it has a significant direct effect on weight

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

```
model2 <- lm(log.mpg. ~ log.weight. + log.acceleration. + model_year + factor(origin), data = cars_log)
summary(model2)
```

```
##
## 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
```

Answer : Yes, based on the p-value weight has a significant direct effect on mpg

B) What is the indirect effect of cylinders on mpg?

```
indirect_effect <- model1$coefficients[2]*model2$coefficients[2]
indirect_effect
```

```
## log.cylinders.
##     -0.7189275
```

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

1) 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(42)
indirect <- replicate(2000,boot_mediation(model1, model2, cars_log))
quantile(indirect, probs=c(0.025, 0.975))
```

```
##       2.5%      97.5%
## -0.7784044 -0.6610106
```

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

```
plot(density(indirect))
abline(v=quantile(indirect, probs=c(0.025, 0.975)),lty=2)
```