Student ID: 112077423

```r
library(ggplot2)
library(compstatslib)
library(data.table)
library(tidyr)
library(lsa)
```

```
## Warning:     'lsa'          R     4.3.3
```

```r
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(weight), log(acceleration),
                                  model_year, origin))
head(cars_log)
```

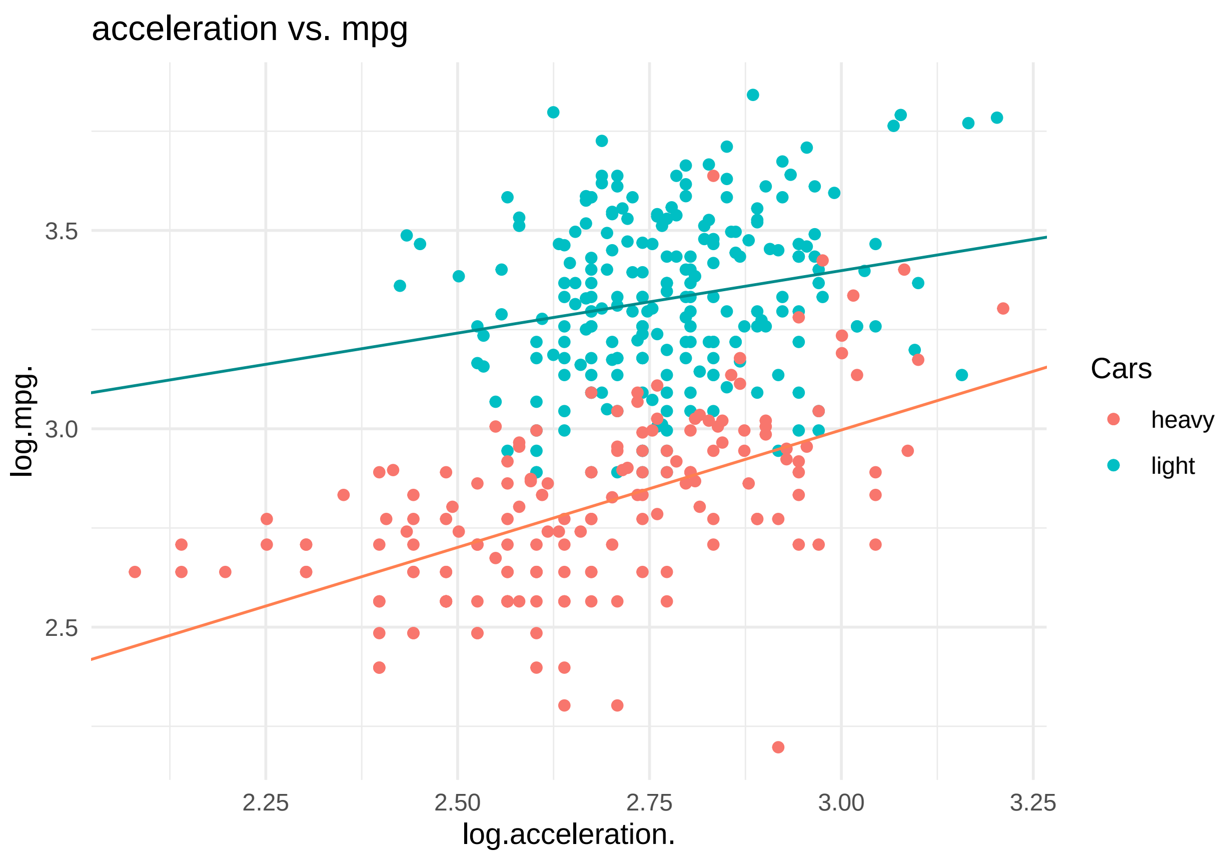
```
##   log.mpg. log.weight. log.acceleration. model_year origin
## 1 2.890372    8.161660          2.484907         70      1
## 2 2.708050    8.214194          2.442347         70      1
## 3 2.890372    8.142063          2.397895         70      1
## 4 2.772589    8.141190          2.484907         70      1
## 5 2.833213    8.145840          2.351375         70      1
## 6 2.708050    8.375860          2.302585         70      1
```

Question 1(a)

```r
mean_wt <- log(mean(cars$weight))
lw <- cars_log[cars_log$log.weight. < mean_wt,]
hw <- cars_log[cars_log$log.weight. > mean_wt,]

lw_regr <- lm(log.mpg. ~ log.acceleration., data=lw)
hw_regr <- lm(log.mpg. ~ log.acceleration., data=hw)

ggplot() +
  geom_point(data = lw, aes(x=log.acceleration., y=log.mpg., color='light')) +
  geom_point(data = hw, aes(x=log.acceleration., y=log.mpg., color='heavy')) +
  geom_abline(slope=lw_regr$coefficients[2], intercept=lw_regr$coefficients[1], color="cyan4") +
  geom_abline(slope=hw_regr$coefficients[2], intercept=hw_regr$coefficients[1], color="coral") +
  labs(title='acceleration vs. mpg', color='Cars') +
  theme_minimal()
```



Question 1(b)

```
full_lw_regr <- lm(log.mpg. ~ . - origin + factor(origin), data=lw)
summary(full_lw_regr)
```

```
##
## Call:
## lm(formula = log.mpg. ~ . - origin + factor(origin), data = lw)
##
## 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
```

The model is significant at alpha=0.01. Adjusted R-squared is 0.72, meaning that ~72% of variation in the dependent variable can be explained by the variation in the independent variables. All the variables are significant at alpha=1% except log.acceleration. and origin.

```r
full_hw_regr <- lm(log.mpg. ~ . - origin + factor(origin), data=hw)
summary(full_hw_regr)
```

```
##
## Call:
## lm(formula = log.mpg. ~ . - origin + factor(origin), data = hw)
##
## 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
```

The model is significant at alpha=0.01. Adjusted R-squared is 0.71, meaning that ~71% of variation in the dependent variable can be explained by the variation in the independent variables. All the variables are significant at alpha=1% except log.acceleration. and origin.

Question 1(c)

- For heavy cars, there are only two countries present;
- mpg is higher for light cars;
- Looking at the scatter plot, heavy cars' mpg rises more as acceleration increases.

Question 2(a)

I think weight is a moderator and acceleration is an independent variable.

Question 2(b)

```r
model1 <- lm(log.mpg. ~ . - origin + factor(origin), data=cars_log)
summary(model1)
```

```
##
## Call:
## lm(formula = log.mpg. ~ . - origin + 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
```

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

```
##
## Call:
## lm(formula = log.mpg. ~ log.weight. + log.acceleration. + log.weight. *
##     log.acceleration. + model_year + factor(origin), 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
```

```
acl_mc <- scale(cars_log$log.acceleration., center=TRUE, scale=FALSE)
wt_mc <- scale(cars_log$log.weight., center=TRUE, scale=FALSE)

model3 <- lm(cars_log$log.mpg. ~ wt_mc + acl_mc +
             wt_mc * acl_mc +
             cars_log$model_year + factor(cars_log$origin))
summary(model3)
```

```
##
## Call:
## lm(formula = cars_log$log.mpg. ~ wt_mc + acl_mc + wt_mc * acl_mc +
##     cars_log$model_year + factor(cars_log$origin))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37807 -0.06868  0.00463  0.06891  0.39857
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.518882   0.132944   3.903 0.000112 ***
## wt_mc                    -0.880393   0.028585 -30.799  < 2e-16 ***
## acl_mc                    0.072596   0.037567   1.932 0.054031 .
## cars_log$model_year       0.033685   0.001735  19.411  < 2e-16 ***
## factor(cars_log$origin)2  0.058737   0.017789   3.302 0.001049 **
## factor(cars_log$origin)3  0.028179   0.018266   1.543 0.123704
## wt_mc:acl_mc             -0.287170   0.123866  -2.318 0.020943 *
## ---
## 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
```

```
wt_x_acl <- cars_log$log.weight. * cars_log$log.acceleration.
interaction_regr <- lm(wt_x_acl ~ cars_log$log.weight. + cars_log$log.acceleration.)
interaction_ortho <- interaction_regr$residuals

model4 <- lm(cars_log$log.mpg. ~ cars_log$log.weight. + cars_log$log.acceleration. +
             cars_log$model_year + factor(cars_log$origin) + interaction_ortho)
summary(model4)
```

```
##
## Call:
## lm(formula = cars_log$log.mpg. ~ cars_log$log.weight. + cars_log$log.acceleration. +
##     cars_log$model_year + factor(cars_log$origin) + interaction_ortho)
##
## 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)                 7.377176   0.311392  23.691  < 2e-16 ***
## cars_log$log.weight.       -0.876967   0.028539 -30.729  < 2e-16 ***
## cars_log$log.acceleration.  0.046100   0.036524   1.262  0.20764
## cars_log$model_year         0.033685   0.001735  19.411  < 2e-16 ***
## factor(cars_log$origin)2    0.058737   0.017789   3.302  0.00105 **
## factor(cars_log$origin)3    0.028179   0.018266   1.543  0.12370
## interaction_ortho          -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
```

Question 2(c)

```r
# raw
cor(cbind(log.mpg = cars_log$log.mpg.,
          log.weight.=cars_log$log.weight., log.acceleration.=cars_log$log.acceleration.,
          intxn = cars_log$log.weight. * cars_log$log.acceleration.))
```

```
##                        log.mpg log.weight. log.acceleration.       intxn
## log.mpg            1.000000000  -0.8744686         0.4640533 0.007445392
## log.weight.       -0.874468594   1.0000000        -0.4256194 0.108305532
## log.acceleration.  0.464053310  -0.4256194         1.0000000 0.852881042
## intxn              0.007445392   0.1083055         0.8528810 1.000000000
```

- The correlation between that interaction term and log.weight. is 0.108
- The correlation between that interaction term and log.acceleration. is 0.852

```r
# mean-centered
tmp <- cor(cbind(log.mpg = cars_log$log.mpg.,
                 wt_mc=wt_mc, acl_mc=acl_mc, intxn = wt_mc * acl_mc))
colnames(tmp) <- c('log.mpg', 'wt_mc', 'acl_mc', 'intxn')
rownames(tmp) <- c('log.mpg', 'wt_mc', 'acl_mc', 'intxn')
tmp
```

```
##            log.mpg      wt_mc     acl_mc      intxn
## log.mpg  1.0000000 -0.8744686  0.4640533  0.2404855
## wt_mc   -0.8744686  1.0000000 -0.4256194 -0.2026948
## acl_mc   0.4640533 -0.4256194  1.0000000  0.3512271
## intxn    0.2404855 -0.2026948  0.3512271  1.0000000
```

- The correlation between that interaction term and mean-centered weight is -0.2

- The correlation between that interaction term and mean-centered accelaration is 0.35

```
# orthogonalized
round(cor(cbind(log.mpg = cars_log$log.mpg.,
                log.weight.=cars_log$log.weight.,
                log.acceleration.=cars_log$log.acceleration., interaction_ortho)), 2)
```

```
##                   log.mpg log.weight. log.acceleration. interaction_ortho
## log.mpg              1.00       -0.87              0.46              0.04
## log.weight.         -0.87        1.00             -0.43              0.00
## log.acceleration.    0.46       -0.43              1.00              0.00
## interaction_ortho    0.04        0.00              0.00              1.00
```

- The correlation between that interaction term and log.weight. is 0
- The correlation between that interaction term and log.acceleration. is 0

Question 3(a)

```
cars_log <- with(cars, data.frame(log(mpg), log(weight), log(acceleration),
                                  log(cylinders), model_year, origin))
head(cars_log)
```

```
##   log.mpg. log.weight. log.acceleration. log.cylinders. model_year origin
## 1 2.890372    8.161660          2.484907       2.079442         70      1
## 2 2.708050    8.214194          2.442347       2.079442         70      1
## 3 2.890372    8.142063          2.397895       2.079442         70      1
## 4 2.772589    8.141190          2.484907       2.079442         70      1
## 5 2.833213    8.145840          2.351375       2.079442         70      1
## 6 2.708050    8.375860          2.302585       2.079442         70      1
```

```
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
```

log.cylinders. is significant at alpha=0.01.

```r
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
```

The weight has a significant direct effect on mpg at alpha=0.01.

Question 3(b)

```r
tmp <- model1$coefficients[2] * model2$coefficients[2]
cat('indirect effect of cylinders on mpg =', tmp, sep=' ')
```

```
## indirect effect of cylinders on mpg = -0.7189275
```

Question 3(c)

```r
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])
}
```

```r
set.seed(645218)

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

```
##       2.5%      97.5%
## -0.7797629 -0.6582639
```

```r
plot(density(indirect), main='distribution of the indirect effect', lwd=1.5)
abline(v=quantile(indirect, probs=c(0.025, 0.975)), lw=1.5, lty='dashed', col='red')
```