Business Analytics using Statistical Modeling
Assignment 12

Create a data.frame called cars_log with log-transformed columns (except model_year and origin – keep them in their original form)

```
cars <- read.table("../10-auto-data.txt", header = FALSE, na.strings = "?")[-9]
names(cars) <- c("mpg", "cylinders", "displacement", "horsepower", "weight",
                 "acceleration", "model_year", "origin")
cars_log <- with(cars, data.frame(log(mpg), log(cylinders), log(displacement),
                                  log(horsepower), log(weight), log(acceleration),
                                  model_year, origin))
```

Question 1

Let's visualize how acceleration is related to mpg.

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)

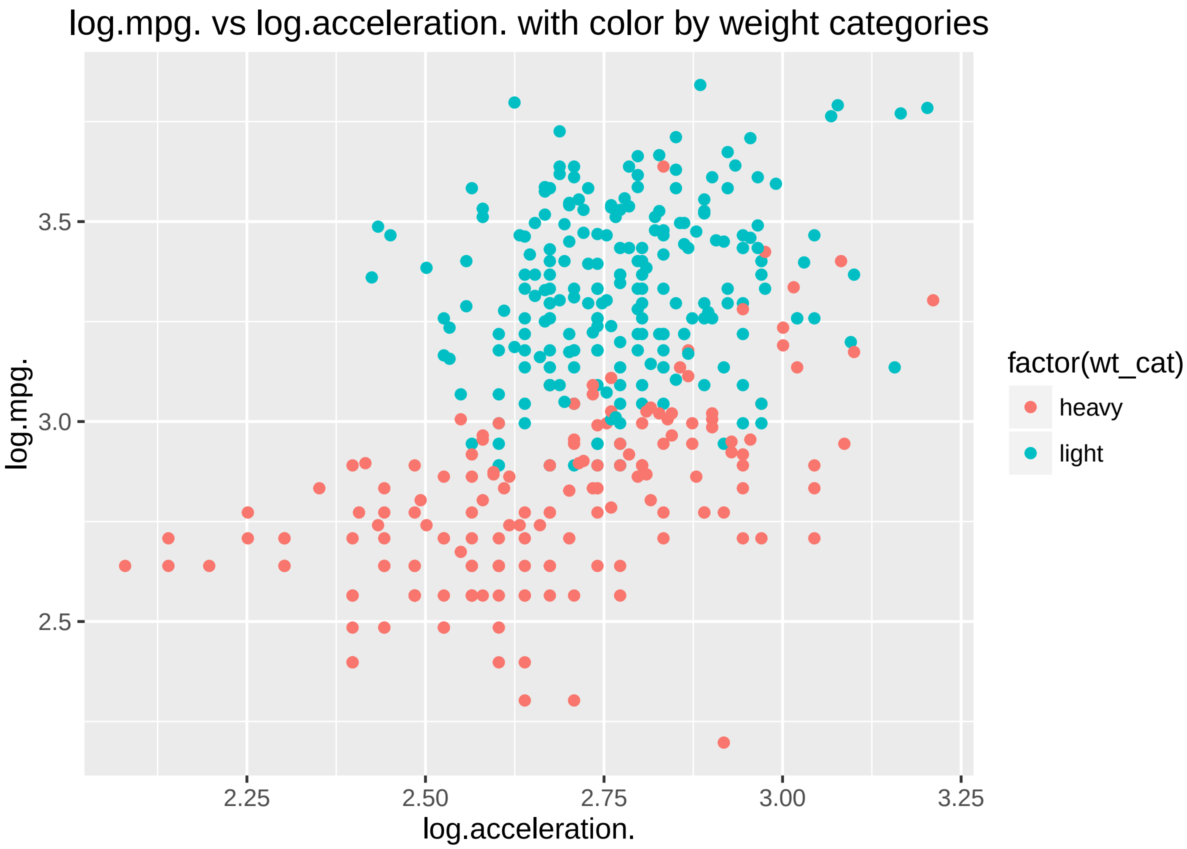
```
cars_light <- cars_log[cars_log$log.weight. <= log(mean(cars$weight)), ]
cars_heavy <- cars_log[cars_log$log.weight. > log(mean(cars$weight)), ]
```

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

```
cars_log$wt_cat <- ifelse(cars_log$log.weight. > log(mean(cars$weight)), 'heavy', 'light')
library(ggplot2)
```

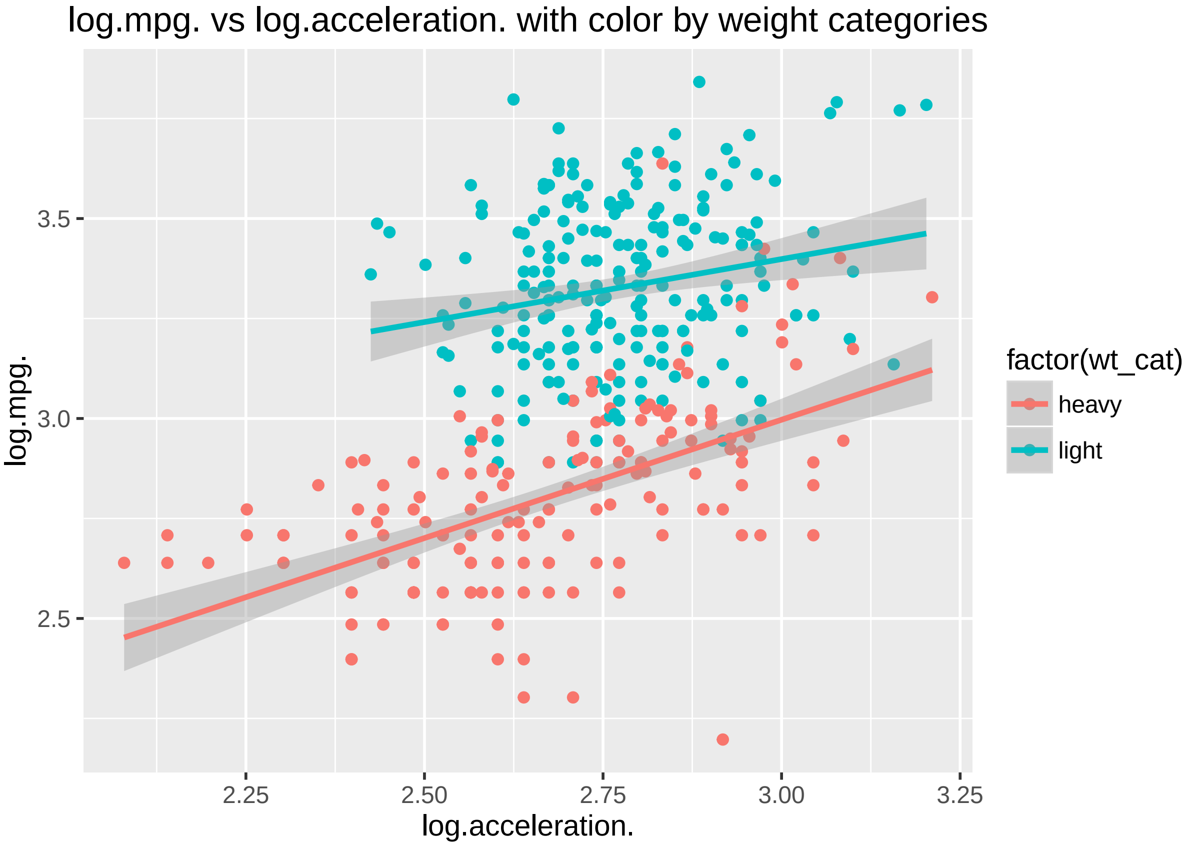
```
## Warning: package 'ggplot2' was built under R version 3.2.5
```

```
ggplot(cars_log, aes_string('log.acceleration.', 'log.mpg.')) +
  geom_point(aes(color = factor(wt_cat))) +
  ggtitle('log.mpg. vs log.acceleration. with color by weight categories')
```



iii. Draw two slopes of acceleration versus mpg over the scatter plot: one for light cars and one for heavy cars (distinguish their appearance)

```
ggplot(cars_log, aes_string('log.acceleration.', 'log.mpg.')) +
  geom_point(aes(color = factor(wt_cat))) +
  geom_smooth(method = 'lm', aes(color = factor(wt_cat))) +
  ggtitle('log.mpg. vs log.acceleration. with color by weight categories')
```



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

```r
summary(lm(log.mpg. ~ log.weight. + log.acceleration. + model_year + factor(origin),
           data = cars_light))
```

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

```r
summary(lm(log.mpg. ~ log.weight. + log.acceleration. + model_year + factor(origin),
           data = cars_heavy))
```

```
##
## Call:
## lm(formula = log.mpg. ~ log.weight. + log.acceleration. + model_year +
##     factor(origin), data = cars_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)        7.188679   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. (not graded)

Question 2

Using our full transformed dataset (cars_log), let's test whether we have moderation.

a. (not graded)

b. Let's use various regression models to test the possible moderation on our full data: (use independent variables log.weight., log.acceleration., model_year and origin)

i. Report a regression without any interaction terms

```
regr_mpg_log1 <- lm(log.mpg. ~ log.weight. + log.acceleration. + model_year +
                      factor(origin), data = cars_log)
summary(regr_mpg_log1)
```

```
##
## 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 a raw interaction between weight and acceleration

```
regr_mpg_log2 <- lm(log.mpg. ~ log.weight. + log.acceleration. + model_year +
                      factor(origin) + log.weight. * log.acceleration., data = cars_log)
summary(regr_mpg_log2)
```

```
##
## 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
```

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

```
wt_log_mc <- scale(cars_log$log.weight., center = TRUE, scale = FALSE)
acc_log_mc <- scale(cars_log$log.acceleration., center = TRUE, scale = FALSE)
mpg_log_mc <- scale(cars_log$log.mpg., center = TRUE, scale = FALSE)
regr_mpg_log3 <- lm(mpg_log_mc ~ wt_log_mc + acc_log_mc + cars_log$model_year +
                      factor(cars_log$origin) + wt_log_mc * acc_log_mc)
summary(regr_mpg_log3)
```

```
##
## Call:
## lm(formula = mpg_log_mc ~ wt_log_mc + acc_log_mc + cars_log$model_year +
##     factor(cars_log$origin) + wt_log_mc * acc_log_mc)
##
## 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)              -2.582502   0.132944 -19.426  < 2e-16 ***
## wt_log_mc                -0.880393   0.028585 -30.799  < 2e-16 ***
## acc_log_mc                0.072596   0.037567   1.932  0.05403 .
## 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
## wt_log_mc:acc_log_mc     -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
```

iv. Report a regression with an orthogonalized interaction term

```
wt_x_acc <- cars_log$log.weight. * cars_log$log.acceleration.
regr_wt_x_acc <- lm(wt_x_acc ~ cars_log$log.weight. + cars_log$log.acceleration.)
regr_mpg_log4 <- lm(cars_log$log.mpg. ~ cars_log$log.weight. +
                      cars_log$log.acceleration. + cars_log$model_year +
                      factor(cars_log$origin) + regr_wt_x_acc$residuals)
summary(regr_mpg_log4)
```

```
##
## Call:
## lm(formula = cars_log$log.mpg. ~ cars_log$log.weight. + cars_log$log.acceleration. +
##     cars_log$model_year + factor(cars_log$origin) + regr_wt_x_acc$residuals)
##
## 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$origin)3    0.028179   0.018266   1.543  0.12370
## regr_wt_x_acc$residuals    -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
```

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.weight. * cars_log$log.acceleration.)
```

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

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

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

Mean-centered

```
cor(wt_log_mc, wt_log_mc * acc_log_mc)
```

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

```
cor(acc_log_mc, wt_log_mc * acc_log_mc)
```

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

Orthogonalized

```
cor(cars_log$log.weight., regr_wt_x_acc$residuals)
```

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

```
cor(cars_log$log.acceleration., regr_wt_x_acc$residuals)
```

```
## [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. Acceleration, model_year, and origin are kept as control variables (see gray variables in diagram).

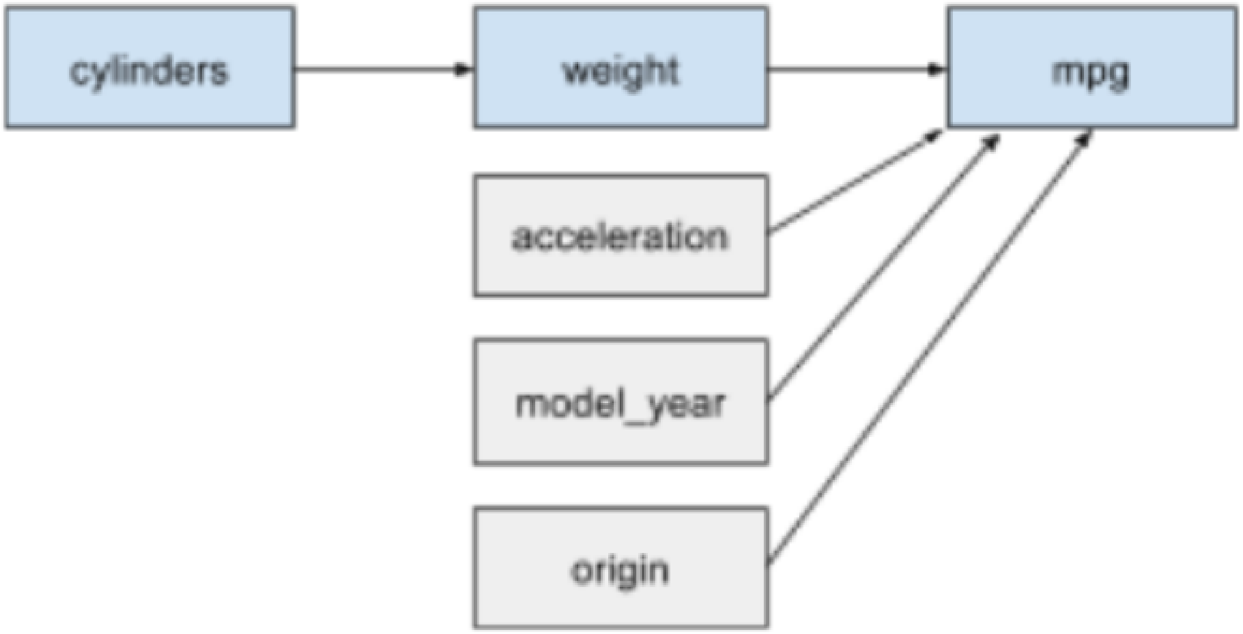


Figure 1: Conceptual Path Diagram of Mediated Model

a. Let's try out the steps of the Baron & Kenny (1984) method for checking for mediation:

i. Regress log.mpg. over log.cylinders. and all control variables

```
regr_mpg_log5 <- lm(log.mpg. ~ log.cylinders. + log.acceleration. + model_year +
                      factor(origin), data = cars_log)
summary(regr_mpg_log5)
```

```
##
## Call:
## lm(formula = log.mpg. ~ log.cylinders. + log.acceleration. +
##     model_year + factor(origin), data = cars_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.56929 -0.09826  0.00206  0.10053  0.48033
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.73840    0.24570   7.075 6.91e-12 ***
## log.cylinders.    -0.68561    0.03849 -17.814  < 2e-16 ***
## log.acceleration.  0.02930    0.05192   0.564  0.57283
## model_year         0.03127    0.00235  13.307  < 2e-16 ***
## factor(origin)2    0.08201    0.02507   3.272  0.00116 **
## factor(origin)3    0.11537    0.02435   4.738 3.02e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.158 on 392 degrees of freedom
## Multiple R-squared:  0.7862, Adjusted R-squared:  0.7835
## F-statistic: 288.4 on 5 and 392 DF,  p-value: < 2.2e-16
```

Does cylinders have a significant direct effect on mpg when weight is not considered?

Yes.

ii. Regress log.weight. over log.cylinders. only

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

```
##
## 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
```

Does cylinders have a significant direct effect on weight itself?

Yes.

iii. Regress log.mpg. over log.weight. and all control variables

```
summary(regr_mpg_log1)
```

```
##
## 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
```

Does weight have a direct effect on mpg?

Yes.

If all steps (i) (ii) and (iii) have been significant, then we at least have “partial mediation”!

We can do one more thing to see if we have full mediation:

iv. Regress log.mpg. on log.weight., log.cylinders., and all control variables

```
regr_mpg_log6 <- lm(log.mpg. ~ log.weight. + log.cylinders. + log.acceleration. +
                      model_year + factor(origin), data = cars_log)
summary(regr_mpg_log6)
```

```
##
## Call:
## lm(formula = log.mpg. ~ log.weight. + log.cylinders. + log.acceleration. +
##     model_year + factor(origin), data = cars_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39866 -0.06888  0.00227  0.06718  0.40603
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.25316    0.34818  20.831   <2e-16 ***
## log.weight.       -0.83628    0.04523 -18.491   <2e-16 ***
## log.cylinders.    -0.05119    0.04438  -1.153   0.2495
## log.acceleration.  0.03997    0.03798   1.053   0.2932
## model_year         0.03240    0.00172  18.838   <2e-16 ***
## factor(origin)2    0.05298    0.01840   2.880   0.0042 **
## factor(origin)3    0.02984    0.01840   1.622   0.1057
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1156 on 391 degrees of freedom
## Multiple R-squared:  0.886,  Adjusted R-squared:  0.8842
## F-statistic: 506.3 on 6 and 391 DF,  p-value: < 2.2e-16
```

Does cylinders have a significant direct effect on mpg when weight is also considered?

No.

If the coefficient of cylinders in step (iv) is not significant, then we have "full mediation"

b. What is the indirect effect of cylinders on mpg?

```
regr_mpg_log7 <- lm(log.mpg. ~ log.weight., data = cars_log)
indirect_effect <- regr_wt_log$coefficients[2] * regr_mpg_log7$coefficients[2]
unname(indirect_effect)
```

```
## [1] -0.8679111
```

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

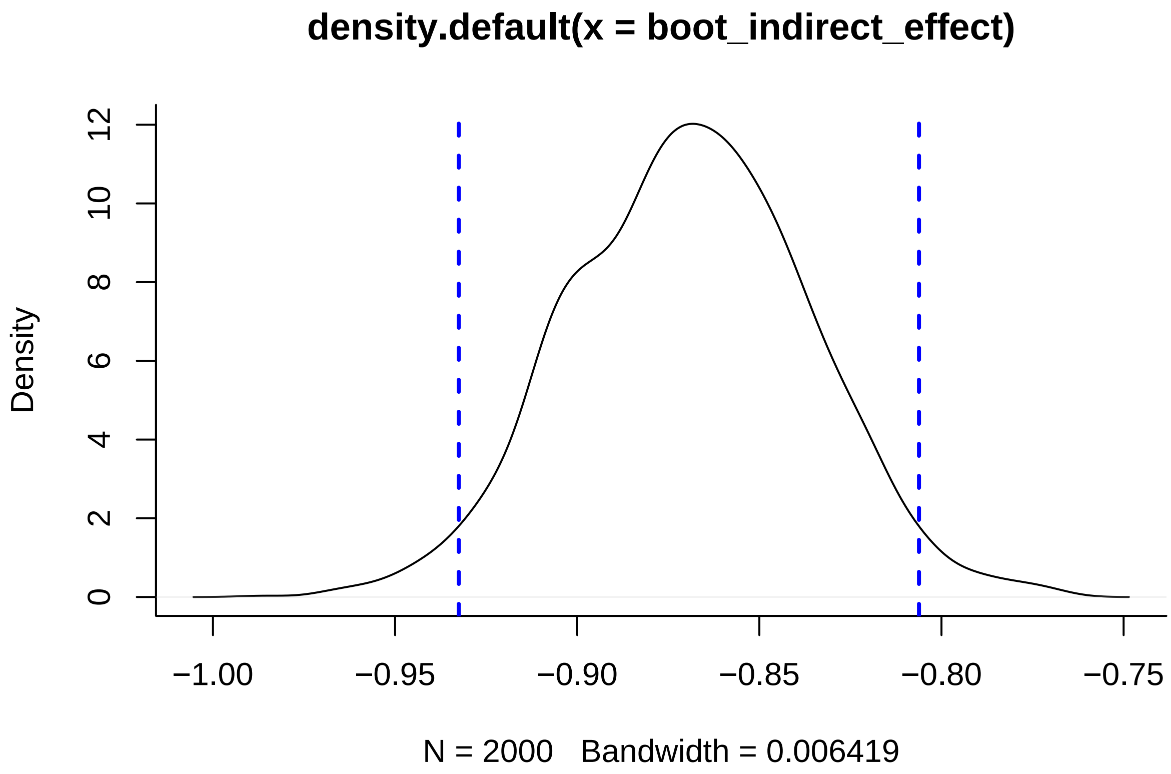
i. Bootstrap regressions (ii) and (iii) to find the range of indirect effects: what is its 95% CI?

```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])
}
set.seed(0)
boot_indirect_effect <- replicate(2000, boot_mediation(regr_wt_log, regr_mpg_log7,
                                                       cars_log))
ci_95 <- quantile(boot_indirect_effect, probs = c(0.025, 0.975))
ci_95
```

```
##       2.5%      97.5%
## -0.9325106 -0.8061844
```

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

```r
plot(density(boot_indirect_effect), bty = 'l')
abline(v = ci_95, col = 'blue', lty = 2, lwd = 2)
```


density.default(x = boot_indirect_effect)
Density
N = 2000   Bandwidth = 0.006419