HW12

Answers are in **black** or red text.

```
auto <- read.table("auto-data.txt", header=FALSE, na.strings = "?", stringsAsFactors = F)
names(auto) <- c("mpg", "cylinders", "displacement", "horsepower", "weight",
"acceleration", "model_year", "origin", "car_name")
cars_log <- with(auto, 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:

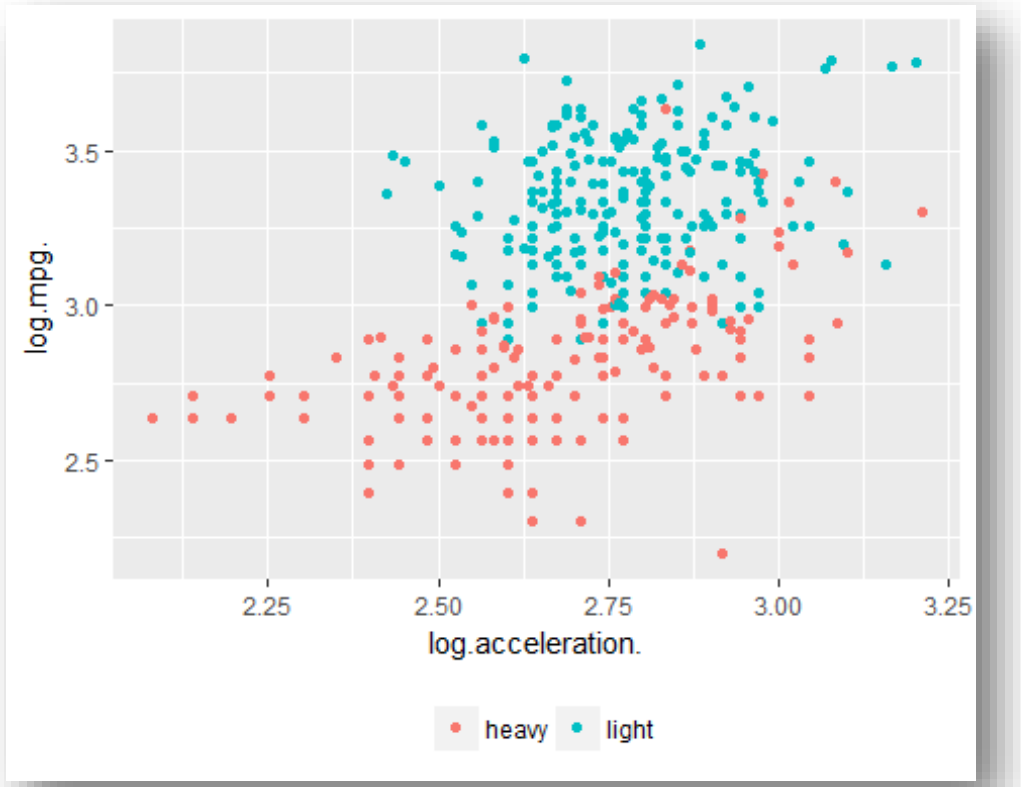
i. Create two subsets of your data, one for light weight cars (less than mean weight) and one for heavy cars (higher than the mean weight) HINT: careful how you compare log weights to mean weight

```
light_weight <- subset(cars_log,log.weight. < log(mean(auto$weight)))
heavy_weight <- subset(cars_log,log.weight. >= log(mean(auto$weight)))
```

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

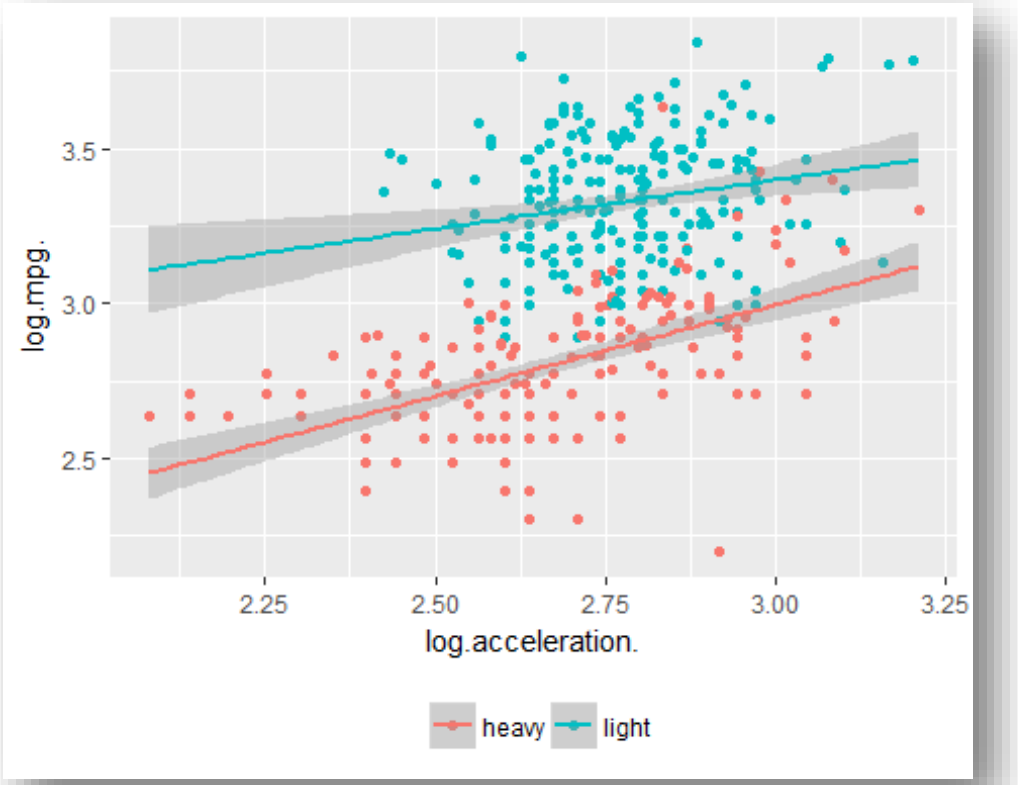
```
library(dplyr)

cars_log <- cars_log %>% mutate(weight_level = ifelse(log.weight.>= log(mean(auto$weight)),
                                                      "heavy", "light"))
library(ggplot2)
plt <- ggplot(data = cars_log, aes(x = log.acceleration., y=log.mpg., col = factor(weight_level)))+
  geom_point()+
  theme(legend.position="bottom", legend.direction="horizontal") +
  theme(legend.title=element_blank())
plt
```



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

```
plt+geom_smooth(method = lm,fullrange = TRUE)
```



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

```
heavy_lm <- lm(data = heavy_weight, log.mpg.~ log.weight. + log.acceleration. + model_year + factor(origin))
light_lm <- lm(data = light_weight, log.mpg.~ log.weight. + log.acceleration. + model_year + factor(origin))

summary(heavy_lm)
```

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

```
summary(light_lm)
```

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

c). (not graded) Using your intuition only: What do you observe about light versus heavy cars so far ?

From the above plot, I think there's no significant diference between **two slopes**. Besides, since *light_wight* have more data point comparing to *heavy_weight*, it can reach the significance level eaisier, even though the regression line seemed to fit better on *heavy_weight* dataset.

Question 2) Using our full transformed dataset (cars_log), let's test whether we have moderation.

a). (not graded) Between weight and acceleration, use your intuition and experience to state which variable might be a moderating versus independent variable, in affecting mpg.

I think acceleration might be a moderating versus independent variable, in affecting mpg.

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

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

```
##
## 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

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

```
##
## 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

```
log.weight.mc <- scale(cars_log$log.weight., center = TRUE, scale = FALSE)
log.acc.mc <- scale(cars_log$log.acceleration., center = TRUE, scale = FALSE)
log.mpg.mc <- scale(cars_log$log.mpg., center = TRUE, scale = FALSE)

summary(lm(log.mpg.mc~log.acc.mc+log.weight.mc+log.acc.mc*log.weight.mc))
```

```
##
## Call:
## lm(formula = log.mpg.mc ~ log.acc.mc + log.weight.mc + log.acc.mc *
##     log.weight.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
## log.acc.mc               0.187500   0.051862   3.615 0.000339 ***
## log.weight.mc           -0.997466   0.031930 -31.239  < 2e-16 ***
## log.acc.mc:log.weight.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
```

iv. Report a regression with an orthogonalized interaction term

```
weight_x_acc <- cars_log$log.weight. * cars_log$log.acceleration.
inter_regr <- lm(weight_x_acc ~ cars_log$log.weight. + cars_log$log.acceleration.)
cor(inter_regr$residuals, cars_log$log.weight.)

## [1] 2.468461e-17

cor(inter_regr$residuals, cars_log$log.acceleration.)

## [1] -6.804111e-17

summary(lm(data = cars_log, log.mpg. ~ log.weight. + log.acceleration. + inter_regr$residuals))

##
## Call:
## lm(formula = log.mpg. ~ log.weight. + log.acceleration. + inter_regr$residuals,
```

```
##     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_regr$residuals   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(log.acc.mc*log.weight.mc, log.weight.mc))
d <- as.vector(cor(log.acc.mc*log.weight.mc, log.acc.mc))

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

```
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
```

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).

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 (does cylinders have a significant direct effect on mpg when weight is not considered?)

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

```
##
## 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
```

Yes, it has significant effect on log.mpg. here.

ii. Regress log.weight. over log.cylinders. only (does cylinders have a significant direct effect on weight itself)

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

```
##
## 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
```

Yes, it has significant effect on log.weight..

iii. Regress log.mpg. over log.weight. and all control variables (does weight have a direct effect on mpg?)

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

##
## 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
```

Yes, it has significant effect on log.mpg. directly.

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 (does cylinders have a significant direct effect on mpg when weight is also considered?) If the coefficient of cylinders in step (iv) is not significant, then we have "full mediation"

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

```
##
## 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
```

The coefficient of cylinders in step (iv) is not significant! We have "full mediation".

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

```
weight_cylinder_regr$coefficients[2] * mpg_weight_regr$coefficients[2]
```

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

The indirect effect coefficient of cylinders on mpg is about -0.719

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?

```
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)
intxns<-replicate(2000, boot_mediation(weight_cylinder_regr, mpg_weight_regr, cars_log))
quantile(intxns, probs=c(0.025, 0.975))
```

```
##        2.5%      97.5% 
## -0.7840590 -0.6578708
```

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

```
plot(density(intxns),col = "lightcoral",lwd=3,main="The 95% CI of the indirect effect")
abline(v=quantile(intxns, probs=c(0.025, 0.975)),lty=2)
```