BACS HW12
Student ID: 110065508

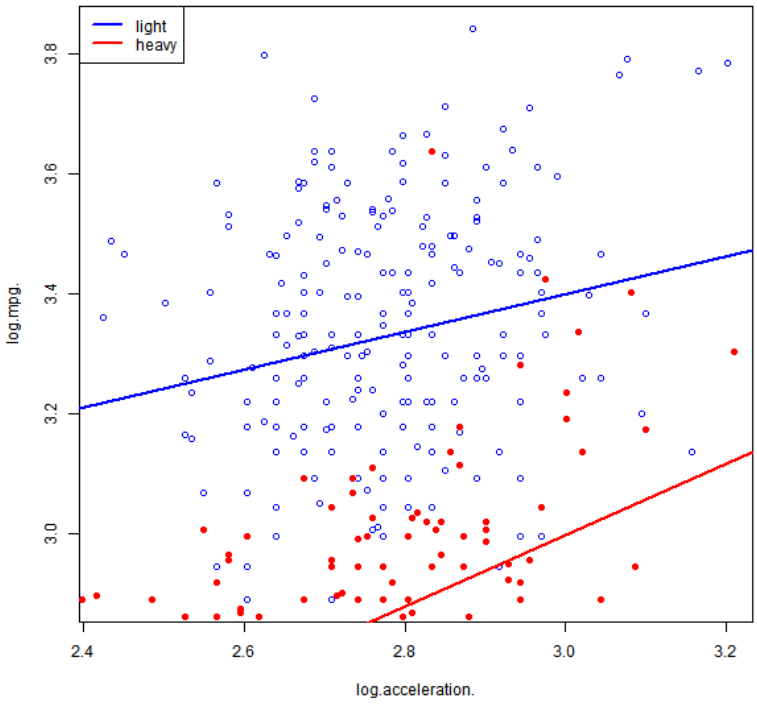
Due: 2022/05/08
Special thanks to 108071001

To begin with some data preprocessing.

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

Question 1. (a) To create a single scatter plot and two slopes, I reused the code in HW11.

```
# Question 1 (a)
weight_mean_log <- log(mean(cars$weight))
cars_light_log <- subset(cars_log, log.weight.<weight_mean_log) # light cars
cars_heavy_log <- subset(cars_log, log.weight.>=weight_mean_log) # heavy cars

regr_light <- with(cars_light_log, lm(log.mpg. ~ log.acceleration.))
regr_heavy <- with(cars_heavy_log, lm(log.mpg. ~ log.acceleration.))

png(filename = "1a.png", width = 600, height = 600) # Subplots
with(cars_light_log, plot(log.acceleration., log.mpg., pch=1, col="blue"))
with(cars_heavy_log, points(log.acceleration., log.mpg., pch=19, col="red"))
abline(regr_light, col="blue", lwd=2)
abline(regr_heavy, col="red", lwd=2)
legend(x = "topleft", legend = c("light", "heavy"),
       col = c("blue","red"), lwd = 2)
dev.off()
```



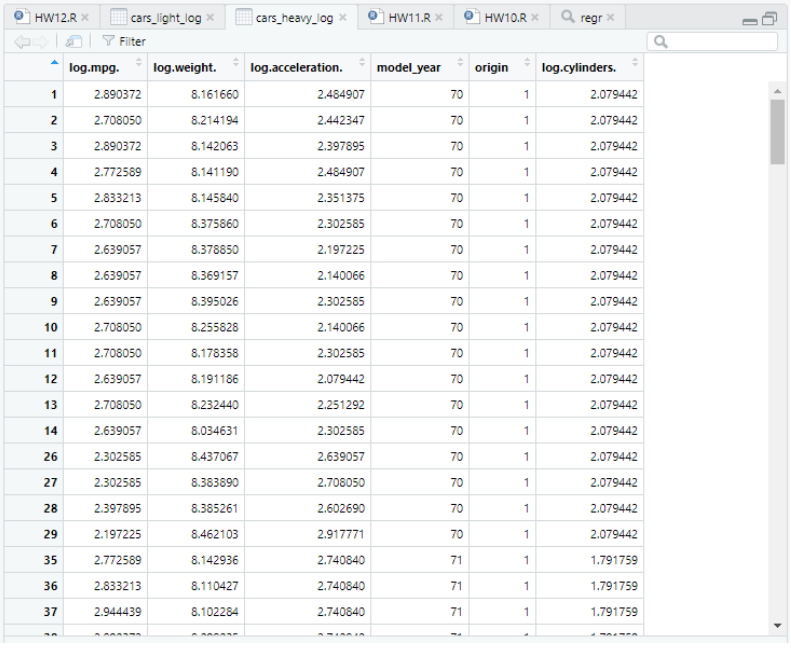
(b) Report the full summaries of two separate regressions.

```r
# Question 1 (b)
regr_light_full <- lm(log.mpg.~ log.weight. + log.acceleration. + model_year +
                      factor(origin), data=cars_light_log)
regr_heavy_full <- lm(log.mpg.~ log.weight. + log.acceleration. + model_year +
                      factor(origin), data=cars_heavy_log)
```

```
> summary(regr_light_full)

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

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

> |
```

```
> summary(regr_heavy_full)

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

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) At the first glance, it seems that two models are not that mush different. All the significant variables are the same. However, I've found that the heavy model depends on origin 2 only! Hence, I take a quick look at `cars_heavy_log`:



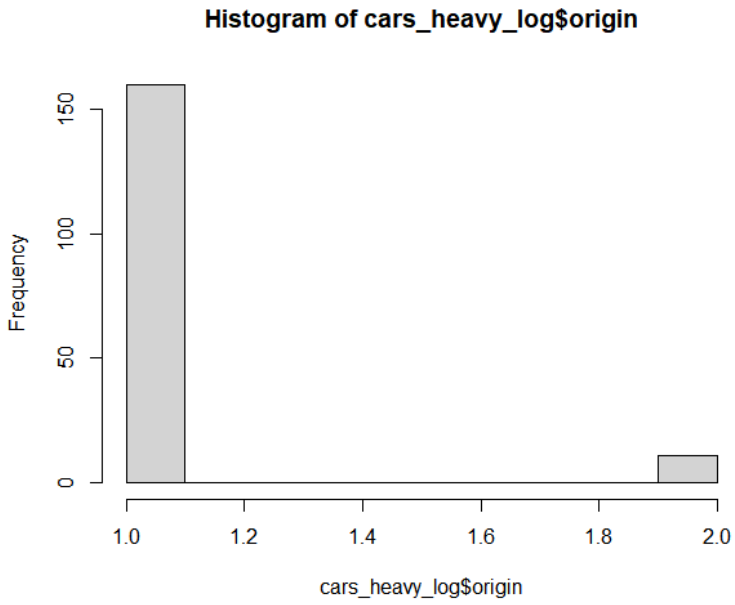
HW12.R | cars_light_log | cars_heavy_log | HW11.R | HW10.R | regr

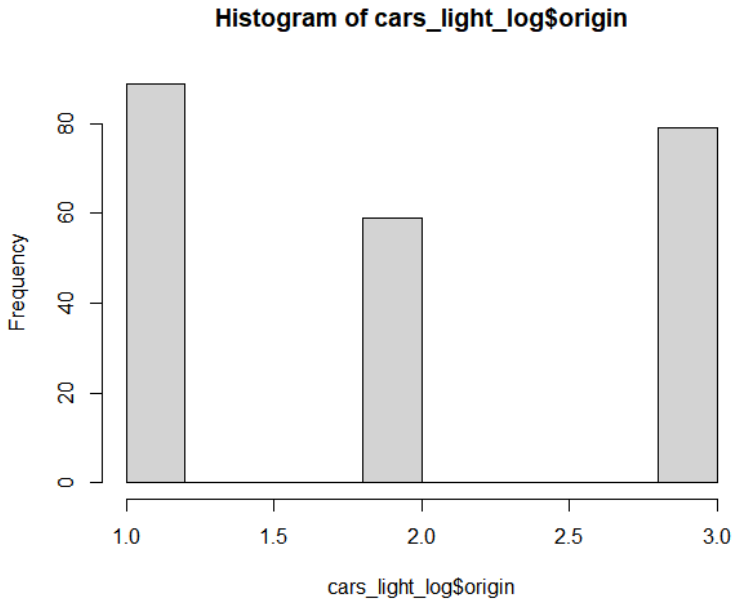
Filter

	log.mpg.	log.weight.	log.acceleration.	model_year	origin	log.cylinders.
1	2.890372	8.161660	2.484907	70	1	2.079442
2	2.708050	8.214194	2.442347	70	1	2.079442
3	2.890372	8.142063	2.397895	70	1	2.079442
4	2.772589	8.141190	2.484907	70	1	2.079442
5	2.833213	8.145840	2.351375	70	1	2.079442
6	2.708050	8.375860	2.302585	70	1	2.079442
7	2.639057	8.378850	2.197225	70	1	2.079442
8	2.639057	8.369157	2.140066	70	1	2.079442
9	2.639057	8.395026	2.302585	70	1	2.079442
10	2.708050	8.255828	2.140066	70	1	2.079442
11	2.708050	8.178358	2.302585	70	1	2.079442
12	2.639057	8.191186	2.079442	70	1	2.079442
13	2.708050	8.232440	2.251292	70	1	2.079442
14	2.639057	8.034631	2.302585	70	1	2.079442
26	2.302585	8.437067	2.639057	70	1	2.079442
27	2.302585	8.383890	2.708050	70	1	2.079442
28	2.397895	8.385261	2.602690	70	1	2.079442
29	2.197225	8.462103	2.917771	70	1	2.079442
35	2.772589	8.142936	2.740840	71	1	1.791759
36	2.833213	8.110427	2.740840	71	1	1.791759
37	2.944439	8.102284	2.740840	71	1	1.791759

It seems like that all most of the heavy cars are designed in from the US! On contrast, the light cars was designed from

all the three places.





■

Question 2. (a) In my intuition, it would be `weight` and `model_year`. Perhaps the car designed in later years pretend to be lighter.

(b) Moderation models:

```
# Question 2 (b-i)
regr_log_i <- lm(log.mpg.~ log.weight. + log.acceleration. +
                   model_year + factor(origin), data=cars_log)

# Question 2 (b-ii)
regr_log_ii <- lm(log.mpg.~ log.weight. + log.acceleration. +
                    model_year + factor(origin) +
                    log.weight.*log.acceleration., data=cars_log)

# Question 2 (b-iii)
log_weight_mc <- scale(cars_log$log.mpg.,
                       center=TRUE,
                       scale=FALSE)
log_acceleration_mc <- scale(cars_log$log.acceleration.,
```

```
                              center=TRUE,
                              scale=FALSE)
regr_log_iii <- lm(log.mpg.~ log_weight_mc + log_acceleration_mc +
                     model_year + factor(origin) +
                     log_weight_mc*log_acceleration_mc, data=cars_log)

# Question 2 (b-iv)
weight_x_acceleration <- cars_log$log.weight. * cars_log$log.acceleration.
interaction_regr <- lm(weight_x_acceleration ~
                         cars_log$log.weight. + cars_log$log.acceleration.)
interaction_ortho <- interaction_regr$residuals
regr_log_iv <- lm(log.mpg. ~ log.weight. + log.acceleration. +
                     model_year + factor(origin) + interaction_ortho,
                   data=cars_log)
```

```
> summary(regr_log_i)

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

>
```

```
> summary(regr_log_ii)

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

>
```

```
> summary(regr_log_iii)

Call:
lm(formula = log.mpg. ~ log_weight_mc + log_acceleration_mc +
    model_year + factor(origin) + log_weight_mc * log_acceleration_mc,
    data = cars_log)

Residuals:
       Min         1Q     Median         3Q        Max
-2.126e-14 -3.870e-17  6.260e-17  1.382e-16  4.645e-16

Coefficients:
                                 Estimate Std. Error   t value Pr(>|t|)
(Intercept)                     3.101e+00  1.470e-15 2.110e+15   <2e-16 ***
log_weight_mc                   1.000e+00  2.598e-16 3.850e+15   <2e-16 ***
log_acceleration_mc             4.243e-16  3.550e-16 1.195e+00   0.2327
model_year                      3.516e-17  1.909e-17 1.842e+00   0.0663 .
factor(origin)2                 1.543e-16  1.706e-16 9.040e-01   0.3665
factor(origin)3                 1.384e-16  1.694e-16 8.170e-01   0.4144
log_weight_mc:log_acceleration_mc 6.878e-17  8.888e-16 7.700e-02   0.9384
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.093e-15 on 391 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 6.393e+30 on 6 and 391 DF,  p-value: < 2.2e-16

> summary(regr_log_iv)

Call:
lm(formula = log.mpg. ~ log.weight. + log.acceleration. + model_year +
    factor(origin) + interaction_ortho, 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)       7.377176   0.311392  23.691  < 2e-16 ***
log.weight.      -0.876967   0.028539 -30.729  < 2e-16 ***
log.acceleration.  0.046100   0.036524   1.262  0.20764
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
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

>
```

(c) I use the following code to compute the required correlations.

```r
# Question 2 (c)
cor_ii_w <- round(cor(cars_log$log.mpg.,
                      cars_log$log.mpg. * cars_log$log.acceleration.), 3)
cor_ii_a <- round(cor(cars_log$log.acceleration.,
                      cars_log$log.mpg. * cars_log$log.acceleration.), 3)
cor_iii_w <- round(cor(log_weight_mc,
                       log_weight_mc*log_acceleration_mc), 3)
cor_iii_a <- round(cor(log_acceleration_mc,
                       log_weight_mc*log_acceleration_mc), 3)
cor_iv_w <- round(cor(cars_log$log.weight., interaction_ortho), 3)
cor_iv_a <- round(cor(cars_log$log.acceleration., interaction_ortho), 3)
```

Let W be the dependent variable (or the dependent variable derived from) `log.mpg.`, A be that of `log.acceleration.`, and AW be the corresponding interaction terms in each model in sub-problem 2(b), we have:

- (ii) $\text{Cor}[W, AW] = 0.924$, $\text{Cor}[A, AW] = 0.764$
- (iii) $\text{Cor}[W, AW] = -0.189$, $\text{Cor}[A, AW] = -0.295$
- (iv) $\text{Cor}[W, AW] = 0$, $\text{Cor}[A, AW] = 0$

■

Question 3. (a) First, I obtain two regression model:

```r
# Question 3 (a)
regr_3ai <- lm(log.weight.~ log.cylinders., data=cars_log)
regr_3aii <- lm(log.mpg.~ log.weight., data=cars_log)
```

```
> summary(regr_3ai)

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

> |
```

```
> summary(regr_3aii)

Call:
lm(formula = log.mpg. ~ log.weight., data = cars_log)

Residuals:
     Min       1Q   Median       3Q      Max
-0.52408 -0.10441 -0.00805  0.10165  0.59384

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.5219     0.2349   49.06   <2e-16 ***
log.weight.  -1.0583     0.0295  -35.87   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.165 on 396 degrees of freedom
Multiple R-squared:  0.7647,    Adjusted R-squared:  0.7641
F-statistic:  1287 on 1 and 396 DF,  p-value: < 2.2e-16

>
```

Hence, the number of cylinders has a significant direct effect on weight. Also, the weight has a significant direct effect on mpg.

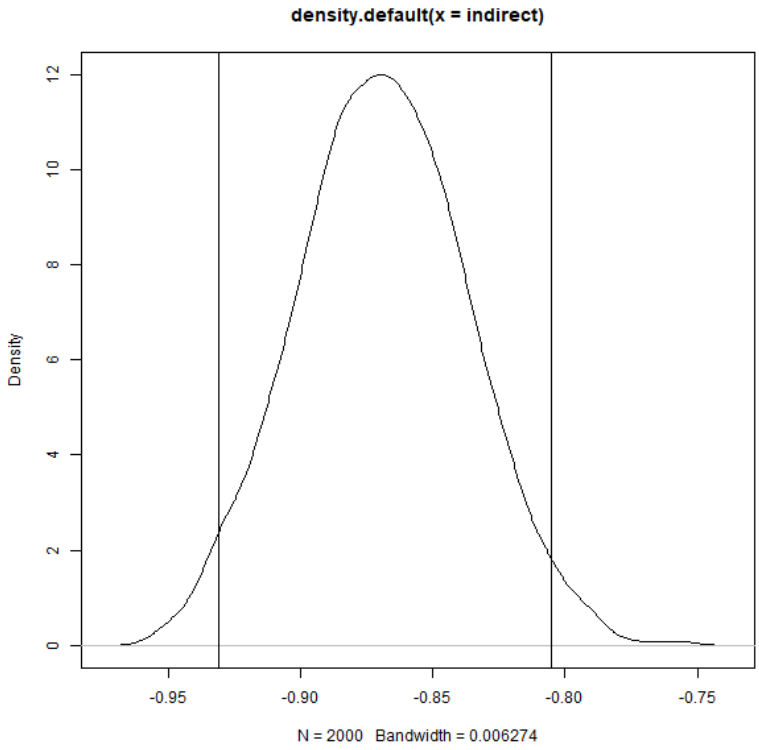
(b) This can be obtained from the command

```
> regr_3ai$coefficients[2] * regr_3aii$coefficients[2]
```

The result is -0.8679111.

(c) The bootstrap and plot code:

```r
# Question 3 (c)
set.seed(42) # Set random seed
boot_mediation <- function(model1, model2, dataset) { # bootstrap
  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])
}

indirect <- replicate(2000, boot_mediation(regr_3ai, regr_3aii, cars_log))
boot_ci <- quantile(indirect, probs=c(0.025, 0.975))

png(filename = "3c.png", width = 600, height = 600) # Subplots
plot(density(indirect))
abline(v=quantile(indirect, probs=c(0.025, 0.975)))
dev.off()
```



The 95% CI is $[-0.931, -0.805]$.

■