BACS HW (Week 12)

108020024

due on 05/07 (Sun)

Question 1) Let's visualize how weight and acceleration are related to mpg.

```
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(weight),log(acceleration),model_year, origin))
```

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 he
```

```
cars_log_light <- subset(cars_log, log.weight. < log(mean(cars$weight)))
cars_log_heavy <- subset(cars_log, log.weight. > log(mean(cars$weight)))
```

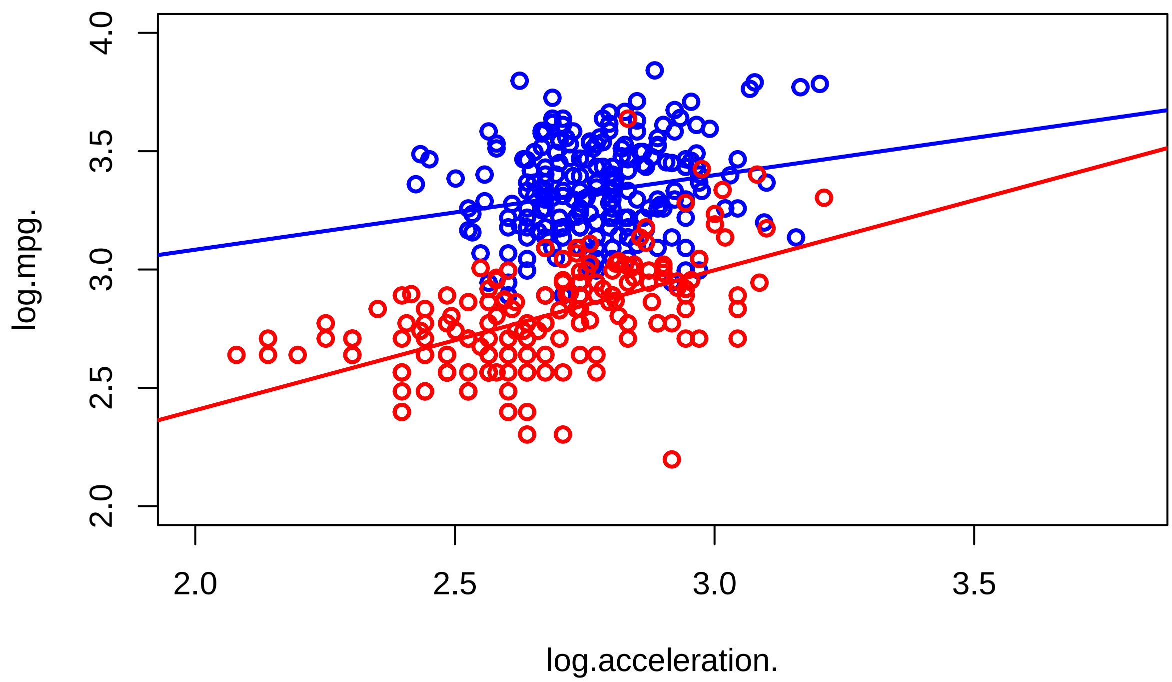
```
ii) Create a single scatter plot of acceleration vs. mpg, with different colors and/or shapes for li
```

```
iii)Draw two slopes of acceleration-vs-mpg over the scatter plot:
one slope for light cars and one slope for heavy cars (distinguish them by appearance)
```

```
cars_log_light_lm <- lm( log.mpg.~log.acceleration. , data=cars_log_light)
cars_log_heavy_lm <- lm( log.mpg.~ log.acceleration., data=cars_log_heavy)


with(cars_log_light,
plot(log.acceleration., log.mpg., col="blue", lwd=2,xlim = c(2,3.8),ylim = c(2,4)))
abline(cars_log_light_lm, col="blue", lwd=2)

with(cars_log_heavy,
points(log.acceleration., log.mpg., col="red", lwd=2))
abline(cars_log_heavy_lm, col="red", lwd=2)
```



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

```
cars_log_light_lm <- lm( log.mpg.~log.weight.+log.acceleration. +model_year+factor(origin), data=cars_lo

cars_log_heavy_lm <- lm( log.mpg.~ log.weight.+log.acceleration. +model_year+factor(origin), data=cars_l

summary(cars_log_light_lm)
```

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

```
summary(cars_log_heavy_lm)
```

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

This are the full summary of two regression model result. log.weight.,model_year,factor(origin)2 are significant under 0.05 significant level.

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

Light cars have a bigger slope on acceleration with mpg.

Question 2) Use the transformed dataset from above (cars_log), to test whether we have moderation.

a) (not graded) 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.

Weight might be a moderating variable I guess.

b) Use various regression models to model the possible moderation on log.mpg.:(use log.weight., log.acceleration., model_year and origin as independent variables)

```
i) Report a regression without any interaction terms
```

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

summary(md)
```

```
##
## 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
```

log.weight.,model_year,factor(origin)2,log.weight. are significant under 0.05 significant level.

```
ii)Report a regression with an interaction between weight and acceleration
```

```
md <- lm( log.mpg.~log.weight.+log.acceleration. +model_year+factor(origin)+log.weight.*log.acceleratio

summary(md)
```

```
##
## 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
```

log.acceleration.,model_year,factor(origin)2,log.weight.,log.acceleration are significant under 0.05 significant level.

```
    iii)Report a regression with a mean-centered interaction term
```

```r
log.mpg._mc <- scale(cars_log$log.mpg., center=TRUE, scale=FALSE)
log.weight._mc <- scale(cars_log$log.weight., center=TRUE, scale=FALSE)
log.acceleration._mc <- scale(cars_log$log.acceleration., center=TRUE, scale=FALSE)
model_year_mc <- scale(cars_log$model_year, center=TRUE, scale=FALSE)



md <- lm( log.mpg._mc~log.weight._mc+log.acceleration._mc +model_year_mc+factor(cars_log$origin)+log.wei

summary(md)
```

```
##
## Call:
## lm(formula = log.mpg._mc ~ log.weight._mc + log.acceleration._mc +
##     model_year_mc + factor(cars_log$origin) + log.weight._mc *
##     log.acceleration._mc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37807 -0.06868  0.00463  0.06891  0.39857
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         -0.022108   0.008442  -2.619  0.00916 **
## log.weight._mc                      -0.880393   0.028585 -30.799  < 2e-16 ***
## log.acceleration._mc                 0.072596   0.037567   1.932  0.05403 .
## model_year_mc                        0.033685   0.001735  19.411  < 2e-16 ***
## factor(cars_log$origin)2             0.058737   0.017789   3.302  0.00105 **
## factor(cars_log$origin)3             0.028179   0.018266   1.543  0.12370
## log.weight._mc:log.acceleration._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
```

log.weight._mc,model_year_mc,factor(cars_log$origin)2, log.weight._mc:log.acceleration._mc are significant under 0.05 significant level.

```
iv)Report a regression with an orthogonalized interaction term
```

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


md <- lm( log.mpg.~log.weight.+log.acceleration. +model_year+factor(origin)+interaction_ortho, data=cars

summary(md)
```

```
##
## 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
```

log.weight.,model_year,factor(origin)2, interaction_ortho are significant.

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?

Mean-Centered Correlation

```r
w_mc = log.weight._mc
acc_mc = log.acceleration._mc
inter = log.weight._mc*log.acceleration._mc

cor( data.frame(w_mc,acc_mc,inter))
```

```
##              w_mc     acc_mc      inter
## w_mc    1.0000000 -0.4256194 -0.2026948
## acc_mc -0.4256194  1.0000000  0.3512271
## inter  -0.2026948  0.3512271  1.0000000
```

Raw Correlation

```r
w_raw = cars_log$log.weight.
acc_raw = cars_log$log.acceleration.
inter = w_raw*acc_raw

cor( data.frame(w_raw,acc_raw,inter))
```

```
##              w_raw    acc_raw     inter
## w_raw    1.0000000 -0.4256194 0.1083055
## acc_raw -0.4256194  1.0000000 0.8528810
## inter    0.1083055  0.8528810 1.0000000
```

orthogonalized Correlation

```r
cor( data.frame(w_raw,acc_raw,interaction_ortho))
```

```
##                           w_raw       acc_raw interaction_ortho
## w_raw              1.000000e+00 -4.256194e-01      2.468461e-17
## acc_raw           -4.256194e-01  1.000000e+00     -6.804111e-17
## interaction_ortho  2.468461e-17 -6.804111e-17      1.000000e+00
```

Question 3)

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

```r
md1 <- lm(log.weight. ~ log.cylinders., data = cars_log)
summary(md1)
```

```
##
## 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
```

The number of cylinders has a significant direct effect on weight

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

```r
md2 <- lm(log.mpg. ~ ., data = cars_log)
summary(md2)
```

```
##
## Call:
## lm(formula = log.mpg. ~ ., data = cars_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41449 -0.06967  0.00040  0.06035  0.39298
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        7.252158   0.363468  19.953  < 2e-16 ***
## log.cylinders.    -0.074879   0.061060  -1.226  0.22083
## log.displacement. -0.008015   0.055532  -0.144  0.88532
## log.horsepower.   -0.296585   0.057548  -5.154 4.09e-07 ***
## log.weight.       -0.554906   0.081716  -6.791 4.26e-11 ***
## log.acceleration. -0.182062   0.059222  -3.074  0.00226 **
## model_year         0.029608   0.001726  17.149  < 2e-16 ***
## origin             0.022419   0.010301   2.176  0.03014 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1132 on 384 degrees of freedom
##   (    6      )
## Multiple R-squared:  0.8912, Adjusted R-squared:  0.8892
## F-statistic: 449.5 on 7 and 384 DF,  p-value: < 2.2e-16
```

We see weight has a significant direct effect on mpg.

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

0.82012 * -0.554906 = -0.4550895

The indirect effect of cylinders on mpg is -0.4550895.

c)Let's bootstrap for the confidence interval of the indirect effect of cylinders on 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(15)
indirect <- replicate(2000, boot_mediation(md1, md2, cars_log))
quantile(indirect, probs=c(0.025, 0.975))
```

```
##        2.5%       97.5%
## -0.16522468  0.03348991
```

The 95% CI of the indirect effect of log.cylinders. on log.mpg. is (-0.16522468, 0.03348991)

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)))
```