Jiayuan Shi
MA684 HW6

1a.

```
> homeprices <- read.csv("homeprices.csv",header=T)
> attach(homeprices)
> cor.test(price,size)

        Pearson's product-moment correlation

data:  price and size
t = 5.5052, df = 23, p-value = 1.344e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5111883 0.8853665
sample estimates:
      cor
0.7540179

> cor.test(price,bedrooms)

        Pearson's product-moment correlation

data:  price and bedrooms
t = 2.8178, df = 23, p-value = 0.009762
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1393426 0.7513232
sample estimates:
      cor
0.5065818

> cor.test(price,age)

        Pearson's product-moment correlation

data:  price and age
t = -2.0947, df = 23, p-value = 0.04741
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.686775119 -0.006094658
sample estimates:
       cor
-0.4002617
```

The correlation between selling price and size of the house is 0.754, and p-value<0.001, so there is a significant association between them. The association is strong positive.
The correlation between selling price and number of bedrooms is 0.5066, and p-value<0.01, so there is a significant association between them. The association is moderate positive.
The correlation between selling price and age of the house is -0.4002, and p-value=0.04741<0.05, so there is a significant association between them. The association is moderate negative.

Size is most strongly correlated with selling price.

1b.

```
> pcor.test(price,bedrooms,size)
   estimate   p.value statistic  n gp  Method
1 0.03388958 0.8736314 0.1590476 25  1 pearson
> pcor.test(price,age,size)
   estimate     p.value statistic  n gp  Method
1 -0.6735942 1.913947e-05   -4.2747 25  1 pearson
```

The correlation between selling price and number of bedrooms, controlling for the size of the house is 0.0339, and p-value=0.8736>0.05, so the association between them is not significant. The association is very weak positive.
The correlation between selling price and age of the house, controlling for the size of the house is -0.6736, and p-value<0.001, so there is a significant association between them. The association is strong negative.

These partial correlations are very different than the correlations reported in 1a, because, taking the number of bedrooms as an example, the partial correlation measures the mutual association between selling price and number of bedrooms, controlling for the size of the house. (So the partial correlation eliminates the association between the size of the house and both selling price and number of bedrooms).

1c.
For age:

$$r_{XY|Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{(1-r_{XZ}^2)(1-r_{YZ}^2)}} = \frac{-0.4002617 - 0.055065 * 0.7540179}{\sqrt{(1-0.055065^2)(1-0.7540179^2)}} = -0.6736$$

The result of my hand calculations match the partial correlation from R in 1b.

1d.

$$(partial\ correlation\ \ r\ (\mathrm{XY|Z}))^2 = \frac{[\mathrm{SSModel(XZ)\text{-}SSModel(Z)}]}{\mathrm{SSResidual(Z)}}$$

$$= \frac{[(11171 + 3846)\text{-}11171]}{8478} = 0.4536447$$

partial r = -0.6735
The partial correlation I calculate matches the partial correlation from 1b.

1A. Model A:

For Lighter Drinkers: the mean birth weight for lighter drinkers is 203.5 grams lower than non-drinkers, controlling for gestational age, child sex, maternal pre-pregnancy weight. The p-value is 0.010 < 0.05, so there is a significant association between lighter alcohol use and birth weight, controlling for gestational age, child sex, maternal pre-pregnancy weight.

For Heavier Drinkers: the mean birth weight for lighter drinkers is 301.2 grams lower than non-drinkers, controlling for gestational age, child sex, maternal pre-pregnancy weight. The p-value is < 0.001, so there is a significant association between heavier alcohol use and birth weight, controlling for gestational age, child sex, maternal pre-pregnancy weight..

1B. Model B:
For Lighter Drinkers:
the mean birth weight for lighter drinkers is 35.3 grams lower than the average mean across all categories, controlling for gestational age, child sex, and pre-pregnancy weight. The p-value is 0.523 > 0.05, so there is not a significant association between lighter alcohol use and birth weight, controlling for gestational age, child sex, and pre-pregnancy weight.

For Heavier Drinkers: the mean birth weight for lighter drinkers is 132.9 grams lower than the average mean across all categories, controlling for gestational age, child sex, and pre-pregnancy weight. The p-value is 0.019 < 0.05, so there is a significant association between heavier alcohol use and birth weight, controlling for gestational age, child sex, and pre-pregnancy weight.

1C.
Model A:

Predicted birth weights = -1559.1+40*107.6+120*2.5 = 3044.9 grams

Model B:
Predicted birth weights = -1727.3+40*107.6+120*2.5-1*(-35.3)-1*(-132.9) = 3044.9 grams
The predicted birth weights between the two versions of the regression model are the same.

1D.
Partial R^2 = 0.26-0.21 = 0.05
Partial F:

$$F_{obs}=\frac{[SSM(Full)- SSM(Reduced)]/m}{MSError(Full)}, \quad m, n-k-1 \text{ df}$$

$$F_{obs}=\frac{[15042453 - 11896782]/2}{176592}=8.91, \quad 2 \text{ and } 241 \text{ df}$$

p-value<0.001
H0: $R^2_{Reduced} = R^2_{Full,}$ H1: H0 is not true
So we can reject the null hypothesis, which is that the Full model explains no more variability than the Reduced model.
Hence, we can conclude birth weights significantly differ for the contribution of the two alcohol variables, controlling for gestational age, child sex, and pre-pregnancy weight.

2a.

```
> reg <- lm(enviroscore1~enviroscore0+SexM+intervention)
> summary(reg)

Call:
lm(formula = enviroscore1 ~ enviroscore0 + SexM + intervention)

Residuals:
    Min      1Q  Median      3Q     Max
-19.3055  -4.6260   0.6834   4.1505  13.1851

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.00385    2.36617   2.115    0.037 *
enviroscore0  0.89399    0.04037  22.147  < 2e-16 ***
SexM         -0.49059    1.33172  -0.368    0.713
intervention  6.66251    1.34209   4.964 2.99e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.649 on 96 degrees of freedom
Multiple R-squared:  0.8514,      Adjusted R-squared:  0.8467
F-statistic: 183.3 on 3 and 96 DF,  p-value: < 2.2e-16
```

The intervention significantly increases the follow-up environmental awareness score, because the p-value for intervention is less than 0.001.

Slope= 6.66251.
If the baseline environmental awareness score and sex are held constant, the environmental awareness score for those randomized to intervention expected to be 6.66251 higher, on average, than the score for those randomized to control.

2b.

```
> reg0i <- lm(enviroscore1~enviroscore0+SexM+intervention+enviroscore0*intervention)
> summary(reg0i)

Call:
lm(formula = enviroscore1 ~ enviroscore0 + SexM + intervention +
    enviroscore0 * intervention)

Residuals:
    Min      1Q  Median      3Q     Max
-17.5498  -3.9526   0.5874   3.2993  14.0299

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           -5.68050    2.67146  -2.126   0.0361 *
enviroscore0           1.09733    0.04791  22.903  < 2e-16 ***
SexM                   0.42031    1.14471   0.367   0.7143
intervention          28.79203    3.80523   7.566 2.44e-11 ***
enviroscore0:intervention -0.42336    0.06943  -6.098 2.31e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.667 on 95 degrees of freedom
Multiple R-squared:  0.8932,      Adjusted R-squared:  0.8887
F-statistic: 198.6 on 4 and 95 DF,  p-value: < 2.2e-16
```

i) If there is an interaction between the baseline score and intervention, it means the effect of the baseline environmental awareness score on the follow-up score may differs depending on the level of the intervention.

ii) The interaction term is significant in this regression.
t-value=-6.098, p-value=2.31e-08<0.001

iii) That depends.
Follow-up environmental awareness score = -5.68050 +1.09733*20 +0.42031*(SexM) +28.79203*(intervention)-0.42336*(20* (intervention))= 16.27+0.42*(SexM)+ 20.32*(intervention)
For students with a baseline environmental awareness score of 20, the environmental awareness score for those randomized to intervention expected to be 20.32 higher, on average, than the score for those randomized to control.

Follow-up environmental awareness score = -5.68050 +1.09733*70 +0.42031*(SexM) +28.79203*(intervention)-0.42336*(70* intervention)= 71.13 +0.42*(SexM)- 0.84*(intervention)
For students with a baseline environmental awareness score of 70, the environmental awareness score for those randomized to intervention expected to be 0.84 lower, on average, than the score for those randomized to control.

2c.

```
> regSi <- lm(enviroscore1~enviroscore0+SexM+intervention+SexM*intervention)
```

```
> summary(regSi)

Call:
lm(formula = enviroscore1 ~ enviroscore0 + SexM + intervention +
    SexM * intervention)

Residuals:
    Min      1Q  Median      3Q     Max
-19.1584  -4.6949   0.5691   4.2752  13.0664

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.21525    2.58304   2.019  0.04630 *
enviroscore0      0.89284    0.04094  21.811  < 2e-16 ***
SexM             -0.79177    1.96359  -0.403  0.68769
intervention      6.37285    1.93099   3.300  0.00136 **
SexM:intervention  0.56693    2.70452   0.210  0.83441
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.683 on 95 degrees of freedom
Multiple R-squared:  0.8515,      Adjusted R-squared:  0.8452
F-statistic: 136.1 on 4 and 95 DF,  p-value: < 2.2e-16
```

For females, SexM=0 and so SexM*intervention=0. The regression equation becomes,
Follow-up environmental awareness score = 5.21 +0.89*(enviroscore0)- 0.79*0 +6.37*(intervention)+0.57*0 = 5.21+0.89*(enviroscore0) + 6.37*(intervention)
=> For females, the environmental awareness score for those randomized to intervention expected to be 6.37 higher, on average, than the score for those randomized to control.

For males, SexM=1 and so SexM*intervention= intervention. The regression equation becomes,
Follow-up environmental awareness score = 5.21 +0.89*(enviroscore0)- 0.79*1 +6.37*(intervention)+0.57*(intervention)= 5.21+0.89*(enviroscore0) + 6.94*(intervention)
=> For males, the environmental awareness score for those randomized to intervention expected to be 6.94 higher, on average, than the score for those randomized to control.

Therefore, the intervention is not more effective for females than for males.

3.
a. without interaction
Based on the Type II tests, body weight significantly relates to systolic blood pressure, after controlling for age and sex, because the p-value for the BMI variables is less than 0.001.

The mean systolic blood pressure for overweight people is 6.7379 units higher than normal weight people, controlling for age and sex.
The mean systolic blood pressure for obese people is 19.0984 units higher than normal weight people, controlling for age and sex.

Age significantly relates to systolic blood pressure, because the p-value for the age variable is less than 0.001.
If the sex and BMI variables are held constant, for each year increase in age, on average, systolic blood pressure will increase in 0.9299 units.

b. with interaction

Based on the Type II tests, the effect of age does not differ by BMI category, because the p-value of the interaction of age and BMI variables is 0.99486, larger than 0.05. So we fail to reject the null hypothesis, which is that the interaction term is not significant.

For the normal weight group:
Systolic blood pressure = 84.70262 + 0.91006*(age) - 5.64157*(sexmale)
So for the normal weight group, if the sex is held constant, for each year increase in age, on average, systolic blood pressure will increase in 0.91 units.

For the overweight group:

Systolic blood pressure = 84.70262 + 0.91006*(age) - 5.64157*(sexmale) + 5.43826 + 0.02659*(age) = 90.14088+ 0.93665*(age) - 5.64157*(sexmale)
So for the overweight group, if the sex is held constant, for each year increase in age, on average, systolic blood pressure will increase in 0.9367 units.

For the obese group:
Systolic blood pressure = 84.70262 + 0.91006*(age) - 5.64157*(sexmale) + 16.83926 + 0.04558*(age) = 101.5419+ 0.95564*(age) - 5.64157*(sexmale)
So for the obese group, if the sex is held constant, for each year increase in age, on average, systolic blood pressure will increase in 0.9556 units.