S631 HW4

Shibi He

1. ALR Problem 2.16

2.16.1 Simple linear regression for log(fertility) and log(ppgdp)

```
library(alr4)
```

```
data(UN11)
reg1 <- lm(log(fertility) ~ log(ppgdp), data=UN11)
summary(reg1)
```

```
##
## Call:
## lm(formula = log(fertility) ~ log(ppgdp), data = UN11)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79828 -0.21639  0.02669  0.23424  0.95596
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.66551    0.12057   22.11   <2e-16 ***
## log(ppgdp)  -0.20715    0.01401  -14.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3071 on 197 degrees of freedom
## Multiple R-squared:  0.526,  Adjusted R-squared:  0.5236
## F-statistic: 218.6 on 1 and 197 DF,  p-value: < 2.2e-16
```

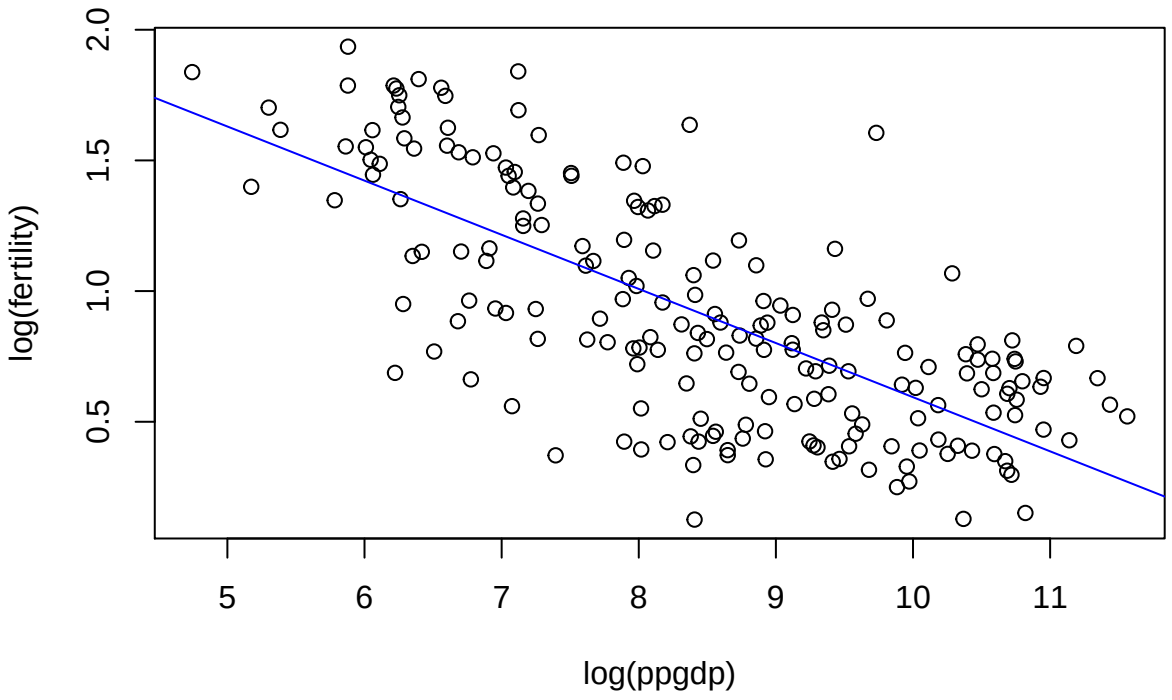
The simple linear regression model is:

$$log(fertility) = 2.6655 - 0.2072 * log(ppgdp)$$

.

2.16.2 Draw a graph and add the fitted line

```
plot(log(fertility) ~ log(ppgdp), UN11)
abline(reg1, col="blue")
```



2.16.3 Hypothesis test:

$H_0 : \beta_1 \geq 0$ $H_1 : \beta_1 < 0$

```
n = length(UN11$fertility)
t = (-0.20715 - 0)/0.01401
pvalue = pt(q=t, df = n-2)
pvalue
```

```
## [1] 4.506246e-34
```

The pvalue=4.506246e-34, which is much smaller than the significance level $\alpha = 0.05$, therefore, we reject the null hypothesis and conclude that the true slope is negative.

2.16.4 Interpret the coefficient of determination

The coefficient of determination $R^2 = 0.526$, suggesting that 52.6% of the variation in the log(fertility) is explained by log(ppgdp).

2.16.5 Predictions

Given ppdgp=1000, the point estimate for log(fertility) is

$$2.6655 - 0.2072 * log(1000) = 1.23$$

.

```
predict(object = reg1,
        newdata = data.frame(ppgdp = 1000),
        interval = "prediction", level=0.95)
```

```
##        fit       lwr      upr
## 1 1.234567 0.6258791 1.843256
```

The 95% prediction interval for log(fertility) is (0.6258791, 1.843256).

The 95% prediction intervale for fertitlity is (1.869889, 6.317073).

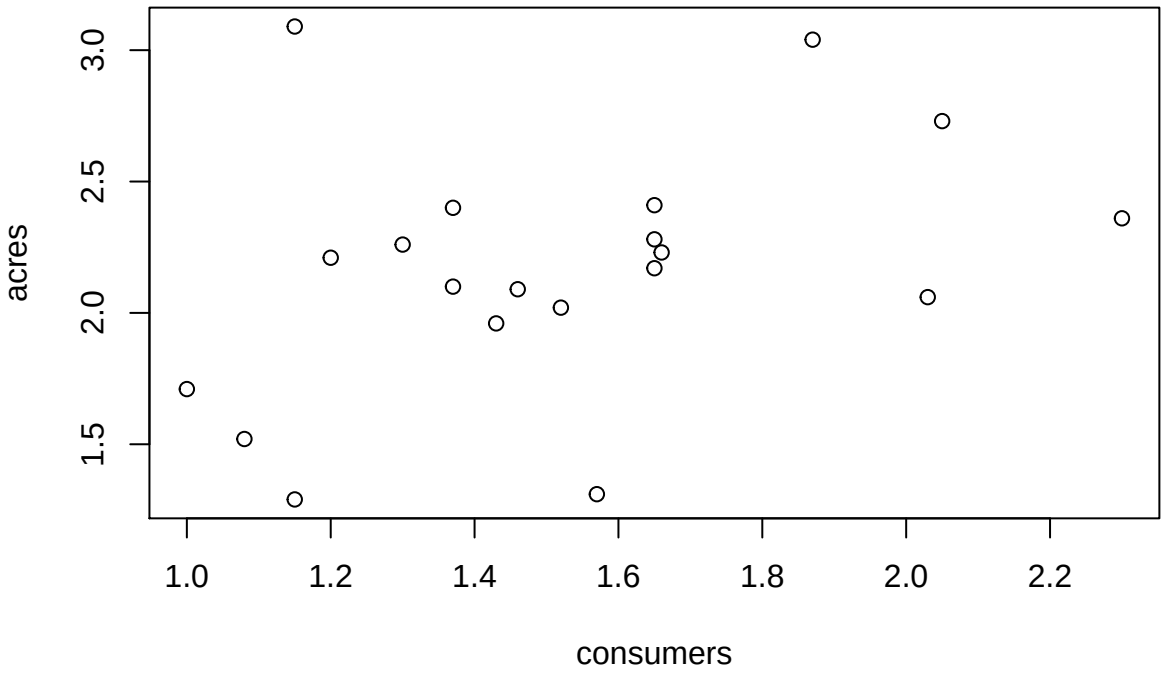
2.16.6 Residuals

```
UN11$residual <- resid(reg1)
```

The locality with the highest fertility is Nigera (fertility=6.925), the locality with the lowest fertility is Bosnia and Herzegovina (fertility=1.134). The two localities with the largest positive residuals when both variables are in log scale are Equatorial Guinea and Angola. The two countries with the largest negative residuals in log scales are Bosnia and Herzegovina and Moldova.

2(a). Draw a scatterplot

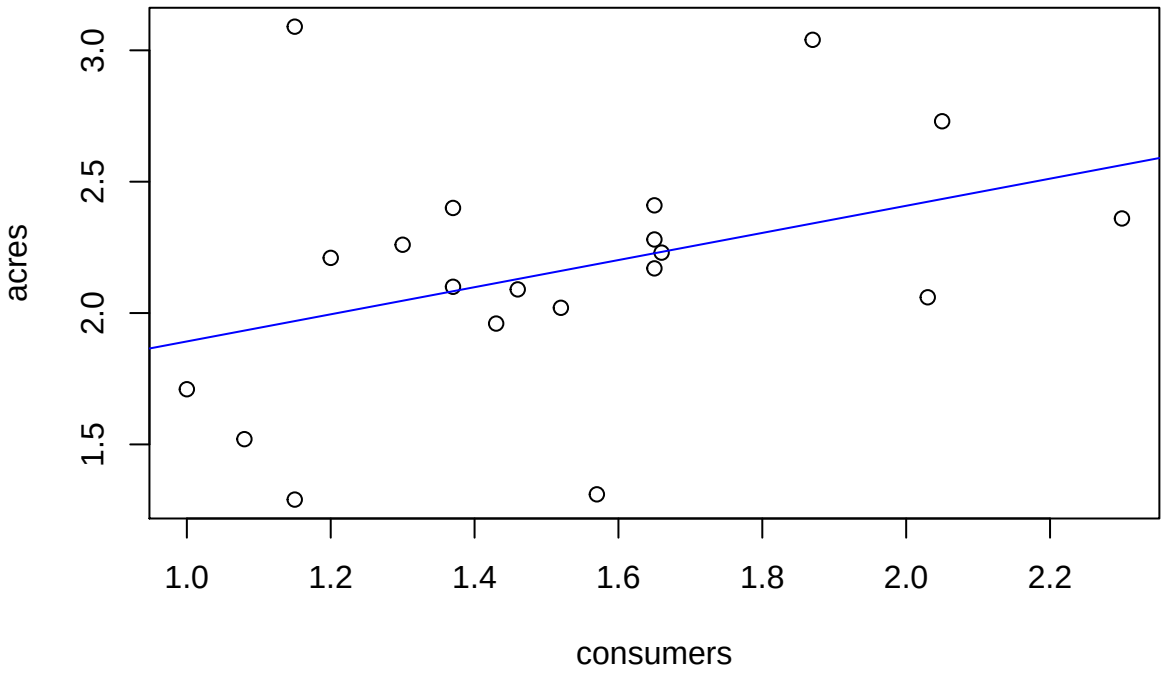
```
sahlins <- read.delim("sahlins.txt", sep=" ")
plot(acres ~ consumers, sahlins)
```



There seems to be a positive linear relationship between consumers/gardener and acres/gardener, however the relationship is rather weak. There are some outliers in the data set. For example, there is a household with the highest labor intensity (Y=3.09), but a relatively low consumption needs (X=1.15).

2(b). Add a linear regression model:

```
plot(acres ~ consumers, sahlins)
m1 <- lm(acres ~ consumers, data=sahlins)
abline(m1, col= "blue")
```




```
summary(m1)$coef
```

```
##              Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 1.3756445  0.4684047 2.936872 0.008813794
## consumers   0.5163201  0.3002335 1.719728 0.102629261
```

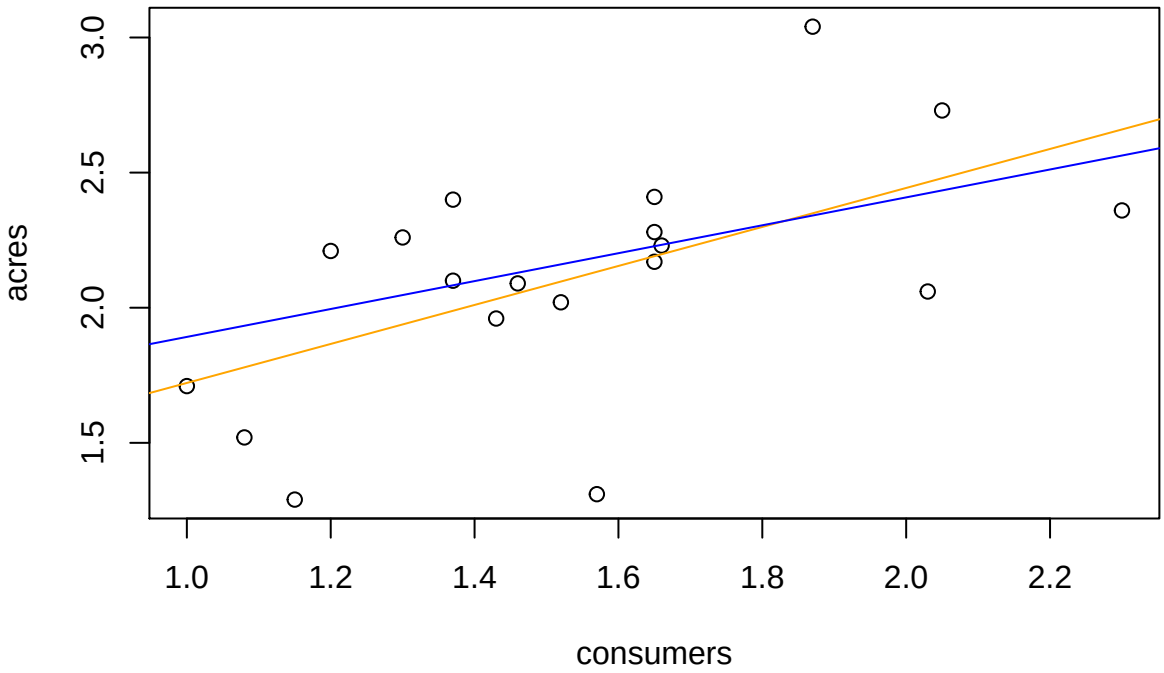
```
Y <- sahlins$acres
X <- cbind(1, sahlins$consumers)
betas.h = qr.solve(t(X)%*%X)%*%t(X)%*%Y
n = length(Y)
p = length(betas.h) - 1
sigma.h.sq = t(Y - X %*% betas.h) %*% (Y - X %*% betas.h)/(n-(p+1))
sigma.h.sq = as.numeric(sigma.h.sq)
sigma.h = sqrt(sigma.h.sq)
sigma.h.sq
```

```
## [1] 0.2064047
```

The regression has an intercept $\hat{\beta}_0 = 1.3756$, meaning that when there is no consumption needs (i.e. X=0), each gardener is expected to cultivate 1.3756 acres of land. The slope $\hat{\beta}_1 = 0.5163$, suggesting that as the consumption needs increase by 1 unit, the gardeners are expected to cultivate 0.5163 more acres of land. Since the regression has a positive slope, it suggests that this village is more likely to redistribute through the market, where each household should work in proportion to its comsumption needs. The estimated variance $\hat{\sigma} = 0.2064$.

Add a linear regression model without the fourth household:

```
new_sahlins <- sahlins[-4, ]
plot(acres ~ consumers, new_sahlins)
m2 <- lm(acres ~ consumers, data=new_sahlins)
abline(m2, col= "orange")
abline(m1, col="blue")
```




```
summary(m2)$coef
```

```
##              Estimate Std. Error  t value   Pr(>|t|)
## (Intercept) 1.0000040  0.3969254 2.519375 0.02205468
## consumers   0.7215941  0.2514140 2.870143 0.01061266
```

The orange line shows the regression line after removing the fourth household. We still have positive intercept and slope, specifically, $\hat{\beta}_0 = 1$ and $\hat{\beta}_1 = 0.7216$. There are still many dots are far away from the fitted line, so neither of this regression does a good job of summarizing the relationship between Acres/gardener and Consumers/gardener.

2(c). Obtain confidence intervals and perform hypothesis tests

Standard errors of intercept and slope

```
var.h.betas.h = sigma.h.sq * qr.solve(t(X) %*% X) #the covariance matrix of betahat
se.betas.h = sqrt(diag(var.h.betas.h)) # The standard errors
se.betas.h
```

```
## [1] 0.4684047 0.3002335
```

95% confidence intervals of intercept and slope

```
beta0.h <- betas.h[1]
se.beta0.h <- se.betas.h[1]
beta0.h + c(-1,1)*qt(0.975,n-(p+1))*se.beta0.h
```

```
## [1] 0.3915628 2.3597263
```

```
#confint(m1, level = 0.95)
```

```
beta1.h <- betas.h[2]
se.beta1.h <- se.betas.h[2]
beta1.h + c(-1,1)*qt(.975, n-(p+1))*se.beta1.h
```

```
## [1] -0.1144471  1.1470872
```

Hypothesis tests:

$H_0 : \beta_0 \leq 0 \qquad H_1 : \beta_0 > 0$

```
t=(beta0.h-0)/se.beta0.h
pvalue= (1-pt(q=abs(t), df=n-(p+1)))
c(beta0.h, se.beta0.h, t, pvalue)
```

```
## [1] 1.375644548 0.468404675 2.936871947 0.004406897
```

$H_0 : \beta_1 \leq 0 \qquad H_1 : \beta_1 > 0$

```
t = (beta1.h - 0)/se.beta1.h
pvalue =  (1-pt(q = abs(t), df = n - (p+1)))

c(beta1.h, se.beta1.h, t, pvalue)
```

```
## [1] 0.51632006 0.30023350 1.71972834 0.05131463
```

The standard errors of intercept and slope are 0.4684 and 0.3002, respectively. The 95% confidence intervals of intercept and slope are (0.3915628, 2.3597263) and (-0.1144471, 1.1470872), respectively. The pvalue for intercept is $0.0044 < \alpha = 0.05$, so we reject the null hypothesis and conclude that the population intercept is greater than zero. The pvalue for slope is $0.0513 > \alpha = 0.05$, so we fail to reject the null hypothesis. That is, there is no evidence to say that the population slope is greater than zero.

Excluding the fourth household

```
Y2 <- new_sahlins$acres
X2 <- cbind(1, new_sahlins$consumers)
betas.h2 = qr.solve(t(X2)%*%X2)%*%t(X2)%*%Y2
n2 = length(Y2)
p2 = length(betas.h2) - 1
sigma.h.sq2 = t(Y2 - X2 %*% betas.h2) %*% (Y2 - X2 %*% betas.h2)/(n2-(p2+1))
sigma.h.sq2 = as.numeric(sigma.h.sq2)
var.h.betas.h2 = sigma.h.sq2 * qr.solve(t(X2) %*% X2)
se.betas.h2 = sqrt(diag(var.h.betas.h2))
```

95% confidence intervals

```
beta0.h2 <- betas.h2[1]
se.beta0.h2 <- se.betas.h2[1]
beta0.h2 + c(-1,1)*qt(0.975,n2-(p2+1))*se.beta0.h2
```

```
## [1] 0.1625647 1.8374433
```

```
#confint(m2, level = 0.95)
```

```
beta1.h2 <- betas.h2[2]
se.beta1.h2 <- se.betas.h2[2]
beta1.h2 + c(-1,1)*qt(.975, n2-(p2+1))*se.beta1.h2
```

```
## [1] 0.191157 1.252031
```

Hypothesis tests

$H_0 : \beta_0 \leq 0$ $H_1 : \beta_0 > 0$

```
t2=(beta0.h2-0)/se.beta0.h2
pvalue2= (1-pt(q=abs(t2), df=n2-(p2+1)))
c(beta0.h2, se.beta0.h2, t2, pvalue2)
```

```
## [1] 1.00000398 0.39692536 2.51937538 0.01102734
```

$H_0 : \beta_1 \leq 0$ $H_1 : \beta_1 > 0$

```
t2 = (beta1.h2 - 0)/se.beta1.h2
pvalue2 =  (1-pt(q = abs(t2), df = n2 - (p2+1)))

c(beta1.h2, se.beta1.h2, t2, pvalue2)
```

```
## [1] 0.721594146 0.251413983 2.870143246 0.005306328
```

Excluding the fourth household, the results of the hypothesis tests changed. The pvalue for intercept is $0.011 < \alpha = 0.05$, so we reject the null hypothesis and conclude that the population intercept is greater than zero. The pvalue for slope is $0.0053 < \alpha = 0.05$, so we also reject the null hypothesis and conclude that the population slope is greater than zero.

2(d). Predictions

Let x=1.5, what is the interval prediction for y?

```
newx = c(1, 1.5)
newy.h = as.numeric(newx%*%betas.h)
newy.h
```

```
## [1] 2.150125
```

```
se.newy.h = sigma.h * sqrt(1 + t(newx)%*%qr.solve(t(X) %*% X)%*%newx)
newy.h + c(-1,1)*qt(.99, n-(p+1)) * as.numeric(se.newy.h)
```

```
## [1] 0.961766 3.338483
```

```
# predict(object = m1,
#         newdata = data.frame(consumers = 1.5),
#         interval = "prediction", level=0.98)
```

I expected the Acres/Gardener ratio to be 2.15 for a household with a Consumer/Gardener ratio equal to 1.5. The prediction interval with a 98% confidence level for the Acres/gardener ratio is (0.961766, 3.338483).

The confidence interval for the mean Acres/Gardener ratio for all households with a Consumers/Gardener ratio equal to 1.5:

```
newx = c(1, 1.5) #Need 1 for the intercept
newy.h = as.numeric(newx%*%betas.h)

se.fit = sigma.h * sqrt(t(newx)%*%solve(t(X) %*% X)%*%newx)
newy.h + c(-1,1)*qt(.99, n-(p+1)) * as.numeric(se.fit)
```

```
## [1] 1.890234 2.410016
```

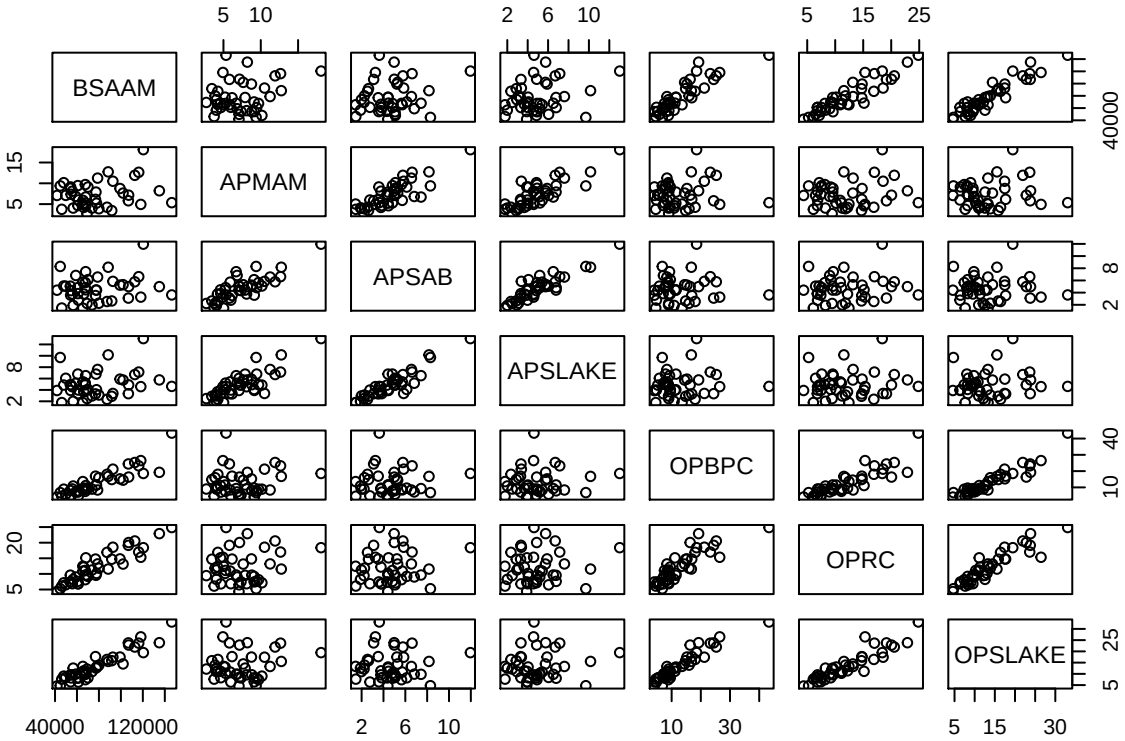
```
# predict(object = m1,
#         newdata = data.frame(consumers = 1.5),
#         interval = "confidence", level=0.98)
```

The estimated mean Acres/Gardener ratio is also 2.15 given the Consumer/Gardener ratio equals to 1.5. However, the 98% confidence interval for the mean Acres/Gardener ratio is (1.890234, 2.410016) for all households with a Consumers/Gardener ratio equal to 1.5. The confidence interval of the mean Acres/Gardener ratio differs from the prediction interval of the new value of Acres/Gardener ratio because their standeard errors are different. More precisely, the new value of Acres/Gardener ratio has an additional source of uncertainty coming from the error term, while the mean Acres/Gardener ratio is devoid of the error term.

3. ALR Problem 1.5 and 3.6

ALR 1.5 Draw the scatterplot matrix

```
data(water)
pairs(~ BSAAM + APMAM + APSAB + APSLAKE + OPBPC + OPRC + OPSLAKE , data=water)
```

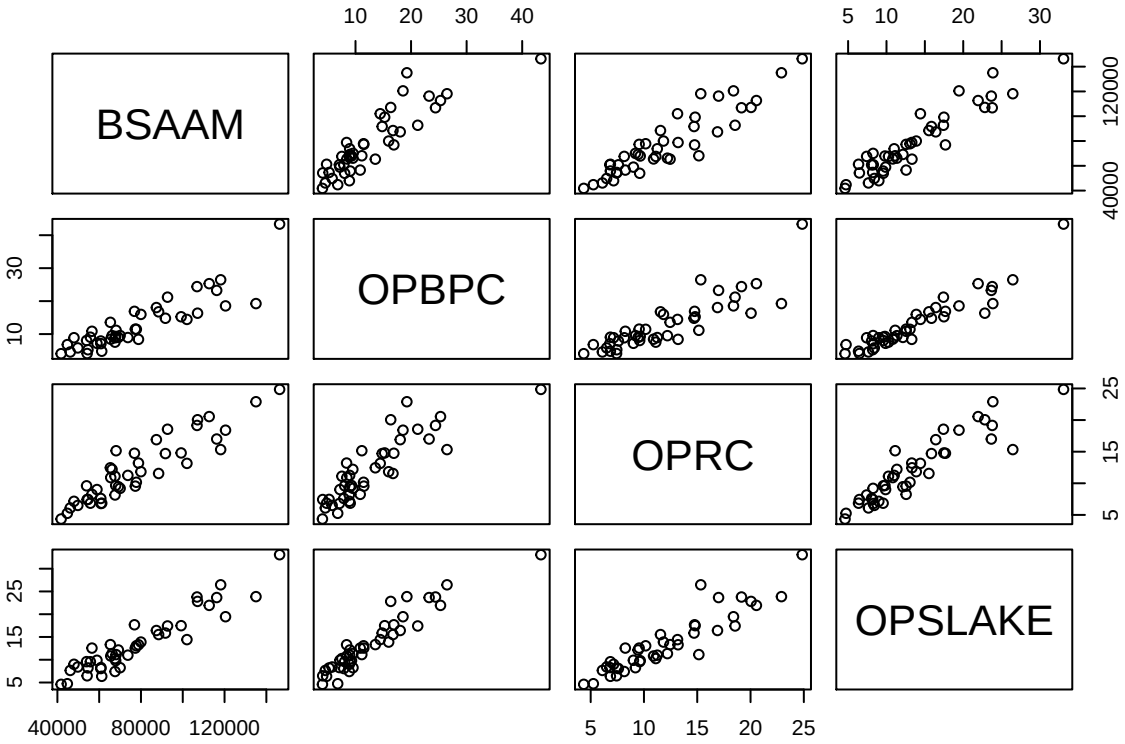


The scatterplot matrix shows that there is a positive linear relationship between BSAAM and OPBPC,

OPRC, and OPSLAKE, repectively. The relationship between APASAB and APSLAKE, between OPRC and OPSLAKE, and between OPBPC and OPSLAKE also seem to be positive and linear.

ALR 3.6.1 Compute the correlation matrix

```
pairs(~ BSAAM + OPBPC + OPRC + OPSLAKE, data=water)
```



According to the above scatterplot matrix, I expect the correlations between BSAAM and OPSLAKE, between BSAAM and OPRC, between OPBPC and OPSLAKE, and between OPRC and OPSLAKE to be large and positive. The correlations between BSAAM and OPBPC and between OPBPC and OPRC also seem to positive, but relatively small.

```
water_sub <- water[5:8]
cor(water_sub)
```

```
##              OPBPC      OPRC   OPSLAKE     BSAAM
## OPBPC   1.0000000 0.8647073 0.9433474 0.8857478
## OPRC    0.8647073 1.0000000 0.9191447 0.9196270
## OPSLAKE 0.9433474 0.9191447 1.0000000 0.9384360
## BSAAM   0.8857478 0.9196270 0.9384360 1.0000000
```

The correlation matrix verifies my conjecture. The correlations between each pair of these variables are large and positive. The correlations between BSAAM and OPBPC and between OPBPC and OPRC are slightly smaller.

ALR 3.6.2 Regession summary

```
reg2 <- lm(BSAAM ~ OPBPC + OPRC + OPSLAKE, data=water)
summary(reg2)
```

```
##
## Call:
## lm(formula = BSAAM ~ OPBPC + OPRC + OPSLAKE, data = water)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15964.1  -6491.8   -404.4   4741.9  19921.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22991.85    3545.32   6.485  1.1e-07 ***
## OPBPC          40.61     502.40   0.081  0.93599
## OPRC         1867.46     647.04   2.886  0.00633 **
## OPSLAKE      2353.96     771.71   3.050  0.00410 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8304 on 39 degrees of freedom
## Multiple R-squared:  0.9017, Adjusted R-squared:  0.8941
## F-statistic: 119.2 on 3 and 39 DF,  p-value: < 2.2e-16
```

The t-value columns shows the t-test statistics for the hypothesis tests that the slopes are zero versus the alternative that the slopes are not zero. The pvalue for the coefficent of OPBPC is 0.936, which is greater than the significance level of 0.05, so we fail to reject the null hypothesis that the true coefficient of OPBPC is zero. That is, the variable OPBPC does not have any impact on the response variable BSAAM. The pvalues for the coefficient of OPRC and OPSLAKE are 0.006 and 0.004, respectively, which are smaller than the significance level of 0.05, suggesting that we can reject the null hypothesis and conclude that the true coefficients of OPRC and OPSLAKE are significantly different from zero. That is, OPRC and POSLAKE have some effects on the response variable BAASM.

ALR 3.6.3 Construct and interpret a 96% confidence interval for a new value for BSAAM.

```
predict(object = reg2,
        newdata = data.frame(OPBPC=5.8, OPRC=6.5, OPSLAKE=16),
        interval = "prediction", level=0.96)
```

```
##        fit      lwr      upr
## 1 73029.15 49905.15 96153.16
```

The 96% confidence interval for a new value of BSAAM is (49905.15, 96153.16). The intepretation is that we are 96% confident that the true BSAAM is between 49905.15 and 96153.16 given the predictors values given above.

ALR 3.6.4 Construct and interpret a 96% confidence interval for the expected value of BSAAM.

```
predict(object = reg2,
        newdata = data.frame(OPBPC=5.8, OPRC=6.5, OPSLAKE=16),
        interval = "confidence", level=0.96)
```

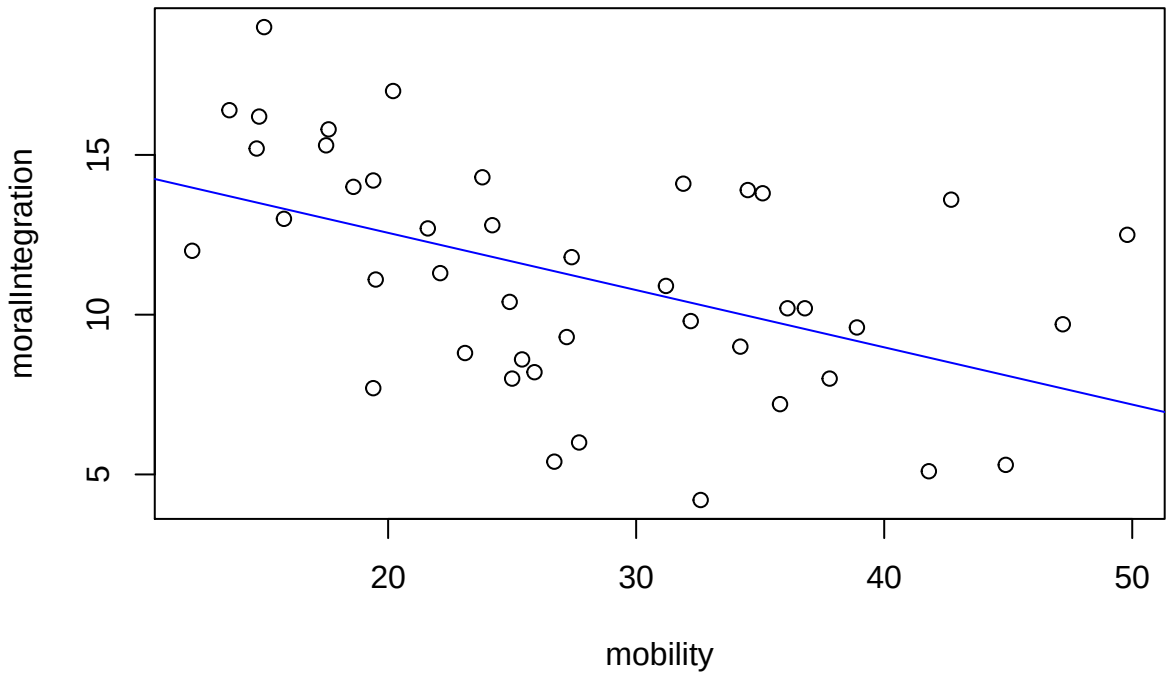
```
##        fit      lwr      upr
## 1 73029.15 58082.93 87975.38
```

The 96% confidence interval for the expected value of BSAAM is (58082.93, 87975.38). The intepretation is that we are 96% confident that the expected value of BSAAM is between 58082.93 and 87975.38 given the

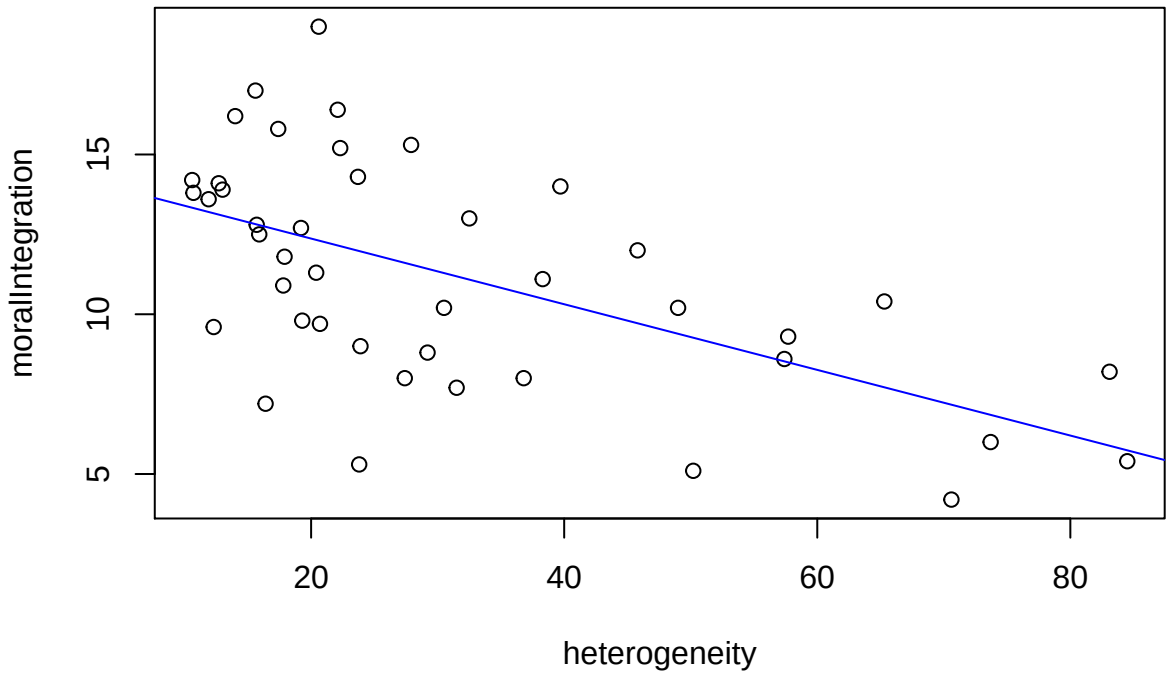
predictors values given above.

4(a). Draw scatterplots and least square lines to show relationship of the response to each predictor:

```
Angell <- read.delim("Angell.txt", header=T, sep="")
plot(moralIntegration ~ mobility, Angell)
m3 <- lm(moralIntegration ~ mobility, data=Angell)
abline(m3, col= "blue")
```




```
plot(moralIntegration ~ heterogeneity, Angell)
m4 <- lm(moralIntegration ~ heterogeneity, data=Angell)
abline(m4, col= "blue")
```



The least-square lines reasonbally summarize the relationship between moralIntegration and the two predictors: mobility and heterogeneity. The relationships are both negative, while the relationship between moralIntegration and mobility seems to be relatively weak.

4(b). Simple linear regression for moralIntegration and heterogeneity.

```
m4 <- lm(moralIntegration ~ heterogeneity, data=Angell)
summary(m4)
```

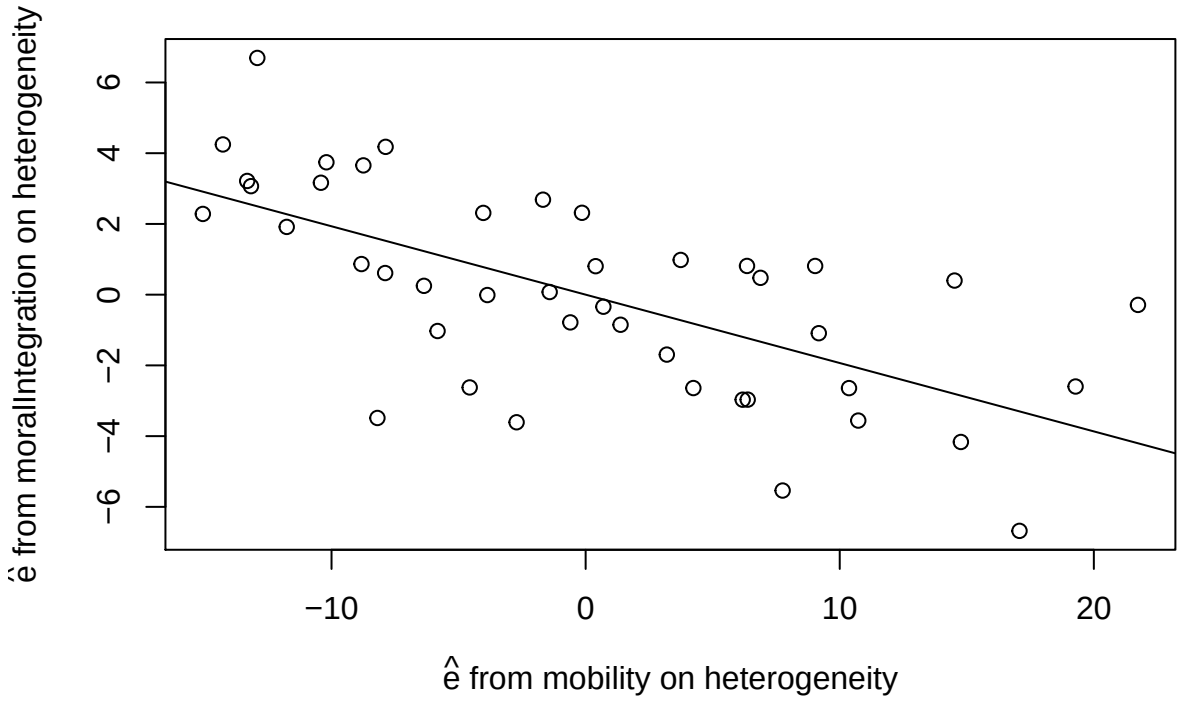
```
##
## Call:
## lm(formula = moralIntegration ~ heterogeneity, data = Angell)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6780 -2.6099  0.2493  2.2971  6.6931
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   14.42355    0.82507  17.482  < 2e-16 ***
## heterogeneity -0.10275    0.02212  -4.645 3.49e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.926 on 41 degrees of freedom
## Multiple R-squared:  0.3448, Adjusted R-squared:  0.3288
## F-statistic: 21.58 on 1 and 41 DF,  p-value: 3.486e-05
```

The estimated intercept $\hat{\beta}_0 = 14.42$, suggesting that moralIntegration is expected to be 14.4236 when the heterogeneity equals zero. The estimated slope $\hat{\beta}_1 = -0.1028$, suggesting that as the heterogeneity increases by 1 unit, the moralIntegration is expected to decrease by 0.1028. The adjusted coefficient of determination $R^2 = 0.3448$, implying that 34.48% of the variation in the moralIntegration can be explained by heterogeneity.

4(c). Multiple linear regression for moralIntegration, heterogeneity, and mobility.

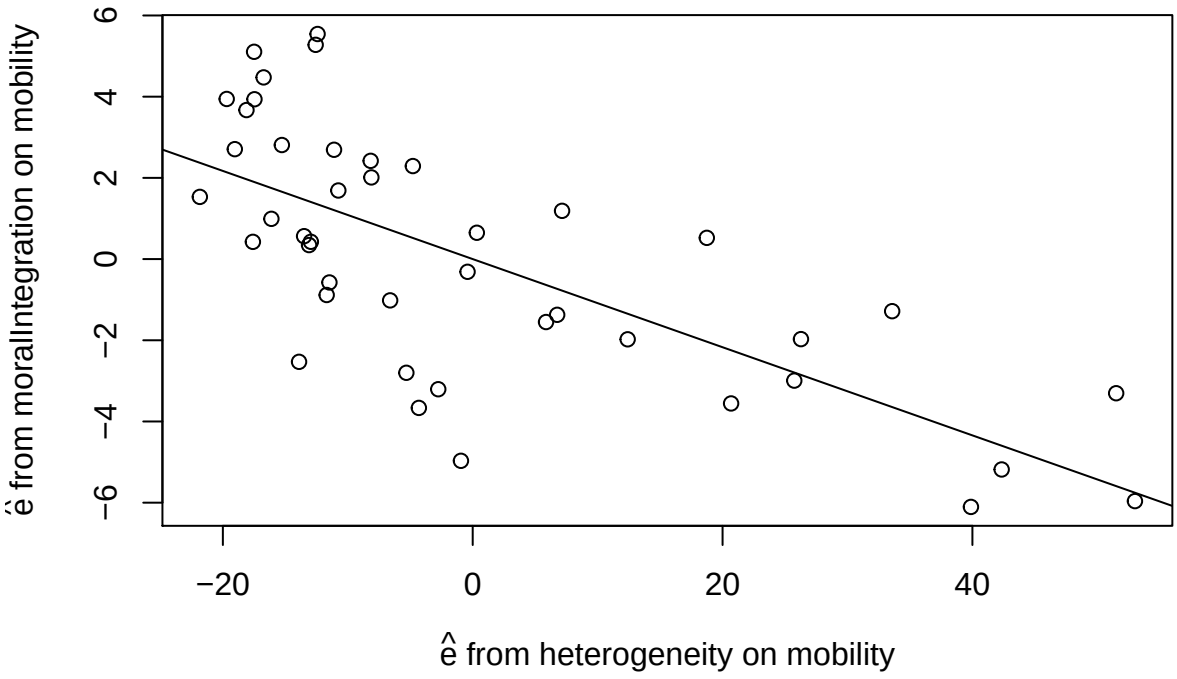
Added-variable plots for mobility:

```
r1 <- residuals(lm(moralIntegration ~ heterogeneity, Angell))
r2 <- residuals(lm(mobility ~ heterogeneity, Angell))
av1 <- lm(r1 ~ r2)
plot(r1 ~ r2,
xlab=expression(paste(hat(e), " from mobility on heterogeneity")),
ylab=expression(paste(hat(e), " from moralIntegration on heterogeneity")))
abline(av1)
```



Added-variable plots for heterogeneity:

```
r3 <- residuals(lm(moralIntegration ~ mobility, Angell))
r4 <- residuals(lm(heterogeneity ~ mobility, Angell))
av2 <- lm(r3 ~ r4)
plot(r3 ~ r4,
xlab=expression(paste(hat(e), " from heterogeneity on mobility")),
ylab=expression(paste(hat(e), " from moralIntegration on mobility")))
abline(av2)
```



Multiple linear regression with both predictors:

```
m5 <- lm(moralIntegration ~ heterogeneity+mobility, data=Angell)
summary(m5)
```

```
##
## Call:
## lm(formula = moralIntegration ~ heterogeneity + mobility, data = Angell)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.071 -1.194 -0.206  1.738  4.195
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   19.94076    1.19265  16.720  < 2e-16 ***
## heterogeneity -0.10856    0.01699  -6.389 1.34e-07 ***
## mobility      -0.19331    0.03543  -5.456 2.74e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.243 on 40 degrees of freedom
## Multiple R-squared:  0.6244, Adjusted R-squared:  0.6056
## F-statistic: 33.25 on 2 and 40 DF,  p-value: 3.126e-09
```

The estimated intercept $\hat{\beta}_0 = 19.9408$, suggesting that the moralIntegration is expected to be 19.9408 when both heterogeneity and mobility equal to zero. The estimated $\hat{\beta}_1 = -0.1086$, meaning that as the heterogeneity increases by 1 unit, the moralIntegration is expected to decrease by 0.1086 unit, with mobility held fixed. The estimated $\hat{\beta}_2 = -0.1933$, suggesting that as the mobility increases by 1 unit, the moralIntegration is expected to decrease by 0.1933 unit, with heterogeneity held fixed. The coefficient of determination $R^2 = 0.6244$, meaning that 62.44% of the variation in moralIntegration can be explained by mobility and heterogeneity.

4(d). Interpret coefficient for heterogeneity

```
m5 <- lm(moralIntegration ~ heterogeneity+mobility, data=Angell)
summary(m5)
```

```
##
## Call:
## lm(formula = moralIntegration ~ heterogeneity + mobility, data = Angell)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.071 -1.194 -0.206  1.738  4.195
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   19.94076    1.19265  16.720  < 2e-16 ***
## heterogeneity -0.10856    0.01699  -6.389 1.34e-07 ***
## mobility      -0.19331    0.03543  -5.456 2.74e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.243 on 40 degrees of freedom
## Multiple R-squared:  0.6244, Adjusted R-squared:  0.6056
## F-statistic: 33.25 on 2 and 40 DF,  p-value: 3.126e-09
```

In the R output, the hypothsis test for the coefficient of heterogeneity is $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 \neq 0$.The t-test statistic equals to -6.389 and the corresponding pvalue is 1.34e-07, which is much smaller than the significance level of 0.05. So we can reject the null hypothesis and conclude that the coefficient of heterogeneity is not zero, i.e. the heterogeneity has some effects on the moralIntegration.

Obtain the 96% confidence interval for the coefficient of heterogeneity.

```
confint(m5, level = 0.96)
```

```
##                       2 %        98 %
## (Intercept)   17.4088640 22.47265117
## heterogeneity -0.1446294 -0.07249023
## mobility      -0.2685253 -0.11810212
```

The 96% confidence interval for the coefficient of heterogeneity is (-0.1446294, -0.07249023), and the interpretation is we are 96% confident that the true coefficient β_1 is between -0.1446 and -0.0725.

4(e). Fitting a diffferent model with variable "social"

```
set.seed(631)
n= dim(Angell)[1]
Angell$social = with(Angell, heterogeneity+mobility+rnorm(n,0,0.1))
mod1 = lm(moralIntegration ~ heterogeneity + mobility + social, data=Angell)
summary(mod1)
```

```
##
## Call:
## lm(formula = moralIntegration ~ heterogeneity + mobility + social,
```

```
##     data = Angell)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2676 -1.4681 -0.1637  1.4847  4.1241
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     20.431      1.224  16.698   <2e-16 ***
## heterogeneity    4.076      2.873   1.419    0.164
## mobility         3.989      2.872   1.389    0.173
## social          -4.191      2.877  -1.457    0.153
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.212 on 39 degrees of freedom
## Multiple R-squared:  0.6438, Adjusted R-squared:  0.6164
## F-statistic: 23.49 on 3 and 39 DF,  p-value: 7.496e-09
```

The t-test statistic for the coefficient of heterogeneity is 1.419 and the corresponding pvalue is 0.164, which is greater than the level of siginificance $\alpha = 0.05$. So we fail to reject the null hypothesis, that is we do not have evidence to say the true coefficient of heterogeneity is different from zero. I think the reason why we get a contadictory result is because of the multicolinearity. The additional variable "social" is just a linear combination of the variable "heterogeneity" and "mobility".