# Question 7, 8

### Question 7

Load the data:

```
property <- read.table('~/Downloads/STA206_FQ2019/property.txt')
names(property) <- c('Y','X1','X2','X3','X4')
```

### (a)

```
n <- length(property[,1])
sample_mean <- c(0,0,0,0,0)
sample_sd <- c(0,0,0,0,0)
for(i in c(1,2,3,4,5)) {
  sample_mean[i] <- mean(property[,i])
  sample_sd[i] <- sd(property[,i])
  print(paste("Sample mean of", names(property)[i], "is" ,sample_mean[i]))
  print(paste("Sample sd of", names(property)[i], "is" ,sample_sd[i]))
}
```

```
## [1] "Sample mean of Y is 15.1388888888889"
## [1] "Sample sd of Y is 1.71958388861957"
## [1] "Sample mean of X1 is 7.8641975308642"
## [1] "Sample sd of X1 is 6.63278426910553"
## [1] "Sample mean of X2 is 9.68814814814815"
## [1] "Sample sd of X2 is 2.58316865066487"
## [1] "Sample mean of X3 is 0.0809876543209877"
## [1] "Sample sd of X3 is 0.134551151409711"
## [1] "Sample mean of X4 is 160633.271604938"
## [1] "Sample sd of X4 is 109098.959608813"
```

Now we perform the correlation transformation.

```
X_star <- as.matrix(cbind(rep(1,n), property[,2:5]))
names(X_star)[1] = "1"
for(i in c(2,3,4,5)) {
  X_star[,i] <- (1/sqrt(n-1))*((X_star[ ,i] - sample_mean[i])/sample_sd[i])
}
```

Check: 1. Sample mean of the transformed variables is zero. 2. Sample sd of the transformed variables is $\frac{1}{\sqrt{n-1}}$.

```
new_sd <- 1/sqrt(n-1)
new_sd
```

```
## [1] 0.1118034
```

```r
for(i in c(2,3,4,5)) {
  print(paste("Sample mean of transformed", names(property)[i], "is" ,mean(X_star[,i])))
  print(paste("Sample sd of transformed", names(property)[i], "is" ,sd(X_star[,i])))
}
```

```
## [1] "Sample mean of transformed X1 is -5.67850887839283e-18"
## [1] "Sample sd of transformed X1 is 0.111803398874989"
## [1] "Sample mean of transformed X2 is 7.51680043226601e-18"
## [1] "Sample sd of transformed X2 is 0.111803398874989"
## [1] "Sample mean of transformed X3 is -6.38073056325728e-18"
## [1] "Sample sd of transformed X3 is 0.111803398874989"
## [1] "Sample mean of transformed X4 is 1.25030010484363e-17"
## [1] "Sample sd of transformed X4 is 0.111803398874989"
```

Based on the results, we can conclude that sample means are 0 and sample sds are $\frac{1}{\sqrt{80}}$ or 0.1118.

### (b)

The standardized model equation is $Y_i = \beta_0^* + \beta_1^* X_{i1}^* + \beta_2^* X_{i2}^* + \beta_3^* X_{i3}^* + \beta_4^* X_{i4}^*$, $i = 1, 2, 3, \ldots, 81$. Then we fit the standardized model and present the regression results

```r
property_star <- as.data.frame(cbind(property[,1], X_star[,2:5]))
names(property_star) <- c('Y','X1','X2','X3','X4')
fit_star <- lm(formula=Y~X1+X2+X3+X4, data=property_star)
summary(fit_star)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4, data = property_star)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1872 -0.5911 -0.0910  0.5579  2.9441
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.1389     0.1263 119.845  < 2e-16 ***
## X1           -8.4262     1.2662  -6.655 3.89e-09 ***
## X2            6.5159     1.4596   4.464 2.75e-05 ***
## X3            0.7454     1.3079   0.570     0.57
## X4            7.7326     1.3513   5.722 1.98e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.137 on 76 degrees of freedom
## Multiple R-squared:  0.5847, Adjusted R-squared:  0.5629
## F-statistic: 26.76 on 4 and 76 DF,  p-value: 7.272e-14
```

The fitted model is $Y = 15.1389 - 8.4262X_1^* + 6.5159X_2^* + 0.7454X_3^* + 7.7326X_4^*$. The fitted regression intercept is 15.1389.

### (c)

Transforming the fitted standardized regression coefficients back to the fitted regression coefficients of the original model yields

```
coeff_star <- summary(fit_star)$coefficients[,1]
original_beta <- c(0,0,0,0,0)
for(i in c(2,3,4,5)) {
  original_beta[i] <- coeff_star[i]/(sqrt(n-1)*sample_sd[i])
}
original_beta[1] <- sample_mean[1] - sum(sample_mean[2:5]*original_beta[2:5])
for(i in c(1,2,3,4,5)) {
  print(paste("Beta_", i, " of the original model is", original_beta[i]))
}
```

```
## [1] "Beta_ 1  of the original model is 12.2005858819743"
## [1] "Beta_ 2  of the original model is -0.142033643508249"
## [1] "Beta_ 3  of the original model is 0.282016529950992"
## [1] "Beta_ 4  of the original model is 0.619343503463947"
## [1] "Beta_ 5  of the original model is 7.92430187605226e-06"
```

The original regression results are

```
fit_original <- lm(formula=Y~X1+X2+X3+X4, data=property)
summary(fit_original)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4, data = property)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1872 -0.5911 -0.0910  0.5579  2.9441
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.220e+01  5.780e-01  21.110  < 2e-16 ***
## X1          -1.420e-01  2.134e-02  -6.655 3.89e-09 ***
## X2           2.820e-01  6.317e-02   4.464 2.75e-05 ***
## X3           6.193e-01  1.087e+00   0.570     0.57
## X4           7.924e-06  1.385e-06   5.722 1.98e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.137 on 76 degrees of freedom
## Multiple R-squared:  0.5847, Adjusted R-squared:  0.5629
## F-statistic: 26.76 on 4 and 76 DF,  p-value: 7.272e-14
```

```
summary(fit_original)$coefficients[,1]
```

```
##   (Intercept)            X1            X2            X3            X4
##  1.220059e+01 -1.420336e-01  2.820165e-01  6.193435e-01  7.924302e-06
```

It is clear from R outputs that both models generate the same regression coefficients.

### (d)

```
sd_star <- summary(fit_star)$coefficients[,2]
sd_original <- sd_star/(sqrt(n-1)*sample_sd)
for(i in c(2,3,4,5)) {
  print(paste("Sample sd of beta_", i,
              " calculated from standardized model ", sd_original[i]))
}
```

```
## [1] "Sample sd of beta_ 2  calculated from standardized model  0.0213426098228305"
## [1] "Sample sd of beta_ 3  calculated from standardized model  0.0631723497451431"
## [1] "Sample sd of beta_ 4  calculated from standardized model  1.08681282876217"
## [1] "Sample sd of beta_ 5  calculated from standardized model  1.38477537679068e-06"
```

```
summary(fit_original)$coefficients[,2]
```

```
##  (Intercept)           X1           X2           X3           X4
## 5.779562e-01 2.134261e-02 6.317235e-02 1.086813e+00 1.384775e-06
```

It is clear from R outputs that both models generate the same standard errors of the fitted regression coefficients of X variables.

### (e)

```
SSTO <- t(property[,1])%*%(diag(1,n)-matrix(rep(1/n, n*n), nrow=n))%*%property[,1]
SSTO_star <- t(property_star[,1])%*%(diag(1,n)-matrix(rep(1/n, n*n), nrow=n))%*%property_star[
X <- as.matrix(cbind(rep(1,n),property[,2:5]))
H <- X%*%solve(t(X)%*%X)%*%t(X)
H_star <- X_star%*%solve(t(X_star)%*%X_star)%*%t(X_star)
SSE <- t(property[,1])%*%(diag(1,n)-H)%*%property[,1]
SSE_star <- t(property_star[,1])%*%(diag(1,n)-H_star)%*%property_star[,1]
SSR <- t(property[,1])%*%(H-matrix(rep(1/n, n*n), nrow=n))%*%property[,1]
SSR_star <-
  t(property_star[,1])%*%
  (H_star-matrix(rep(1/n, n*n),nrow=n))%*%
  property_star[,1]
```

SSTO, SSE, SSR under the original model are

```
c(SSTO, SSE, SSR)
```

```
## [1] 236.55750  98.23059 138.32691
```

SSTO, SSE, SSR under the standardized model are

```r
c(SSTO_star, SSE_star, SSR_star)
```

```
## [1] 236.55750  98.23059 138.32691
```

As we can see, they are the same in both models.

### (f)

$R^2, R_a^2$ under the original model are $0.58447, 0.5629$ respectively. $R^2, R_a^2$ under the standardized model are also $0.58447, 0.5629$ respectively.

## Question 8

### (a)

```r
X_star_sq <- t(X_star)%*%X_star
r_inverse <- solve(X_star_sq)[2:5,2:5]
r_sq_1 <- summary(lm(formula=X1~X2+X3+X4, data=property))$r.squared
r_sq_2 <- summary(lm(formula=X2~X1+X3+X4, data=property))$r.squared
r_sq_3 <- summary(lm(formula=X3~X1+X2+X4, data=property))$r.squared
r_sq_4 <- summary(lm(formula=X4~X1+X2+X3, data=property))$r.squared
c(r_inverse[1,1], r_inverse[2,2], r_inverse[3,3], r_inverse[4,4])
```

```
## [1] 1.240348 1.648225 1.323552 1.412722
```

```r
c(1/(1-r_sq_1),1/(1-r_sq_2),1/(1-r_sq_3),1/(1-r_sq_4))
```

```
## [1] 1.240348 1.648225 1.323552 1.412722
```

The same results from two methods confirm $VIF_k = \frac{1}{1-R_k^2}$, $k = 1, 2, 3, 4$. All four VIF values are a little bit higher than 1 and far less than 10, so we can conclude that there is not much multicollinearity in the model.

### (b)

```r
fit_X4 <- lm(formula=Y~X4, data=property)
summary(fit_X4)$coefficients[2,1]
```

```
## [1] 8.436639e-06
```

```r
fit_X3X4 <- lm(formula=Y~X3+X4, data=property)
summary(fit_X3X4)$coefficients[3,1]
```

```
## [1] 8.406741e-06
```

The estimated regression coeffcients of X4 in these two models are almost the same.

```r
anova(fit_X4)
```

```
## Analysis of Variance Table
##
```

```
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X4         1  67.775  67.775  31.723 2.628e-07 ***
## Residuals 79 168.782   2.136
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(fit_X3X4)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X3         1   1.047   1.047  0.4842    0.4886
## X4         1  66.858  66.858 30.9213 3.626e-07 ***
## Residuals 78 168.652   2.162
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(fit_X4)[1,2]
```

```
## [1] 67.7751
```

```r
anova(fit_X3X4)[2,2]
```

```
## [1] 66.85829
```

$SSR(X_4) = 67.7751$ and $SSR(X_4|X_3) = 66.8583$, they are quite similar. This is expected, since the correlation matrix shows that there is almost no correlation between $X_3$ and $X_4$, the marginal effect of adding $X_4$ into the model which already has $X_3$ is very closed to the explaining ablility of $X_4$ alone.

### (c)

```r
fit_X2 <- lm(formula=Y~X2, data=property)
summary(fit_X2)$coefficients[2,1]
```

```
## [1] 0.2754531
```

```r
fit_X4X2 <- lm(formula=Y~X4+X2, data=property)
summary(fit_X4X2)$coefficients[3,1]
```

```
## [1] 0.1469682
```

Two estimated regression coefficients of $X_2$ are quite different.

```r
anova(fit_X2)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X2         1  40.503  40.503  16.321 0.0001231 ***
```

```
## Residuals 79 196.054   2.482
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(fit_X4X2)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X4         1  67.775  67.775 33.1457 1.611e-07 ***
## X2         1   9.291   9.291  4.5438   0.03619 *
## Residuals 78 159.491   2.045
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(fit_X2)[1,2]
```

```
## [1] 40.50333
```

```r
anova(fit_X4X2)[2,2]
```

```
## [1] 9.290987
```

$SSR(X_2) = 40.5033 > SSR(X_2|X_4) = 9.2910$. The correlation matrix shows that there $X_2$ and $X_4$ are moderately correlated, so the marginal effect of adding $X_2$ into the model which already has $X_4$ is expected to less effective compared to the explaining ability of $X_2$ alone.