**Bios 7345, Fall 2017**
**Lab #6, Sums of Squares**

Suppose researchers conduct a study measuring the concentration of abnormal toxins in a specific kind of human tissue. They are interested in knowing whether differences exist based on gender or race. Data are taken from 90 individuals, in which each observation includes a response (continuous $y_i$), gender (male, female), and race (White, Black, or Other).

1. Use the following R code to simulate data using a balanced design.

```
n=90
gender = as.factor(c(rep("male",n/2),rep("female",n/2)))
race = as.factor( rep(c(rep("W",n/6), rep("B",n/6), rep("O",n/6)),2))
Xbeta = 50 + 0.5*(gender=="male") + .5*(race=='B') + 0*(race=='O') +
  1*(gender=='male')*(race=='B') +
          2*(gender=='male')*(race=='O')
set.seed(1)
y = rnorm(n,mean=Xbeta,sd=2)
```

   (a) Calculate the estimated regression coefficients and corresponding p-values for the model y=race+gender+race*gender using the lm() function.

   (b) Calculate the estimated regression coefficients and corresponding p-values for the model y=gender+race+race*gender (i.e. switch the order of main effects). How do the estimates and p-values compare to the above model?

   (c) Calculate the Type I SS and corresponding p-values for models 1(a) and 1(b). How do they compare?

   (d) For the model in 1(a), calculate the Type II and Type III Sums of squares and corresponding p-values using the 'car' package (pay attention to the "type=" option in the `Anova` function). How do they compare to the Type I SS?

   (e) Is gender associated with the concentration of toxins?

2. Use the following R code to simulate data using an unbalanced design

```
gender = as.factor(c(rep("male",n/2),rep("female",n/2)))
race = as.factor( c(rep("W",10), rep("B",20),
                    rep("O",15), rep("W",5), rep("B",10),
                    rep("O",30) ))
Xbeta = 50 + 0.5*(gender=="male") + .5*(race=='B') +
  0*(race=='O') + 1*(gender=='male')*(race=='B') +
  2*(gender=='male')*(race=='O')
```

```
set.seed(345)
y = rnorm(n,mean=Xbeta,sd=2)
```

(a) Calculate the estimated regression coefficients and corresponding p-values for the model y=race+gender+race*gender

(b) Calculate the estimated regression coefficients and corresponding p-values for the model y=gender+race+race*gender (i.e. switch the order of main effects). How do the estimates and p-values compare to the above model?

(c) Calculate the Type I SS and corresponding p-values for models 1(a) and 1(b). How do they compare?

(d) For the model in 1(a), calculate the Type II and Type III Sums of squares and corresponding p-values using the 'car' package. How do they compare to the Type I SS?

(e) Is gender associated with the concentration of toxins?

3. Specify the appropriate $\boldsymbol{C}$ matrix to test whether gender is associated with the concentration of toxins using $H_0 : \boldsymbol{C\beta} = \mathbf{0}$. This is the strategy used in anova.rms() with the test='Chisq' option.