BIOS7345 Lab # 6

Sums of squares

Sarah Lotspeich

18 Oct 2019

Introduction

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, n which each observation includes a response (continuous y_i), gender (male/ female), and race (White/ Black/ Other).

Balanced design

Use the following `R` code to simulate data using a balanced design:

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

Model 1(a)

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

```
mod1a <- lm(y ~ race*gender)
summary(mod1a)
```

```
##
## Call:
## lm(formula = y ~ race * gender)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6311 -1.0660  0.0192  1.0951  4.4431
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        50.8601     0.4702 108.165   <2e-16 ***
## raceO              -0.7791     0.6650  -1.172    0.245
## raceW              -0.5096     0.6650  -0.766    0.446
## gendermale          0.2681     0.6650   0.403    0.688
## raceO:gendermale    1.3315     0.9404   1.416    0.161
## raceW:gendermale    0.0831     0.9404   0.088    0.930
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.821 on 84 degrees of freedom
## Multiple R-squared:  0.08002,    Adjusted R-squared:  0.02526
## F-statistic: 1.461 on 5 and 84 DF,  p-value: 0.2112
```

Model 1(b)

Calculate the estimated regression cofficients and corresponding p-values for the model `y \sim gender + race + gender*race` (i.e. switch the order of the main effects) using the `lm()` function.

```
mod1b <- lm(y ~ gender*race)
summary(mod1b)
```

```
##
## Call:
## lm(formula = y ~ gender * race)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6311 -1.0660  0.0192  1.0951  4.4431
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       50.8601     0.4702 108.165   <2e-16 ***
## gendermale         0.2681     0.6650   0.403    0.688
## raceO             -0.7791     0.6650  -1.172    0.245
## raceW             -0.5096     0.6650  -0.766    0.446
## gendermale:raceO   1.3315     0.9404   1.416    0.161
## gendermale:raceW   0.0831     0.9404   0.088    0.930
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.821 on 84 degrees of freedom
## Multiple R-squared:  0.08002,    Adjusted R-squared:  0.02526
## F-statistic: 1.461 on 5 and 84 DF,  p-value: 0.2112
```

How do the estimates and p-values for Models 1(a) and 1(b) compare?

> They are the same. Thus, the ordering of main effects did not matter. These p-values are for the individual levels of the factor variables (e.g. Black versus Other and Black versus White) rather than the collective effect of gender.

Calculate the Type I SS and corresponding p-values for Models 1(a) and 1(b).

```
anova(mod1a)
```

```
## Analysis of Variance Table
##
## Response: y
##             Df  Sum Sq Mean Sq F value  Pr(>F)
## race         2   3.577  1.7884  0.5393 0.58519
## gender       1  12.307 12.3073  3.7110 0.05744 .
## race:gender  2   8.346  4.1728  1.2582 0.28946
## Residuals   84 278.580  3.3164
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod1b)
```

```
## Analysis of Variance Table
##
## Response: y
##             Df  Sum Sq Mean Sq F value  Pr(>F)
## gender       1  12.307 12.3073  3.7110 0.05744 .
## race         2   3.577  1.7884  0.5393 0.58519
## gender:race  2   8.346  4.1728  1.2582 0.28946
## Residuals   84 278.580  3.3164
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How do they compare?

> Once again, reordering the main effects between Models 1(a) and 1(b) did not affect the Type I SS. However, it could affect the interpretations as these are "sequential" sums of squares, but since this is a balanced design it does not. We would interpret the SS from Model 1(a) as follows: the p-value for the main effect of race is 0.586 (so we fail to reject), the p-value for the main effect of gender after the main effect of race is 0.057 (so we would fail to reject except at $\alpha = 0.10$), and the p-value for the interaction effect after the main effects of race and gender is 0.289 (so we fail to reject).

For Model 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.

```
car::Anova(mod1a, type = 2)
```

```
## Anova Table (Type II tests)
##
## Response: y
##              Sum Sq Df F value  Pr(>F)
## race          3.577  2  0.5393 0.58519
## gender       12.307  1  3.7110 0.05744 .
## race:gender   8.346  2  1.2582 0.28946
## Residuals   278.580 84
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
car::Anova(mod1b, type = 3)
```

```
## Anova Table (Type III tests)
##
## Response: y
##             Sum Sq Df    F value Pr(>F)
## (Intercept)  38801  1 11699.7149 <2e-16 ***
## gender           1  1     0.1625 0.6879
## race             5  2     0.7080 0.4955
## gender:race      8  2     1.2582 0.2895
## Residuals      279 84
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How do they compare to the Type I SS?

> The F statistics and p-values are the same for the Type I, II, and III SS based on Model 1(a).

Type II sums of squares tests for each main effect after the other main effect. Type III sums of squares tests for the presence of a main effect after the other main effect and interaction. When data are balanced, the factors are orthogonal, and types I, II and III all give the same results.

Is gender associated with the concentration of toxins? You may use `rms::anova()`.

```
anova(rms::ols(y~gender+race + gender*race), test='Chisq')
```

```
##                 Wald Statistics          Response: y
##
##  Factor                                             d.f. Partial SS MS
##  gender  (Factor+Higher Order Factors)               3    20.652900 6.884300
##   All Interactions                                  2     8.345598 4.172799
##  race  (Factor+Higher Order Factors)                4    11.922419 2.980605
##   All Interactions                                  2     8.345598 4.172799
##  gender * race  (Factor+Higher Order Factors)       2     8.345598 4.172799
##  REGRESSION                                         5    24.229720 4.845944
##  ERROR                                             84   278.579707 3.316425
##  Chi-Square P
##  6.23       0.1011
##  2.52       0.2842
##  3.59       0.4636
##  2.52       0.2842
##  2.52       0.2842
##  7.31       0.1989
##
```

With a χ^2 test statistic of 6.23 and corresponding p-value of 0.1011, we fail to reject the null hypothesis that gender is not associated with the concentration of toxins.

Unbalanced design

Use the following `R` code to simulate data using an unbalanced design:

```
n <- 90
gender <- as.factor(rep(c("male","female"), each = n/2))
race <- as.factor(c(rep("W",10),rep("B",20),rep("O",15),rep("W",5),rep("B",10),rep("O",30) ))
mu <- 50 + 0.5*(gender == "male") + 0.5*(race == "B") + 0*(race == "O") + 1*(gender =="male")*(race ==
set.seed(345)
y <- rnorm(n, mean = mu, sd = 2)
```

Model 2(a)

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

```
mod2a <- lm(y ~ race*gender)
summary(mod2a)
```

```
##
## Call:
## lm(formula = y ~ race * gender)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -4.7627 -1.5052 -0.3098  1.2780  5.8309
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       49.2202     0.6297  78.168   <2e-16 ***
## raceO              0.5855     0.7271   0.805   0.4229
## raceW              1.5846     1.0906   1.453   0.1500
## gendermale         1.3784     0.7712   1.787   0.0775 .
## raceO:gendermale   0.3877     0.9956   0.389   0.6979
## raceW:gendermale  -1.2770     1.3357  -0.956   0.3418
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.991 on 84 degrees of freedom
## Multiple R-squared:  0.1289, Adjusted R-squared:  0.07702
## F-statistic: 2.485 on 5 and 84 DF,  p-value: 0.03774
```

Model 2(b)

Calculate the estimated regression cofficients and corresponding p-values for the model `y \sim gender + race + gender*race` (i.e. switch the order of the main effects) using the `lm()` function.

```
mod2b <- lm(y ~ gender*race)
summary(mod2b)
```

```
##
## Call:
## lm(formula = y ~ gender * race)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7627 -1.5052 -0.3098  1.2780  5.8309
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       49.2202     0.6297  78.168   <2e-16 ***
## gendermale         1.3784     0.7712   1.787   0.0775 .
## raceO              0.5855     0.7271   0.805   0.4229
## raceW              1.5846     1.0906   1.453   0.1500
## gendermale:raceO   0.3877     0.9956   0.389   0.6979
## gendermale:raceW  -1.2770     1.3357  -0.956   0.3418
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.991 on 84 degrees of freedom
## Multiple R-squared:  0.1289, Adjusted R-squared:  0.07702
## F-statistic: 2.485 on 5 and 84 DF,  p-value: 0.03774
```

How do the estimates and p-values for Models 2(a) and 2(b) compare?

> As with the balanced design, reordering does not effect estimates or p-values from the individual t-tests.

Calculate the Type I SS and corresponding p-values for Models 2(a) and 2(b).

```
anova(mod2a)
```

```
## Analysis of Variance Table
##
## Response: y
##             Df Sum Sq Mean Sq F value   Pr(>F)
## race         2   5.38   2.690  0.6784 0.510211
## gender       1  36.96  36.961  9.3222 0.003032 **
## race:gender  2   6.93   3.466  0.8742 0.420964
## Residuals   84 333.05   3.965
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod2b)
```

```
## Analysis of Variance Table
##
## Response: y
##             Df Sum Sq Mean Sq F value   Pr(>F)
## gender       1  32.66  32.658  8.2369 0.005192 **
## race         2   9.68   4.841  1.2210 0.300112
## gender:race  2   6.93   3.466  0.8742 0.420964
## Residuals   84 333.05   3.965
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How do they compare?

> The F test statisics and p-values are different for the main effects in Model 2(a) versus Model 2(b). Unlike with the balanced design, with an unbalanced design the sequential sums of squares will vary for different specified orderings of the main effects. We would conclude that the effect of gender is significantly different from 0 after the effect of race, but nothing else.

For Model 2(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.

```
car::Anova(mod2a, type = 2)
```

```
## Anova Table (Type II tests)
##
## Response: y
##             Sum Sq Df F value   Pr(>F)
## race          9.68  2  1.2210 0.300112
## gender       36.96  1  9.3222 0.003032 **
## race:gender   6.93  2  0.8742 0.420964
## Residuals   333.05 84
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
car::Anova(mod2b, type = 3, contrasts=list(gender=contr.sum, race=contr.sum))
```

```
## Anova Table (Type III tests)
##
## Response: y
##              Sum Sq Df   F value  Pr(>F)
## (Intercept) 24226.3  1 6110.3004 < 2e-16 ***
## gender         12.7  1    3.1947 0.07748 .
## race            8.4  2    1.0596 0.35117
```

```
## gender:race     6.9  2    0.8742 0.42096
## Residuals     333.0 84
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How do they compare to the Type I SS?

> The Type II SS for gender was the same as the Type I SS, as was the Type II SS for the interaction between race and gender. However, only the Type III SS for the interaction was consistent with the Type I SS.

Is gender associated with the concentration of toxins?

```
anova(rms::ols(y~gender+race + gender*race), test='Chisq')
```

```
##                 Wald Statistics          Response: y
##
##  Factor                                          d.f. Partial SS MS
##  gender  (Factor+Higher Order Factors)             3    43.892895 14.630965
##   All Interactions                                2     6.931948  3.465974
##  race  (Factor+Higher Order Factors)              4    16.614146  4.153536
##   All Interactions                                2     6.931948  3.465974
##  gender * race  (Factor+Higher Order Factors)     2     6.931948  3.465974
##  REGRESSION                                      5    49.271979  9.854396
##  ERROR                                          84   333.045172  3.964823
##  Chi-Square P
##  11.07      0.0114
##   1.75      0.4172
##   4.19      0.3809
##   1.75      0.4172
##   1.75      0.4172
##  12.43      0.0294
##
```

> With a χ^2 test statistic of 11.07 and corresponding p-value of 0.0114, we reject the null hypothesis that gender is not associated with the concentration of toxins at the $\alpha = 0.05$ significance level.

Tests for association

Specify the appropriate $\boldsymbol{C}$ matrix to test whether gender is associated with the concentration of toxins using $H_0 : \boldsymbol{C\beta} = \boldsymbol{0}$.

```
C <- matrix(c(0,1,0,0,0,0,
              0,0,0,0,1,0,
              0,0,0,0,0,1),
            byrow=TRUE,nrow=3)
gmodels::estimable(mod2b,cm=C,joint.test=TRUE)
```

```
##    X2.stat DF Pr(>|X^2|)
## 1 11.07058  3 0.01135028
```

Specify the appropriate $\boldsymbol{C}$ matrix to test whether race is associated with the concentration of toxins using $H_0 : \boldsymbol{C\beta} = \boldsymbol{0}$.

```
C <- matrix(c(0,0,1,0,0,0,
              0,0,0,1,0,0,
              0,0,0,0,1,0,
```

```
              0,0,0,0,0,1),
            byrow=TRUE,nrow=4)
gmodels::estimable(mod2b,cm=C,joint.test=TRUE)
```

```
##    X2.stat DF Pr(>|X^2|)
## 1 4.190387  4  0.3808525
```

References

This is where I got the interpretations of the different SS.