Math 158, Spring 2018
Jo Hardin
Multiple Regression II – R code
Decomposition of Sums of Squares (and F-tests)

NC Births, ANOVA & F-tests

A description of the data is given at `http://pages.pomona.edu/~jsh04747/courses/math58/Math58Data/NCBIRTH800_description.pdf`

```
births <- read.table("http://pages.pomona.edu/~jsh04747/courses/math58/Math58Data/NCBIRTH800.csv", head
dim(births)

## [1] 800  14

names(births)

##  [1] "plural"  "sex"     "mage"    "weeks"   "marital" "racemom" "hispmom"
##  [8] "gained"  "smoke"   "drink"   "tounces" "tgrams"  "low"     "premie"
```

Notice that the variable names we'll use are `mage`, `tounces`, `gained`, and `smoke`.

1. We're interested in predicting a baby's size from the mother's weight gain.

```
oz.g.lm <- lm(tounces ~ gained, data=births)
summary(oz.g.lm)$coef

##                Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 105.7024598 1.90526159 55.479237 4.125120e-272
## gained        0.3611576 0.05689578  6.347706  3.717947e-10

anova(oz.g.lm)

## Analysis of Variance Table
##
## Response: tounces
##            Df Sum Sq Mean Sq F value    Pr(>F)
## gained      1  18856   18856  40.293 3.718e-10 ***
## Residuals 775 362671     468
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2. What if we include smoking status as a variable?

```
oz.gs.lm <- lm(tounces ~ gained + smoke, data=births)
summary(oz.gs.lm)$coef
```

```
##                Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 106.9480736 1.92006170 55.700332 5.260646e-273
## gained        0.3586615 0.05644457  6.354225  3.573523e-10
## smoke        -8.0400596 2.18391060 -3.681497  2.479683e-04
```

```
anova(oz.gs.lm)
```

```
## Analysis of Variance Table
##
## Response: tounces
##           Df Sum Sq Mean Sq F value    Pr(>F)
## gained     1  18856 18855.8  40.946 2.709e-10 ***
## smoke      1   6241  6241.4  13.553  0.000248 ***
## Residuals 774 356429   460.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3. What if we let smoking and weight gain *interact*?

```
oz.gis.lm <- lm(tounces ~ gained * smoke, data=births)
summary(oz.gis.lm)$coef
```

```
##                Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)  105.9627761 2.10092125 50.436339 1.089682e-246
## gained         0.3908076 0.06293174  6.210024  8.644328e-10
## smoke         -3.0713887 4.82712503 -0.636277  5.247842e-01
## gained:smoke  -0.1641078 0.14219035 -1.154141  2.487989e-01
```

```
anova(oz.gis.lm)
```

```
## Analysis of Variance Table
##
## Response: tounces
##               Df Sum Sq Mean Sq F value    Pr(>F)
## gained         1  18856 18855.8  40.964 2.688e-10 ***
## smoke          1   6241  6241.4  13.559 0.0002472 ***
## gained:smoke   1    613   613.1   1.332 0.2487989
## Residuals    773 355816   460.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. What happens to the model if we add in another quantitative variable?

```
oz.gsm.lm <- lm(tounces ~ gained + smoke + mage, data=births)
summary(oz.gsm.lm)$coef
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
```

```
## (Intercept) 91.3340048 3.86669135 23.620713 2.754802e-93
## gained       0.3517505 0.05573292  6.311359 4.657244e-10
## smoke       -7.2060335 2.16310784 -3.331333 9.053582e-04
## mage         0.5826345 0.12576562  4.632701 4.234351e-06
```

```r
anova(oz.gsm.lm)
```

```
## Analysis of Variance Table
##
## Response: tounces
##            Df Sum Sq Mean Sq F value    Pr(>F)
## gained      1  18856 18855.8  42.029 1.603e-10 ***
## smoke       1   6241  6241.4  13.912 0.0002056 ***
## mage        1   9629  9628.7  21.462 4.234e-06 ***
## Residuals 773 346801   448.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5. What happens to the model if we add in another quantitative variable with the interaction?

```r
oz.gism.lm <- lm(tounces ~ gained * smoke + mage, data=births)
summary(oz.gism.lm)$coef
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  90.6736580 3.92785662 23.0847678 4.552785e-90
## gained        0.3781869 0.06219363  6.0807970 1.879687e-09
## smoke        -3.1379061 4.76585424 -0.6584142 5.104683e-01
## mage          0.5771142 0.12590426  4.5837546 5.325282e-06
## gained:smoke -0.1346252 0.14053214 -0.9579670 3.383793e-01
```

```r
anova(oz.gism.lm)
```

```
## Analysis of Variance Table
##
## Response: tounces
##               Df Sum Sq Mean Sq F value    Pr(>F)
## gained         1  18856 18855.8 42.0240 1.607e-10 ***
## smoke          1   6241  6241.4 13.9102 0.0002058 ***
## mage           1   9629  9628.7 21.4596 4.240e-06 ***
## gained:smoke   1    412   411.8  0.9177 0.3383793
## Residuals    772 346389   448.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6. What is the F-test to see if `mage` should be added to the model with interaction? (yes, we need `mage`, p-value is .0000053 for $H_0 : \beta_{mage}$)

```r
anova(oz.gis.lm, oz.gism.lm)

## Analysis of Variance Table
##
## Model 1: tounces ~ gained * smoke
## Model 2: tounces ~ gained * smoke + mage
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    773 355816
## 2    772 346389  1    9427.3 21.011 5.325e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

7. Can we drop both the interaction and `mage`? (No, the p-value is still very small. Now the test is $H_0 : \beta_{gainxsmk} = \beta_{mage} = 0$.)

```r
anova(oz.gs.lm, oz.gism.lm)

## Analysis of Variance Table
##
## Model 1: tounces ~ gained + smoke
## Model 2: tounces ~ gained * smoke + mage
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    774 356429
## 2    772 346389  2     10040 11.189 1.621e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

8. Calculate the coefficient of partial determination for `mage` given `gained` and `smoke`

```r
anova(oz.gsm.lm)

## Analysis of Variance Table
##
## Response: tounces
##            Df Sum Sq Mean Sq F value    Pr(>F)
## gained      1  18856 18855.8  42.029 1.603e-10 ***
## smoke       1   6241  6241.4  13.912 0.0002056 ***
## mage        1   9629  9628.7  21.462 4.234e-06 ***
## Residuals 773 346801   448.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

9629 / (9629 + 346801)

## [1] 0.02701512
```

The variability (in ounces of baby weight) remaining after modeling with `gained` and `smoke` is reduced by a further 2.7% when additionally adding `mage`.

9. On a different note... consider the two different multiple regression model with race as a quantitative variable (a) and race as a categorical variable (b):

(a)
```r
oz.gr.lm <- lm( tounces ~ gained + racemom, data=births)
summary(oz.gr.lm)$coef

##                Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 106.4712456 2.18548709 48.717398 5.684528e-238
## gained        0.3620035 0.05692569  6.359228  3.464657e-10
## racemom      -0.5852568 0.81426689 -0.718753  4.725100e-01

anova(oz.gr.lm)

## Analysis of Variance Table
##
## Response: tounces
##            Df Sum Sq Mean Sq F value    Pr(>F)
## gained      1  18856 18855.8 40.2682 3.766e-10 ***
## racemom     1    242   241.9  0.5166    0.4725
## Residuals 774 362429   468.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(b)
```r
oz.grc.lm <- lm( tounces ~ gained + factor(racemom), data=births )
summary(oz.grc.lm)$coef

##                     Estimate  Std. Error    t value      Pr(>|t|)
## (Intercept)      108.0562919  1.92956943 56.0002092 9.707504e-274
## gained             0.3505041  0.05601502  6.2573240  6.496008e-10
## factor(racemom)2  -9.7503634  1.87326646 -5.2050061  2.490632e-07
## factor(racemom)3  -6.6229286  6.19722645 -1.0686924  2.855432e-01
## factor(racemom)4   2.1710217 15.07063645  0.1440564  8.854936e-01
## factor(racemom)7 -22.6229286 21.26516998 -1.0638489  2.877306e-01
## factor(racemom)8  11.8906620  6.46736549  1.8385635  6.636439e-02

anova(oz.grc.lm)

## Analysis of Variance Table
##
## Response: tounces
##                  Df Sum Sq Mean Sq F value    Pr(>F)
## gained            1  18856 18855.8 41.7699 1.821e-10 ***
## factor(racemom)   5  15077  3015.4  6.6799 4.202e-06 ***
## Residuals       770 347594   451.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```