R Codes for Chapter-7

Extra Sums of Squares in R

We will work again with the data from Problem 6.9, "Grocery Retailer." You can obtain the ANOVA table using the function "anova(model)". Here you get sum of squares for each predictor variable in the model:

```
> anova(lrm)
Analysis of Variance Table

Response: Retailer
          Df  Sum Sq Mean Sq F value    Pr(>F)
Cases      1  136366  136366  6.6417   0.01309 *
Costs      1    5726    5726  0.2789   0.59987
Holiday    1 2034514 2034514 99.0905 2.941e-13 ***
Residuals 48  985530   20532
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ANOVA table given by R provides the extra sum of squares for each predictor variable, *given that* the previous predictors are already in the model. Thus the Sum of Squares given for "Cases" is $\boldsymbol{SSR(X_1) = 136366}$, while the Sum of Squares given for "Costs" is $\boldsymbol{SSR(X_2 \mid X_1) = 5726}$, and the Sum of Squares given for "Holiday" is $\boldsymbol{SSR(X_3 \mid X_1, X_2) = 2034514}$. This corresponds to Table 7.3 on p.261 of the text.

Now you have $SSR(X_1)$, $SSR(X_2 \mid X_1)$, and $SSR(X_3 \mid X_1, X_2)$ and their corresponding degrees of freedom and mean squares. If you sum them together you get $SSR(X_1, X_2, X_3)$, which has 3 degrees of freedom. Divide $SSR(X_1, X_2, X_3)$ by 3 to get $MSR(X_1, X_2, X_3)$. To get $SSE(X_1, X_2, X_3)$, its degrees of freedom, and $MSE(X_1, X_2, X_3)$, use the line beginning with "Residuals." To calculate and store these in R, use the commands

```
> SSR = sum( anova(lrm)[1:3,2] )
> SSR
[1] 2176606
> MSR = SSR / 3
> MSR
[1] 725535.4
> SSE = anova(lrm)[4,2]
> SSE
[1] 985529.7
> MSE = anova(lrm)[4,3]
> MSE
[1] 20531.87
```

You can obtain alternate decompositions of the regression sum of squares into **extra sum of squares** by running new linear models with the predictors entered in a different order. For an example, if we want $SSR(X_3)$, $SSR(X_1|X_3)$ and $SSR(X_2|X_1,X_3)$, we could try:

```
> Model2 <- lm( Retailer ~ Holiday+Cases+Costs)
> anova(Model2)
Analysis of Variance Table

Response: Retailer
          Df  Sum Sq Mean Sq  F value    Pr(>F)    
Holiday    1 2077646 2077646 101.1913 2.086e-13 ***
Cases      1   92285   92285   4.4947    0.0392 *  
Costs      1    6675    6675   0.3251    0.5712    
Residuals 48  985530   20532                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

GENERAL LINEAR TEST

If we are considering dropping **Costs (X_2)** from the **lrm** model, we run a reduced model which uses only the other two predictors **Cases** and **Holiday**:

```
> Reduced <- lm( Retailer ~ Holiday+Cases) #fitting the reduced model
> Reduced

Call:
lm(formula = Retailer ~ Holiday + Cases)

Coefficients:
(Intercept)      Holiday        Cases  
  4.058e+03    6.196e+02    7.704e-04  
```

Then to perform the F test, just type

```
> anova(Reduced, Retailer)
```

To get the ANOVA comparison:

```
> anova(Reduced, lrm)
Analysis of Variance Table

Model 1: Retailer ~ Holiday + Cases
Model 2: Retailer ~ Cases + Costs + Holiday
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     49 992204                           
2     48 985530  1    6674.6 0.3251 0.5712
```

Note that the first argument to the **anova()** function must be the **reduced model**, and the second argument must be the full model (the one with all the original predictors).

General Linear Test for the other reduced models:

Now suppose we want to test H_0: $\beta_2 = 0$, $\beta_3 = 600$ against its alternative. In this case the reduced model, corresponding to H_0, is $Y_i = \beta_0 + \beta_1 X_{i1} + 600X_{i3} + \varepsilon_i$, which may be rewritten as $Y_i - 600X_{i3} = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$. To obtain the reduced model in R, use the formulation:

```
> RetailerN=Retailer-600*Holiday
> Reduced2 <- lm(RetailerN ~ Cases)
> Reduced2
```

```
Call:
lm(formula = RetailerN ~ Cases)

Coefficients:
(Intercept)        Cases
  4.059e+03    7.756e-04



> anova(Reduced2)
Analysis of Variance Table

Response: RetailerN
          Df Sum Sq Mean Sq F value  Pr(>F)
Cases      1  93738   93738   4.714 0.03469 *
Residuals 50 994244   19885
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

However, you will get an error message if you attempt to use the anova() function to compare this model with the full model, because the two models do not have the same response variable. Instead, you will need to obtain the *SSE* for this reduced model, along with its degrees of freedom, from its ANOVA table, and the *SSE* from the full model, along with its degrees of freedom, from the ANOVA table for the full model, then calculate F^* using equation **(2.9)** in the textbook.

```
> SSE_R=anova(Reduced2)[2,2]
> #SSE_R
> DF_R=anova(Reduced2)[2,1]
> #DF_R
>
> SSE_F = anova(lrm)[4,2]
> #SSE_F
> DF_F=anova(lrm)[4,1]
> #DF_R
>
> #Test atatistics
> F=((SSE_R-SSE_F)/(DF_R-DF_F))/(SSE_F/DF_F)
> F
[1] 0.2122226


> #P-value
> Pvalue=1-pf(F,DF_R-DF_F,DF_F)
> Pvalue
[1] 0.8095395
```

Coefficients of Partial Determination

To obtain the coefficients of partial determination, you will need to use formulae like those in section 7.4. You may also need to run several different models, with the predictors in various different orders, in order to obtain values for the needed forms of *SSE* and the extra sums of squares.

To calculate $R^2_{Y1|23}$, first the following model with X_2 and $X3$ should be fit to calculate SSE(X_2, X_3)

```
> M1=lm(Retailer~ Costs+Holiday)
> SSEX_2X_3=anova(M1)[3,2]
> SSEX_2X_3
[1] 1081237
```

A model with X_1, X_2 and X_3 should be fit to calculate SSE(X_1, X_2, X_3)

```
> M2=lm(Retailer~ Cases+Costs+Holiday)
> SSEX_1X_2X_3=anova(M2)[4,2]
> SSEX_1X_2X_3
[1] 985529.7
```

Then $R^2_{Y1|23}$

```
> RSQ_Y1_23=(SSEX_2X_3-SSEX_1X_2X_3)/SSEX_2X_3
> RSQ_Y1_23
[1] 0.08851609
```