1 General linear hypothesis

- the linear multiple linear regression model is $y = X\beta + \epsilon$.
- the general linear hypothesis is
  $H_0 : T\beta = c$, vs $H_A : T\beta \neq c$.
- where T is an m by p matrix having r independent rows, and c is a $m \times 1$ vector of constants.
- The test statistic is given by

$$F = \frac{(T\hat{\beta} - c)' \left[T(X'X)^{-1}T'\right]^{-1} (T\hat{\beta} - c)/r}{SSE_{full}/(n-p)}$$

- with p-value $P(F_{r,n-p} > F_{obs})$.

2 a general example

Suppose $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$.

- to test $\beta_2 = \beta_3$, use $T = [0, 0, 1, -1]$ with $c = 0$.
- to test $\beta_2 = \beta_3$ and $\beta_1 = 2$, use a matrix T with two rows and four columns. First row of T is $[0, 0, 1, -1]$ and second row of T is $[0, 1, 0, 0]$. c is $[0, 2]'$.
  - T has two linearly indepedent rows.
  - the F statistic will have 2 numerator, and n-4 denominator degrees of freedom.

3 Application to oneway analysis of variance

Here is a slightly more complicated example. In the one way analysis of variance example with four diets A, B, C, and D, consider the model

$$Y = \beta_1 I_A + \beta_2 I_B + \beta_3 I_C + \beta_4 I_D + \epsilon$$

Here we have dropped the intercept and used all 4 indicator variables.

Note that the mean of Y is β_1 for diet A, β_2 for diet B, β_3 for diet C, and β_4 for diet D. In this respect, this form of the model is really the most natural way to write the oneway ANOVA model using multiple regression.

Consider the hypothesis $H_0 : \beta_1 = \beta_2, \beta_2 = \beta_3, \beta_3 = \beta_4$. This is clearly equivalent to the usual ANOVA hypothesis $H_0 : \mu_A = \mu_B = \mu_C = \mu_D$.

where T is a 3X4 matrix with first row (1,-1,0,0), 2nd row (0,1,-1,0), 3rd row (0,0,1,-1), the hypothesis is equivalent to $H_0 : T\beta = \mathbf{0}$, which is in the form of the general linear hypothesis.

Clearly there are many matrices T which give an equivalent to the usual ANOVA hypothesis.

The following evaluates the general linear hypothesis test statistic F.

```
> weight=c(3.42 ,3.17 ,3.34 ,3.65 ,
+ 3.96 ,3.63 ,3.72 ,3.93 ,
+ 3.87 ,3.38 ,3.81 ,3.77 ,
+ 4.19 ,3.47 ,3.66 ,4.18 ,
+ 3.58 ,3.39 ,3.55 ,4.21 ,
+ 3.76 ,3.41 ,3.51 ,3.88 ,
+ 3.84 ,3.55       ,3.96 ,
+      3.44        ,3.91 )
>  diet=as.factor(c(rep(c("A","B","C","D"),6),"A","B","D","B","D"))
>  IA=ifelse(diet=="A",1,0)
>  IB=ifelse(diet=="B",1,0)
>  IC=ifelse(diet=="C",1,0)
>  ID=ifelse(diet=="D",1,0)
>  X=cbind(IA,IB,IC,ID)
>  bhat=solve(t(X)%*%X)%*%t(X)%*%weight
>  fits=X%*%bhat
>  resids=weight-fits
>  SSE=sum(resids^2)
>  dfe=length(weight)-dim(X)[2]
>  MSE=SSE/dfe
>  sqrt(MSE) # RESIDUAL STANDARD ERROR

[1] 0.1892526

> #fit a model include all 4 indicators, but no intercept
> lm.out= lm(weight~IA+IB+IC+ID-1)
> summary(lm.out) #note the point estimates

Call:
lm(formula = weight ~ IA + IB + IC + ID - 1)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.38286 -0.05625 -0.00625  0.12000  0.38714 

Coefficients:
   Estimate Std. Error t value Pr(>|t|)    
IA  3.80286    0.07153   53.16   <2e-16 ***
IB  3.43000    0.06691   51.26   <2e-16 ***
IC  3.59833    0.07726   46.57   <2e-16 ***
ID  3.93625    0.06691   58.83   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 0.1893 on 25 degrees of freedom
Multiple R-squared:  0.9977,	Adjusted R-squared:  0.9974 
F-statistic:  2771 on 4 and 25 DF,  p-value: < 2.2e-16 

>                  #and that the usual F test is testing the wrong thing
> 
> #now carry out F TEST FOR GENERAL LINEAR HYPOTHESIS
> T=matrix(c(1,-1,0,0,
+            0,1,-1,0,
+            0,0,1,-1),byrow=T,ncol=4); T

     [,1] [,2] [,3] [,4]
[1,]    1   -1    0    0
[2,]    0    1   -1    0
[3,]    0    0    1   -1

> b=c(0,0,0)
> numSS=t(T%*%bhat-b)%*%solve(T%*%solve(t(X)%*%X)%*%t(T))%*%(T%*%bhat-b)
> numeratorMS=numSS/dim(T)[1]
> Fobs=numeratorMS/MSE
> c(numSS,numeratorMS,Fobs)

[1]  1.1649036  0.3883012 10.8413906

> #CHECK
> summary(lm(weight~IA+IC+ID))

Call:
lm(formula = weight ~ IA + IC + ID)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.38286 -0.05625 -0.00625  0.12000  0.38714 
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.43000    0.06691  51.262  < 2e-16 ***
IA           0.37286    0.09795   3.807 0.000813 ***
IC           0.16833    0.10221   1.647 0.112079    
ID           0.50625    0.09463   5.350 1.51e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1893 on 25 degrees of freedom
Multiple R-squared:  0.5654,	Adjusted R-squared:  0.5132 
F-statistic: 10.84 on 3 and 25 DF,  p-value: 9.502e-05
```