Testing hypotheses using the lm and anova commands

- We are comparing two models, a *full model*, and a *reduced model*.
- The reduced model is obtained from the full model by leaving out certain predictor variables, or equivalently, setting certain parameters to 0.

  example:

  - full model

    $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{k-2} x_{k-2} + \beta_{k-1} x_{k-1} + \beta_k x_k + \epsilon.$
  - reduced model

    $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{k-2} x_{k-2} + \epsilon.$
  - The reduced model is obtained from the full model by removing the predictors x_{k-1} and x_k, or equivalently, setting $\beta_{k-1} = \beta_k = 0$.
- Fit the models using the *lm* command.

  - full model: $lmfull = lm(y \sim x_1 + x_2 + ... + x_{k-2} + x_{k-1} + x_k)$
  - reduced model: $lmred = lm(y \sim x_1 + x_2 + ... + x_{k-2})$
- Use the *anova* command to compare the full and reduced models. Formally, this is a test of the $H_0 : \beta_{k-1} = \beta_k = 0$ against the alternative, H_A : at least one of β_{k-1}, β_k is non-zero.

$$anova(lmfull, lmred)$$

- report the associated F statistic and $p - value$.

```
  rm(list=ls())
  data=read.csv("http://chase.mathstat.dal.ca/~bsmith/stat3340/Data/cement.csv")
  attach(data)
> summary(data)
       i             y               x_1              x_2             x_3       
 Min.   : 1   Min.   : 72.50   Min.   : 1.000   Min.   :26.00   Min.   : 4.00  
 1st Qu.: 4   1st Qu.: 83.80   1st Qu.: 2.000   1st Qu.:31.00   1st Qu.: 8.00  
 Median : 7   Median : 95.90   Median : 7.000   Median :52.00   Median : 9.00  
 Mean   : 7   Mean   : 95.42   Mean   : 7.462   Mean   :48.15   Mean   :11.77  
 3rd Qu.:10   3rd Qu.:109.20   3rd Qu.:11.000   3rd Qu.:56.00   3rd Qu.:17.00  
 Max.   :13   Max.   :115.90   Max.   :21.000   Max.   :71.00   Max.   :23.00  
      x_4       
 Min.   : 6.00  
 1st Qu.:20.00  
 Median :26.00  
 Mean   :30.00  
 3rd Qu.:44.00  
 Max.   :60.00  


#Full model includes x_1,x_2,x_3,x_4
#Example 1: reduced model leaves out x_3, testing H0:beta3=0

lm.out=lm(y~x_1+x_2+x_3+x_4)
lm.out2=lm(y~x_1+x_2+x_4)
anova(lm.out,lm.out2)

Analysis of Variance Table

Model 1: y ~ x_1 + x_2 + x_3 + x_4
Model 2: y ~ x_1 + x_2 + x_4
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      8 47.864                           
2      9 47.973 -1  -0.10909 0.0182 0.8959


#Example 2: reduced model leaves out x_2 and x_3, testing H0:beta2=beta3=0
```

```
lm.out3=lm(y~x_1+x_4)
anova(lm.out,lm.out3)

Analysis of Variance Table

Model 1: y ~ x_1 + x_2 + x_3 + x_4
Model 2: y ~ x_1 + x_4
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      8 47.864
2     10 74.762 -2   -26.898 2.2479  0.168



#Example3: reduced model leaves out x_1, x_2, x_3, and x_4.
#testing H0: beta1 = beta2 = beta3 = beta4
#Reduced model only includes an intercept.

lm.out4=lm(y~1)
anova(lm.out,lm.out4)

 Analysis of Variance Table

Model 1: y ~ x_1 + x_2 + x_3 + x_4
Model 2: y ~ 1
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1      8   47.86
2     12 2715.76 -4   -2667.9 111.48 4.756e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
summary(lm.out)
Call:
lm(formula = y ~ x_1 + x_2 + x_3 + x_4)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1750 -1.6709  0.2508  1.3783  3.9254

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.4054    70.0710   0.891   0.3991
x_1           1.5511     0.7448   2.083   0.0708 .
x_2           0.5102     0.7238   0.705   0.5009
x_3           0.1019     0.7547   0.135   0.8959
x_4          -0.1441     0.7091  -0.203   0.8441
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.446 on 8 degrees of freedom
Multiple R-squared:  0.9824,    Adjusted R-squared:  0.9736
F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07
```

R code for a few related examples:

```
data=read.csv("http://chase.mathstat.dal.ca/~bsmith/stat3340/Data/cement.csv")
y=data$y
x=data$x_3
z=data$x_4

#outcome variable is y
#2 predictor variables, x and z
lm0=lm(y~1) #fit model with no predictors
summary(lm0)

lm1=lm(y~x) #fit model with x only
summary(lm1)
anova(lm0,lm1)
lm2=lm(y~z) #z only
lm3=lm(y~x+z) #x+z
lm4=lm(y~x+z+x:z) #x+z+interaction
lm5=lm(y~x+z+I(x^2)+I(z^2)) #with quadratic terms
lm6=lm(y~x+z++x:z+I(x^2)+I(z^2)) #with quadratic terms and interaction of linear t

anova(lm3,lm0) #overall F test for model with x and z
anova(lm4,lm3) #F test of interaction in linear model
anova(lm4,lm6) #test of quadratic effects
anova(lm5,lm6) #test of interaction in quadratic model
anova(lm6,lm0) #overall F test for quadratic model with interaction
```