6 F-tests

If you look at `summary(outBC)` you'll see the lines

```
Residual standard error: 0.4931 on 42 degrees of freedom
F-statistic: 45.47 on 5 and 42 DF,  p-value: 6.974e-16.
```

What does that mean?

Remember that $F_{\ell,k}$ is the distribution of $(S_\ell^2/\ell)/(S_k^2/k)$ when $S_k^2 \sim \chi_k^2$ independently of $S_\ell^2 \sim \chi_\ell^2$.

For $F_{5,42}$ you already know where the χ_{42}^2 comes from: it is RSS/σ^2.

Under the model $y \sim N(\mu, \sigma^2 I_n)$ with $\mu \in \mathcal{X}$, a p-dimensional subspace, the fitted vector $\widehat{y}$ is independent of RSS. Let H_0 denote the matrix for orthogonal projection onto the $(p-1)$-dimensional subspace $\mathcal{X}_0$ of $\mathcal{X}$ that is orthogonal to $\mathbb{1}$. The component of y in $\mathcal{X}_0$ equals

$$H_0 y = \widehat{y} - \overline{y}\mathbb{1} = H_0\mu + H_0\xi \sim N(H_0\mu, \sigma^2 H_0).$$

If $H_0\mu = 0$ then $\|H_0 y\|^2/\sigma^2 \sim \chi_{p-1}^2$ and

$$\text{F-stat} = \frac{\|H_0 y\|^2/(p-1)}{RSS/(n-p)} \sim F_{p-1,n-p}.$$

For the BC data,

```r
extra <- outBC$fit - mean(BC$rate)
RSS <- sum((outBC$res)^2)
Xdim <- outBC$rank
nn <- length(outBC$res)
Fratio <- ( sum(extra^2)/(Xdim-1) )/ ( RSS/(nn-Xdim) )
pvalue <- 1 - pf(Fratio,Xdim - 1, nn - Xdim)
print(c(Fratio,pvalue))

## [1] 4.547233e+01 6.661338e-16
```

Hmmm! The p-value is a bit different from the value in the summary. Before we get too anguished about a difference in two very small quantities, let me look at the actual F-statistics:

```r
print(summary(outBC)$fstat)

##    value    numdf    dendf
## 45.47233  5.00000 42.00000
```

I have little interest in quibbling about how close a quantity of order 10^{-16} is to zero.

It would be highly surprising if the `F-statistic` were not very close to zero. If it were not significant we would be inclined to think that the data were just random noise around some constant level. Other F-statistics are sometimes more interesting.

For the general case, the same logic works for any subspace of $\mathcal{X}$. If H_0 projected $\mathbb{R}^n$ orthogonally onto some k-dimensional subspace $\mathcal{X}_0$ of $\mathcal{X}$, and if $H_0\mu$ were zero, then we would have

$$\|H_0 y\|^2 / \sigma^2 \sim \chi^2_k$$

and

$$\frac{\|H_0 y\|^2 / k}{RSS/(n-p)} \sim F_{k,n-p}.$$

For example, for the BC data, suppose $\mathcal{X}_0$ were the 3-dimensional subspace of $\mathcal{X}$ orthogonal to span(F). A zero component of μ in that subspace would suggest that the `Hp` factor was not contributing anything significant to the `outBC$fit`. For the BC data, such an idea is not particularly plausible, but for the sake of illustration let me show you the relevant F-statistic.

To make things a bit more interesting I'll kill the first six observations, to make sure the data set is not balanced. **R** has a special function for dealing with this sort of F-test:

```
anova( lm(rate ~ Ht + Hp, BC, subset = -(1:6)) )

## Analysis of Variance Table
##
## Response: rate
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Ht         3 28.5203  9.5068  38.801 2.257e-11 ***
## Hp         2 26.6403 13.3201  54.365 1.328e-11 ***
## Residuals 36  8.8204  0.2450
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova( lm(rate ~ Hp + Ht, BC, subset = -(1:6)) )

## Analysis of Variance Table
```

```
##
## Response: rate
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Hp         2 38.456 19.2280  78.478 7.503e-14 ***
## Ht         3 16.705  5.5682  22.726 1.978e-08 ***
## Residuals 36  8.820  0.2450
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The difference in the anova summaries arises because of the lack of balance that I created. The summaries depend on which factor appears first in the model formula. The first anova summary refers to the effect of `Hp` after `1+Ht` the second to the effect of `Ht` after `1+Hp`. With a balanced design, it makes no difference.

```
anova( lm(rate ~ Ht +Hp, BC) )

## Analysis of Variance Table
##
## Response: rate
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Ht         3 20.414  6.8048  27.982 4.192e-10 ***
## Hp         2 34.877 17.4386  71.708 2.865e-14 ***
## Residuals 42 10.214  0.2432
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova( lm(rate ~ Hp +Ht, BC)    )

## Analysis of Variance Table
##
## Response: rate
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Hp         2 34.877 17.4386  71.708 2.865e-14 ***
## Ht         3 20.414  6.8048  27.982 4.192e-10 ***
## Residuals 42 10.214  0.2432
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In any case, both factors appear to be making a significant contribution to the prediction.