CALCULATING POWER OF F-TEST

1. Some Theory

We have our usual linear model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where it is assumed that $\boldsymbol{\varepsilon}$ is independent of $\boldsymbol{X}$ and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I)$.

We want to consider testing

$$H_0 : \beta_{r+1} = \cdots = \beta_k = 0\,. \tag{1}$$

Let

$$V_0 = \mathscr{L}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r)$$
$$V_1 = \mathscr{L}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)\,.$$

and if $\widehat{\boldsymbol{Y}^{(i)}} = \Pi_{V_i}\boldsymbol{Y}$, then

$$\mathrm{RSS}_i = \sum_{j=1}^{n}(Y_j - \widehat{Y_j^{(i)}})^2 = \|\boldsymbol{Y} - \widehat{\boldsymbol{Y}^{(i)}}\|^2\,.$$

We know that under H_0,

$$F = \frac{\|\widehat{\boldsymbol{Y}^{(0)}} - \widehat{\boldsymbol{Y}^{(1)}}\|^2/(k-r)}{\|\boldsymbol{Y} - \widehat{\boldsymbol{Y}^{(1)}}\|^2/(n-k)} = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(k-r)}{\mathrm{RSS}_1/(n-k)}$$

has an F-distribution with $k-r, n-k$ degrees of freedom.

To carry out this test in R, fit two models and use `anova` to compute the F-statistic. See Table 1 for an example.

```
library(xtable)
boston = read.table(url("http://pages.uoregon.edu/dlevin/DATA/BostonB.txt"),
                    header=T)
g1 = lm(medv~.,data=boston)
g2 = lm(medv~.-age-indus,data=boston)
xtable(anova(g2,g1),caption="F test on coefficients of age and industry",
       label="Tab:F")
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	494	11081.36				
2	492	11078.78	2	2.58	0.06	0.9443

Table 1. F test on coefficients of age and industry

If we want to compute probabilities for F without the assumption of H_0 (e.g., to compute the power), we need:

Theorem 1. *The distribution of F is a* non-central *F-distribution with degrees of freedom $k-r$ and $n-k$ and with* non-centrality parameter

$$\delta = \frac{\|\Pi_{V_1 \cap V_0^\perp} \boldsymbol{\theta}\|^2}{\sigma^2},$$

where $\boldsymbol{\theta} = \mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}]$.

Note that

$$\Pi_{V_1 \cap V_0^\perp} \boldsymbol{\theta} = \sum_{j=r+1}^{k} \beta_j (\boldsymbol{x}_j - \Pi_{V_0} \boldsymbol{x}_k) = \sum_{j=j+1}^{k} \beta_k \boldsymbol{x}_j^\perp,$$

where $\boldsymbol{x}_j^\perp = \boldsymbol{x}_j - \Pi_{V_0} \boldsymbol{x}_k$.

Thus,

$$\delta = \frac{1}{\sigma^2} \left[\sum_{j=r+1}^{k} \beta_j^2 \|\boldsymbol{x}_j^\perp\|^2 + 2 \sum_{r+1 \le i < j \le k} \beta_i \beta_j \langle \boldsymbol{x}_i^\perp, \boldsymbol{x}_j^\perp \rangle \right].$$

Question 1. The parameter δ is "unitless", i.e. it does not depend on the units of any of the variables. Why?

To calculate δ, we need to calculate $\boldsymbol{x}_j^\perp$. There are several ways to do this. Note that $\boldsymbol{x}_j^\perp$ is the vector of residuals when performing OLS of $\boldsymbol{x}_j$ against $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r$. (Why?)

Suppose $\beta_{\text{indus}} = 0.05$ and $\beta_{\text{age}} = 0.001$. Let us find δ:

```
ageperp = residuals(lm(age~.-medv-indus,data=boston))
indusperp = residuals(lm(indus~.-medv-age,data=boston))
de = (0.05^2*sum(ageperp^2) + 0.001^2*sum(indusperp^2)
      + 2*0.001*0.05*sum(ageperp*indusperp))/summary(g1)$sigma^2
```

Now we find the cut-off for the F-test, say with significance level 0.05:

```
fstar = qf(0.95,2,492)
```

To find the power:

```
1 - pf(fstar,2,492,ncp = de)
## [1] 0.9322011
```

Question 2. The design matrix $\boldsymbol{X}$ is said to have *collinearity* if there are *near* linear relationships among the columns. Why is the power of the F-test limited when the variables $\boldsymbol{x}_j$ for $j > r$ are nearly collinear with the variables $\boldsymbol{x}_j$ for $j \le r$.

2. An example

This example concerns the `crime` dataset, available via

```
crime = read.table(url("http://pages.uoregon.edu/dlevin/DATA/crime.txt"),header=T)
```

A description of this data set is at
http://pages.uoregon.edu/dlevin/DATA/USCrimeDatafile.html
Here, the mortality rate `R` is modelled as a function of the other variables.
First, use OLS to fit a model including all the variables.

Question 3. Run a `summary` of the OLS model. Note the most of the coefficients are "not significant". Test the hypothesis that none of the "not significant" coefficients are non-zero. What do you find? Discuss.

Âŕ

Question 4. In the first `summary` of the full OLS model, the coeffiecients of `Ex0` and `Ex1` have different signs. Looking at the description of the variables, does this make sense? How would you explain this?

Question 5. In the above question, running the `summary` command performed several tests. What were these tests? Afterwards, the question asked you to perform another test, *based on the results* given in `summary`. Does this matter in interpreting the results of the second test?

Now, consider the test that the coefficients of both `Ex1` and `U1` are zero.

Question 6. What is the power of this test when $\beta_{\mathrm{Ex1}} = -1$ and $\beta_{\mathrm{U1}} = -1$.

First, carry out the F-test that these two coefficients are zero.