Testing for groups of variables with the partial F-test

Stats 203, Winter 2024

Lecture 9 [Last update: February 11, 2024]

1 Testing for a significant group of variables

The T-test tests the null hypothesis $H_0 : \beta_j = 0$ which can be interpreted as saying that the jth predictor does not add anything to a linear model with all the other predictors included. The F-test tests the null hypothesis $H_0 : \beta_1 = \ldots = \beta_p = 0$ which can be interpreted as saying that all of the predictors collectively add nothing to a model with only the intercept included. The partial F-test tests hypotheses between these two extremes.

Specifically, for $0 \leq q < p$, suppose we are interested in testing hypotheses of the form

$$H_0 : \beta_{q+1} = \ldots = \beta_p = 0, \quad \text{vs.} \quad H_1 : \beta_j \neq 0 \ \text{ for some } j \in q+1, \ldots, p. \tag{1}$$

The null hypothesis is that a group of variables do not "add" to the model, once the rest of the variables are accounted for. An example application is multiomics data, where the response is some phenotype of interest – say, recovery time after surgery – and the predictors are various biological markers that can be divided into different categories: genomics, proteomics, transcriptomics, etc. Predictors in a certain category (say, genomic data) may be expensive to collect and we want to know if they add anything once predictors from other categories are included.

For notational convenience, we assume that the variables with 0 slope under the null hypothesis belong to the "last" columns of $\mathbf{X}$. But everything we will say extends to testing whether any group of predictors have zero slope, simply by relabeling the columns. So, for example, we can use the partial F-test to test whether $\beta_1 = \beta_3 = 0$.

2 A more general null model

The null hypothesis in (1) is that we have independent observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ for which

$$Y_i = \gamma_0 + \gamma_1 X_{i1} + \ldots + \gamma_q X_{iq} + \epsilon_i, \epsilon_i \perp\!\!\!\perp X_i, \epsilon_i \sim N(0, \sigma^2). \tag{2}$$

The ordinary least squares estimates of the parameters in (2) are

$$\hat{\gamma} = (\mathbf{X}_{0:q}^\top \mathbf{X}_{0:q})^{-1} \mathbf{X}_{0:q}^\top Y.$$

where $\mathbf{X}_{0:q} \in \mathbb{R}^{n \times (q+1)}$ picks out the first $(q+1)$ columns of $\mathbf{X}$. The fitted values can be written as the projection of Y onto the column space of $\mathbf{X}_{0:q}$:

$$\mathbf{X}_{0:q}\hat{\gamma} = \mathbf{H}_{0:q} Y,$$

where $\mathbf{H}_{0:q} = \mathbf{X}_{0:q}(\mathbf{X}_{0:q}^\top \mathbf{X}_{0:q})^{-1}\mathbf{X}_{0:q}^\top$. The sum of the squared residuals in this null model is $\|Y - \mathbf{H}_{0:q}Y\|^2$. This is called the **partial sum of squares**.

Two special cases. Suppose $q = 0$. Then the null hypothesis is $H_0 : \beta_1 = \ldots = \beta_p = 0$. This is the null hypothesis tested by the (full) F-test, and in this case the partial F-statistic, defined below, reduces to just the (full) F-statistic.

Suppose $q = p - 1$. Then the null hypothesis is $H_0 : \beta_p = 0$. This is the added variable null hypothesis for predictor p. In this case some linear algebra will show that the partial F-statistic is the T-statistic *squared*.

3 Partial F test

As in the case of the (full) F-test, we evaluate the evidence against the null by comparing how much the sum of squared residuals decreases when the remaining predictors are included. Mathetmatically, the **partial F** statistic is

$$F_{0:q} = \frac{\|Y - \mathbf{H}_{0:q}Y\|^2 - \|Y - \mathbf{H}Y\|^2}{\hat{\sigma}^2} \cdot \frac{1}{p-q}.$$

To derive the distribution of the partial F statistic one proceeds similarly as in the case of the F statistic. I am not going through the details of this derivation as they are very similar to those for the F-statistic. But broadly speaking, there is a Pythogorean relation

$$\|\mathbf{H}Y - \mathbf{H}_{0:q}Y\|^2 = \|Y - \mathbf{H}_{0:q}Y\|^2 - \|Y - \mathbf{H}Y\|^2$$

Under the null hypothesis in (2), the left hand side can be rewritten as

$$\epsilon^\top(\mathbf{H} - \mathbf{H}_{0:q})\epsilon,$$

which has expectation $\sigma^2(p-q)$ and is distributed χ^2_{p-q}. So under the null hypothesis $F_{0:q}$ is the ratio of two independent chi-squared random variables, each divided by their degrees of freedom, and is thus distributed $F_{p-q,n-p-1}$. We reject the null hypothesis if the observed partial F-statistic is greater than the $(1-\alpha)$th quantile of this distribution.

Data analysis example: fuel consumption data

We try out the F-test with the fuel consumption data. Again we assume the linear model

$$\text{Fuel}_i = \beta_0 + \beta_1 \text{Tax}_i + \beta_2 \text{Dlic}_i + \beta_3 \text{Income}_i + \beta_4 \log(\text{Miles}_i) + \epsilon_i,$$

and compute the OLS coefficients using `lm`.

```
# linear regression
fuel.lm = lm(Fuel ~ Tax + Dlic + Income + logMiles, data = fuel)
summary(fuel.lm)
```

```
##
## Call:
## lm(formula = Fuel ~ Tax + Dlic + Income + logMiles, data = fuel)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -163.145  -33.039    5.895   31.989  183.499
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 154.1928   194.9062   0.791 0.432938
## Tax          -4.2280     2.0301  -2.083 0.042873 *
## Dlic          0.4719     0.1285   3.672 0.000626 ***
## Income       -6.1353     2.1936  -2.797 0.007508 **
## logMiles     18.5453     6.4722   2.865 0.006259 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64.89 on 46 degrees of freedom
## Multiple R-squared:  0.5105, Adjusted R-squared:  0.4679
## F-statistic: 11.99 on 4 and 46 DF,  p-value: 9.331e-07
```

The last line computes the F-statistic for the null hypothesis that an intercept-only model is sufficient, and $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$. The degrees of freedom are $p = 4$ and $n - p - 1 = 46$. Since this is a very restrictive null hypothesis, it is not a surprise that the p-value is extremely small.

Two predictors

We saw in Lecture 7 that when proportion of people with bachelor's degrees (`GradFreq`) was included in the model, neither that variable nor `Income` had significant T-statistics, due to their high degree of correlation. Now let us try running an F-test to see what happens when we drop both variables. The relevant command is `anova(fomega.lm,fOmega.lm)` where `fomega.lm` and `fOmega.lm` are the results of fitting two different nested linear models. This produces a table called an **ANOVA** table. ANOVA stands for analysis of variance since this is all the (partial) F-test is doing: comparing how much of the variance in Y can be "explained'' by different models. The table presents the components of the partial F statistic (partial and residual sums of square, and degrees of freedom) and test in a purportedly easy-to-read manner.

```
full.lm = update(fuel.lm, . ~ . + GradFreq) # full model with GradFreq
null.lm = update(fuel.lm, . ~ . - Income)   # null model with neither GradFreq nor Income
anova(null.lm,full.lm)
```

```
## Analysis of Variance Table
##
## Model 1: Fuel ~ Tax + Dlic + logMiles
## Model 2: Fuel ~ Tax + Dlic + Income + logMiles + GradFreq
```

```
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     47 226640
## 2     45 193244  2     33396 3.8884 0.02769 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Most important for our purposes is the ultimate p-value .028. We see that even though neither variable appears significant individually (after including the other), the combined effect of the two is significant.