Chapter 6: Test and Analysis of Variance

This chapter presents a different approach to testing based on comparing the fit of mean functions rather than comparing parameter estimates to hypothesized values. In linear regression, this

leads to F-tests. These are also called analysis of variance.

1 F-Tests

Suppose we have a response Y and a vector of p' regressors $\boldsymbol{X}' = (\boldsymbol{X}_1', \boldsymbol{X}_2')$ that we partition into two parts so that $\boldsymbol{X}_2$ has q regressors and $\boldsymbol{X}_1$ has the remaining $p' - q$ regressors. The general hypothesis test we consider is

$$NH : E(Y|\boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X}_2 = \boldsymbol{x}_2) = \boldsymbol{x}_1'\boldsymbol{\beta}_1,$$

$$AH : E(Y|\boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X}_2 = \boldsymbol{x}_2) = \boldsymbol{x}_1'\boldsymbol{\beta}_1 + \boldsymbol{x}_2'\boldsymbol{\beta}_2.$$

The regular formula for the test is

$$F = \frac{(RSS_{NH} - RSS_{AH})/(df_{NH} - df_{AH})}{RSS_{AH}/df_{AH}},$$

where RSS_{NH} and RSS_{AH} denote the residual sum of squares under the null model and the alternative model, respectively.

For the wool data example, we can consider testing

$$NH: log(cycles) \sim len + amp + load + len:amp + len:load,$$

$$AH: log(cycles) \sim len + amp + load + len:amp + len:load + amp:load.$$

Under NH, we have RSS=0.181 with df=12. Under AH, we have RSS=0.166 with df=8.

Therefore

$$F = \frac{(0.181 - 0.166)/(12 - 8)}{0.166/8} = 0.18.$$

Compared with the F(4,8) distribution, we have a p-value of 0.94.

The F-tests are applications of likelihood ratio tests to linear models with normal errors.

2 The Analysis of Variance

Suppose we fit the following model:

$$Y \sim A+B+C+A:B+A:C+B:C+A:B:C,$$

where each of A, B, or C could represent a continuous predictor with a single df, or a factor, polynomial, or spline basis with more than 1 df. An interaction like A:B can have many df.

The approach to testing we adopt follows from the *marginality principle*. A lower-order term, such as the A main effect, is never tested in models that include any of its higher-order relatives like A:B, A:C, or A:B:C. All regressors that are not high-order relatives of the regressor of interest, such as B, C, and B:C, are always included in both NH and AH.

Based on the marginality principle, testing should begin with the highest-order interaction first:

$$NH: Y \sim A + B + C + A:B + A:C + B:C,$$

$$AH: Y \sim A + B + C + A:B + A:C + B:C + A$$

If the A:B:C interaction is judged to be nonzero, no further testing is called for, since A:B:C is a higher-order relative of all remaining regressors in the mean function.

If the A:B:C interaction is judged nonsignificant, then proceed to examine the two-factor interactions, such as

$$NH : Y \sim A + B + C + A : C + B : C,$$

$$AH : Y \sim A + B + C + A : B + A : C + B : C,$$

which tests the interaction A:B.

Tests for a main-effect like A would be carried out only if all its higher-order relatives, A:B:C, A:B, and A:C, are judged to be unimportant. One would then test

$$NH : Y \sim B + C + B : C,$$

$$AH : Y \sim A + B + C + B : C,$$

where B:C is included in both the NH and the AH. Table 1 shows the analysis of variance for the UN data.

Table 1: ANOVA for the UN11 Data

```
Analysis of Variance Table

Response: lifeExpF
                           Df  Sum Sq Mean Sq  F value Pr(>F)
factor(group)               2 12563.0  6281.5 238.7563 <2e-16 ***
log(ppgdp)                  1  2639.8  2639.8 100.3376 <2e-16 ***
factor(group):log(ppgdp)    2    12.7     6.3   0.2409 0.7862
Residuals                 193  5077.7    26.3
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

An analysis of variable table derived under the marginality principle has the unfortunate name of Type II analysis of variance. At least two other types of ANOVA are commonly available in software packages:

Type I ANOVA, also called sequential ANOVA, fits model according to the order that the regressors are entered into the mean function. For ex-

ample, if we fit the model

$$Y \sim A+B+C+A:B+A:C+B:C+A:B:C,$$

then the sequence of models that would be represented in the ANOVA table would have regressors $\{A\}$, $\{A, B\}$, $\{A, B, C\}$, $\{A, B, C, A : B\}$, $\{A, B, C, A : B, A : C\}$, $\{A, B, C, A : B, A : C, B : C\}$ and $\{A, B, C, A : B, A : C, B : C, A : B : C\}$.

If the terms were written in a different order, then the analysis would have different conditioning. Type I ANOVA generally has only pedagogical interest and should not be used.

Type III ANOVA violates the marginally principle. It computes the test for every regressor adjusted for every other regressor. For example, the test for the A main effect would include the interactions A:B, A:C, and A:B:C in both NH and AH. There is a justification for this testing paradigm, called the marginal means method, but some of these tests depend on the parameterization used for the regressors and so they are not recommended

for general use.

The wool data is from a designed experiment in which all the factors are orthogonal to each other. Table 2 shows the ANOVA table for the full second-order model. Because the regressors are orthogonal, Type I, Type II, and Type III tests are identical.

Table 2: ANOVA for the wool data

```
Analysis of Variance Table
Response: log(cycles)
                          Df  Sum Sq Mean Sq  F value    Pr(>F)
factor(len)                2 12.516  6.258 301.744 2.930e-08 ***
factor(amp)                2  7.167  3.584 172.799 2.620e-07 ***
factor(load)               2  2.802  1.401  67.551 9.767e-06 ***
factor(len):factor(amp)    4  0.401  0.100   4.836   0.02806 *
factor(len):factor(load)   4  0.136  0.034   1.636   0.25620
factor(amp):factor(load)   4  0.015  0.004   0.176   0.94456
Residuals                  8  0.166  0.021
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

3 Power and Non-Null Distributions

For the full-reduced model test, the test statistic is

$$F = \frac{(RSS_{NH} - RSS_{AH})/(df_{NH} - df_{AH})}{RSS_{AH}/df_{AH}},$$

For a fixed significance level, the probability of rejecting a false NH is called the power of the test:

$$\begin{aligned} power &= Prob(\text{detect a false NH}) \\ &= Prob(F > f^* | \text{AH is true}), \end{aligned}$$

where f^* denotes the critical value of the test.

When AH is true, the numerator and denominator of the test statistic remain independent. The denominator estimates σ^2 under both the NH and the AH. The distribution of the numerator sum of squares is different under the NH and the AH. Apart from df, the numerator under the AH is distributed as σ^2 times a noncentral χ^2.

The expected value of the numerator will be

$$\sigma^2(1+\lambda),$$

where λ is called the noncentrality parameter and can be expressed as

$$\lambda = \frac{\boldsymbol{\beta}_2' \boldsymbol{X}_2' (I - \boldsymbol{X}_1 (\boldsymbol{X}_1 \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1') \boldsymbol{X}_2 \boldsymbol{\beta}_2}{q\sigma^2}.$$

In the case that X_2 consists of a single variable, i.e., $q = 1$, $\boldsymbol{\beta}_2$ is a scalar and

$$\lambda = (n-1)\left(\frac{\beta_2}{\sigma}\right)^2 SD_2^2(1 - R^2_{x_2,x_1}),$$

where SD_2 is the standard deviation of X_2, and $R^2_{x_2,x_1}$ is the value of R^2 for the OLS regression with response X_2 and regressor $\boldsymbol{X}_1$.

Power increases with λ, so it increases with sample size n, the size of the parameter relative to the error standard deviation $(\beta_2/\sigma)^2$, and SD_2^2.

In most designed experiments, interesting tests concern effects that are orthogonal. Then

$$\lambda = (n-1)\frac{\boldsymbol{\beta}_2'\boldsymbol{S}_2\boldsymbol{\beta}_2}{q\sigma^2},$$

where $\boldsymbol{S}_2$ is the sample covariance matrix for $\boldsymbol{X}_2$.

4 Wald Tests

Wald tests about regression coefficients are based on the distribution of the estimate $\hat{\boldsymbol{\beta}}$. In most regression problems, the estimator is at least approximately normally distributed,

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \boldsymbol{V}).$$

Generally, $\boldsymbol{V}$ is unknown, but an estimate $\hat{\boldsymbol{V}}$ is available. For OLS estimators, we have

$$\hat{\boldsymbol{V}} = \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

4.1 One Coefficient

To test a hypothesis, say NH: $\beta_j = \beta_{j0}$ versus AH: $\beta_j \neq \beta_{j0}$, compute

$$t = (\hat{\beta}_j - \beta_{j0})/\sqrt{\hat{v}_{jj}},$$

where $\hat{v}_{jj}$ is the (j,j)th element of $\hat{\boldsymbol{V}}$. This test is compared with the t-distribution with df equal to the df in estimating σ^2 to get p-values.

In problems like logistic regression in which there is no σ^2 to estimate, the Wald test is compared with the standard normal distribution.

4.2 One Linear Combination

Suppose $\boldsymbol{a}$ is a vector. Then the linear combination $l = \boldsymbol{a}'\boldsymbol{\beta}$ has estimate $\hat{l} = \boldsymbol{a}'\hat{\boldsymbol{\beta}}$ and

$$\hat{l} \sim N(l, \boldsymbol{a}'\boldsymbol{V}\boldsymbol{a}).$$

Therefore, for NH: $l = l_0$, the statistic is

$$t = (\hat{l} - l_0)/\sqrt{\boldsymbol{a}'\hat{\boldsymbol{V}}\boldsymbol{a}},$$

which is compared with the t-distribution with df given by the df for $\hat{\sigma}^2$.

Suppose we wish to test NH: $\boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{c}$ versus AH: $\boldsymbol{L}\boldsymbol{\beta} \neq \boldsymbol{c}$, where $\boldsymbol{L}$ is an $q \times p'$ matrix of constants.

The test statistic is

$$F = \frac{(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c})'(\boldsymbol{L}\hat{\boldsymbol{V}}\boldsymbol{L}')^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c})}{q}.$$

Under NH and normality, this statistic can be compared with an $F(q, n - p')$ distribution to get significance levels.

4.4 Equivalence of Wald and Likelihood-Ratio Tests

For linear regression, the Wald tests and the likelihood ratio tests give the same answer for any fixed hypothesis test. This equality does not carryover to other regression settings like logistic regression. Wald and likelihood ratio tests for logistic regression are equivalent, in the sense that for large enough samples they will give the same inference, but not equal, as the computed statistics

generally have different values. Likelihood ratio tests are generally preferable.

5 Interpreting Tests

5.1 Interpreting p-values

Under the appropriate assumptions, the p-value is the conditional probability of observing a value of the computed statistic as extreme or more extreme than the observed value, given that the NH is true. A small p-value provides evidence against the NH.

In many research areas it has become traditional to adopt a fixed significance level when examining p-values. The most common choice for the significance level is $\alpha = 0.05$, which would mean that, were the NH to be true, we would incorrectly find evidence against it about 5% of the time, or about one test in 20.

There is an important distinction between statistical significance, the observation of a sufficiently small p-value, and scientific significance, observing an effect of sufficient magnitude to be meaningful. Judgment of the latter usually will require examination of more than just the p-value.

5.2 Multiple Testing

Multiple testing is one of the most important problems with interpreting tests. If 100 independent tests are done, each at level $\alpha = 0.05$, even if NH is true in all 100 tests, then about $100 \times 0.05 = 5$ tests are expected to be "significant at the 5% level" and therefore false discoveries.

Traditional methods of surviving multiple testing are to control the *family-wise error rate* rather than the *per-test error rate*, but recent methodology is based on controlling the *false discovery rate*. Except for testing for outliers, we leave discussion and application of multiple testing methods to other sources.