SOME ANALYSES OF VARIANCE

The General Linear Hypothesis

Let $Y = X\beta + W$ with Y, X, β, W of dimensions $n\times 1$, $n\times p$, $p\times 1$, $n\times 1$ respectively and the entries of W being $IN(0,\sigma^2)$ Suppose $r(X) = r$.

Consider the (composite) null hypothesis $H_0 : P^T\beta = 0$ with $r(P) = q$.

Let $\hat\beta_*$ denote the estimate of β when H_0 holds. By a Lagrange multiplier argument it is seen to satisfy

$$X^T X\hat\beta_* + P\lambda = X^T Y$$
$$P^T\hat\beta_* = 0$$

Because $X\hat\beta_*$, $X(\hat\beta-\hat\beta_*)$ and $Y-X\hat\beta$ are mutually orthogonal one has the anova identity

$$|\, Y^T Y \,|^2 = |\, X\hat\beta_* \,|^2 + |\, X(\hat\beta - \hat\beta_*) \,|^2 + |\, Y - X\hat\beta \,|^2 \qquad (1)$$

with degrees of freedom breakdown

$$n = (r-q) + q + (n-r)$$

The components in (1) might be called "Total", "Reduced model", "Hypothesis", "Residual" respectively. The null hypothesis may be examined via

$$F = \frac{|\, X(\hat\beta - \hat\beta_*) \,|^2/q}{|\, Y - X\hat\beta \,|^2/(n-r)}$$

which, under H_0, has an F distribution with degrees of freedom q and $n-r$.

Some particular cases.

1. Single Sample

Consider the model $Y_i = \mu + W_i$ for $i=1,...,n$ and the null hypothesis $H_0 : \mu = 0$.

The identities become

$$\sum_i Y_i^2 = n\overline{Y}^2 + \sum_i (Y_i - \overline{Y})^2 \qquad (2)$$

$$n = 1 + (n-1)$$

and the F-statistic

$$F = \frac{n\overline{Y}^2}{S^2} = t^2$$

2. d-sample/Single Factor

Consider the model $Y_{ki} = \mu_k + W_{ki}$ for $i=1,...,n_k$ and $k=1,...,d$ and the null hypothesis $H_0 : \mu_k = \mu$ for all k. Writing $\mu_k = \mu + \beta_k$ for $k=1,...,d$ with $\beta_1=0$ the hypothesis become $H_0 : \beta_k = 0$ for $k=2,...,d$.

The identities become

$$\sum_k \sum_i Y_{ki}^2 = \sum_k \sum_i \overline{Y}^2 + \sum_k \sum_i (\overline{Y}_k - \overline{Y})^2 + \sum_k \sum_i (Y_{ki} - \overline{Y}_k)^2 \qquad (3)$$

$$n = 1 + (d-1) + (n-d)$$

and the F-statistic

$$F = \frac{\sum_k \sum_i (\overline{Y}_k - \overline{Y})^2 / (d-1)}{\sum_k \sum_i (Y_{ki} - \overline{Y}_k)^2 / (n-d)}$$

3. Single Factor With Covariate

Consider the model $Y_{ki} = \mu_k + \gamma(x_{ki} - \overline{x}) + W_{ki}$ for $i=1,...,n_k$ and $k=1,...,d$. Consider the null hypothesis $H_0 : \mu_k = \mu$ for $k=1,...,d$.

The identities become

$$\sum_k \sum_i Y_{ki}^2 = n\overline{Y}^2 + \hat{\gamma}_*^2 \sum_k \sum_i (X_{ki} - \overline{X})^2 + \sum_k \sum_i [\overline{Y}_k + \hat{\gamma}(X_{ki} - \overline{X}_k) - \overline{Y} - \hat{\gamma}_* (X_{ki} - \overline{X})]^2 + \sum_k \sum_i [Y_{ki} - \overline{Y}_k - \hat{\gamma}(X_{ki} - \overline{X}_k)]^2$$

$$n = 1 + 1 + (d-1) + (n-d-1)$$

The F-statistic is the ratio of the last two terms in (4) standardized by their degrees of freedom.

4. Two Factors With Equal Replicates

Consider the model $Y_{jki} = \mu_{jk} + W_{jki}$ for $j=1,...,J$ $k=1,...,K$ and $i=1,...,I$. A basic identity here is

$$\sum_j\sum_k\sum_i Y_{jki}^2 = \sum_j\sum_k\sum_i \overline{Y}_{..}^2 + \sum_j\sum_k\sum_i(\overline{Y}_{j.}-\overline{Y}_{..})^2 + \sum_j\sum_k\sum_i(\overline{Y}_{.k}-\overline{Y}_{..})^2 + \sum_j\sum_k\sum_i(\overline{Y}_{jk}-\overline{Y}_{j.}-\overline{Y}_{.k}+\overline{Y}_{..})^2 + \sum_j\sum_k\sum_i(Y_{jki}-\overline{Y}_{jk})^2$$

Also

$$IJK = n = 1 + (J-1) + (K-1) + (J-1)(K-1) + I(J-1)(K-1)$$

Writing $\mu_{jk}=\mu+\alpha_j+\beta_k+\gamma_{jk}$ with $\alpha_1,\beta_1,\gamma_{j1},\gamma_{1k} = 0$ the null hypothesis of no interaction is $H_0 : \gamma_{jk} = 0$. The F-statistic is the ratio of the last two terms in (5) standardized by their degrees of freedom.

5. Two Factors, One Replicate

Consider the model $Y_{jk} = \mu+\alpha_j+\beta_k+W_{jk}$ for $j=1,..,J$ $k=1,...,K$ with $\alpha_1,\beta_1=0$. Consider the null hypotheses $H_0 : \alpha_j=0$ and $H_0' : \beta_k=0$. A basic identity is

$$\sum_j\sum_k Y_{jk}^2 = \sum_j\sum_k \overline{Y}_{..}^2 + \sum_j\sum_k(\overline{Y}_{j.}-\overline{Y}_{..})^2 + \sum_j\sum_k(\overline{Y}_{.k}-\overline{Y}_{..})^2 + \sum_j\sum_k(Y_{jk}-\overline{Y}_{j.}-\overline{Y}_{.k}+\overline{Y}_{..})^2 \qquad (6)$$

Also

$$JK = 1 + (J-1) + (K-1) + (J-1)(K-1)$$

The F-statistic for H_0' is the ratio of the last two terms in (6) standardized by their degrees of freedom.

References

Fraser, D. A. S. (1958). Statistics: An Introduction. Wiley, New York.

John, P. W. M. (1971). Statistical Design and Analysis of Experiments. MacMillan, New York.

Kendall, M. G. and Stuart, A. (1966). The Advanced Theory of Statistics, Vol. 3. Griffin, London.

Neter, J. and Wasserman, W. (1974). Applied Linear Statistical Models. Irwin, Homewood.

Rao, C. R. (1973). Linear Statistical Inference and Its Applications. Wiley, New York.

Scheffe, H. (1959). The Analysis of Variance. Wiley, New York.

Searle, S. R. (1971). Linear Models. Wiley, New York.

Sen, A. K. and Srivastava, M. (1990). Regression Analysis. Springer, New York.

Wilks, S. S. (1962). Mathematical Statistics. Wiley, New York.