Handout on Factors and Contrasts, STAT 770

Eric Slud Spring 2024

Factors are objects in `R` with **contrasts** as attributes that specify the coding of dummy columns in the design matrices of linear and generalized-linear models. They rise particularly in the context of loglinear models because of the difference between the default 'side-conditions' in the loglinear model fitting function `loglin` versus the generalized-linear fitting function `glm` that can be used to fit the same models.

The `factors` themselves are simple: they are vectors of character entries, drawn from a set of K ordered labels called `levels`, coded numerically as `1:K`. Thus, we could define a factor to specify a sequence of 8 color-labels `c("blue","green","orange","green","red","blue","orange","orange")`, as

```
> colfac = factor(c("blue","green","orange","green","red","blue","orange","orange"))
```

with structure given by

```
> levels(colfac)
[1] "blue"   "green"  "orange" "red"
> as.numeric(colfac)
[1] 1 2 3 2 4 1 3 3
```

So the idea is that the numeric values are pointers to the ordered labels in `levels`.

Linear and categorical regressions can be fitted in `R` using factors as explanatory variables, e.g.

```
> yval = rnorm(8, 1 + 2*(0:7), 1.5)
> lmfit = lm(yval ~ colfac)
   lmfit
Call:
lm(formula = yval ~ colfac)

Coefficients:
 (Intercept)   colfacgreen  colfacorange     colfacred
      6.2698       -0.1001        4.1006        2.4058
```

What does this mean ? The `levels` (other than the first, which is by default associated with the intercept) are labels for the vector of color-category indicators, and these indicator vectors are 'dummy' predictor variables for the regression, with coefficients then fitted (in this example) by least squares. The convention that the intercept is the coefficient of the first category ("blue") and the other fitted coefficients are the least-squares coefficients for the dummy predictors is encoded in the `contrasts` associated with the `factor`:

```
 contrasts(colfac)
       green orange red
blue       0      0   0
green      1      0   0
orange     0      1   0
red        0      0   1
```

and the design matrix for the least-squares fit is `model.matrix(lmfit)`, identical to
`cbind(1, contrasts(colfac)[as.numeric(colfac),]`, and the estimates equivalently expressed as

```
> as.numeric(lm(yval ~ model.matrix(lmfit) - 1)$coef)
[1]  6.2698281 -0.1001303  4.1005673  2.4058138
```

1 Linear-Model Coding of Main Effects

We continue by explaining the relationship between factors, contrasts, and side-conditions, first for parameterization of linear models in terms of group means. Consider the group means

$$\mu_k = E(y_i \,|\, \texttt{fac}_i = \texttt{levels(fac)[k]}) \qquad \text{for} \quad k = 1, \ldots, K, \qquad \text{in linear model} \qquad y \sim \texttt{fac}$$

where `fac` is a K-level factor and observations y_i and factor values $\texttt{fac}_i$ are indexed by individuals i. The default 'treatment' contrasts encode the Intercept as the coefficient β_0 of the $\underline{1}$ vector and the coefficients $\beta_k,\ k = 2, \ldots, \beta_K$ as multiplying the respective indicator dummy-vectors `I[as.numeric(fac`$_i$`) ==k]` for $k = 2, \ldots, K$ with $\beta_1 = 0$ and with $\beta_0 = \mu_1$, $\beta_k - \beta_0 = \mu_k$ for $k \geq 2$. Here $\beta = (\beta_0, \beta_1, \ldots, \beta_K)$ is a redundant parameter vector with the 'side-condition' $\beta_1 = 0$; in the example $K = 4$. The relation between the coefficients and group means in this case can be seen, with `Group.1 = fac` and `x =`μ, in

```
> cbind.data.frame(aggregate(yval, by=list(colfac), mean),beta=lmfit$coef)
             Group.1         x       beta
(Intercept)     blue  6.269828  6.2698281
colfacgreen    green  6.169698 -0.1001303
colfacorange  orange 10.370395  4.1005673
colfacred        red  8.675642  2.4058138
```

Next consider the redundant $K+1$ dimensional parameter-vector γ with 'sum' side-condition $\sum_{k=1}^K \gamma_k = 0$ which implies that the parameter γ_1 is determined from the later coordinates by $\beta_1 = -\sum_{k=1}^K \beta_k$. The model-matrix that implements the model `y` $\sim$ `fac` can still be written with the same syntax using the $(K-1) \times K$ contrasts matrix `Smat = rbind(rep(-1,K), diag(K-1))`, which in our example with $K = 4$ is given as

```
> fac2 = colfac
   contrasts(fac2) = A = contr.sum(levels(colfac))[c(4,1:3),]
   A
  [,1] [,2] [,3]
4   -1   -1   -1
1    1    0    0
2    0    1    0
3    0    0    1
> lm(yval ~ fac2)$coef
(Intercept)       fac21       fac22       fac23
  7.8713908  -1.7016930   2.4990046   0.8042511
```

The coefficients shown are $\gamma_0, \gamma_2, \gamma_3, \gamma_4$, where the γ vector with side-condition $\sum_{k=1}^K \gamma_k$ is determined by $\gamma_k + \gamma_0 = \mu_k = \beta_0 + \beta_k$ for $k = 1, \ldots, K$, where $\beta_1 = \sum_{k=1}^K \gamma_k = 0$, so that $\gamma_k - \beta_k = \beta_0 - \gamma_0 = -1.601563$. In addition, by summing the relations $\mu_k = \gamma_0 + \gamma_k$ over $k = 1, \ldots, K$, we find $\gamma_0 = \mu_1 + \cdots + \mu_k$ and $K\gamma_k = (K-1)\mu_k - \sum_{j:j\neq k} \mu_j$.

One further example of contrast coding is **Helmert contrasts**, given in `R` for $K = 4$ levels by

```
> contr.helmert(4)
  [,1] [,2] [,3]
1   -1   -1   -1
2    1   -1   -1
3    0    2   -1
4    0    0    3
```

with an analogous pattern for larger K. Let the model matrix H_0 be defined with i'th row consisting of first element 1 and remaining $K-1$ elements `contr.helmert(K)[fac`$_i$`, ]`. Then using this matrix H as the dummy columns in a linear main-effect-only regression model `y` $\sim$ `fac` implies the group-mean coding

$\mu_k = H(h_0, h_2, \ldots, h_K)^{tr}$, where H is the $K \times K$ matrix `cbind(1, contr.helmert(K))`. It is easy to check by inverting H that $4h_0 = \mu_1 + \mu_2 + \mu_3 + \mu_4$ and

$$\mu_2 - \mu_1 = 2h_2, \;\; \mu_3 - \frac{1}{2}(\mu_1 + \mu_2) = 3h_3, \;\; \mu_4 - \frac{1}{3}(\mu_1 + \mu_2 + \mu_3) = 4h_4, \quad \cdots$$

Define the redundant parameter $h_1 = \mu_1$. Then it is easy to check that

$$\mu_1 = h_1, \;\; \mu_1+\mu_2 = 2(h_1+h_2), \;\; \mu_1+\mu_2+\mu_3 = 3(h_1+h_2+h_3), \;\; \mu_1+\mu_2+\mu_3+\mu_4 = 4(h_1+h_2+h_3+h_4), \;\; \cdots$$

and both in the special case $K = 4$ and for general case, $K\,h_0 = K\,(h_1 + h_2 + h_3 + h_4)$. Thus the general side-condition for the Helmert parameterization of dummy columns is $h_0 = \sum_{k=1}^{K} h_k$.

2 Contrasts specifying Linear Hypotheses for Group Means

In all the one-main-effect linear models discussed above with `y` $\sim$ `fac` and K-level factor, we found that a parameterization of categorical dummy columns could be motivated either with a $K \times (K-1)$ contrast matrix or a single linear side-condition connecting the entries of a $K+1$-entry parameter $(\theta_0, \theta_1, \ldots, \theta_K)$. The motivation for such a construction, in classical statistics as taught in STAT 740, is the convenience of being able to test one-or-two-sided linear hypotheses expressed in terms of group-means $\mu_1, \ldots, \mu_K$ in terms of a single parameter. Thus, in the treatment-contrasts case, the linear hypotheses of interest are the pairwise-equality hypotheses $H_{0,k}: \; \mu_k = \mu_1$ for $k = 2, \ldots, K$, and we can test these in the form $H_{0,k}: \; \beta_k = 0$. In the sum-contrasts case, we are interested in the linear hypotheses $H^*_{0,k}: \; \mu_k = \sum_{j:j\neq k} \mu_j/(K-1)$, and these are equivalently expressed as $H^*_{0,k}: \; \gamma_k = 0$. Finally, in the Helmert contrasts case, the linear hypotheses $H^\dagger_{0,k}: \; \mu_k = k^{-1} \sum_{j=1}^{k} \mu_j$ can be re-espressed as $H^\dagger_{0,k}: \; h_k = 0$.

The way to see which linear hypotheses a contrast matrix M parameterizes with successive parameters $\underline{\lambda} = (\lambda_0, \lambda_2, \ldots, \lambda_K)$ set to 0 is simply to translate the equalities

$$\underline{\mu} \;=\; \left(\underline{1} \,\middle|\, M\right) \underline{\lambda} \qquad \Longleftrightarrow \qquad \underline{\lambda} \;=\; \left(\underline{1} \,\middle|\, M\right)^{-1} \underline{\mu}$$

This is the connecting thread in the derivations of the parameters β, γ, h above: each can be found for the treatment, sum and Helmert contrast matrices M by the inversion above. For example,

```
> solve(cbind(1,contr.sum(4)))
         1     2     3     4
[1,]  0.25  0.25  0.25  0.25
[2,]  0.75 -0.25 -0.25 -0.25
[3,] -0.25  0.75 -0.25 -0.25
[4,] -0.25 -0.25  0.75 -0.25

> solve(cbind(1,contr.helmert(4)))
               1           2           3    4
[1,]  0.25000000  0.25000000  0.25000000 0.25
[2,] -0.50000000  0.50000000  0.00000000 0.00
[3,] -0.16666667 -0.16666667  0.33333333 0.00
[4,] -0.08333333 -0.08333333 -0.08333333 0.25
```

Now in fact, any of the linear hypotheses described above could be tested in terms of any of the parameterizations, within a linear model with independent mean-0 equal-variance errors, either with exact reference distributions when the errors are normally distributed or in the large-sample setting with general identically distributed errors. Any of the parameterizations imply (either exactly or via large-sample limit theorems) jointly normally distributed unbiased estimators $(\hat{\mu}_1, \ldots, \hat{\mu}_K)$ for $\underline{\mu}$ with variance matrix of the form $\sigma^2 V$ where V is known. Accordingly, linear hypotheses of the form $L^{tr}\underline{\mu} = 0$ can be tested using the standardized form of the statistic $L^{tr}\underline{\hat{\mu}}$, which involves the estimated variance $\hat{\sigma}^2\, L^{tr}\, V\, L$. So the reparameterization discussed in Section 1 is not necessary, but it can be very convenient.

3 Generalized-Linear Model Extensions of Categorical Dummy-Variable Coding

In categorical data analysis using generalized linear models, we are dealing with multinomial or Poisson data ($Y_a,\ a = 1, \ldots, A$) with a total of n trials and category probabilities π_a or rates $n_0\pi_a$ expressed through a link function g (usually the logarithm or logit) as $g^{-1}(X\,\beta)$ for some parameter β, where X is often a categorical random variable. Suppose for definiteness that $g = \log$ as in Poisson regression. Linear hypotheses expressed in terms of π_a are sometimes of interest, but more often the interesting hypotheses are expressed in terms of contrasts among $\log(\pi_a) = X\beta$. In that case, the considerations of the previous sections still suggest convenient reparameterizations in terms of treatment or sum contrasts, as long as the model in question contains main-effect terms only. However, when interactions are present, as they are almost always considered to be in loglinear models, only treatment or sum side-conditions are generally used, and it is convenient and important to understand how to transform between the parameterizations for the two types of side-conditions.

4 Treatment and Sum Side-Conditions for Interactions

As in the Lecture 20 Slide-deck for STAT 770, we consider multinomial data ($Y_a,\ a = 1, \ldots, A$) with $a \leftrightarrow (i, j, k, \ldots)$ corresponding to some multi-index for categorical factors $X, Z, W, \ldots$. We assume $\log \pi_a$ can be expanded linearly in terms of parameters $\lambda_i^X, \lambda_j^Z, \lambda_{ij}^{XZ}$, etc. with parameter-values structurally equal to 0 for all index-values 1. Equivalently, the same $\log \pi_a$ values can be expanded using parameters $\gamma_i^X, \gamma_j^Z, \gamma_{ij}^{XZ}$, etc. with each type of parameter γ summing to 0 across all possible values of any any of its indices, e.g. $\gamma_+^X = \gamma_+^Z = \gamma_{i+}^{XZ} = 0$.

The rule for transforming from γ parameters to λ parameters was indicated in a special case in the Lecture 20 slide-deck. It can be summarized more generally as follows:

- if X appears in no higher interaction, then $\lambda_i^X = \gamma_i^X - \gamma_1^X$
- if X appears in an interaction XZ, then λ_i^X includes an additional term $\gamma_{i1}^{XZ} - \gamma_{11}^{XZ}$
- if X appears in an interaction XZW, then λ_i^X includes an additional term $-\gamma_{i11}^{XZW} + \gamma_{11}^{XZW}$,

and so on. Similarly,

- if X, Z appears in an interaction XZ and nothing higher, then $\lambda_{ij}^{XZ} = \gamma_{ij}^{XZ} - \gamma_{1j}^{XZ} - \gamma_{i1}^{XZ} + \gamma_{11}^{XZ}$
- if X, Z appear in XZW, then λ_{ij}^{XZ} has an additional term $\gamma_{ij1}^{XZW} - \gamma_{i11}^{XZW} - \gamma_{1j1}^{XZW} + \gamma_{111}^{XZW}$,

and so on. This would be very laborious to code in general, but for loglinear models with interactions of order no higher than 3, it is do-able.