Linear Combinations

Linear combinations play a central role in linear models.

What are they?

Why are they important?

The mean of a linear combination.

The variance of a linear combination.

The covariance of two linear combinations.

Contrasts.

Weighted averages.

Orthogonal linear combinations.

A *linear combination* of $Y_1, \ldots, Y_n$ is a quantity of the form $a_1 Y_1 + \ldots + a_n Y_n$. The constants $a_1, \ldots, a_n$ are called coefficients or weights. A (scalar) *linear function* of the vector Y is a linear combination of $Y_1, \ldots, Y_n$. It is also common to call the linear combination $a_1 Y_1 + \ldots + a_n Y_n$ a *weighted sum* of $Y_1, \ldots, Y_n$.

The OLS estimator in the classical multiple linear regression model is

$$\hat{\beta} = (X^\top X)^{-} X^\top Y.$$

If we define $A = (X^\top X)^{-} X^\top$, we can write $\hat{\beta} = AY$. That is, each $\hat{\beta}_j$ is a linear function of Y, a linear combination of $Y_1, \ldots, Y_n$.

The fitted values are $\hat{\mu} = X\hat{\beta} = (XA)Y$. Each fitted value $\hat{\mu}_i$ is a linear combination of $Y_1, \ldots, Y_n$. The matrix $H = XA$ is the hat matrix.

The residuals are $R = Y - \hat{\mu} = (I - H)Y$, again, a linear function of Y. Each residual R_i is a linear combination of $Y_1, \ldots, Y_n$.

A special case is when the coefficients $\{a_i\}$ are non-negative, and sum to 1. In that case, the linear combination is also called a *weighted average* or a *weighted mean* of $Y_1, \ldots, Y_n$. Naturally, a weighted average must fall between the smallest and largest Y_i inclusive.

Another special case is when the coefficients sum to 0. In that case, the linear combination is called a *contrast*.

In standard multiple linear regression (intercept and covariates, OLS), each $\hat{\beta}_j$ except the intercept is a contrast. i.e. if expressed in the form $a_1 Y_1 + \ldots + a_n Y_n$, it will be found that $a_1 + \ldots + a_n = 0$. The intercept is not a contrast.

Some important types of contrasts:
contrasts in random variables, e.g. $(Y_2 + Y_3)/2 - Y_1$,
contrasts in observed values, e.g. $(y_2 + y_3)/2 - y_1$,
contrasts in observation means, e.g. $(\mu_2 + \mu_3)/2 - \mu_1$,
contrasts in parameters, e.g. $(\beta_4 + \beta_2)/2 - \beta_1$,
contrasts in parameter estimates, e.g. $\hat{\beta}_4 + \hat{\beta}_2 - 2\hat{\beta}_1$.

In matrix-vector notation, the linear combination $a_1 Y_1 + \ldots + a_n Y_n$, can be written as $a^\top Y$ or $Y^\top a$.

If vectors a and b satisfy the relationship $a^\top b = 0$, we say that a is *orthogonal* to b, b is orthogonal to a or a and b are orthogonal vectors (the angle between a and b is 90 degrees).

We say that the linear combinations $a^\top Y$ and $b^\top Y$ are *orthogonal linear combinations* if $a^\top b = 0$, i.e. if a is orthogonal to b. Orthogonal contrasts play a key role in ANOVA.

Exercise: In what case can two weighted averages be orthogonal?

Exercise: Compute $\mathrm{cov}(a^\top Y, b^\top Y)$ in the case of a orthogonal to b and $Y_1, \ldots, Y_n$ independent and with a common variance. What if $Y_1, \ldots, Y_n$ have different variances?

Computing the matrix $A = (X^\top X)^- X^\top$ above in the case of X being of full column rank:

In SAS:

```
proc iml;
  X = {1 1, 1 2, 1 3, 1 4};
  A = solve(t(X) * X, t(X));
  print A;
```

In R:

```
X = matrix(c(1, 1, 1, 2, 1, 3, 1, 4), byrow=TRUE, ncol=2);
A = solve(t(X) %*% X, t(X));
print(A);
```

The matrix A is useful in interpreting parameters and their estimates. The j-th row in A contains the coefficents of the linear combination that defines β_j and $\hat{\beta}_j$.

Recall that *only* estimable parameters can be interpreted. Non-estimable parameter *can not* be interpreted. Example: One-way ANOVA with 3 groups and model

$$\eta_i = \mu + \alpha_i \qquad i = 1, 2, 3.$$

There are four parameters in this model, $(\mu, \alpha_1, \alpha_2, \alpha_3)$, none of which is estimable. The design matrix X is:

```
1 1 0 0
1 0 1 0
1 0 0 1
```

This matrix can be plugged into the code given above, with "solve" replaced by "ginv". In SAS: `A = ginv(t(X) * X) * t(X)`. In R: `A = ginv(t(X) %*% X) %*% t(X)`. But now we have to remember that A is *not unique*, and the resulting A can't be used to interpret the individual parameters $(\mu, \alpha_1, \alpha_2, \alpha_3)$.

To summarize, regardless of how we compute A, the rule is that a row in A can be used to interpet a parameter *only if that parameter is estimable*, and we have to verify that estimability by other means.

Example: The second row in A, to be denoted $r_2^\top$, corresponds to α_1, while the third row, to be denoted $r_3^\top$, corresponds to α_2. Neither row by itself is usable for interpretation. However, since we know that $\alpha_1 - \alpha_2 = \mu_1 - \mu_2$ is estimable, $r_2 - r_3$ can be used to interpret $\alpha_1 - \alpha_2$, and $(r_2 - r_3)^\top$ is the row vector $(1, -1, 0)$, as expected.