## The matrix formulation for regression and ANOVA (Neter *et al.* 1996).

Both regression and ANOVA can be described using the general linear model $Y = X\beta + \epsilon$, where

Y = an $n \times 1$ column vector of values of the response variable Y
There are n observations.
X = an $n \times p$ matrix with columns corresponding to the $p$ predictor variables $X_i$
$\beta$ = an $p \times 1$ column vector of parameters, with row numbers corresponding to the column numbers in X
$\epsilon$ = an $n \times 1$ column vector of errors

In regression, the columns in X are fairly straightforward. Most regression models contain an intercept ($\beta_o$), which is fit by setting the first column of X to a dummy variable $X_o$ with value=1 for all observations. One column is added to the X matrix for each of the predictor variables, and if there are interaction terms or polynomial terms, the appropriate products or powers of the predictor variables are added as additional columns. For example, in simple linear regression, we use:

$$X = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ 1 & x_{31} \\ \vdots & \vdots \\ 1 & x_{n1} \end{bmatrix}$$

In multiple regression with two predictors and an interaction, we use:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{11} * x_{12} \\ 1 & x_{21} & x_{22} & x_{21} * x_{22} \\ 1 & x_{31} & x_{32} & x_{31} * x_{32} \\ \vdots & \vdots & \vdots & \vdots \quad \vdots \\ 1 & x_{n1} & x_{n2} & x_{n1} * x_{n2} \end{bmatrix}$$

In ANOVA, the X matrix contains qualitative indicator variables indicating membership in treatment groups. If there are $m$ groups, there are $m$ columns in X. There are an infinite number of ways to define the qualitative variables, but one way is to calculate the overall mean for Y (using the same approach described for $\beta_o$, above) together with deviations of particular treatments from this overall mean. This involves assigning a column in X to all but one treatment group; because the overall mean is already known, the deviation for the last group is determined from the sum of the other deviations. The indicator variables are set to 1 when an observation (row) is in the group that corresponds to that X variable, -1 if the observation is in the treatment group without its own column, and 0 otherwise. For example, suppose there were four treatment groups and three observations per group. The X matrix for the models described in Table 1 might look like:

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \end{bmatrix}$$

Regardless of how X is formulated, the equation $Y = X\,\beta + \epsilon$ is solved for $\beta$ using the normal equations, giving the parameter estimates

$$\hat{b} = (X'\ X)^{-1}\, X' Y.$$

Once the parameters are estimated, we partition the overall variance in the data as follows, given that $p$ is the number of columns in X (ie $p$=2 for a simple linear regression, $p$=4 for a two-factor regression, and $p$=m, the total number of treatments, for any ANOVA).

| Source | SS | df | MS | F |
|---|---|---|---|---|
| Model (M) | $b'\,X'\,Y - n\bar{Y}^2$ | p-1 | MSM | MSM/MSE |
| Error (E) | $Y'\,Y - b'\,X'\,Y$ | n-p | MSE | |
| Corrected Total (T) | $Y'\,Y - n\bar{Y}^2$ | n-1 | | |

We also determine the percent of variability explained by the model, $R^2$, as SSM/SST.

## Reference

Neter J, Kutner MH, Nachtsheim CJ, and Wasserman W. 1996. Applied Linear Statistical Models. Chicago: Richard D Irwin, Inc.

## Lack of fit tests

RR designs provide an underappreciated opportunity to test whether a particular regression model is appropriate for the data using lack-of-fit tests (Draper and Smith 1998). These tests are particularly good at diagnosing deviations from the linear model that may be difficult to detect by eye. Lack of fit tests work by partitioning residual variation around a regression line into two components: that due to variability among replicates within a treatment (the "pure error") and that due to deviations of the treatment means from the fitted curve (the "lack of fit"; Table 2A). The pure error is obtained from an ANOVA that uses the predictor variable(s) as a classification factor rather than as a quantitative one, and the lack-of-fit component is estimated from the difference between the error SS from the regression model and the error SS from the ANOVA model. There is significant lack-of-fit when the ratio of the mean squares lack-of-fit ($MS_L$) to mean squares pure error ($s_e^2$) exceeds a critical F-statistic. If there is no significant lack-of-fit, then the regression model is appropriate for the data and conclusions can be drawn accordingly. If, however, there is significant lack-of-fit, remedial action is required. In some cases, the regression model can be modified to be more appropriate for the data, for example by adding polynomial terms (which will reduce the power somewhat due to the additional parameters). However, in other cases, there is no appropriate linear model for the data. In this case, researchers can switch to non-linear regression or "fall back" to drawing conclusions using ANOVA.

We note briefly that lack-of-fit tests are also available for nonlinear regression, although we do not develop them here (see Draper and Smith 1998 instead), providing another argument for the use of RR designs in ecological research.

**Table 2A. ANOVA table for a replicated regression**
*N* indicates the total number of experimental units, *p* is the number of columns of *X*, *m* indicates the number of treatments with replicates, and $n_j$ is the number of replicates for treatment *j*.

| Source | | SS | df | MS | |
|---|---|---|---|---|---|
| Model | Regression SS | SSR | *p-1* | | |
| Error | Lack of fit | SSLOF | *m-p* | $MS_L$ | $s^2$ |
| | Pure error | SSPE | *N-m* | $s_e^2$ | |
| Total (corrected) | | | *N-1* | | |

## Using RR designs to see the parallels between regression and ANOVA

Replicated regression provides a currency for relating the model and residual sums-of-squares (SS) for regression and ANOVA models fit to the same data (Table 2B). The alternative partitioning of sums of squares and degrees of freedom has some interesting implications. Most importantly, the lack-of-fit SS (SSLOF) are part of the error SS in regression (SSE), but part of the model SS in ANOVA (SSA). As a result, we expect changes in $R^2$ (and thus effect size) between regression and ANOVA models applied to the same dataset. $R^2$ will always be bigger for ANOVA than for regression by the amount SSLOF/SST.

**Table 2B. Partitioning variability in a RR dataset according to the regression, RR, and ANOVA models**

| | As a regression | | As a replicated regression | | As an ANOVA | |
|---|---|---|---|---|---|---|
| **Source** | **SS** | **df** | **SS** | **df** | **SS** | **df** |
| Model | SSR | 1 | SSR | 1 | SSA | m-1 |
| Error | SSE | N-2 | SSLOF | m-2 | | |
| | | | SSPE | N-m | SSPE | N-m |
| Total | SST | N-1 | SST | N-1 | SST | N-1 |

## How Figure 3 was created

To create the scenarios in Figure 3, we started with power curves (as explained in Statistical Panel 2) for experimental designs with 24 (Figure 1a), 36 (not shown), and 48 (Figure 1b) experimental units. We then selected a minimum power for the ANOVA (0.8, following convention).

1. Left panel: Minimum $R^2$ vs the number of treatments. On the power curves for experiment size, we drew a line horizontally across the figure at the target power level. At each intersection of this "minimum power" line with a power curve, we dropped down to the X-axis and recorded $R^2$ at that point, which is the minimum $R^2$ required to produce that power for that experimental design. We then plotted this minimum $R^2$ versus the number of treatments in that design in Figure 3a.

2. Right panel: Maximum allowable ratio of SSPE/SST vs number of treatments
   Statistical Panel 3 introduces several abbreviations for the sum-of-squares terms in a replicated regression:
   - SSR = sums-of-squares due to regression
   - SSPE = sums-of-squares due to pure error, the variability around the mean for each level of the predictor variable(s)
   - SSLOF = sums-of-squares due to lack of fit, the deviation from the regression line not explained by the ANOVA (determined as SSR-SSPE).

   From Table 2B in Web-only Appendix 2, we also know that $R^2_{anova}$ = (SSR+SSLOF) / SST.

   Therefore, we can define
   $1\text{-}R^2_{anova}$ = SST/SST – (SSR+SSLOF)/SST = SSPE/SST, which provided us with a formula to convert the minimum $R^2$ obtained in Step 1 to the fraction of the total variability that is explained by the pure error, or variability among replicates within a treatment.

   Estimates of SSPE/SST are closely related to those used to calculate power analyses in *t*-tests and straightforward ANOVA models, and so are frequently estimable from past experiments (eg Case Study Panel 3).

## Cottingham – Web-onlWeb-only appendix 4. A Matlab program for calculating power. ____

Cut and paste the code for use in Matlab. The raw data file used for the simulation is available from the authors.

```
% calculatepower.m
% determine power for a series of potential one- and two-way experimental designs specified by the user
% author KL Cottingham (cottingham@dartmouth.edu)
% created 19 Dec 03 from compareRvsA_vsf2.m;
% last modified 23 December 2004 for Frontiers website


% 888888888888888888888888888888888888888888888
% give the necessary info
% 888888888888888888888888888888888888888888888

clear;
lookpowerfigs=0; % toggle figures on and off
lookthresholds=0; % toggle evaluating thresholds on and off

% setups
output=[];
thresholds=[];

% specify the target p-value
alpha=0.05;

% prepare figures (if desired)
if lookpowerfigs,
        figure(1); clf; orient tall;
end;

% set constraints
minnr=2; % minimum number of replicates per treatment
maxnr=5; % maximum number of replicates per treatment


% 888888888888888888888888888888888888888888888
% looping structure
% 888888888888888888888888888888888888888888888

% specify number of levels of factor A
for Alevels=2:4, %input('Number of levels of factor A? ');

% specify number of levels of factor B
for Blevels=1:4, %input('Number of levels of factor B? ');

% specify number of replicates of each cell
for nreps=minnr:maxnr, %input('Number of replicates per cell? ');

        % 888888888888888888888888888888888888888888888
        % determine df for regression & for ANOVA
        % 888888888888888888888888888888888888888888888

        % calculate number of EU
        N=Alevels*Blevels*nreps;

        % assume we're fitting a regression with three parameters: effects of A & B
        % and their interaction
        if Blevels==1,
  DFM_reg=1;
    else DFM_reg=3;
  end;
    DFE_reg = N - DFM_reg - 1;

    % assume we're fitting an ANOVA with main effects and interactions
    DFM_anova=(Alevels-1) + (Blevels-1) + (Alevels-1)*(Blevels-1);
    DFE_anova=N - DFM_anova - 1;

    % 888888888888888888888888888888888888888888888
    % calculate the power of each design, based on case 0 of Cohen Ch 9
    % delta = (effect size)squared * (u+v+1)
    % 888888888888888888888888888888888888888888888

    % determine critical value of F needed to reject Ho: no difference for each design
    Fcrit_reg=finv(1-alpha,DFM_reg,DFE_reg);
    Fcrit_anova=finv(1-alpha,DFM_anova,DFE_anova);

    % list of R2 to compare
    R2 = (0 : 0.01 : 0.99)';

    % list of effect sizes that go with those R2 values
    % f2 = R2 / (1 - R2)
    ES = R2 ./ (1-R2);

    % calculate delta as f2 * (u+v+1)
    delta = N.*ES;

    %calculate the power for each design here following other program
    power_reg=1-ncfcdf(Fcrit_reg,DFM_reg,DFE_reg,delta);
    power_anova=1-ncfcdf(Fcrit_anova,DFM_anova,DFE_anova,delta);

    output=[output; ones(length(ES),1)*[Alevels Blevels nreps] ES power_reg power_anova];

% 888888888888888888888888888888888888888888888
    % plot power vs. effect size
    % 888888888888888888888888888888888888888888888

if lookpowerfigs,
            sb=sb+1;
            if sb>8, sb=1; figno=figno+1; figure(figno); clf; orient tall; end;
            subplot(4,2,sb);
            semilogx(ES,power_reg,'r-',ES,power_anova,'k:');

            if sb==1, legend('Regression','ANOVA',2); end;
            ylabel('power');
            xlabel('Effect Size');
            title([num2str(Alevels) ' x ' num2str(Blevels) ' x ' num2str(nreps) ' design']);
end;

    % 888888888888888888888888888888888888888888888
    % determine thresholds of interest
    % 888888888888888888888888888888888888888888888

    reggtpt8=min(ES(find(power_reg>=0.8)));
    anovagtpt8=min(ES(find(power_anova>=0.8)));
    reggtanova=min(ES(find(power_reg>=power_anova)));
```

```
            reggtanovaandpt8=min(ES(find(power_reg>0.8 & power_reg>=power_anova)));

            % 88888888888888888888888888888888888888888888
            % use algebra to determine what SSR & SSPE need to be to exceed these
            % thresholds
            % 88888888888888888888888888888888888888888888

            % regression power > 0.8
            minpctSSR=reggtpt8./(reggtpt8+1);

            % anova power > 0.8
            maxpctSSPE=1./(anovagtpt8+1);

            % (regression power > anova power) & (regr power > 0.8) -> works out to
            minpctSSRforRtowin=reggtanovaandpt8./(reggtanovaandpt8+1);
            maxpctSSPEforRtowin=1./(reggtanovaandpt8+1);


            % 88888888888888888888888888888888888888888888
            % collect these thresholds for particular designs: are there patterns?
            % 88888888888888888888888888888888888888888888

            thresholds=[thresholds; Alevels Blevels nreps minpctSSR maxpctSSPE minpctSSRforRtowin maxpctSSPEforRtowin reggtanovaandpt8];

      end; % for nreps

   end; % for Blevels
   end; % for Alevels


   save powervsESinfo.dat output /ascii;
   save thresholds.dat thresholds /ascii;
```