Important Formulas for DATA ANALYSIS: A Model Comparison Approach

Charles M. Judd and Gary H. McClelland

April 27, 1993

This is intended as a short guide to important formulas used in the book. It does not include everything and should not be substitutued for careful study of the text. There are other formulas not included here that may be useful in specific situations. The formulas are organized by concept instead of by chapter.

1 Basics

$$\textbf{DATA} = \textbf{MODEL} + \textbf{ERROR} \tag{1}$$

DATA: The set of scores on the variable which we are predicting.

MODEL: A set of rules or formulas that make a specific prediction for each observation in **DATA**

ERROR: The amount by which **MODEL** mispredicts **ERROR**

$$Y_i = \Psi_i + \epsilon_i \tag{2}$$

Y_i The observed **DATA** value for the i^{th} observation.

Ψ_i The true but unknowable value for the i^{th} observation.

ϵ_i The true but unknowable amount by which Y_i has been disturbed relative to Ψ_i.

$$Y_i = \hat{Y}_i + e_i \tag{3}$$

$\hat{Y}_i$ The value of Y_i predicted by the **MODEL**.

e_i The amount by which **DATA** is mispredicted by **MODEL**; that is,

$$e_i = Y_i - \hat{Y}_i \tag{4}$$

ERROR = Some aggregate function of the e_i. (5)

MODEL C A compact model which uses PC parameters estimated from **DATA** to make predictions $\hat{Y}_{iC}$.

MODEL A An augmented model which uses PA parameters estimated from **DATA** to make predictions $\hat{Y}_{iA}$.

NOTE: $PC < PA$ and *all* the parameters of **MODEL C** must be included in **MODEL A**.

Proportional reduction in error due to **MODEL A** instead of **MODEL C**:

$$PRE = \frac{ERROR(C) - ERROR(A)}{ERROR(C)} = 1 - \frac{ERROR(A)}{ERROR(C)} \tag{6}$$

Proportional reduction in error per extra parameter relative to potential proportional reduction in error per remaining parameter:

$$F^*_{PA-PC,n-PA} = \frac{PRE/(PA - PC)}{(1 - PRE)/(n - PA)} \tag{7}$$

If $PA - PC = 1$ then

$$t^*_{n-PA} = \sqrt{F^*_{1,n-PA}}. \tag{8}$$

Sum of Squared Errors:

$$SSE = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{9}$$

Mean Squared Error:

$$MSE = \frac{SSE}{n - PA} \tag{10}$$

Analysis of Variance (ANOVA) decomposition of sums of squares:

$$SSE(C) = SSR + SSE(A) \tag{11}$$

Table 1: Generic ANOVA Layout

Source	SS	df	MS	F^*	PRE
Reduce, Model A	SSR	PA-PC	$MSR = \frac{SSR}{PA-PC}$	$\frac{MSR}{MSE}$	$\frac{SSR}{SSE(C)}$
Error for Model A	SSE(A)	n-PA	$MSE = \frac{SSE(A)}{n-PA}$		
Total	SSE(C)	n-PC			

Sum of Squares Reduced:

$$SSR = SSE(C) - SSE(A) = \sum_{i=1}^{n} (\hat{Y}_{iC} - \hat{Y}_{iA})^2. \tag{12}$$

Formula for PRE using Sum of Squared Errors:

$$PRE = \frac{SSE(E) - SSE(A)}{SSE(C)} = \frac{SSR}{SSE(C)} \tag{13}$$

Alternative but Equivalent ANOVA Formula for F^*:

$$F^*_{PA-PC,n-PA} = \frac{SSR/(PA-PC)}{SSE(A)/(n-PA)} = \frac{MSR}{MSE} \tag{14}$$

Table 1 summarizes the ANOVA decomposition of sums of squares and the relationships betweeen SSE's, MSE's, and F's.

2 Statistical Inference and Power

F^* has an F-distribution and the standard statistical tables for F and PRE may be used if

Independence: The ϵ_i are independent of one another

Normality: The ϵ_i are sampled from a Normal distribution

Homogeneity of Variance: The ϵ_i are sampled from a single distribution with a constant variance.

Statistical inference is always an uncertain decision with the possibility of two types of errors:

Type I Error: The probability that **MODEL C** will be rejected in favor of **MODEL A** when in fact **MODEL C** is correct. This probability is represented by α. Normally, α is chosen before the data are analyzed.

Type II Error: The probability that **MODEL C** will *not* be rejected in favor of **MODEL A** when in fact **MODEL A** is correct. This probability is represented by β. **Power** $= 1 - \beta$. Power is not chosen but instead is a function of α, n, experimental design, and the MODEL's used. Increasing α and/or n improves power. Controlling for more factors in the design or including them in the **MODEL** generally increases power.

PRE is a biased estimate of η^2, the true proportional reduction in error if all the parameters were known exactly. An unbiased estimate is given by:

$$\eta^2 = 1 - (1 - PRE)\left[\frac{n - PC}{n - PA}\right]. \tag{15}$$

Small Effect: $\eta^2 = .03$

Moderate Effect: $\eta^2 = .1$

Big Effect: $\eta^2 = .3$

If guestimates of the parameter values and the variance are available, then the expected proportional reduction in error is

$$\eta^2 = \frac{SSR}{SSE(A) + SSR} = \frac{1}{\frac{SSE(A)}{SSR} + 1} \tag{16}$$

3 General Linear Model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i. \tag{17}$$

$$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_{p-1} X_{i,p-1} \tag{18}$$

where $b_0, b_1, \ldots, b_{p-1}$ are p parameters estimated from **DATA** so as to minimize the **ERROR** defined as

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{19}$$

Interpretation of Parameters:

b_0 The predicted value $\hat{Y}_i$ when $X_{ij} = 0$ for all j.

b_j The amount by which $\hat{Y}_i$ is adjusted, on average, for each unit increase in the predictor variable X_{ij}, controlling for all the other predictor variables $X_{ik}, k \neq j$.

For the Simple **MODEL**: $\hat{Y}_i = b_0$

$$b_0 = \frac{\sum_{i=1}^{n} Y_i}{n} = \bar{Y}. \tag{20}$$

and the sample variance of Y is defined as

$$s^2 = MSE = \frac{SSE}{n-1} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}. \tag{21}$$

For the Simple Regression **MODEL**: $\hat{Y}_i = b_0 + b_1 X_i$

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{22}$$

$$b_0 = \bar{Y} - b_1 \bar{X} \tag{23}$$

For more complex **MODEL**s, always use computer programs to estimate $b_0, b_1, b_2, \ldots$

Redundancy of the j^{th} predictor variable with the other predictor variables is assessed by TOL_j, tolerance:

$$TOL_j = 1 - PRE_j \tag{24}$$

where PRE_j is the PRE (relative to the simple model) obtained when using all the other predictors to predict X_{ij}.

$(1-\alpha)\%$ Confidence Interval for $\beta_j, j \neq 0$:

$$b_j \pm \sqrt{\frac{(F_{1,n-PA;\alpha})(MSE)}{(SSX_j)(TOL_j)}} \tag{25}$$

where

$$SSX_j = \sum_{i=1}^{n}(X_{ij} - \bar{X}_{ij})^2 \tag{26}$$

and MSE is the mean squared error for the **MODEL** that includes β_j as a parameter. $F_{1,n-PA;\alpha}$ is the critical value for the F-distribution obtained from a table. For the simple regression **MODEL**, simply let $TOL_j = 1$ and use the same formula for the confidence interval.

$(1-\alpha)\%$ Confidence Interval for β_0 *when and only when* all predictors are in mean deviation form, that is, $\bar{X}_j = 0$ for all $j > 0$:

$$b_0 \pm \sqrt{\frac{(F_{1,n-PA;\alpha})(MSE)}{n}} \tag{27}$$

Note that when the predictors are in mean deviation form it will always be the case that $b_0 = \bar{Y}$. Hence, this confidence interval is useful for making the most powerful inferences about $\bar{Y}$.

4 Models with Product Terms

Models with Interaction between Two Predictors.

$$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + b_3 X_{i1} X_{i2} \tag{28}$$

which can be rewritten to represent the "simple" relationship between Y and X_2 as

$$\hat{Y}_i = [b_0 + b_1 X_{i1}] + [b_2 + b_3 X_{i1}] X_{i2} \tag{29}$$

or can be written to represent the "simple" replationship between Y and X_1 as

$$\hat{Y}_i = [b_0 + b_2 X_{i2}] + [b_1 + b_3 X_{i2}] X_{i1}. \tag{30}$$

In both cases, the terms inside the first set of brackets produce the "intercept" for the simple relationship and those inside the second set generate the "slope."

Polynomial regression for power 2 (i.e., curvilinear).

$$\hat{Y}_i = b_0 + b_1 X_i + b_2 X_i^2. \tag{31}$$

which can be rewritten to represent a "simple" relationship between Y and X as

$$\hat{Y}_i = [b_0 - b_2 X_i^2] + [b_1 + 2 b_2 X_i] X_i + \epsilon_i \tag{32}$$

where the simple slope is the first derivative of the function with respect to X_i.

5 Asking Questions of DATA

In this section are listed typical questions that can be asked with two predictor variables. These represent all the usual questions and they can be easily generalized to situations with more predictor variables. However, they by no means represent all the questions that can be asked.

Prior Hypothesis. Do pre-specified parameter values adequately describe the **DATA**?

$$\begin{aligned} \textbf{MODEL A} &: Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \\ \textbf{MODEL C} &: Y_i = \mu_0 + \mu_1 X_{i1} + \mu_2 X_{i2} + \epsilon_i \\ H_0 &: \beta_0 = \mu_0, \beta_1 = \mu_1, \beta_2 = \mu_2. \end{aligned}$$

Simple Regression. Relative to making the same prediction for all observations, is it useful to make predictions conditional on X_1? That is, is there a reliable relationship between Y and X_1?

$$\begin{aligned} \textbf{MODEL A} &: Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i \\ \textbf{MODEL C} &: Y_i = \beta_0 + \epsilon_1 \\ H_0 &: \beta_1 = 0. \end{aligned}$$

The square root of the PRE from this test is known as the *correlation coefficient* and represented by r_{YX_1}.

Partial Regression. If predictions are already conditional on X_1, is it useful to make them also conditional on X_2? That is, is X_2 useful *over and above* X_1? In still other words, *controlling for* X_1 or *holding* X_1 *constant*, is there a reliable relationship between Y and X_2? This is equivalent to adding a predictor to an existing model.

$$\begin{aligned} \textbf{MODEL A} &: Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \\ \textbf{MODEL C} &: Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i \\ H_0 &: \beta_2 = 0. \end{aligned}$$

The PRE from this test is known as the *Coefficient of Partial Determination* and its square root is the partial correlation coefficient, represented by $r_{YX_2.X_1}$. This is the correlation between Y and X_2, controlling for X_1. This test is the fundamental one-degree-of-freedom test in regression analysis.

Multiple Regression. Instead of making the same prediction for all observations, is it useful to make predictions jointly conditional on X_1 and X_2?

$$\begin{aligned} \textbf{MODEL A} &: Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \\ \textbf{MODEL C} &: Y_i = \beta_0 + \epsilon_i \\ H_0 &: \beta_1 = 0 \textit{ and } \beta_2 = 0. \end{aligned}$$

The PRE from this test is known as the *Coefficient of Multiple Determination* and is represented by $R^2_{YX_1X_2}$. The square root of this PRE is the *Multiple R*. If **MODEL C** is rejected, it is ambiguous whether the $\beta_1 = 0$ or the $\beta_2 = 0$ part of the null hypothesis is incorrect, or both.

Interaction. Does the simple relationship between Y and X_1 depend on the level of X_2? That is, is there an interaction between X_1 and X_2?

$$\begin{aligned} \textbf{MODEL A} &: Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i \\ \textbf{MODEL C} &: Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \\ H_0 &: \beta_3 = 0. \end{aligned}$$

Note: The product of X_1X_2 represents the interaction only if both components of the product also appear separately in the model. Also, putting X_1' and X_2' in mean deviation form does *not* ensure that their product, $X_1'X_2'$, is in mean deviation form.

Polynomial Regression. Does the simple relationship between Y and X_1 depend on the level of X_1? That is, is the slope constant or does it change depending on the value of X_1.

$$\begin{aligned} \textbf{MODEL A} &: Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \epsilon_i \\ \textbf{MODEL C} &: Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i \\ H_0 &: \beta_2 = 0. \end{aligned}$$

Deviation Form. To shift the focus of an equation, it is often useful to transform the predictor variables to *deviation form*:

$$X'_{ij} = X_{ij} - k_j \tag{33}$$

where k_j is a specific constant. For example, if $k_j = \bar{X}_j$ for all j, then b_0, the predicted value when all predictors are at their mean zero, is equal to

$\hat{Y}$. Thus, questions about the mean can be asked with

$$\begin{aligned} \textbf{MODEL A} &: Y_i = \beta_0 + \beta_1 X'_{i1} + \beta_2 X'_{i2} + \epsilon_i \\ \textbf{MODEL C} &: Y_i = \mu_0 + \beta_1 X'_{i1} + \beta_2 X'_{i2} + \epsilon_i \\ H_0 &: \beta_0 = \mu_0 \text{ or, equivalently, } \hat{Y} = \mu_0. \end{aligned}$$

6 Analysis of Variance

Contrast Codes. If the categorical predictors define J groups or cells, then in a complete analysis there must be J-1 contrast codes. Contrast codes must satisfy two rules:

$$1. \sum_{j=1}^{J} \lambda_{ij} = 0. \tag{34}$$

$$2. \sum_{j=1}^{J} \lambda_{ij}\lambda_{kj} = 0 \text{ for all i, k.} \tag{35}$$

Parameter Estimates. The value of b that estimates β for a contrast-coded predictor is given by

$$b_i = \frac{\sum_{j=1}^{J} \lambda_{ij}\bar{Y}_j}{\sum_{j=1}^{J} \lambda_{ij}^2} \tag{36}$$

where λ_i is the i^{th} contrast code, $\bar{Y}_j$ is the mean for the j^{th} group, and J is the number of groups defined by the categorical variable(s).

Sum of Squares. The SSR for a particular contrast is given by:

$$\text{SSR for } \lambda_i = \frac{(\sum_{j=1}^{J} \lambda_{ij}\bar{Y}_j)^2}{\sum_{j=1}^{J}(\lambda_j^2/n_j)} \tag{37}$$

where n_j is the number of observations in the j^{th} group. If b_j is known instead of the group means, then

$$\text{SSR for } \lambda_i = \frac{b_i^2(\sum_{j=1}^{J} \lambda_{ij}^2)^2}{\sum_{j=1}^{J}(\lambda_j^2/n_j)}. \tag{38}$$

One-Way ANOVA. If, for example, there is one categorical predictor variable with three levels (which therefore define three groups), then there will be two contrast-coded predictors X_1 and X_2. In the One-Way ANOVA we will always have

$$\textbf{MODEL A: } Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1} + \epsilon_i.$$

The form of **MODEL C** will depend on which question is being asked. Using

$$\textbf{MODEL C: } Y_i = \beta_0 + \epsilon_i$$

asks whether the means of the three groups are different from each other. Using a **MODEL C** that leaves out all but one of the contrast-coded predictors asks whether that specific contrast equals 0.

Two-Way ANOVA. If, for example, there are two categorical variables, each with exactly two levels, then four groups will be defined in a factorial design. Each categorical variable will have its own contrast-coded predictor X_{i1} and X_{i2}, respectively. In the Two-Way ANOVA we will always have

$$\textbf{MODEL A: } Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \epsilon_i.$$

The form of **MODEL C** will depend on which question is being asked. Using

$$\textbf{MODEL C: } Y_i = \beta_0 + \epsilon_i$$

asks whether the means of the four groups are different from each other. Using a **MODEL C** that leaves out just $\beta_1 X_{i1}$ asks whether there is a "main effect" for the first categorical variable. Leaving out just $\beta_2 X_{i2}$ similarly asks whether there is a "main effect" for the second categorical variable. Leaving out just $\beta_3 X_{i1} X_{i2}$ asks whether there is an "interaction" between the two categorical variables.

Analysis of Covariance. If X_{i1} is a continuous predictor variable and X_{i2} is a categorical predictor variable with two levels, then the simple analysis of covariance (ANCOVA) will always have

$$\begin{aligned} \textbf{MODEL A} &: Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i. \\ \textbf{MODEL C} &: Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i. \end{aligned}$$

If the F^* for a simple regression of Y_i on X_{i1} is greater than one, then ANCOVA will provide a more powerful test of $H_0 : \beta_2 = 0$ than would ANOVA.