Algebraic Aspects of Multiple Regression

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Algebraic Aspects of Multiple Regression

Introduction

- In this module, we quickly review some fundamental aspects of the algebra of multiple regression.

Key Matrix Formulas

- We already saw in our treatment of the two-sample independent sample t-test how additional regressors can be tested for signficance using the partial F-test for nested models, implemented in the R command `anova`.
- Now we present the formulas for the model, estimated coefficients, and standard errors.

Key Matrix Formulas

The Model and its Coefficients

- The multiple regression model can be written

$$E(Y|\mathbf{X}) = \mathbf{X}\beta \tag{1}$$
$$\mathrm{Var}(Y|\mathbf{X}) = \sigma^2 \tag{2}$$

- As in the simple regression model, the first column of $\mathbf{X}$ is a column of 1's, and the first element of $\boldsymbol{\beta}$ is typically labeled β_0, with subsequent elements labeled $\beta_1 \ldots \beta_p$.
- Given a set of n criterion ("response") scores in $\mathbf{y}$ and an $n \times p+1$ set of predictor scores (including the intercept) in the matrix $\mathbf{X}$, the ordinary least squares estimates of $\boldsymbol{\beta}$ may be calculated as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{3}$$

- The predicted scores are calculated as

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{P}_x\mathbf{y} \tag{4}$$

- The residual scores are, of course,

$$\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{P}_x)\mathbf{y} = \mathbf{Q}_x\mathbf{y} \tag{5}$$

Multiple Regression of Fuel Data

Introduction

- Suppose we use *Dlic*, *Income*, *logMiles*, and *Tax* to predict *Fuel*.
- We begin by analyzing the scatterplot matrix.
- As we can see in the next slide, the potential predictors are only moderatey related to *Fuel*.

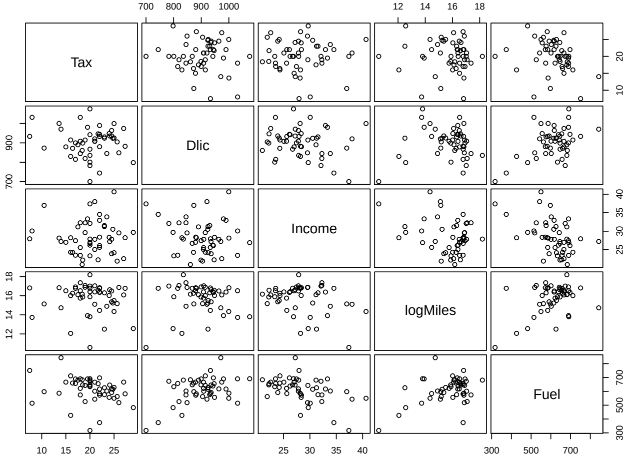
Multiple Regression of Fuel Data

Scatterplot Matrix Code

```
> data(fuel2001)
> fuel2001$Dlic <- 1000*fuel2001$Drivers/fuel2001$Pop
> fuel2001$Fuel <- 1000*fuel2001$FuelC/fuel2001$Pop
> fuel2001$Income <- fuel2001$Income/1000
> fuel2001$logMiles <- logb(fuel2001$Miles,2)
> f <- fuel2001[,c(7,9,3,10,9)]
> pairs(f,gap=0.4,cex.labels=1.5)
```

Multiple Regression of Fuel Data

Scatterplot Matrix



Multiple Regression of Fuel Data

Correlation Matrix

What we see in the scatterplot matrix is reflected in the matrix of intercorrelations.

```
> round(cor(f),4)

            Tax    Dlic  Income logMiles    Fuel
Tax      1.0000 -0.0858 -0.0107  -0.0437 -0.2594
Dlic    -0.0858  1.0000 -0.1760   0.0306  0.4685
Income  -0.0107 -0.1760  1.0000  -0.2959 -0.4644
logMiles -0.0437 0.0306 -0.2959   1.0000  0.4220
Fuel    -0.2594  0.4685 -0.4644   0.4220  1.0000
```

Multiple Regression of Fuel Data

Multiple Regression Output

- The next slide shows the output from the multiple regression for predicting *Fuel* from the 4 predictors.
- The far right column is the two-sided p-value for the t-statistic for each coefficient of a model term.
- In specifying the model, I use the specialized language used by R for setting up linear models. Each included term is assumed to have a coefficient, and the 1 explicitly indicates the intercept. R assumes an intercept is present. If you wish to specify a model with no intercept, you must include a −1 term.

Multiple Regression of Fuel Data

Multiple Regression Output

```
> attach(fuel2001)
> fuel.fit.all <- lm(Fuel~1 + Tax + Dlic + Income + logMiles)
> summary(fuel.fit.all)

Call:
lm(formula = Fuel ~ 1 + Tax + Dlic + Income + logMiles)

Residuals:
     Min       1Q   Median       3Q      Max
-163.145  -33.039    5.895   31.989  183.499

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 154.1928   194.9062   0.791 0.432938
Tax          -4.2280     2.0301  -2.083 0.042873 *
Dlic          0.4719     0.1285   3.672 0.000626 ***
Income       -6.1353     2.1936  -2.797 0.007508 **
logMiles     18.5453     6.4722   2.865 0.006259 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 64.89 on 46 degrees of freedom
Multiple R-squared:  0.5105,        Adjusted R-squared:  0.4679
F-statistic: 11.99 on 4 and 46 DF,  p-value: 9.331e-07
```

Multiple Regression of Fuel Data

ANOVA for Model Comparison

- ANOVA is a key tool for comparing models.
- Define p' to be the number of terms in the regression model, including the intercept.
- As before, SYY is the sum of squared Y deviation scores, and RSS is the sum of squared residuals. Then

$$SSreg = SSY - RSS \tag{6}$$

- To assess the overall significance of the prediction equation with 4 predictors, we follow the table shown below.

Source	df	SS	MS	F	p-value
Regression	p	$SSreg$	$MSreg = SSreg/p$	$MSreg/\hat{\sigma}^2$	
Residual	$n - p'$	RSS	$\hat{\sigma}^2 = RSS/(n - p')$		
Total	$n - 1$	SYY			

Multiple Regression of Fuel Data

ANOVA for Model Comparison

- The overall test for the *combined* significance of β_1, β_2, β_3, and β_4 compares a model with only an intercept β_0 against a model with the intercept and all other terms.

```
> fuel.fit.intercept.only <- lm(Fuel~1)
> anova(fuel.fit.intercept.only,fuel.fit.all)

Analysis of Variance Table

Model 1: Fuel ~ 1
Model 2: Fuel ~ 1 + Tax + Dlic + Income + logMiles
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     50 395694
2     46 193700  4    201994 11.992 9.331e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple Regression of Fuel Data

Partial F-Tests: A General Approach

- Actually, the F-tests we've been discussing so far are a special case of a general procedure for generating *partial F-tests* on a nested sequence of models.
- Suppose Model A includes Model B as a special case. That is, Model B is a special case of Model A where some terms have coefficients of zero. Then Model B is nested within Model A.
- If we define SS_a to be the sum of squared residuals for Model A, SS_b the sum of squared residuals for Model B, df_a to be $n - p_a$, where p_a is the number of terms in Model A including the intercept, and $df_b = n - p_b$, then to compare Model B against Model A, we compute the partial $F-$statistic as follows.

$$F_{df_b - df_a, df_a} = \frac{MS_{comparison}}{MS_{res}} = \frac{(SS_b - SS_a)/(p_a - p_b)}{SS_a/df_a} \tag{7}$$

Multiple Regression of Fuel Data

Testing Significance of a Single Term

- R does this model comparison for us using the `anova` function.
- Suppose we wish to test the significance of the *Tax* term when all the other 3 predictors are already in the model (along with the intercept).
- There are several ways we can do this in R.
- A direct way is to specify a second model without the *Tax* term and compare it to the model with the *Tax* term.

```
> Fuel.Fit.Without.Tax <- lm(Fuel ~ 1 +  Dlic + Income + logMiles)
> Fuel.Fit.With.Tax <- lm(Fuel ~ 1 +  Dlic + Income + logMiles + Tax)
> anova(Fuel.Fit.Without.Tax,Fuel.Fit.With.Tax)

Analysis of Variance Table

Model 1: Fuel ~ 1 + Dlic + Income + logMiles
Model 2: Fuel ~ 1 + Dlic + Income + logMiles + Tax
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     47 211964
2     46 193700  1     18264 4.3373 0.04287 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Multiple Regression of Fuel Data

Automatic Sequential Testing of Single Terms

- R will automatically perform a sequence of term-by-term tests on the terms in your model, *in the order they are listed in the model specification*.
- Just use the `anova` command on the single full model.
- You can prove for yourself (C.P.!) that the order of testing matters. The significance level for a term depends on the terms entered before it.

```
> anova(Fuel.Fit.With.Tax)

Analysis of Variance Table

Response: Fuel
          Df Sum Sq Mean Sq F value    Pr(>F)
Dlic       1  86854   86854 20.6262 4.019e-05 ***
Income     1  59576   59576 14.1481 0.0004765 ***
logMiles   1  37300   37300  8.8581 0.0046399 **
Tax        1  18264   18264  4.3373 0.0428733 *
Residuals 46 193700    4211
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Standard Errors for Coefficients

- In a formula that is virtually identical in form to the simpler one for bivariate regression, the covariance matrix of the estimated regression coeffients is given by

$$\mathrm{Var}(\hat{\beta}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \tag{8}$$

- The unbiased estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n - p'} = \frac{\mathrm{RSS}}{n - (p+1)} \tag{9}$$

- Consequently, the typical estimate for $\mathrm{Var}(\hat{\beta}|X)$ is

$$\widehat{\mathrm{Var}}(\hat{\beta}|X) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1} \tag{10}$$

Standard Errors for Predicted and Fitted Values

Introduction

- You recall from our earlier discussion that there are two distinctly different standard errors that we can compute in connection with the regression line.
- One standard error, sepred, deals with the situation where we have a new set of $\mathbf{x}$ values, and we wish to compute the standard error for the value of $\hat{y}$ computed from these values.
- Another standard error, sefit, deals with the situation where we would like to compute a set of standard errors for the (population) fitted values on the regression line.

Standard Errors for Predicted and Fitted Values

Key Formulas

- Suppose we have observed, or will in the future observe, a new case with its own set of predictors that result in a vector of terms $\mathbf{x}^*$.
- We would like to predict the value of the response given $\mathbf{x}^*$.
- As in simple regression, the point prediction is $\tilde{y}^* = \mathbf{x}^{*\prime}\hat{\beta}$, and the standard error of prediction, $\text{sepred}(\tilde{y}^*|\mathbf{x}^*)$, is

$$\text{sepred}(\tilde{y}^*|\mathbf{x}^*) = \hat{\sigma}\sqrt{1 + \mathbf{x}^{*\prime}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}^*} \tag{11}$$

- Similarly, the estimated average of all possible units with a value $\mathbf{x}$ for the terms is given by the estimated mean function at $\mathbf{x}$, $\hat{E}(Y|\mathbf{X} = \mathbf{x}) = \hat{y} = \mathbf{x}'\hat{\beta}$., with standard error given by

$$\text{sefit}(\hat{y}|\mathbf{x}) = \hat{\sigma}\sqrt{\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}} \tag{12}$$

Standard Errors for Predicted and Fitted Values

Key Formulas

- A given software package may not produce all these estimates.
- If a program produces sefit but not sepred, the latter can be computed from the former from the result

$$\text{sepred}(\tilde{y}^*|\mathbf{x}^*) = \sqrt{\hat{\sigma}^2 + \text{sefit}(\tilde{y}^*|\mathbf{x}^*)^2} \tag{13}$$