Multiple Linear Regression

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Multilevel Regression Modeling, 2009

Multiple Linear Regression

Introduction

In this lecture we discuss the multiple linear regression model, variable selection, and statistical testing.

The Multiple Regression Model

Multiple linear regression is similar in many respects to bivariate regression, except that there are several X variables. The multiple regression model states that the conditional distribution of y given X is normal, and that the conditional mean is a linear function of the predictors, i.e.,

$$y = X\beta + \epsilon = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon \tag{1}$$

$$E(y|X) = X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p \tag{2}$$

and

$$Var(y|X) = \sigma^2 \tag{3}$$

The Multiple Regression Model

Note that

- The conditional variance is not a function of X, so again the distribution of regression residuals is normal with constant variance and mean zero
- The intercept can be incorporated into the above specification by including a column of 1's in X, putting the intercept in the corresponding (usually the first) position in β

The Multiple Regression Model

Note that

- The conditional variance is not a function of X, so again the distribution of regression residuals is normal with constant variance and mean zero
- The intercept can be incorporated into the above specification by including a column of 1's in X, putting the intercept in the corresponding (usually the first) position in β

Calculating Beta

Ordinary least squares (OLS) regression chooses β to minimize the sum of squared errors. β estimates are calculated as

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

The $\hat{\beta}$ estimates are unbiased with a variance of

$$Var(\hat{\beta}) = \sigma^2(X'X)^{-1} \tag{5}$$

The Multiple Correlation R

The correlation between the predicted scores $\hat{y} = X\hat{\beta}$ and the criterion scores is called the *multiple correlation coefficient*, and is almost universally denoted with the value R.

Since R is always positive, and R^2 is the percentage of variance in y accounted for by the predictors. (in the colloquial sense), most discussions center on R^2 rather than R.

When it is necessary for clarity, one can denote the squared multiple correlation as $R^2_{y|x_1 x_2}$ to indicate that variates x_1 and x_2 have been included in the regression equation.

Bias of the Sample R^2

When a population correlation is zero, the sample correlation is hardly ever zero. As a consequence, the R^2 value obtained in an

analysis of sample data is a biased estimate of the corresponding population value.

An unbiased estimator exists (Olkin and Pratt, 1958), but is not available in standard statistics packages. As a result, most packages compute an approximate *shrunken* (or *adjusted*) estimate and report it alongside the uncorrected value. The adjusted estimator when there are k predictors is

$$\tilde{R}^2 = 1 - (1 - R^2)\frac{N-1}{N-k-1} \tag{6}$$

A Host of Challenges

Specifying a multiple regression model has all the challenges of bivariate regression, and more. These include:

- Significance tests and confidence intervals for R^2
- Methods for assessing model fit
- Selecting input variables and predictors
- Choosing appropriate transforms to achieve linearity
- Dealing with collinearity
- Deciding whether to include interactions between input variables
- Detecting outliers in the multivariate framework

Some of these issues are unique to the multivariate arena, while others are a more challenging version of issues we also confront in bivariate regression.

A Host of Challenges

Specifying a multiple regression model has all the challenges of bivariate regression, and more. These include:

- Significance tests and confidence intervals for R^2
- Methods for assessing model fit
- Selecting input variables and predictors
- Choosing appropriate transforms to achieve linearity
- Dealing with collinearity
- Deciding whether to include interactions between input variables
- Detecting outliers in the multivariate framework

Some of these issues are unique to the multivariate arena, while others are a more challenging version of issues we also confront in bivariate regression.

A Host of Challenges

Specifying a multiple regression model has all the challenges of bivariate regression, and more. These include:

- Significance tests and confidence intervals for R^2
- Methods for assessing model fit
- Selecting input variables and predictors
- Choosing appropriate transforms to achieve linearity
- Dealing with collinearity
- Deciding whether to include interactions between input variables
- Detecting outliers in the multivariate framework

Some of these issues are unique to the multivariate arena, while others are a more challenging version of issues we also confront in bivariate regression.

A Host of Challenges

Specifying a multiple regression model has all the challenges of bivariate regression, and more. These include:

- Significance tests and confidence intervals for R^2
- Methods for assessing model fit
- Selecting input variables and predictors
- Choosing appropriate transforms to achieve linearity
- Dealing with collinearity
- Deciding whether to include interactions between input variables
- Detecting outliers in the multivariate framework

Some of these issues are unique to the multivariate arena, while others are a more challenging version of issues we also confront in bivariate regression.

A Host of Challenges

Specifying a multiple regression model has all the challenges of bivariate regression, and more. These include:

- Significance tests and confidence intervals for R^2
- Methods for assessing model fit
- Selecting input variables and predictors
- Choosing appropriate transforms to achieve linearity
- Dealing with collinearity
- Deciding whether to include interactions between input variables
- Detecting outliers in the multivariate framework

Some of these issues are unique to the multivariate arena, while others are a more challenging version of issues we also confront in bivariate regression.

A Host of Challenges

Specifying a multiple regression model has all the challenges of bivariate regression, and more. These include:

- Significance tests and confidence intervals for R^2
- Methods for assessing model fit
- Selecting input variables and predictors
- Choosing appropriate transforms to achieve linearity
- Dealing with collinearity
- Deciding whether to include interactions between input variables
- Detecting outliers in the multivariate framework

Some of these issues are unique to the multivariate arena, while others are a more challenging version of issues we also confront in bivariate regression.

A Host of Challenges

Specifying a multiple regression model has all the challenges of bivariate regression, and more. These include:

- Significance tests and confidence intervals for R^2
- Methods for assessing model fit
- Selecting input variables and predictors
- Choosing appropriate transforms to achieve linearity
- Dealing with collinearity
- Deciding whether to include interactions between input variables
- Detecting outliers in the multivariate framework

Some of these issues are unique to the multivariate arena, while others are a more challenging version of issues we also confront in bivariate regression.

Test of $R^2 = 0$

A routine test of the hypothesis that $R^2 = 0$ is performed with an F statistic.

$$F_{k,N-k-1} = \frac{R^2/k}{(1-R^2)/n-k-1} \tag{7}$$

$$= \frac{SS_{\hat{y}}/k}{SS_{\epsilon}/(N-k-1)} \tag{8}$$

where

$$SS_{\hat{y}} = \sum_{i=1}^{N} (\hat{y}_i - \overline{y})^2 \tag{9}$$

and

$$SS_{\epsilon} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} \epsilon_i^2 \tag{10}$$

Confidence Interval for R^2

Most major statistical packages do not report an exact confidence interval for R^2, although one is available. This

confidence interval can be quite revealing when precision of estimate for R^2 is inadequate.

Partial F Test of R^2 change

Suppose you have k predictors $x_1, x_2, \ldots, x_k$, and you add a new predictor w. The test that this new predictor has significantly improved R^2 (the null hypothesis is that there is no change) is:

$$F_{1,N-k-2} = \frac{R^2_{new} - R^2_{old}}{R^2_{new}/(N-k-2)} \quad (11)$$

$$= \frac{SS_{\hat{y}(new)} - SS_{\hat{y}(old)}}{SS_{\epsilon(new)}} \quad (12)$$

Selecting Input Variables and Predictors

Selecting the input variables is often not an issue — there are only a few variables, and they were pre-selected because of their relevance. Many of the examples of Gelman & Hill start with a small set of input variables.

However, in some "exploratory" situations, there is a large list of potential X variables. A number of different techniques for selecting input variables are standard in major statistics packages.

Forward Selection

Forward selection proceeds as follows:

1. You select a group of independent variables to be examined
2. The variable with the highest squared correlation with the criterion is added to the regression equation
3. The partial F statistic for each possible remaining variable is computed
4. If the variable with the highest F statistic passes a criterion, it is added to the regression equation, and R^2 is recomputed
5. Keep going back to step 3, recomputing the partial F statistics until no variable can be found that passes the criterion for significance

Forward Selection

Forward selection proceeds as follows:

1. You select a group of independent variables to be examined
2. The variable with the highest squared correlation with the criterion is added to the regression equation
3. The partial F statistic for each possible remaining variable is computed
4. If the variable with the highest F statistic passes a criterion, it is added to the regression equation, and R^2 is recomputed
5. Keep going back to step 3, recomputing the partial F statistics until no variable can be found that passes the criterion for significance

Forward Selection

Forward selection proceeds as follows:

1. You select a group of independent variables to be examined
2. The variable with the highest squared correlation with the criterion is added to the regression equation
3. The partial F statistic for each possible remaining variable is computed
4. If the variable with the highest F statistic passes a criterion, it is added to the regression equation, and R^2 is recomputed
5. Keep going back to step 3, recomputing the partial F statistics until no variable can be found that passes the criterion for significance

Forward Selection

Forward selection proceeds as follows:

1. You select a group of independent variables to be examined
2. The variable with the highest squared correlation with the criterion is added to the regression equation
3. The partial F statistic for each possible remaining variable is computed
4. If the variable with the highest F statistic passes a criterion, it is added to the regression equation, and R^2 is recomputed
5. Keep going back to step 3, recomputing the partial F statistics until no variable can be found that passes the criterion for significance

Forward Selection

Forward selection proceeds as follows:

1. You select a group of independent variables to be examined
2. The variable with the highest squared correlation with the criterion is added to the regression equation
3. The partial F statistic for each possible remaining variable is computed
4. If the variable with the highest F statistic passes a criterion, it is added to the regression equation, and R^2 is recomputed
5. Keep going back to step 3, recomputing the partial F statistics until no variable can be found that passes the criterion for significance

Backward Selection

Backward elimination:

1. You start with all the variables you have selected as possible predictors included in the regression equation
2. You then compute partial F statistics for each of the variables remaining in the regression equation
3. Find the variable with the lowest F
4. If this F is low enough to be below a criterion you have selected, remove it from the model, and go back to step 2
5. Continue until no partial F is found that is sufficiently low

Backward Selection

Backward elimination:

1. You start with all the variables you have selected as possible predictors included in the regression equation
2. You then compute partial F statistics for each of the variables remaining in the regression equation
3. Find the variable with the lowest F
4. If this F is low enough to be below a criterion you have selected, remove it from the model, and go back to step 2
5. Continue until no partial F is found that is sufficiently low

Backward Selection

Backward elimination:

1. You start with all the variables you have selected as possible predictors included in the regression equation
2. You then compute partial F statistics for each of the variables remaining in the regression equation
3. Find the variable with the lowest F
4. If this F is low enough to be below a criterion you have selected, remove it from the model, and go back to step 2
5. Continue until no partial F is found that is sufficiently low

Backward Selection

Backward elimination:

1. You start with all the variables you have selected as possible predictors included in the regression equation
2. You then compute partial F statistics for each of the variables remaining in the regression equation
3. Find the variable with the lowest F
4. If this F is low enough to be below a criterion you have selected, remove it from the model, and go back to step 2
5. Continue until no partial F is found that is sufficiently low

Backward Selection

Backward elimination:

1. You start with all the variables you have selected as possible predictors included in the regression equation
2. You then compute partial F statistics for each of the variables remaining in the regression equation
3. Find the variable with the lowest F
4. If this F is low enough to be below a criterion you have selected, remove it from the model, and go back to step 2
5. Continue until no partial F is found that is sufficiently low

Stepwise Selection

Stepwise regression works like forward regression except that you examine, at each stage, the possibility that a variable entered at a previous stage has now become superfluous because of additional variables now in the model that were not in the model when this variable was selected.

To check on this, at each step a partial F test for each variable in the model is made as if it were the variable entered last. We look at the lowest of these Fs and if the lowest one is sufficiently low, we remove the variable from the model, recompute all the partial Fs, and keep going until we can remove no more variables.

Multiple Regression in R

The "Kids Data" data set contains heights, weights, and ages for 12 children.

```
> kids.data ← read.table("KidsData.txt",header=T)
> kids.data
```

	WGT	HGT	AGE
1	64	57	8
2	71	59	10
3	53	49	6
4	67	62	11
5	55	51	8
6	58	50	7
7	77	55	10
8	57	48	9
9	56	42	10
10	51	42	6
11	76	61	12
12	68	57	9

Multiple Regression in R

We'll try fitting 3 models. We'll start with just the intercept, then add the HGT input variable, and next add AGE.

```
> attach(kids.data)
> m0 <- lm(WGT~1)
> m1 <- lm(WGT~HGT)
> m2 <- lm(WGT~HGT+AGE)
> m0

Call:
lm(formula = WGT ~ 1)

Coefficients:
(Intercept)
      62.75

> m1

Call:
lm(formula = WGT ~ HGT)

Coefficients:
(Intercept)          HGT
      6.190        1.072

> m2

Call:
lm(formula = WGT ~ HGT + AGE)

Coefficients:
(Intercept)          HGT          AGE
      6.553        0.722        2.050
```

Multiple Regression in R

Comparing the models is often done by analysis of variance.

```
> anova(m0,m1,m2)

Analysis of Variance Table

Model 1: WGT ~ 1
Model 2: WGT ~ HGT
Model 3: WGT ~ HGT + AGE
  Res.Df    RSS Df Sum of Sq       F    Pr(>F)
1     11 888.25
2     10 299.33  1    588.92 27.1216 0.0005582 ***
3      9 195.43  1    103.90  4.7849 0.0564853 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Forward Selection

Weisberg gives an example of forward regression in Chapter 10.

R uses the AIC (Akaike Information Criterion) instead of the F statistic in its `step` command.

Forward Selection

TABLE 10.5 Definition of Terms for the Highway Accident Data

Variable	Description
log(*Rate*)	Base-two logarithm of 1973 accident rate per million vehicle miles, the response
log(*Len*)	Base-two logarithm of the length of the segment in miles
log(*ADT*)	Base-two logarithm of average daily traffic count in thousands
log(*Trks*)	Base-two logarithm of truck volume as a percent of the total volume
Slim	1973 speed limit
Lwid	Lane width in feet
Shld	Shoulder width in feet of outer shoulder on the roadway
Itg	Number of freeway-type interchanges per mile in the segment
log(*Sigs1*)	Base-two logarithm of (number of signalized interchanges per mile in the segment + 1)/(length of segment)
Acpt	Number of access points per mile in the segment
Hwy	A factor coded 0 if a federal interstate highway, 1 if a principal arterial highway, 2 if a major arterial, and 3 otherwise

Forward Selection

```
> data(highway)
> a ← highway
> a$logADT ← logb(a$ADT,2)
> a$logTrks ← logb(a$Trks,2)
> a$logLen ← logb(a$Len,2)
> a$logSigs1 ← logb((a$Sigs*a$Len+1)/a$Len,2)
> a$logRate ← logb(a$Rate,2)
> # set the contrasts to the R default
> options(contrasts=c(factor="contr.treatment",ordered="contr.p
> a$Hwy ← if(is.null(version$language) == FALSE) factor(a$Hwy,
> attach(a)
> names(a)

 [1] "ADT"     "Trks"     "Lane"    "Acpt"     "Sigs"     "Itg"
 [7] "Slim"    "Len"      "Lwid"    "Shld"     "Hwy"      "Rate"
[13] "logADT"  "logTrks"  "logLen"  "logSigs1" "logRate"

> cols ← c(17,15,13,14,16,7,10,3,4,6,9,11)
> m1 ← lm(logRate ~ logLen+logADT+logTrks+logSigs1+Slim+Shld+
+                  Lane+Acpt+Itg+Lwid+Hwy)
```

Forward Selection

```
> summary(m1)

Call:
lm(formula = logRate ~ logLen + logADT + logTrks + logSigs1 +
    Slim + Shld + Lane + Acpt + Itg + Lwid + Hwy)

Residuals:
      Min        1Q    Median        3Q       Max
-0.646354 -0.147045 -0.009977  0.176454  0.607610

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.704639   2.547137   2.240   0.0342 *
logLen      -0.214470   0.099986  -2.145   0.0419 *
logADT      -0.154625   0.111893  -1.382   0.1792
logTrks     -0.197560   0.239812  -0.824   0.4178
logSigs1     0.192322   0.075367   2.552   0.0172 *
Slim        -0.039327   0.024236  -1.623   0.1172
Shld         0.004291   0.049281   0.087   0.9313
Lane        -0.016061   0.082264  -0.195   0.8468
Acpt         0.008727   0.011687   0.747   0.4622
Itg          0.051536   0.350312   0.147   0.8842
Lwid         0.060769   0.197391   0.308   0.7607
Hwy1         0.342705   0.576821   0.594   0.5578
Hwy2        -0.412295   0.393960  -1.047   0.3053
Hwy3        -0.207358   0.336809  -0.616   0.5437
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3761 on 25 degrees of freedom
Multiple R-squared: 0.7913,     Adjusted R-squared: 0.6828
F-statistic: 7.293 on 13 and 25 DF,  p-value: 1.247e-05
```

Forward Selection

```
> m0 <- lm(logRate ~logLen,data=a)
> ansf1 <- step(m0,scope=list(lower=~logLen,
+           upper=~logLen+logADT+logTrks+logSigs1+Slim+Shld+
+                   Lane+Acpt+Itg+Lwid+Hwy),
+                   direction="forward", data=a)
Start:  AIC=-43.92
logRate ~ logLen

           Df Sum of Sq    RSS     AIC
+ Slim      1     5.302  6.112 -66.278
+ Acpt      1     4.374  7.040 -60.767
+ Shld      1     3.553  7.861 -56.464
+ logSigs1  1     2.001  9.413 -49.437
+ Hwy       3     2.789  8.625 -48.848
+ logTrks   1     1.515  9.898 -47.477
+ logADT    1     0.892 10.522 -45.094
<none>                  11.414 -43.921
+ Lane      1     0.547 10.867 -43.835
+ Itg       1     0.452 10.962 -43.496
+ Lwid      1     0.385 11.029 -43.259

Step:  AIC=-66.28
logRate ~ logLen + Slim

           Df Sum of Sq    RSS     AIC
+ Acpt      1     0.600  5.512 -68.310
+ logTrks   1     0.548  5.564 -67.940
<none>                   6.112 -66.278
+ logSigs1  1     0.305  5.807 -66.277
+ Hwy       3     0.700  5.412 -65.024
+ Shld      1     0.068  6.044 -64.714
+ logADT    1     0.053  6.059 -64.620
+ Lwid      1     0.035  6.078 -64.500
+ Lane      1     0.007  6.105 -64.324
+ Itg       1     0.006  6.107 -64.313

Step:  AIC=-68.31
logRate ~ logLen + Slim + Acpt

           Df Sum of Sq    RSS     AIC
+ logTrks   1     0.360  5.152 -68.944
<none>                   5.512 -68.310
+ logSigs1  1     0.250  5.262 -68.120
+ Shld      1     0.072  5.440 -66.823
+ logADT    1     0.032  5.480 -66.534
+ Lane      1     0.031  5.481 -66.530
+ Itg       1     0.028  5.484 -66.509
+ Lwid      1     0.026  5.485 -66.497
+ Hwy       3     0.453  5.059 -65.652

Step:  AIC=-68.94
logRate ~ logLen + Slim + Acpt + logTrks

           Df Sum of Sq    RSS     AIC
<none>                   5.152 -68.944
+ Shld      1     0.136  5.016 -67.987
+ logSigs1  1     0.105  5.047 -67.749
+ logADT    1     0.065  5.087 -67.439
+ Hwy       3     0.540  4.612 -67.263
+ Lwid      1     0.040  5.112 -67.245
+ Itg       1     0.023  5.129 -67.117
+ Lane      1     0.007  5.145 -66.996
```

Setting Up Interaction Terms

```
> Slim.centered ← Slim - mean(Slim)
> two.var.fit ← lm(logRate ~ logLen + Slim.centered)
> summary(two.var.fit)

Call:
lm(formula = logRate ~ logLen + Slim.centered)

Residuals:
     Min       1Q   Median       3Q      Max
-0.63450 -0.30111  0.01509  0.29034  1.05981

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.92531    0.28230  10.363 2.38e-12 ***
logLen        -0.32122    0.07964  -4.033 0.000274 ***
Slim.centered -0.06621    0.01185  -5.588 2.47e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.412 on 36 degrees of freedom
Multiple R-squared: 0.6394,	Adjusted R-squared: 0.6194
F-statistic: 31.92 on 2 and 36 DF,  p-value: 1.062e-08
```

Setting Up Interaction Terms

In the model specification language, two way interactions are set up as follows:

```
> interaction.fit <- lm(logRate ~ logLen + Slim.centered +
+  logLen:Slim.centered)
> summary(interaction.fit)

Call:
lm(formula = logRate ~ logLen + Slim.centered + logLen:Slim.centered)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.63451 -0.29502  0.01204  0.28903  1.05641 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           2.935651   0.295048   9.950 9.67e-12 ***
logLen               -0.323567   0.082375  -3.928 0.000384 ***
Slim.centered        -0.060553   0.040994  -1.477 0.148589    
logLen:Slim.centered -0.001711   0.011856  -0.144 0.886090    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.4178 on 35 degrees of freedom
Multiple R-squared: 0.6396,	Adjusted R-squared: 0.6087 
F-statistic: 20.71 on 3 and 35 DF,  p-value: 6.847e-08
```