Introduction to Multiple Regression

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

Introduction to Multiple Regression

1. The Multiple Regression Model
2. Some Key Regression Terminology
3. The Kids Data Example
   - Visualizing the Data – The Scatterplot Matrix
   - Regression Models for Predicting Weight
4. Understanding Regression Coefficients
5. Statistical Testing in the Fixed Regressor Model
   - Introduction
   - Partial *F*-Tests: A General Approach
   - Partial *F*-Tests: Overall Regression
   - Partial *F*-Tests: Adding a Single Term
6. Variable Selection in Multiple Regression
   - Introduction
   - Forward Selection
   - Backward Elimination
   - Stepwise Regression
   - Automatic Single-Term Sequential Testing in R
7. Variable Selection in R
   - Problems with Statistical Testing in the Variable Selection Context
8. Information-Based Selection Criteria
   - The Active Terms
   - Information Criteria
9. (Estimated)Standard Errors
10. Standard Errors for Predicted and Fitted Values

The Multiple Regression Model

The simple linear regression model states that

$$\begin{aligned} \mathrm{E}(Y|X=x) &= \beta_0 + \beta_1 x && (1) \\ \mathrm{Var}(Y|X=x) &= \sigma^2 && (2) \end{aligned}$$

In the multiple regression model, we simply add one or more predictors to the system. For example, if we add a single predictor X_2, we get

$$E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \qquad (3)$$

More generally, if we incorporate the intercept term as a 1 in $\mathbf{x}$, and place all the β's (including β_0) in a vector we can say that

$$\begin{aligned} \mathrm{E}(Y|\mathbf{x}=\mathbf{x}^*) &= \mathbf{x}^{*\prime}\boldsymbol{\beta} && (4) \\ \mathrm{Var}(Y|\mathbf{x}=\mathbf{x}^*) &= \sigma^2 && (5) \end{aligned}$$

Challenges in Multiple Regression

Dealing with multiple predictors is considerably more challenging than dealing with only a single predictor. Some of the problems include:

- *Choosing the best model.* In multiple regression, often several different sets of variables perform equally well in predicting a criterion. Which set should you use?
- *Interactions between variables*. In some cases, independent variables interact, and the regression equation will not be accurate unless this interaction is taken into account.

Challenges in Multiple Regression

- *Much greater difficulty visualizing the regression relationships*. With only one independent variable, the regression line can be plotted neatly in two dimensions. With two predictors, there is a *regression surface* instead of a regression line, and with 3 predictors and one criterion, you run out of dimensions for plotting.
- *Model interpretation becomes substantially more difficult*. The multiple regression equation changes as each new variable is added to the model. Since the regression weights for each variable are modified by the other variables, and hence depend on what is in the model, the substantive interpretation of the regression equation is problematic.

Some Key Regression Terminology

Introduction

In Section 3.3 of ALR, Weisberg introduces a number of key ideas and nomenclature in connection with a regression model of the form

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{6}$$

Some Key Regression Terminology

Predictors vs. Terms

- Regression problems start with a collection of potential predictors.
- Some of these may be continuous measurements, like the height or weight of an object.
- Some may be discrete but ordered, like a doctor's rating of overall health of a patient on a nine-point scale.
- Other potential predictors can be categorical, like eye color or an indicator of whether a particular unit received a treatment.
- All these types of potential predictors can be useful in multiple linear regression.
- A key notion is the distinction between *predictors* and *terms in the regression equation*.
- In early discussions, these are often synonymous. However, we quickly learn that they need not be.

Some Key Regression Terminology

Types of Terms

Many types of *terms* can be created from a group of predictors. Here are some examples

- *The intercept.* We can rewrite the mean function on the previous slide as
$$E(Y|X) = \beta_0 X_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{7}$$
where X_0 is a term that is always equal to one. Mean functions without an intercept would not have this term included.
- *Predictors.* The simplest type of term is simply one of the predictors.

Some Key Regression Terminology

Types of Terms

- *Transformations of predictors.* Often we will transform one of the predictors to create a term. For example, X_1 in a previous example was the logarithm of one of the predictors.
- *Polynomials.* Sometimes, we fit curved functions by including polynomial terms in the predictor variables. So, for example, X_1 might be a predictor, and X_2 might be its square.
- *Interactions and other Combinations of Predictors.* Combining several predictors is often useful. An example of this is using body mass index, given by height divided by weight squared, in place of both height and weight, or using a total test score in place of the separate scores from each of several parts. Products of predictors called *interactions* are often included in a mean function along with the original predictors to allow for joint effects of two or more variables.

Some Key Regression Terminology

Types of Terms

- *Dummy Variables and Factors.* A categorical predictor with two or more levels is called a factor. Factors are included in multiple linear regression using dummy variables, which are typically terms that have only two values, often zero and one, indicating which category is present for a particular observation. We will see in ALR, Chapter 6 that a categorical predictor with two categories can be represented by one dummy variable, while a categorical predictor with many categories can require several dummy variables.

Comment. A regression with k predictors may contain fewer than k terms or more than k terms.

Kids Data

Example (The Kids Data)

As an example consider the following data from the Kleinbaum, Kupper and Miller text on regression analysis. These data show weight, height, and age of a random sample of 12 nutritionally deficient children. The data are available online in the file `KidsDataR.txt`.

Kids Data

WGT(y)	**HGT**(x_1)	**AGE**(x_2)
64	57	8
71	59	10
53	49	6
67	62	11
55	51	8
58	50	7
77	55	10
57	48	9
56	42	10
51	42	6
76	61	12
68	57	9

The Scatterplot Matrix

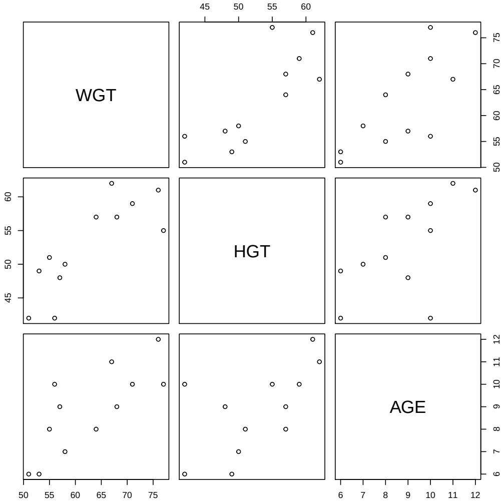
- The scatterplot matrix on the next slide shows that both HGT and AGE are strongly linearly related to WGT. However, the two potential predictors are also strongly linearly related to each other.
- This is corroborated by the correlation matrix for the three variables.

```
> kids.data <- read.table("KidsDataR.txt", header = T, sep = ",")
> cor(kids.data)

       WGT    HGT    AGE
WGT 1.0000 0.8143 0.7698
HGT 0.8143 1.0000 0.6138
AGE 0.7698 0.6138 1.0000
```

The Scatterplot Matrix

```
> pairs(kids.data)
```



Potential Regression Models

- The situation here is relatively simple.
- We can see that height is the best predictor of weight.
- Age is also an excellent predictor, but because it is also correlated with height, it may not add too much to the prediction equation.
- We fit the two models in succession. The first model has only height as a predictor, while the second adds age.
- In the following slides, we'll perform the standard linear model analysis, and discuss the results, after which we'll comment briefly on the theory underlying the methods.

```
> attach(kids.data)
> model.1 <- lm(WGT ~ HGT)
> model.2 <- lm(WGT ~ HGT + AGE)
> summary(model.1)
> summary(model.2)
```

Fitting the Models

```
Call:
lm(formula = WGT ~ HGT)

Residuals:
   Min     1Q Median     3Q    Max 
 -5.87  -3.90  -0.44   2.26  11.84 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)    6.190     12.849    0.48   0.6404   
HGT            1.072      0.242    4.44   0.0013 **
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.47 on 10 degrees of freedom
Multiple R-squared:  0.663,Adjusted R-squared:  0.629 
F-statistic: 19.7 on 1 and 10 DF,  p-value: 0.00126 

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

Residuals:
   Min     1Q Median     3Q    Max 
-6.871 -1.700  0.345  1.464 10.234 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)    6.553     10.945    0.60    0.564  
HGT            0.722      0.261    2.77    0.022 *
AGE            2.050      0.937    2.19    0.056 .
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.66 on 9 degrees of freedom
Multiple R-squared:   0.78,Adjusted R-squared:  0.731 
F-statistic:   16 on 2 and 9 DF,  p-value: 0.0011 
```

Comparing the Models with ANOVA

```
> anova(model.1, model.2)

Analysis of Variance Table

Model 1: WGT ~ HGT
Model 2: WGT ~ HGT + AGE
  Res.Df RSS Df Sum of Sq    F Pr(>F)
1     10 299
2      9 195  1       104 4.78  0.056 .
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Squared Multiple Correlation Coefficient R^2

- The correlation between the predicted scores and the criterion scores is called the *multiple correlation coefficient*,and is almost universally denoted with the value R.
- Curiously, many writers use this notation whether a sample or a population value is referred to, which creates some problems for some readers.
- We can eliminate this ambiguity by using either ρ^2 or R^2_{pop} to signify the population value.
- 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 $R^2_{y|x_1x_2}$ to indicate that variates x_1 and x_2 have been included in the regression equation.

The Partial Correlation Coefficient

- The *partial correlation coefficient* is a measure of the strength of the linear relationship between two variables after the contribution of other variables has been “partialled out” or “controlled for” using linear regression.
- We will use the notation $r_{yx|w_1,w_2,\ldots w_p}$ to stand for the partial correlation between y and x with the w's partialled out.
- This correlation is simply the Pearson correlation between the regression residual $\varepsilon_{y|w_1,w_2,\ldots w_p}$ for y with the w's as predictors and the regression residual $\varepsilon_{x|w_1,w_2,\ldots w_p}$of x with the w's as predictors.

Partial Regression Coefficients

- In a similar approach to calculating partial correlation coefficients, we can also calculate partial regression coefficients.
- For example, the partial regression between y and x_j with the other x's partialled out is simply the slope of the regression line for predicting the residual of y with the other x's partialled out from that of x_j with the other x's partialled out.

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 population value.
- An unbiased estimator is available (Olkin and Pratt, 1958), but requires very powerful software like *Mathematica* to compute, and consequently is not available in standard statistics packages. As a result, these packages compute an approximate "shrunken" (or "adjusted") estimate and report it alongside the uncorrected value. The adusted estimator when there are k predictors is

$$\widetilde{R}^2 = 1 - (1 - R^2)\frac{N-1}{N-k-1} \tag{8}$$

Understanding Regression Coefficients

- The β weights in a regression equation can change when a new predictor term is added.
- This is because the regression weights are, in fact, partial regression weights.
- That is, the β weight for predicting y from x_j is the regression coefficient for predicting the residual of y after partialling out all other predictors from the residual of x_j after partialling out all the other predictors.
- Some authors discuss this using Venn diagrams of "overlapping variance." (See next slide.)
- With modern graphics engines, we can quickly examine the actual scatterplots of the partial regressions. R will construct them automatically with the `av.plots` command.

Venn Diagrams

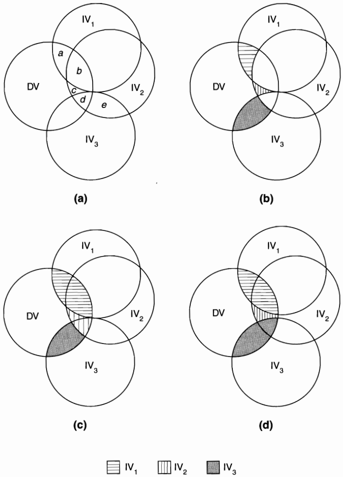
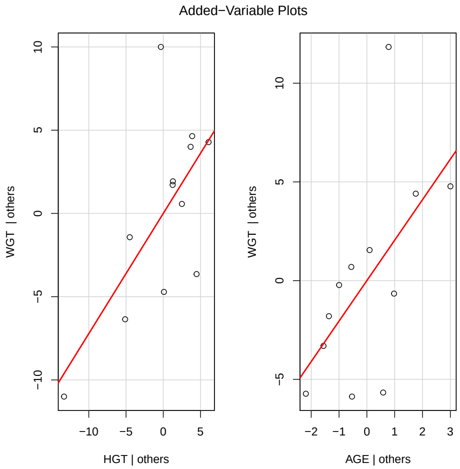


FIGURE 5.2 Venn diagrams illustrating (a) overlapping variance sections; and allocation of overlapping variance in (b) standard multiple regression, (c) sequential regression, and (d) stepwise regression.

Added Variable Plots

```
> library(car)
> av.plots(model.2)

Warning:  'av.plots' is deprecated.
Use 'avPlots' instead.
See help("Deprecated") and help("car-deprecated").
```



Introduction

- In connection with regression models, we've seen two different F tests.
- One is the test of significance for the overall regression. This tests the null hypothesis that, for the current model, $R^2 = 0$ in the population.
- The other test we frequently see is a model comparison F, which tests the hypothesis that the R^2 for the more complex model (which has all the terms of the previous model *and* some additional ones) is statistically significantly larger than the R^2 for the less complex model.

Partial F-Tests: A General Approach

- Actually, the F-tests we've seen are a special case of a general procedure for generating *partial F-tests* on a *nested sequence* of models.
- Consider a sequence of J models $M_j, j = 1, \ldots, J$. Suppose Model M_k includes Model M_j as a special case for all pairs of values of $j < k \leq J$. That is, Model M_j is a special case of Model M_k where some terms have coefficients of zero. Then Model M_j is nested within Model M_k for all these values, and we say the set of models is a *nested sequence*.
- Define t_j and t_k respectively as the number of terms *including the intercept term* in models M_j and M_k.
- As a mnemonic device, associate k with *complex*, because model M_k is more complex (has more terms) than model M_j.

Partial F-Tests: A General Approach

- Consider pairs of models in this nested sequence. If we define SS_j to be the sum of squared residuals for less complex model Model M_j, SS_k the sum of squared residuals for more complex Model M_k, df_j to be $n - t_j$ and $df_k = n - t_k$, then SS_k will always be less than or equal to SS_j, because, as the more complex nesting model, model M_k can always achieve identical fit to model M_j simply by setting estimates for all its additional parameters to zero. To statistically compare Model M_j against Model M_k, we compute the partial F-statistic as follows.

$$F_{df_j - df_k, df_k} = \frac{MS_{comparison}}{MS_{res}} = \frac{(SS_j - SS_k)/(t_k - t_j)}{SS_k/df_k} \tag{9}$$

- The statistical null hypothesis is that the two models fit equally well, that is, the more complex model M_k has no better fit than M_j.

Partial F-Tests: Overall Regression

- The overall F test in linear regression is routinely reported in regression output when testing a model with one or more predictor terms in addition to an intercept. It tests the hypothesis that $R^2_{pop} = 0$, against the alternative that $R^2_{pop} > 0$.
- The overall F test is simply a partial F test comparing a regression model M_k with t_k terms (including an intercept) with a model M_1 that has only one *intercept* term.
- Now, a model that has only an intercept term must, in least squares regression, define the intercept coefficient β_0 to be $\overline{y}$, the mean of the y scores, because it is well known that the sample mean is that value around which the sum of squared deviations is a minimum.
- So SS_j for the model with only an intercept term becomes SS_y, the sum of squared deviations around the mean for the dependent variable.

Partial F-Tests: Overall Regression

- Since Model M_k has $p = t_k - 1$ predictor terms, and Model M_j has one (i.e., the intercept), the degrees of freedom for regression become $t_k - t_j = (p+1) - 1 = p$, and we have, for the test statistic,

$$F_{k,n-p-1} = \frac{(SS_y - SS_k)/(p)}{SS_k/(n-p-1)} = \frac{SS_{\hat{y}}/p}{SS_e/(n-p-1)} \tag{10}$$

- Now, in traditional notation, SS_k, being the sum of squared errors for the regression model we are testing, is usually called SS_e.
- Since $SS_y = SS_{\hat{y}} + SS_e$, we can replace $SS_y - SS_k$ with $SS_{\hat{y}}$.
- Remembering that $R^2 = SS_{\hat{y}}/SS_y$, we can show that the F statistic is also equal to

$$F_{p,n-p-1} = \frac{R^2/p}{(1-R^2)/(n-p-1)} \tag{11}$$

Partial F-Tests: Adding a Single Term

- If we are *adding* a single term to a model that currently has p predictors plus an intercept, the model comparison test becomes

$$F_{1,n-p-2} = \frac{(SS_k - SS_j)/(1)}{SS_k/(n-p-2)} = \frac{SS_{\hat{y}k} - SS_{\hat{y}j}}{SS_k/(n-p-2)} \tag{12}$$

- Remembering that $R^2 = SS_{\hat{y}}/SS_y$, and that $SS_y = SS_{\hat{y}} + SS_e$, we can show that the F statistic is also equal to

$$F_{1,n-p-2} = \frac{R_k^2 - R_j^2}{(1-R_k^2)/(n-p-2)} \tag{13}$$

Introduction

- When there are only a few potential predictors, or theory dictates a model, selecting which variables to use as predictors is relatively straightforward.
- When there are many potential predictors, the problem becomes more complex, although modern computing power has opened up opportunities for exploration.

Forward Selection

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 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 Regression

- This 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.

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 that the order of testing matters, and significance level for a term's model comparison test depends on the terms entered before it.

Automatic Sequential Testing of Single Terms

- For example, for the `kids.data`, we entered *HGT* first, then *AGE*, and so our model was

```
> model.2 <- lm(WGT ~ HGT + AGE)
```

- Here is the report on the sequential tests.

```
> anova(model.2)

Analysis of Variance Table


Response: WGT
          Df Sum Sq Mean Sq F value  Pr(>F)
HGT        1    589     589   27.12 0.00056 ***
AGE        1    104     104    4.78 0.05649 .
Residuals  9    195      22
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Notice that *HGT*, when entered first, has a *p*-value of .0005582.

Automatic Sequential Testing of Single Terms

- Next, try a model with the same two variables listed in reverse order.
- R will test the terms with sequential difference tests, and now the *p*-value for *HGT* will be higher.
- In colloquial terms, *HGT* is "less significant" when entered after *AGE*, because *AGE* can predict much of the variance predicted by *HGT* and so *HGT* has much less to add after *AGE* is already in the equation.

```
> model.2b <- lm(WGT ~ AGE + HGT)
> anova(model.2b)

Analysis of Variance Table

Response: WGT
          Df Sum Sq Mean Sq F value  Pr(>F)
AGE        1    526     526   24.24 0.00082 ***
HGT        1    166     166    7.66 0.02181 *
Residuals  9    195      22
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Automatic Sequential Testing of Single Terms

```
> summary(model.2b)


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

Residuals:
   Min     1Q Median     3Q    Max
-6.871 -1.700  0.345  1.464 10.234

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.553     10.945    0.60    0.564
AGE            2.050      0.937    2.19    0.056 .
HGT            0.722      0.261    2.77    0.022 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.66 on 9 degrees of freedom
Multiple R-squared:  0.78,Adjusted R-squared:  0.731
F-statistic:   16 on 2 and 9 DF,  p-value: 0.0011
```

Automatic Sequential Testing of Single Terms

- Notice also that the difference test *p*-value for the last variable entered is the same as the *p*-values reported in the overall output for the full model, but, in general, the other *p*-values will not be the same.

```
> anova(model.2b)

Analysis of Variance Table

Response: WGT
          Df Sum Sq Mean Sq F value  Pr(>F)
AGE        1    526     526   24.24 0.00082 ***
HGT        1    166     166    7.66 0.02181 *
Residuals  9    195      22
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary(model.2b)


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

Residuals:
   Min     1Q Median     3Q    Max
-6.871 -1.700  0.345  1.464 10.234

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.553     10.945    0.60    0.564
AGE            2.050      0.937    2.19    0.056 .
HGT            0.722      0.261    2.77    0.022 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.66 on 9 degrees of freedom
Multiple R-squared:  0.78,Adjusted R-squared:  0.731
F-statistic:   16 on 2 and 9 DF,  p-value: 0.0011
```

Problems with Statistical Testing

- Frequently multiple regressions are at least partially exploratory in nature.
- You gather data on a large number of predictors, and try to build a model for explaining (or predicting) y from a number of x's.
- A key aspect of this is choosing which x's to retain.
- A key problem is that, especially when n is small and the number of x's is large, there will be a number of *spuriously large correlations* between the criterion and the x's.
- You can capitalize on chance, as it were, and build a regression equation using variables that have high correlations with the criterion, but this equation will not generalize to any new situation.

Problems with Statistical Testing

- There are a number of statistical tests available in multiple regression, and they are printed routinely by statistical software such as SPSS, SAS, Statistica, SPLUS, and R.
- It is important to realize that these test do *not* in general correct for post hoc selection.
- So, for example, if you have 90 potential predictors that all actually correlate zero with the criterion, you can choose the predictor with the highest absolute correlation with the criterion in your current sample, and invariably obtain a "significant" result.
- Strangely, this fact is seldom brought to the forefront in textbook chapters on multiple regression.
- Consequently, people actually believe that the F statistics and associated probability values somehow determine whether the regression equation is significant in the sense most relatively naive users would expect.

Demonstration: Forward Regression with Random UncorrelatedPredictors

- We can demonstrate how Forward Regression or Stepwise Regression can produce wrong results.
- We begin by creating a list of names for our variables.

```
> names <- c("Y", paste("X", 1:90, sep = ""))
```

- Then we create a data matrix of order 50 × 91 containing totally independent normal random numbers.

```
> set.seed(12345)  #so we get the same data
> data <- matrix(rnorm(50 * 91), 50, 91)
```

- Then I add the column names to the data and turn the data matrix into a dataframe.

```
> colnames(data) <- names
> test.data <- data.frame(data)
> attach(test.data)
```

- Note that these data simulate samples from a population where $R^2 = 0$, as all the variables are uncorrelated.

Demonstration: Forward Regression with Random UncorrelatedPredictors

- We start the forward selection procedure (which is fully automated by SPSS) by looking for the X predictor that correlates most highly with the criterion variable Y. We can examine all the predictor-criterion correlations, sorted, using the following command, which grabs the first column and sorts its entries, then restrict the output to the largest 3 values:

```
> sort(cor(test.data)[, 1])[88:91]

   X48    X77    X53      Y
0.2568 0.3085 0.3876 1.0000
```

- Since we have been privileged to examine all the data and select the best predictor, the probability model on which the F-test for overall regression is based is no longer valid. We can see that $X53$ has a correlation of .388, despite the fact that the population correlation is zero. Here is the evaluation of model fit:

```
> summary(lm(Y ~ X53))


Call:
lm(formula = Y ~ X53)

Residuals:
   Min     1Q Median     3Q    Max
-2.171 -0.598  0.106  0.601  1.998

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.189      0.144    1.31   0.1979
X53            0.448      0.154    2.91   0.0054 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.02 on 48 degrees of freedom
Multiple R-squared:  0.15,Adjusted R-squared:  0.133
F-statistic: 8.49 on 1 and 48 DF,  p-value: 0.00541
```

- The regression is “significant” beyond the .01 level.

Demonstration: Forward Regression with Random UncorrelatedPredictors

- The next largest correlation is X77. Adding that to the equation produces a "significant" improvement, and an R^2 value of 0.26.

```
> summary(lm(Y ~ X53 + X77))


Call:
lm(formula = Y ~ X53 + X77)

Residuals:
   Min     1Q Median     3Q    Max
-2.314 -0.626  0.100  0.614  1.788

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.211      0.136    1.55   0.1284
X53            0.471      0.145    3.24   0.0022 **
X77            0.363      0.137    2.65   0.0110 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.963 on 47 degrees of freedom
Multiple R-squared:  0.261,Adjusted R-squared:  0.229
F-statistic: 8.28 on 2 and 47 DF,  p-value: 0.00083
```

- It is precisely because F tests perform so poorly under these conditions that alternative methods have been sought. Although R implements stepwise procedures in its `step` library, it does not use the F-statistic, but rather employs information-based criteria such as the AIC.
- In its `leaps` procedure, R implements an "all-possible-subsets" search for the best model.
- We shall examine the performance of some of these selection procedures in Homework 5.

The Active Terms

- One way of conceptualizing variable selection is to parse the available terms in the analysis into active and inactive groups.
- A simple notation for describing this is as follows:
- Given a response Y and a set of terms X, the idealized goal of variable selection is to divide X into two pieces, i.e., $X = (X_{\mathcal{A}}, X_{\mathcal{I}})$, where $X_{\mathcal{A}}$ is the set of active terms, and $X_{\mathcal{I}}$ is the set of inactive terms not needed to specify the mean function.
- $\mathrm{E}(Y|X_{\mathcal{A}}, X_{\mathcal{I}})$ and $\mathrm{E}(Y|X_{\mathcal{A}})$ would give the same results.

The Active Terms

- We could write
$$\mathrm{E}(Y|X=\mathbf{x}) = \boldsymbol{\beta}'_{\mathcal{A}}\mathbf{x}_{\mathcal{A}} + \boldsymbol{\beta}'_{\mathcal{I}}\mathbf{x}_{\mathcal{I}} \tag{14}$$

- If we have specified the model correctly, then to a close approximation, we should see $\boldsymbol{\beta}'_{\mathcal{I}} = \mathbf{0}$.

Information Criteria

- Suppose that we have a candidate subset $X_{\mathcal{C}}$, and that the selected subset is actually equal to the entire set of active terms $X_{\mathcal{A}}$.
- Then, of course (depending on sample size) the fit of the mean function including only $X_{\mathcal{C}}$ should be similar to the fit of the mean function including all the non-active terms.
- If $X_{\mathcal{C}}$ misses important terms, the residual sum of squares should be increased.

Information Criteria

The Akaike Information Criterion (AIC)

- Criteria for comparing various candidate subsets are based on the lack of fit of a model in this case, as assessed by the residual sum of squares (RSS), and its complexity, assessed by the number of terms in the model.
- Ignoring constants that are the same for every candidate subset, the AIC, or *Akaike Information Criterion*, is

$$AIC_{\mathcal{C}} = n \log(RSS_{\mathcal{C}}/n) + 2p_{\mathcal{C}} \tag{15}$$

- According to the Akaike criterion, the model with the smallest AIC is to be preferred.

Information Criteria

The Schwarz Bayesian Criterion

- This criterion is

$$BIC_{\mathcal{C}} = n \log(RSS_{\mathcal{C}}/n) + p_{\mathcal{C}} \log(n) \tag{16}$$

Information Criteria

- Mallows C_p criterion is defined as

$$C_{p\mathcal{C}} = \frac{RSS_{\mathcal{C}}}{\hat{\sigma}^2} + 2p_{\mathcal{C}} - n \tag{17}$$

  where $\hat{\sigma}^2$ is obtained from the fit of the model with all terms included.

- Note that, for a fixed number of parameters, all three criteria are monotonic in RSS.

Estimated Standard Errors

- Along with other statistical output, statistical software typically can provide a number of "standard errors."
- Since the estimates associated with these standard errors are asymptotically normally distributed, the standard errors can be used to construct Wald Tests and/or confidence intervals for the hypothesis that a parameter is zero in the population.
- Typically, software does not provide a confidence interval for R^2 itself, or even a standard error.
- The calculation of an exact confidence interval for R^2 is possible, and Rachel Fouladi and I provided the first computer program to do that, in 1992.

Standard Errors for Predicted and Fitted Values

- Recall that there are two related but distinct goals in regression analysis. One goal is *estimation*: from the data at hand, we wish to determine an optimal predictor set and accurately estimate β weights and R^2_{pop}.
- Another goal is *prediction*, and one variant of that involves estimation of $\hat{\beta}$ followed by the use of that $\hat{\beta}$ with a new set of observations x_*. We would like to be able to gauge how accurate our estimate of the (not yet observed) y_* will be. Following Weisberg, we will refer to those predictions as $\tilde{y}_*$.

Standard Errors for Predicted and Fitted Values

- In keeping with the above considerations, there are two distinctly different standard errors that we can compute in connection with the regression line.
- One standard error, sefit, deals with the *estimation* situation where we would like to compute a set of standard errors for the (population) fitted values on the regression line. This estimation of the conditional means does not require a new $\mathbf{x}_*$.
- Another standard error, sepred, deals with the *prediction* situation where we have a new set of predictor values $\mathbf{x}_*$, and we wish to compute the standard error for the predicted value of $\mathbf{y}$, i.e., $\tilde{y}_*$, computed from these values.

Standard Errors for Predicted and Fitted Values

Key Formulas

Section 3.6 of ALR gives the following:

> 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 $\mathbf{x}_*$. In exactly the same way as was done in simple regression, the point prediction is $\tilde{y}_* = \mathbf{x}_*'\hat{\boldsymbol{\beta}}$, and the standard error of prediction, $\text{sepred}(\tilde{y}_*|\mathbf{x}_*)$, using Appendix A.8, is
>
> $$\text{sepred}(\tilde{y}_*|\mathbf{x}_*) = \hat{\sigma}\sqrt{1 + \mathbf{x}_*'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_*} \tag{3.23}$$
>
> Similarly, the estimated average of all possible units with a value $\mathbf{x}$ for the terms is given by the estimated mean function at $\mathbf{x}$, $\hat{E}(Y|X = \mathbf{x}) = \hat{y} = \mathbf{x}'\hat{\boldsymbol{\beta}}$ with standard error given by
>
> $$\text{sefit}(\hat{y}|\mathbf{x}) = \hat{\sigma}\sqrt{\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}} \tag{3.24}$$
>
> Virtually all software packages will give the user access to the fitted values, but getting the standard error of prediction and of the fitted value may be harder. If a program produces sefit but not sepred, the latter can be computed from the former from the result
>
> $$\text{sepred}(\tilde{y}_*|\mathbf{x}_*) = \sqrt{\hat{\sigma}^2 + \text{sefit}(\tilde{y}_*|\mathbf{x}_*)^2}$$