Multiple Regression I:
Mathematics and Interpretation

Alan B. Gelder
06E:071, The University of Iowa[1]

1 The Least-Squares Problem

In simple regression, we had one x variable predicting a y variable. We generalize this set-up in multiple regression by allowing for several x variables (we still only have one y variable). The *population* regression equation with k predictor variables (or regressors) is given as follows:

$$y_i = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \ldots + x_{i,k}\beta_k + \varepsilon_i$$

We assume that ε_i is normally distributed with mean 0 and variance σ^2.

In the multiple regression framework, the *sample* regression model for predicting y based on x_1, x_2, ...,x_k is

$$\hat{y}_i = b_0 + x_{i,1}b_1 + x_{i,2}b_2 + \ldots + x_{i,k}b_k$$

Our goal in multiple regression is to estimate our parameters β_0, β_1, β_2, $\ldots \beta_k$ in such a way that we are able to minimize the squared difference between what we actually observe in the data y and what we predict $\hat{y}$. This is a multi-dimensional problem that can be very difficult to visualize. However, the same underlying principle applies from the simple regression case: we are looking for the line that best fits the data.

How is this done mathematically? Let's start with the error term (after all, we want to minimize the sum of squared errors). Our data has N total observations. The error for the i^{th} observation is

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + x_{i,1}b_1 + x_{i,2}b_2 + \ldots + x_{i,k}b_k)$$

We want to choose b_0, b_1, b_2, ..., b_k to minimize the sum of squared errors over our N total observations:

$$\min_{b_0,b_1,\ldots,b_k} \sum_{i=1}^{N} (y_i - b_0 - x_{i,1}b_1 - x_{i,2}b_2 - \ldots - x_{i,k}b_k)^2$$

[1]The notes for this class follow the general format of Blake Whitten's lecture notes for the same course.

It will be simpler to write this problem in matrix notation. Define $\mathbf{y}$, $\mathbf{b}$, and $\mathbf{X}$ as follows:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,k} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \dots & x_{N,k} \end{pmatrix}$$

The minimization problem then becomes:

$$\min_{\mathbf{b}} (\mathbf{y} - \mathbf{X}\mathbf{b})^2$$

We can rewrite $(\mathbf{y} - \mathbf{X}\mathbf{b})^2$ as follows:

$$(\mathbf{y} - \mathbf{X}\mathbf{b})^2 = (\mathbf{y} - \mathbf{X}\mathbf{b})^T (\mathbf{y} - \mathbf{X}\mathbf{b})$$

$$= \mathbf{y}^T\mathbf{y} - (\mathbf{X}\mathbf{b})^T \mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{b} + (\mathbf{X}\mathbf{b})^T \mathbf{X}\mathbf{b}$$

Note that $(\mathbf{X}\mathbf{b})^T \mathbf{y} = \mathbf{b}^T\mathbf{X}^T\mathbf{y}$ is a 1×1 matrix, so $\left(\mathbf{b}^T\mathbf{X}^T\mathbf{y}\right)^T = \mathbf{y}^T\mathbf{X}\mathbf{b}$. Hence,

$$\mathbf{y}^T\mathbf{y} - (\mathbf{X}\mathbf{b})^T \mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{b} + (\mathbf{X}\mathbf{b})^T \mathbf{X}\mathbf{b} = \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\mathbf{b} + \mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{b}$$

The problem can now be written as:

$$\min_{\mathbf{b}} \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\mathbf{b} + \mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{b}$$

Differentiating with respect to $\mathbf{b}$ we obtain:

$$-2\mathbf{y}^T\mathbf{X} + 2\mathbf{X}^T\mathbf{X}\mathbf{b}$$

In order to solve our minimization problem, we need to set the derivative equal to zero (this is actually a $k \times 1$ matrix of zeros):

$$-2\mathbf{y}^T\mathbf{X} + 2\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{0}$$

$$\Rightarrow \mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{y}^T\mathbf{X}$$

$$\Rightarrow \mathbf{b}^* = (\mathbf{X}^\mathbf{T}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^\mathbf{T}\mathbf{y}$$

Note that in order for $\mathbf{(X^TX)^{-1}}$ to exist the x variables must be linearly independent (Assumption **A.5**). Additionally, the second order conditions verify that $\mathbf{b}^*$ does indeed minimize our least squares problem.

Another way to see this derivation is to note that the error terms must be independent of our x variables (by Assumption **A.2**). Then,

$$(\mathbf{Xb})^T \mathbf{e} = \mathbf{0}$$

$$\Rightarrow \mathbf{b}^T\mathbf{X}^T (\mathbf{y} - \mathbf{Xb}) = \mathbf{0}$$

$$\Rightarrow \mathbf{b}^T\mathbf{X}^T\mathbf{y} - \mathbf{b}^T\mathbf{X}^T\mathbf{Xb} = \mathbf{0}$$

$$\Rightarrow \mathbf{b}^T \left[\mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{Xb}\right] = \mathbf{0}$$

$$\Rightarrow \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{Xb} = \mathbf{0}$$

$$\Rightarrow \mathbf{X}^T\mathbf{Xb} = \mathbf{X}^T\mathbf{y}$$

$$\Rightarrow \mathbf{b}^* = (\mathbf{X^TX})^{-1}\mathbf{X^Ty}$$

Minitab and other statistical programs use this formula to calculate the coefficients for multiple regression. Completely written out, the formula for the sample coefficients in a multiple regression equation is the following:

$$\begin{pmatrix} b_0^* \\ b_1^* \\ b_2^* \\ \vdots \\ b_k^* \end{pmatrix} =$$

$$\left(\begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{1,1} & x_{2,1} & \dots & x_{N,1} \\ x_{1,2} & x_{2,2} & \dots & x_{N,2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,k} & x_{2,k} & \dots & x_{N,k} \end{pmatrix} \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,k} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,k} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{N,1} & x_{N,2} & \dots & x_{N,k} \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{1,1} & x_{2,1} & \dots & x_{N,1} \\ x_{1,2} & x_{2,2} & \dots & x_{N,2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,k} & x_{2,k} & \dots & x_{N,k} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

In the simple regression case, all of these matrices can be simplified to give us the slope and intercept formula:

$$b_1 = r\frac{s_y}{s_x} \qquad b_0 = \bar{y} - b_1\bar{x}$$

For multiple regression, we will rely on Minitab to compute the slopes of our different predictor variables, as well as the intercept.

2 Direct and Indirect Effects

In multiple regression we need to keep track of several x variables. Fundamentally, we want to quantify how a particular x variable is related to a y variable. However, any particular x variable is also interacting with the other x variables, and those variables in turn are influencing the y variable. Multiple regression helps us to isolate the *direct* effect that a particular x variable has on the y variable, independent of the other x variables. We can also identify the *indirect* effect that a particular x variable has on y through the ripple effects that go through the other x variables and then to y.

The following theorem essentially states for a regression of y on $x_1, x_2, ..., x_k$, the coefficients (or slopes) that we obtain are the same as if we were to carefully take away all of the indirect effects. Hence, multiple regression just leaves us with the direct effect that each x variable has on y.

Theorem. *(Frisch-Waugh)*
The multiple regression coefficient b_i for a predictor variable x_i is the same coefficient that is obtained by

1. *Regressing y on all of the other x variables and keeping the residuals;*
2. *Regressing x_i on all of the other x variables and keeping the residuals;*
3. *Regressing the residuals from (1) on the residuals from (2).*

Recall that the residuals, $e_i = y_i - \hat{y}_i$, constitute the portion of the data that *cannot* be explained by the regression. In step (1), we regress y on all of the other x variables. This leaves us with the part of y that cannot be explained by any of the other x variables. In step (2), we regress x_i on all of the other x variables. Here we are left with the portion of x_i which cannot be explained by any of the other

x variables. Finally, we regress the leftover portion of y on the leftover portion of x_i.

Interpretation of coefficients: From this theorem, we can interpret b_i as follows: For every 1 unit increase in x_i, y increases by b_i units on average, holding the other variables constant. (Make sure to specify your units.)

We will not formally prove this theorem in this class. However, it will be illustrative to do an example.

2.1 Housing Prices

[The following example is from Dr. Whitten's notes. Use the data from Table 2.13.]

The sales price of a house is related to both the area of the house and the age of the house. However, the age of a home (in years) and the area of a home (in square feet) are also related. We want to know how the area of a house is related to the sales price of a house *after* accounting for the house's age. How can we do this?

1. Remove the effect of age from the sales price of a house:
    - Stat > Regression > Regression > (Response: Sales Price; Predictor: Age) > Storage > (Select Residuals) > OK > OK
    - Re-title "RESI1" as "Sales Price Residuals"
2. Remove the effect of age from the area of a house:
    - Stat > Regression > Regression > (Response: Area; Predictor: Age) > Storage > (Select Residuals) > OK > OK
    - Re-title "RESI2" as "Area Residuals"
3. Regress the sales price residuals on the area residuals:
    - Stat > Regression > Fitted Line Plot > (Response: Sales Price Residuals; Predictor: Area Residuals) > OK

Interpret the slope of the area residuals in this last regression:
For every additional square foot of area, the price of a house increases by \$74.24 on

average, holding the age of the house constant.

A similar process can be used to identify the *direct* effect that age has on the price of a house.

Now lets look at the multiple regression equation for determining the price of a house based on both the area of the house and the age of the house. The Minitab commands are as follows:

- Stat > Regression > Regression > (Response: Sales Price; Predictors: Area Age) > OK

Notice that the coefficient on the area of a house is the same as the one that we found by taking away the effect that age has on the sales price and the area of the house. If we did the same exercise for the age of a house, we would also find that the coefficient is the same in that case. (This would be a good thing to try to do yourself just to solidify the concept.)

2.2 Total and Indirect Effects

We have seen that the multiple regression coefficients identify the *direct* effect that a predictor variable x_i has on the response variable y. We may also want to quantify the *indirect* effect that x_i has on y through the other x variables.

Housing Prices Example Continued: Find the indirect effect that age has on mean sales prices through area.

Recall that the second step in the housing prices example was to regress area on age. The regression equation was as follows:

$$\text{Area} = 1713 - 7.154 \times \text{Age}$$

This implies that for every additional year older a house is, it will have 7.154 fewer square feet, on average.

The third step was to regress the sales price residuals on the area residuals:

$$\text{Sales Price Residuals} = -\ 0 + 74.27 \times \text{Area Residuals}$$

We can see the indirect effect of age on sales price as follows:
Increase the age of a house by one year, then the area of the house will be 7.154 fewer square feet, on average. A decrease in 7.154 square feet leads to a drop in the sales price of

$$-7.154 \times 74.27 = -531.33$$

Hence, the indirect effect of age on sales price (through the area of the house) is $-\$531.33$.

From the multiple regression output, the direct effect of age on sales price is $-\$803.2$. Define the total effect as the sum of the direct and the indirect effects. That is,

$$\text{Direct Effect} + \text{Indirect Effect} = \text{Total Effect}$$

Therefore, the *total* effect of the age of a house on the average sales price is

$$(-\$531.33) + (-\$803.2) = -\$1334.53$$

Now find the simple regression equation for predicting sales price with the age of a house.

$$\text{Sales Price} = 189226 - 1334.5 \times \text{Age}$$

The total effect is then equal to the coefficient from simple regression. **What is the intuition for this?**

In summary, the total effect is the coefficient from *simple* regression, the direct effect is the coefficient from *multiple* regression, and the indirect effect is the difference of the coefficients (i.e. total effect − direct effect).

3 F-statistics

There are two F-statistics that we will discuss for multiple regression. The first F-statistic is given in the regression output (in the ANOVA table). The second F-statistic is one that you will need to calculate from a formula (which is included on the formula sheet).

3.1 The ANOVA F-statistic

Remember that in ANOVA, the null hypothesis is that all of the means are equal:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$$

The F-statistic in the ANOVA table is based on a very similar hypothesis. The ANOVA F-statistic for regression tests the null hypothesis that all of the slopes are the same and equal to zero:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

If we fail to reject this null hypothesis, then none of the predictor variables in our regression model are significant. However, if we reject H_0, then at least one of our x variables is a significant predictor of our response variable y. Hence, the ANOVA F-statistic is a gateway test. It tells us whether or not *any* of our predictor variables are significant, but it does not tell us which ones are significant (if any).

3.2 Following-up with t-tests

If the F-statistic is significant, then we know that at least one of our predictor variables is significant. We now need to determine which ones are significant. This can be accomplished by doing t-tests on the different coefficients.

Just as in simple regression, there are t-statistics that are associated with each of the coefficients in multiple regression (b_0, b_1, b_2, ..., b_k). These t-statistics are included in the regression output.

Minitab (as well as most other statistical programs) always does the two-tailed hypothesis test:

$$H_0 : \beta_i = 0$$

$$H_A : \beta_i \neq 0$$

(We can, of course, find the p-value for a one-tailed hypothesis test based on the t-statistic and the two-tailed p-value.)

While the hypothesis test looks exactly the same as in simple regression, the t-test in multiple regression takes on a different meaning. In multiple regression, we are testing to see if a particular x variable is a significant predictor of y *after accounting for the other x variables in our model*. Hence, it is possible for an x variable to be

a significant predictor of y in simple regression, but that same x variable may not be a significant predictor of y when it is included with other x variables (because it does not add enough unique information for predicting y above and beyond the information that the other x variables are already contributing).

In fact, it is possible to have a significant F-statistic, and have insignificant t-statistics for all of the predictor variables in the *full model* (i.e. the model with all of the x variables). This does not contradict the F-statistic, however. A significant F-statistic means that at least one of the x variables by itself (and possibly after accounting for other x variables) is a significant predictor of y.

Example: Apply the ANOVA F-test to the regression model predicting housing prices based on the area and age of houses. Use $\alpha = 0.05$. If appropriate, follow-up with t-tests for the individual predictor variables. Interpret the results of the hypothesis tests.

3.3 The F-statistic for a Group of Variables

The second F-statistic that we will study in multiple regression tests to see whether a group of predictor variables significantly improves the predictive power of our regression. Multiple regression allows us to have any number of predictor variables. However, in building regression models there is a fine balance between having a model that is complex enough to account for all of the relevant information, yet simple enough that it is useful.

We can always add more variables to our regression model, but that does not necessarily improve our model. Why is that?

1. Adding more variables will never decrease R^2 and will typically increase it– even if the added variables do not really contribute to our model. (This result follows from how R^2 is defined mathematically.) So we cannot say that a model with more variables is better just because it has a higher R^2.

2. Adding unnecessary variables will increase the variance of the coefficients (which makes accurate predictions more difficult).

3. Some variables may contain essentially the same information as other variables. If our model already can account for most of the information that additional variables would bring to our model, we could probably simplify our model and do without the extra variables.

From a business standpoint, additional data may be expensive to obtain. We may want to test to see if our model would significantly improve by including a set of additional variables. In other words, would the benefit of having the extra variables outweigh the cost of purchasing or gathering the data?

The F-test for a group of variables, sometimes referred to as the restricted F-test, gives us a formal way of determining if a set of variables statistically improves a regression model.

Let p be the total number of explanatory variables in a full regression model. We can test to see if a group of q predictor variables contributes significantly to our model by using the following hypothesis:

$$H_0 : \beta_{j+1} = \beta_{j+2} = \ldots = \beta_{j+q} = 0$$

(Here we assume that the group of variables being tested are listed in order in the full model after the j^{th} variable not being tested.) If we fail to reject the null hypothesis, then we can discard the group of q variables from our full regression model and keep the remaining j variables ($p - q = j$).

The F-statistic is calculated as follows:

$$F_{(df=\,q,\,n-p-1)} = \frac{\left(R_1^2 - R_2^2\right)/q}{(1 - R_1^2)/(n - p - 1)}$$

where p = # of regressors in full model, q = # of variables tested as a group, R_1^2 is from the full model, and R_2^2 is from the reduced model.

3.4 Architectural Firms

Table 1.3 on page 13 of the text lists the billings (in million of dollars) for 25 architectural firms in the Indianapolis area. Consider two possible models: one where the architectural billings in 1998 (ABill98) is predicted by architectural billings in 1997 (ABill97) and the number of architects on staff (Arch); the second model is the same as the first except that the number of engineers on staff (Eng), the total size of the staff (Staff), and the year the firm was founded (Yr) are also included as predictor variables.

Is the regression model significantly improved by adding the variables Eng, Staff, and Yr? Test using $\alpha = 0.10$. Once you have calculated the F-statistic, you will need to use Minitab to find the p-value.

```
ABill98 = 0.277 + 0.524 ABill97 + 0.179 Arch


Predictor      Coef  SE Coef      T      P
Constant     0.2774   0.4724   0.59  0.563
ABill97      0.5239   0.2954   1.77  0.090
Arch        0.17857  0.06995   2.55  0.018


S = 1.24294   R-Sq = 82.3%   R-Sq(adj) = 80.7%


Analysis of Variance

Source           DF       SS      MS      F      P
Regression        2  158.158  79.079  51.19  0.000
Residual Error   22   33.988   1.545
Total            24  192.146
```

```
ABill98 = - 10.0 + 0.404 ABill97 + 0.133 Arch - 0.0896 Eng + 0.0277 Staff
          + 0.0051 Yr


Predictor       Coef  SE Coef       T      P
Constant     -10.03    30.64   -0.33  0.747
ABill97      0.4039   0.3166    1.28  0.217
Arch        0.13323  0.07607    1.75  0.096
Eng        -0.08961  0.05445   -1.65  0.116
Staff       0.02771  0.01525    1.82  0.085
Yr          0.00512  0.01550    0.33  0.745


S = 1.23300   R-Sq = 85.0%   R-Sq(adj) = 81.0%


Analysis of Variance

Source           DF       SS      MS      F      P
Regression        5  163.261  32.652  21.48  0.000
Residual Error   19   28.886   1.520
Total            24  192.146
```

4 Best Conservative Model

There are several different criteria that can be used to determine a "good" regression model. If our primary goal is to accurately estimate y (especially if we want prediction intervals and confidence intervals for $\hat{y}$), then the best conservative model does a decent job of selecting the optimal combination of x variables to include in a model for predicting y.

Definition. *(Best Conservative Model)*
For a given set of predictor variables $\{x_1, x_2, ..., x_k\}$ and a response variable y, the best conservative model is the regression model satisfying the following conditions:
1. All of the predictor variables are significant;
2. It has the highest R^2 of the different models in which all of the predictor variables are significant.

The best conservative model is also called the parsimonious model. Parsimonious is defined by Merriam-Webster's online dictionary as "frugal to the point of stinginess"—and that is exactly what we are going to be. We are being statistically frugal in only allowing significant variables into our model. We are also going to be stingy in that we want the biggest bang for our buck. Not only do we want all of the variables that we keep in our model to be significant, we also want to have the highest R^2 possible.

Example: Return to the architectural firms data (Table 1.3). Starting with the model that uses ABill97, Arch, and Eng to predict ABill98, find the best conservative model.

5 Quadratic Regression

Up to this point in our study of regression we have insisted that our x variables are linearly related to our y variable. Our standard linear model with one x variable and one y variable is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Instead of a linear relationship, suppose that our x variable has a quadratic relationship with y. Multiple regression provides us with a unique way of capturing this quadratic pattern. The population regression equation for quadratic regression with one x variable is the following:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

For each observation i, we simply square the value for x, and then the squared term becomes another variable. We can then proceed with the standard multiple regression tools that we have developed.

Checking significance: Since x and x^2 are correlated, both terms may not be significant when they are used together. In quadratic regression, the x^2 term dominates. If x^2 is significant, then you do not need to be concerned if x is not significant.

Oxygen and RPM:

[This example is adapted from Dr. Whitten's notes.]

The RPM of an engine was measured when the combustion chamber contained different amounts of oxygen. Here are the data:

Oxygen (%)	RPM
10	222.3
20	264.1
30	271.2
40	273.2
50	238.6
60	233.9
70	167.1
80	95.3

- Do a fitted line plot of the regression equation where oxygen is predicting RPM
- Check the residuals
- Do a fitted line plot of the quadratic regression
- Check the residuals
- Predict the engine's RPM when the combustion chamber contains 45% oxygen

Squaring variables in Minitab: Begin by naming a new column (Oxygen2). Then use the following commands: Calc > Calculator > Store result in variable: Oxygen2 > Expression: Oxygen*Oxygen > OK

Additional reference for these notes:
Greene, William H. *Econometric Analysis* 6th ed. Upper Saddle River, NJ: Prentice Hall, 2008.