Lectures on Multiple Linear Regression
Stat 431, Summer 2012

Hyunseung Kang

July 18-30, 2012

Last Updated: July 25, 2012 11:59PM

1 Introduction

Multiple regression is an extension to simple regression in several ways. First, instead of examining a linear relationship bewteen two measurements, X_i and Y_i, multiple regression aims to explain relationships between p measurements of an object i, denoted as $(X_{i,1}, ..., X_{i,p})$ and Y_i. Based on a sample of n points, with each point having measurements $(Y_i, X_{i,1}, ..., X_{i,p})$, multiple regression aims to estimate the true underlying relationship of this type

$$Y = \beta_0 + \beta_1 X_{,1} + ... + \beta_p X_{,p} \tag{1}$$

Here, each β_j represents a *partial correlation* between Y and $X_{,j}$. This is *not* the same as sample correlation between Y and $X_{,j}$, which we'll see in later sections and we'll examine the meaning behind each β_j.

Second, instead of dealing primarily with numerical variables in simple regression, the multiple regression can handle categorical and numerical variables. Specifically, $X_{,j}$ can either be categorical or numerical variables. Ys still have to be numerical. But, this restriction will be relaxed in future lectures on *generalized linear models*.

Third, even though X's are linear combinations of Y and we are, in essence, studying linear relationships between Y and $X_{,1}, ..., X_{,p}$, multiple regression framework is flexible to handle non-linear relationships. This is commonly known as *polynomial regression* and will be discussed in a future section.

2 Estimation and Interpretation

Suppose we obtain a sample $(Y_i, X_{i,1}, ..., X_{i,p})$ of n individuals and we want to use this sample to discover underlying relationships described in equation (1). In the least squares framework, we achieve this goal by minimizing the distance between Y_i and the linear combination of Xs (including the intercept. Mathematically, we attempt to minimize

$$\min_{\beta_0, \beta_1, ..., \beta_p} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i,1} + ... + \beta_p X_{i,p}))^2 \tag{2}$$

To minimize this quantity, we take partial derivatives with respect to each β_j, set all of them equal to zero, and solve for β_js.

$$\text{for all } j = 1, ..., p\text{:}\ \frac{\delta}{\delta \beta_j} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{i,1} + ... + \beta_p X_{i,p}))^2 = 0$$

The β_js we solve, denoted as $\hat{\beta}_j$s are the global minimizers of equation (2) because the function is convex over a convex domain.

2.1 Categorical Data and Reformatting Xs

When some of the $X_{,j}$s are categorical, you must reformat the $X_{,j}$s to work within the regression framework.You reformat the categorical data from, say $X_{,j} = "A", "B", "C"$ to 1s and 0s. Specifically,

1. Count the number of choices/factors in a categorical measurement. Call this K. For example, if the categorical variable is class year, there are four choices, freshmen, sophomore, junior, and senior.

2. Create $K-1$ Xs. Each $X_{i,j}$ will represent $K-1$ choices out of K possible choices by assigning a binary value. In particular,

$$X_{i,j} = \begin{cases} 1 & \text{if the } i\text{th individual chose } k \text{ amongst } 1, ..., K \text{ choices} \\ 0 & \text{otherwise} \end{cases}$$

   For example, for class year, we will have three new Xs, $X_{i,1}$, $X_{i,2}$ and $X_{i,3}$. $X_{i,1}$ is 1 if the ith individual was a freshmen and 0 if he wasn't. $X_{i,2}$ is 1 if the ith individual was a sophomore and 0 if he wasn't. $X_{i,3}$ is 1 if the ith individual was a junior and 0 if he wasn't. Notice that if $X_{i,1} = X_{i,2} = X_{i,3} = 0$, then the ith individual must be a senior.

Thankfully, R automatically does this and you don't need to worry about it. However, if you decide to not work with R, be mindful that you have to convert a categorical data into binary format.

2.2 Missing Data

Often in real data, we have missing values for some of the measurements. For example, in a survey with 10 questions, given out to 500 people, some individuals may decide to not answer certain questions on the survey for a variety of reasons. If there is a substantial minority of individuals who decide to not answer certain questions, we may wonder whether there is a characteristic that links those individuals who answered the question and those who did not. The method we'll introduce here will attempt to capture this "pattern of missingness". That is, it will attempt to capture the behavior of those who answered the question vs. those who didn't. This method only works if there is a sizable minority of missing observations per measurement.[1]

Suppose you have a numerical $X_{,j}$ where some of the observations in the jth measurements are missing. To capture the "pattern of missingness", we can create a new a variable $X'_{,j}$ in lieu with $X_{,j}$ where

$$X'_{i,j} = \begin{cases} 1 & \text{if the } i\text{th individual has a missing value for the } j\text{th measurement} \\ 0 & \text{otherwise} \end{cases}$$

In addition, for the original $X_{,j}$ which contains the missing values, you would replace them with the mean of the measurements, $\bar{X}_{,j}$.

If $X_{,j}$ is categorical with K choices, you would treat this variable as if they had $K+1$ choices, with the extra choice being the choice for missing. Then, you would follow the procedure outlined above.

Unfortunately, R's default behavior with missing data is to drop missing observations. Although this is another way to handle missingness at the expense of sample size, to incorporate the above method, you would have to manually create these new X's.

2.3 Interaction Terms

Interaction terms are terms that are included in the regression if you want to study the combined effects of two measurements. For example, suppose your Y is the fever temperature of a child. $X_{,1}$ is whether the child has taken drug A which claims to lower fever and $X_{,2}$ is whether the child has taken drug B which also claims it lowers fever.

[1] There are many methods to deal with missing values and these methods are context-specific. Here, we introduce one technique that is popular in the field of observational studies

Each β_1 and β_2 will represent individual effects from drug 1 and drug 2. However, it is possible that a combination of these drugs may have effects that we want to measure. A method to incorporate this is by using interaction terms.

Interaction terms work by "multiplying" the $X_{,j}$s that are under consideration for interaction. You can technically "multiply" more than two variables (e.g. you may have drug C and you want to study the combined effects of drug A,B, and C), but the expression gets complicated very quickly and therefore, we'll only deal with *two-way* interactions. Here are a couple of things to remember when you "multiply" various measurements

1. If you are studying the interaction between two categorical variables, you will have a new categorical X for each possible combination the choices from two categorical variables. This variable For example, if you are studying the effects of drug A and B, you will have four possible choices of taking A and B: took A + took B, took A + not took B, not took A + took B, and not took A + not took B. Of course, you will have to reformat this newly formed categorical variable using the procedure outlined above. Thankfully, R automatically does this and you don't have to worry about it.

2. If you are studying the interaction between a categorical variable and a numerical variable, you will have a

3.

2.4 Interpretation of Estimates

Once your data is reformatted, you can solve the optimization problem outlined in equation (2) and obtain $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p$. Each $\hat{\beta}_j$ states that a one-unit increase (in numerical variables) or choosing a particular option (in categorical variables) will increase (if the sign of $\hat{\beta}_j$ is positive) or decrease (if the sign of $\hat{\beta}_j$ is negative) Y by $|\hat{\beta}_j|$ amount, *under the condition that all the other Xs remain constant or fixed.* Another way to say this is that *controlling for all other measurements that are not j*, one-unit increase (in numerical variables) or choosing a particular option (in categorical variables) will increase (if the sign of $\hat{\beta}_j$ is positive) or decrease (if the sign of $\hat{\beta}_j$ is negative) Y by $|\hat{\beta}_j|$ amount.

3 Inference

Inference in multiple regression revolves around testing hypotheses for a single $\hat{\beta}_j$ or a group of $\hat{\beta}_j$s. The latter is more applicable to categorical X's. Inference on the slope coefficients requires a significant use of the ANOVA tables. In addition, there is also inference on prediction. Here, our focus will mostly be on constructing CIs and PIs, like we did in simple regression.

3.1 Assumption

To make inference possible in the case of multiple regression, we have to make some assumptions about our population in which we drew the samples $(Y_i, X_{i,1}, ..., X_{i,p})$ from. The assumptions for multiple linear regression are identical to those from simple linear regression and they are summarized in the following statement below.

Assumption 1. A linear regression model assumes the following about $X_{i,1}, ..., X_{i,p}$ and Y_i

1. $X_{i,1}, ..., X_{i,p}$ are assumed to be fixed, non-random quantities. Y_i's are the only random quantities.

2. Y_i is related to $X_{i,1}, ..., X_{i,p}$ by the following *linear* relationship

$$Y_i = \beta_0 + \beta_1 X_{i,1} + ... + \beta_p X_{i,p} + \epsilon_i \tag{3}$$

   where ϵ_i are i.i.d random variables, commonly referred to as errors. Here, β_js are the parameters that characterize the true underlying relationship between $X_{i,j}$ and Y_i

3. ϵ_i is Normally distributed

4. ϵ_i have the same variance, σ^2, for any value of $X_{i,1}, ..., X_{i,p}$. Another way to say this is that Y_i's have *homoscedastic* errors.

We can validate these assumptions using the diagnostic tools from simple linear regression.

1. *Homoscedasticity:* We can use a residual plot and look for any "spreading" behavior along the x-axis.
2. *Linearity:* We can use a residual plot and look for any non-linear patterns along the x-axis. T
3. *Normality of the errors:* We can use a QQ plot of the residuals and see if the points "hug" the $y = x$ line.
4. *Outliers:* For leverage points and influential points, we can use H_{ii} and D_i values to evaluate the magnitude of leverage and influence, respectively. For regression outliers (i.e. outliers in the y axis), we resort to residual plots. In particular, if there are points in the residual plot that has large deviations in the y direction, these points are likely to be regression outliers.
5. *Collinearity:* We'll talk about this when we get to model selection. It states that all the $X_{,j}$ measurements should not be correlated with each other. While this is not a violation of assumptions 1, a severe degree of collinearity may cloud our interpretation when we do inference.

We can deal with any problems that may come up from these diagnostics the same way de dealt with them in simple linear regression. For example, if there is a strong reason to believe heteroscedasticity maybe present, then transformation of Y is not a bad idea. If there is a strong reason to believe nonlinearity may exist, then it is generally advised that you transform one or several Xs until you get rid of the non-linear pattern in your residual practice. From practice, the Xs that need transforming are those that already exhibit non-linear patterns if you plot Y and $X_{,j}$ in a separate scatterplot.

Note that you *cannot* use a scatterplot of X and Ys, simply because there is no way to plot all of $X_{,1}, ..., X_{,p}$ and Y in a scatterplot; you would need a $p + 1$ dimensional scatterplot, which is impossible!

3.2 ANOVA Table

For every regression we fit, we have an ANOVA table associated with the regression fit. Just like in simple linear regression, mutiple regression's ANOVA tables have the same interpretations. However, we must be careful with the degrees of freedom. While DFT has the same degrees of freedom ($n - 1$), no matter what regression we fit (since it is only the degrees of freedom associated with the sample variance of Y), DFE and DFR have different degrees of freedom depending on the number of categorical variables, the number of numerical variables, and the number of interaction terms. If we can count the number of slope terms, p, the task is very easy.

Sum of Squares (SS)	Mean SS(MS)	Degrees of Freedom (DF)
SSE	MSE	$DFE = n - p + 1$
SSR	MSR	$DFR = p$
SST	MST	$DFT = n - 1$

Table 1: ANOVA Table. This table is useful if you know the exact number of $X_{,j}$s in your model, after reformatting your categorical variables and taking interaction terms into account.

3.3 Inference on Slope Coefficients

First, we start off with testing individual β_js.Suppose you want to test

$$H_0 : \beta_j = 0 \text{ vs. } H_a : \beta_j \neq 0 \tag{4}$$

These tests use the t-distribution as their sampling distriubtion with degrees of freedom that is equal to that associated with the SSE. The procedure to conduct this test is identical to the ones in simple regression. In particular, the p-value for this test is

$$\max_{\beta_j \in H_0} P(\text{reject } H_0 | H_0 \text{ is true}) = \max_{\beta_j \in H_0} P\left(\left| \frac{\hat{\beta}_j - \beta_j}{\sqrt{Var(\hat{\beta}_j)}} \right| > \left| \frac{\hat{\beta}_{j,obs} - \beta_j}{\sqrt{Var(\hat{\beta}_j)}} \right| |H_0 \text{ is true} \right)$$
$$= P\left(|t_{DFE}| > \left| \frac{\hat{\beta}_{j,obs} - 0}{\sqrt{Var(\hat{\beta}_j)}} \right| \right)$$

The p-values for each hypothesis regarding β_j are on the R output. Also, if the test calls for $H_0 : \beta_j = \beta$ vs. $H_a : \beta_j \neq \beta$ where $\beta \neq 0$ (i.e. the boundary is not zero), you can use the estimates of β_j and $Var(\hat{\beta}_j)$ (written in R as standard errors, which are $\sqrt{Var(\hat{\beta}_j)}$) to obtain the necessary p-values; the only difference in calculating the p-value would be the maximization step, which would be replaced by $\frac{\hat{\beta}_{j,obs} - \beta}{\sqrt{Var(\hat{\beta}_j)}}$ instead of $\frac{\hat{\beta}_{j,obs} - 0}{\sqrt{Var(\hat{\beta}_j)}}$

While the mechanics of testing is identical to that from simple linear regresison, the interpretation of the hypothesis are drastically different in the multiple regression setting. Here, the test specified in equation (4) tests the importance of the jth measurement in prediciting Y *controlling for all the other measurements*. Specifically, *given that all the other measurements besides $X_{,j}$ are fixed*, the test in equation (4) is testing whether the jth measurement, $X_{,j}$, adds any value in explaining the variation in Y. Notice the similarity in interpetation between the estimates of each $\hat{\beta}_j$ and the inference for single β_js.

Next, we can expand our testing framework to test for multiple β_js in one single hypohtesis testing. For example, suppose you want to test whether any of the β_1, β_2, and β_3 are useful terms in explaining the variation of Y, *controlling for all the other X's not present in the testing* (here, we assume $p > 3$). Then, our hypothesis is written as follows.

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \text{ vs. } H_a : \text{at least one } \beta_j \neq 0 \tag{5}$$

Again, we have to always remember that we're controlling for all the other X's that are not in the hypothesis. That is, given that the other $X_{,4}, X_{,5}, ..., X_{,p}$ explain some variation in Y, is $X_{,1}$,$X_{,2}$, and $X_{,3}$ still useful in explaining the variation we find in Y?

The best way to test this is to fit two regression models. First regression model contains all the $X_{,j}$ (denoted as the full model) and the second regression model contains only $X_{,4}, X_{,5}, ..., X_{,p}$ (denoted as the reduced model). From each regression model, we obtain ANOVA tables shown below.

Sum of Squares (SS)	Mean SS(MS)	Degrees of Freedom (DF)
SSE_{full}	MSE_{full}	$DFE_{full} = n - p_{full} + 1$
SSR_{full}	MSR_{full}	$DFR_{full} = p_{full}$
SST_{full}	MST_{full}	$DFT_{full} = n - 1$

Table 2: ANOVA Table for the full regression model with all the $X_{,j}$s in it: $Y = \beta_0 + \beta_1 X_{,1} + ... + \beta_p X_{,p}$

Sum of Squares (SS)	Mean SS(MS)	Degrees of Freedom (DF)
$SSE_{reduced}$	$MSE_{reduced}$	$DFE_{reduced} = n - p_{reduced} + 1$
$SSR_{reduced}$	$MSR_{reduced}$	$DFR_{reduced} = p_{reduced}$
$SST_{reduced}$	$MST_{reduced}$	$DFT_{reduced} = n - 1$

Table 3: ANOVA Table for the reduced regression model with all the $X_{,j}$s in it: $Y = \beta_0 + \beta_4 X_{,4} + \beta_5 X_{,5} + ... + \beta_p X_{,p}$. In this example with the first three X's missing, $p_reduced = p_{full} - 3$

A couple of points about comparing two ANOVA tables

1. $SST_{reduced} = SST_{full}$, $MST_{reduced} = MST_{full}$, and $DFT_{reduced} = DFT_{full} = n - 1$ since SST is measuring the variance of Y. None of the X's play a role in determining any of the "..T" values in the ANOVA table. This fact holds true for *any* ANOVA table

2. $SSE_{full} \leq SSE_{reduced}$ since the reduced model is a subset of the full model. Specifically, with the full model, we have more X variables to minimze the error between the observed Y_i and $\hat{Y}_i$, even if some of them are completely useless.

3. $DFE_{full} \leq DFE_{reduced}$ since there are less β's in the reduced model we have to estimate.

Once we have the two tables, a test statistic to test hypothesis (5) is

$$F = \frac{\frac{SSE_{reduced} - SSE_{full}}{DFE_{reduced} - DFE_{full}}}{\frac{SSE_{full}}{DFE_{full}}} \tag{6}$$

Intuitively, equation (6) measures improvement in the full model in comparison to the reduced model, without the X's under testing; the numerator measures the magnitude of the improvement while the denominator "calibrates" the magnitude from the numerator.

Under H_0 specified in equation (5), the proposed test statistic in equation (6) would be zero since there won't be any differences between $SSE_{reduced}$ and SSE_{full}. Under H_a specified in equation (5), the proposed statistic in equation (6) would be away from zero. Hence, we can reformulate our hypothesis as

$$H_0 : F = 0 \quad \text{vs.} \quad H_a : F > 0$$

Finally, equation (6) has an F distribution with $DFE_{reduced} - DFE_{full}$ and DFE_{full} degrees of freedom. Then, we can compute our p-value the same way we computed p-values for any other sampling distribution.

3.4 Inference on Prediction

Regression allows us to naturally estimate $\hat{Y}_i$ given a set of X's, $X_{i,1}, ..., X_{i,p}$. This estimate, $\hat{Y}_i$, is actually an estimate of the conditional mean of Y at the $X_{i,1}, ..., X_{i,p}$

$$E(Y | X_{i,1}, ..., X_{i,p})$$

Hence, $\hat{Y}_i$ is an estimator for a parameter, the conditional mean. This also means that we can make inference about this parameter, much like we made inference about β_j's. Here, we'll focus on obtaining CIs and PIs.

CIs are the confidence interval for the conditional mean, $E(Y | X_{i,1}, ..., X_{i,p})$, and its interpretaion is identical to that from simple linear regression. PIs are prediction intervals for the conditional mean, $E(Y | X_{i,1}, ..., X_{i,p})$ and its interpretation is the same as that from simple linear regression.

The formula for CIs and PIs are very complicated. Hence, we'll let R deal with it. To obtain CI and PIs, simply use the "predict()" function. The usage is described in the R cheat sheet.

4 Examples

4.1 Polynomial Regression

Polynomial regression is multiple regression where we take powers of the measurement to obtain a polynomial fit between Y and X.

4.2 ANOVA

ANOVA is basically regression with categorical Xs.

4.3 ANCOVA

ANCOVA is regression with categorical and numerical Xs.