Topic 5 – Multiple Linear Regression

Basics

Essentially, we extend the SLR model (though, MLR is harder to visualize). Always keep in mind that, now, we deal with "multiple" models, too. Each model must be dealt with, case by case. Thus, focus is on regression coefficients and model effects (effects of explanatory variables on the response).

MLR Model (General)

An MLR model specifies that the mean response is a linear function of k explanatory variables ($X_1, \ldots, X_k$).

$$\text{Model: } Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$$
$$\text{or } \quad \mu(Y \mid X_1, \ldots, X_k) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

Or, "the mean response (μ_Y) at some value of the explanatory variables ($X_1, \ldots, X_k$) is equal to $\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$".

The equation depends on $k + 1$ statistical parameters (the coefficients): $\beta_0, \beta_1, \ldots, \beta_k$.

What is linear?

The following are all examples of possible MLR models:

$$\mu(Y \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$
$$\mu(Y \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$
$$\mu(Y \mid X_1, X_2) = \beta_0 + \beta_1 X_1^2 + \beta_2 \log(X_2)$$

They are all "linear" in the regression coefficients.

Examples that are NOT linear in the coefficients are NOT considered MLR models.

$$\mu(Y \mid X_1) = \frac{1}{\beta_0 + \beta_1 X_1}$$

$$\mu(Y \mid X_1, X_2) = \frac{1}{\beta_0 + \beta_1 X_1 + \beta_2 X_2}$$

Aside: The following topics will not be covered in class, so they won't be testable. Yet they may be needed for the lab. Focus on definitions, concepts, and rules, not the formulas or the theory.

- Variable Selection Techniques
- Identifying Influential Observations using SPSS
- Collinearity
- Checking Model Assumptions based on Graphical Displays

Model Assumptions

It is assumed, for any ($X_1 = x_1, \ldots, X_k = x_k$), that the errors in the response are independent and that $\varepsilon \sim N(\mu_\varepsilon = 0,\ \sigma_\varepsilon^2)$, where σ_ε is independent of ($X_1, \ldots, X_k$).

1. Linearity ($\mu_\varepsilon = 0$): The mean response is a linear function of ($X_1, \ldots, X_k$).
2. Constant Variance (σ_ε is independent of any X): The variability in the response for any fixed ($X_1 = x_1, \ldots, X_k = x_k$) is constant.
3. Normality: For any fixed ($X_1 = x_1, \ldots, X_k = x_k$), the response is normally distributed.
4. Independence (errors are independent): The responses are all independent.

Estimating the Model

Data: We observe a random sample of n sets of observations

($x_{i1}, \ldots, x_{ik}, y_i$), for $i = 1, \ldots, n$

We find the least-squares estimates for the model parameters ($\beta_0, \beta_1, \ldots, \beta_k$) the same way as with SLR by minimizing *SS*(Residual). ALWAYS done by computer. Once determined, the estimates combine to form the estimated regression function

$$\hat{Y} = \hat{\mu}(Y \mid X_1, \ldots, X_k) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k$$

Estimating Mean Response

As with SLR model, we can estimate a mean response or a single predicted response, but now these come from the set of ($X_1 = x_1, \ldots, X_k = x_k$) being placed into the estimated regression model. $\hat{Y}$ is still also called the fitted value of the response.

Approximate PI for Single Responses

A useful approximation to the margin of error for a $(1 - \alpha)100\%$ PI for single response is

$$M.E.(\hat{Y}) \approx t_{n-(k+1),\,\alpha/2} \times s_\varepsilon$$

Thus, an approximate $(1 - \alpha)100\%$ PI for a response at ($X_1 = x_1, \ldots, X_k = x_k$) is

$$(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k) \pm t_{n-(k+1),\,\alpha/2} \times s_\varepsilon$$

Caution: This approximation has no "extrapolation penalty", so use it only for "reasonable" values of the explanatory variables.

Error Variability

As in SLR, we also want to estimate $Var(\varepsilon) = \sigma_\varepsilon^2$. Extending to MLR is simple.

$$res_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \ldots + \hat{\beta}_k X_{ik})$$

$$SS(\text{Residuals}) = \sum_{i=1}^{n} (res_i)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

Degrees of freedom associated with a set of residuals
= (# of observations) – (# of parameters in the model for the means)
$= n - (k + 1) = n - k - 1$

Thus, $\sigma_\varepsilon^2 \approx \hat{\sigma}_\varepsilon^2 = s_\varepsilon^2 = \dfrac{SS(\text{Residuals})}{n - k - 1} = MSE$

Model Effects

Effects of Numerical X's on Mean Response
The effect of a numerical X is measured as the change in the mean Y that is associated with a 1-unit increase in that X, holding all other X's fixed.

In SLR, the effect of the (only) X on the response was simply measured as the slope (or coefficient of X). This result was derived from the definition above. Since SLR has only one specific model, we didn't have to refer to the definition from case to case.

In MLR, however, every model is potentially different. As a result, the above definition is extremely important in order to apply it to any given situation.

Suggestion: Find effect in terms of regression coefficients first; then, substitute in their appropriate estimates to estimate the effect. Also, when asked for the effect of an X, your answer should be in terms of the model parameters only. If asked to estimate the effect of an X, then you will need to substitute in the parameter estimates.

Basic Observed Variables vs. Derived Variables
There is a distinction between these two types of variables. In the 3 MLR models mentioned earlier, there are 2 basic observed variables, X_1 and X_2, in all 3 cases. The derived variables are functions of the observed variables. Thus, X_1X_2 (in the 2nd model) is a derived variable (i.e. a product of the 2 observed variables).

Interaction Term (cross-product term)
Two X's are said to interact if the effect of one of them on the mean response depends on the level of the other. In MLR, a derived explanatory variable for interaction between X_1 and X_2 is constructed as the product (X_1X_2) of the two variables.

Categorical Explanatory Variables
Quite often, the response variable is related to both numerical and categorical variables. The latter is usually referred to as factors and the level of the factors as treatments. Since the treatment levels are not numerical, we cannot simply assign numbers to the levels and enter it as a numerical variable. In order to properly introduce categorical variables into a regression model, we need to use indicator variables to indicate which level of the factor we are referring to.

Suppose a categorical variable has I levels; then an indicator variable is as follows:

Def'n: An indicator variable is a variable that takes on one of two values: '1' or '0'. The indicator variable for treatment level i takes on a value of '1' if the observation is measured from treatment i; otherwise, it will take the value '0'.

Since each observation is assigned to exactly one level of a categorical variable, exactly one of the indicators will be '1'. As a result, we only need to introduce $I - 1$ of the indicator variables into the model to indicate which treatment level we are discussing.

For example, suppose we want to consider gender as a predictor in a regression model for average blood pressure. We can define the following indicator variables:

$$X_1 = female = \begin{cases} 1, \text{ if a female} \\ 0, \text{ if not a female} \end{cases}, \quad X_2 = male = \begin{cases} 1, \text{ if a male} \\ 0, \text{ if not a male} \end{cases}$$

Using any one of the two (let's pick X_1), we can define a regression model for average blood pressure given gender as

$$\mu(Y \mid X_1) = \beta_0 + \beta_1 X_1 = \beta_0 + \beta_1 female$$

Then, average blood pressure for a female is described as $\beta_0 + \beta_1$, and the average blood pressure for male is described as β_0.

Effects of Categorical Explanatory Variables on the Mean Response
Categorical variable effects are measured as the mean difference for each pair of treatment levels. The effect of say, level 2 vs. level 1 of a categorical variable is the difference in mean response at level 2 vs. level 1.

Comparing SLR Models: Separate Lines, Parallel Lines, and Equal Lines
Consider an MLR model where we have exactly one numerical explanatory variable (X_1) and exactly one categorical variable (X_2 with I treatment levels), plus their interaction. In this situation, we are essentially considering several SLR models, one for each level of X_2. To enter the categorical variable, we enter $I - 1$ of the I treatment level indicator variables.

$$\text{Define the indicators for } X_2 \text{ as: } trt_i = \begin{cases} 1, \text{ if in treatment } i \\ 0, \text{ else} \end{cases}, \text{ for } i = 1, \ldots, I.$$

There are $I - 1$ interaction terms, one for the product of each of the $I - 1$ indicators and the numerical explanatory variable. To illustrate this, suppose the categorical variable had three levels. Thus, we need to include two indicators and two interaction terms.

1. **The Separate-Lines (SL) (or intersecting lines) model: There is interaction.**
   $\mu(Y \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 trt_1 + \beta_3 trt_2 + \beta_4 (X_1 \times trt_1) + \beta_5 (X_1 \times trt_2)$

   This model can be rewritten as 3 *separate* SLR models, one for each level of X_2:
   i. $\mu(Y \mid X_1, X_2 = 1) = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) X_1$
   ii. $\mu(Y \mid X_1, X_2 = 2) = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) X_1$
   iii. $\mu(Y \mid X_1, X_2 = 3) = \beta_0 + \beta_1 X_1$

   They are said to be three *separate* or intersecting lines because they have different intercepts and different slopes.

SL Model Effects:
The effect of X_1 on μ_Y is: $\mu(Y \mid X_1 + 1, X_2) - \mu(Y \mid X_1, X_2) = \beta_1 + \beta_4 trt_1 + \beta_5 trt_2$
The effects of X_2 on μ_Y are:

i. trt_1 vs. trt_2: $\mu(Y \mid X_1, X_2 = 1) - \mu(Y \mid X_1, X_2 = 2) = (\beta_2 - \beta_3) + (\beta_4 - \beta_5)X_1$
ii. trt_1 vs. trt_3: $\mu(Y \mid X_1, X_2 = 1) - \mu(Y \mid X_1, X_2 = 3) = \beta_2 + \beta_4 X_1$
iii. trt_2 vs. trt_3: $\mu(Y \mid X_1, X_2 = 2) - \mu(Y \mid X_1, X_2 = 3) = \beta_3 + \beta_5 X_1$

2. **The Parallel-Lines (PL) model: There is no interaction.**
$\mu(Y \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 trt_1 + \beta_3 trt_2$

This model can be rewritten as 3 *parallel* SLR models, one for each level of X_2:

i. $\mu(Y \mid X_1, X_2 = 1) = (\beta_0 + \beta_2) + \beta_1 X_1$
ii. $\mu(Y \mid X_1, X_2 = 2) = (\beta_0 + \beta_3) + \beta_1 X_1$
iii. $\mu(Y \mid X_1, X_2 = 3) = \beta_0 + \beta_1 X_1$

They are *parallel* lines because they have equal slopes but different intercepts.

PL Model Effects:
The effect of X_1 on μ_Y is: $\mu(Y \mid X_1 + 1, X_2) - \mu(Y \mid X_1, X_2) = \beta_1$
The effects of X_2 on μ_Y are:

i. trt_1 vs. trt_2: $\mu(Y \mid X_1, X_2 = 1) - \mu(Y \mid X_1, X_2 = 2) = (\beta_2 - \beta_3)$
ii. trt_1 vs. trt_3: $\mu(Y \mid X_1, X_2 = 1) - \mu(Y \mid X_1, X_2 = 3) = \beta_2$
iii. trt_2 vs. trt_3: $\mu(Y \mid X_1, X_2 = 2) - \mu(Y \mid X_1, X_2 = 3) = \beta_3$

3. **The Equal-Lines (EL) model: There is no categorical variable effect.**
$\mu(Y \mid X_1, X_2) = \beta_0 + \beta_1 X_1$

There is one common line (*equal lines*) for all levels of the categorical variable.

EL Model Effects:
The effect of X_1 on μ_Y is: $\mu(Y \mid X_1 + 1, X_2) - \mu(Y \mid X_1, X_2) = \beta_1$
The effects of X_2 on μ_Y are: There are no X_2 effects.

Inferential Tools for MLR

Data analysis involves finding a good-fitting model whose parameters relate to the questions of interest. Once the model has been established, questions of interest can be investigated through the parameter estimates, with uncertainty expressed through *p*-values and confidence intervals, depending on the nature of the questions asked.

Which Model Fits Best?

Need to balance finding the best model that describes the data while not overcomplicating the model. The more complicated model (more parameters) will, almost always, appear better. Our goal is to determine if it is significantly better. If not, we will prefer the simpler model (less parameters).

Inferential Procedures for MLR

When comparing models or model effects, it is recommended you set up the hypotheses in terms of the coefficients. To answer questions about model effects or to compare models, we will consider the following situations:

1) Test for ANY linear significance → ANOVA *F*-test

2) Effect or comparison involves a single regression coefficient → *t*-tools as in SLR

3) Effect or comparison involves a subset of regression coefficients → ESS *F*-test

(In-class examples look at these tests and also incorporate model effects in more concept-based questions.)