Multiple Linear Regression (MLR)

1 Introduction

Linear regression models are used to explore the relationship between variables as well as make predictions. Simple linear regression (SLR) concerns the study of only one predictor variable with one response variable. However, given the context, it may be clear there are multiple predictors that relate to the response variable. In such a context, we want to:

- Improve predictions on the response variable by including more useful predictors.
- Assess how a predictor relates to the response variable when controlling for other predictors.

Multiple linear regression (MLR) models allow us to examine the effect of multiple predictors on the response variable simultaneously.

There are a couple of ways to think about MLR:

- Extension of SLR to MLR.
- SLR as a special case of MLR.

As a motivating example, we look at data regarding black cherry trees. The data, `cherry` come from the `openintro` package. Researchers want to understand the relationship between the volume of these trees and their diameter and height. Data come from 31 trees in the Allegheny National Forest, Pennsylvania.

```
library(openintro)
Data<-openintro::cherry
```

From this context, we know that volume of a tree is influenced by its diameter and height, so we have more than one predictor in this study.

As you read this set of notes, take note of the similarities and differences between SLR and MLR.

2 Notation in MLR

We write the **MLR model** as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i. \tag{1}$$

In this setup, we have k quantitative predictors. The notation in (1) are as follows:

- y_i: value of response variable for observation i,
- β_0: **intercept** for MLR model,
- β_j: **coefficient** (or slope) for predictor j, for $j = 1, 2, \cdots, k$. We have k predictors, each with its corresponding coefficient.
- x_{ij}: observation i's value for predictor j. Notice there are two numbers in the subscript. The first number denotes which observation, and the second denotes which predictor.
- ϵ_i : **error** for observation i.

The **assumptions** in MLR are identical to SLR:

$$\epsilon_1, \ldots, \epsilon_n \ i.i.d. \sim N(0, \sigma^2). \tag{2}$$

Let us use the `cherry` data from `openintro` as an example:

```
head(Data)
```

```
## # A tibble: 6 x 3
##    diam height volume
##   <dbl>  <int>  <dbl>
## 1   8.3     70   10.3
## 2   8.6     65   10.3
## 3   8.8     63   10.2
## 4  10.5     72   16.4
## 5  10.7     81   18.8
## 6  10.8     83   19.7
```

- $y_2 = 10.3$ cubic feet, the volume for observation 2
- $x_{41} = 10.5$ inches, observation 4's diameter (predictor 1)
- $x_{22} = 65$ feet, observation 2's height (predictor 2)

The MLR model in (1) is often expressed using matrices, which is a lot neater:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & & & \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix},$$

or

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{3}$$

The notation in (3) are as follows:

- $\boldsymbol{y}$: **vector of responses** (length n),
- $\boldsymbol{\beta}$: **vector of parameters** (length $p = k + 1$, where p denotes the number of regression parameters),
- $\boldsymbol{X}$: **design matrix** (dimension $n \times p$),
- $\boldsymbol{\epsilon}$: **vector of residuals** (length n).

The formulation in (3) is the basis for calling the model a "linear" regression. The model is **linear in the parameters**, not the predictors. A common misconception is that the model is linear in the predictors.

Following (1), the **MLR equation** can be written as:

$$E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k. \tag{4}$$

And in turn, the **estimated MLR equation** can be written as:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k. \tag{5}$$

2.1 Interpreting coefficients in MLR

The interpretation of estimated coefficients are similar with SLR, with a small caveat: $\hat{\beta}_j$ denotes the change in the predicted response per unit change in x_j, **when the other predictors are held constant.** There are other common ways to state the bold part:

- when controlling for the other predictors.
- when the other predictors are taken into account.

- after adjusting for the effect of the other predictors.

If you are familiar with how a partial derivative is interpreted in multivariate calculus, you will realize that the interpretation of estimated coefficients in MLR sound like how a partial derivative is interpreted.

Let us look at the estimated regression equation for the `cherry` data:

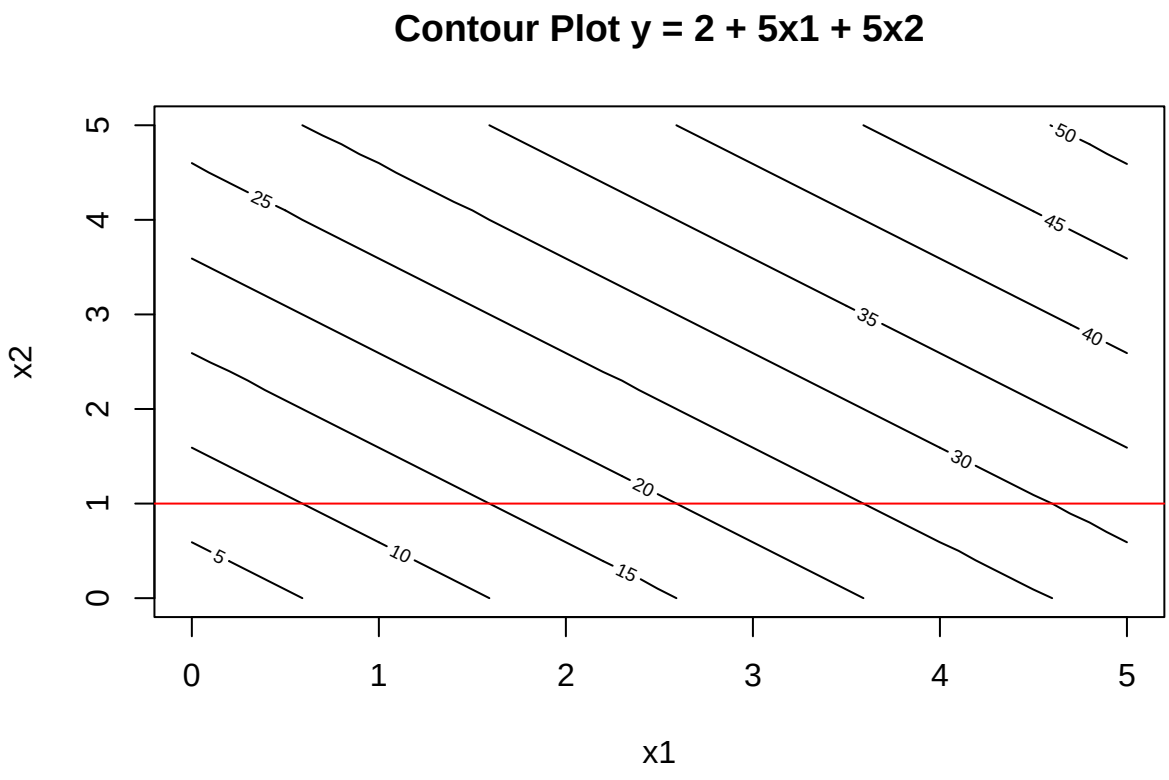
```
result<-lm(volume~., data=Data)
result
```

```
##
## Call:
## lm(formula = volume ~ ., data = Data)
##
## Coefficients:
## (Intercept)         diam       height
##    -57.9877       4.7082       0.3393
```

The estimated MLR equation is $\hat{y} = -57.9877 + 4.7082x_1 + 0.3393x_2$. The estimated coefficient for diameter inform us that for each additional inch in diameter, the predicted volume of a cherry tree increases by 4.7082 cubic feet, while holding height constant.

2.2 Visualizing MLR

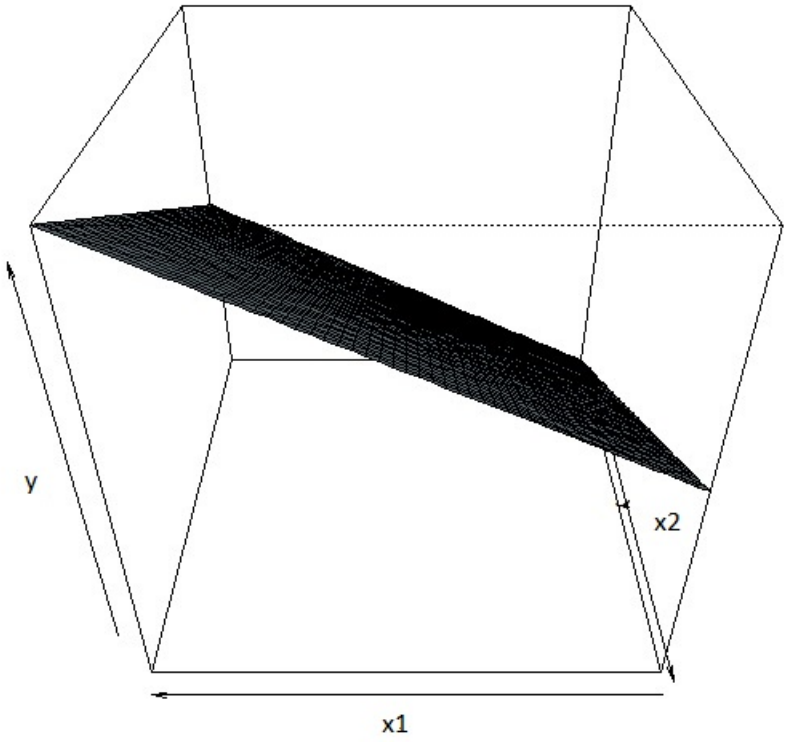
How should we visualize an MLR equation? Suppose we have two predictor variables, x_1 and x_2, with MLR equation $E(y|x_1, x_2) = 2 + 5x_1 + 5x_2$. A **contour plot** can be used to visualize a response variable with two predictors:



The contour plot creates an axis for each predictor. The value of the response variable is denoted by the

contour lines with the actual value displayed on the line. In this toy example, $\beta_1 = 5$. This means that if we hold x_2 constant (e.g. set $x_2 = 1$ per the red horizontal line), increasing x_1 by 1 unit increases the mean of y by 5 units.

The regression equation $E(y|x_1, x_2) = 2 + 5x_1 + 5x_2$ is sometimes called a **regression plane**, instead of a regression line, since we have more than 1 predictor. We can visualize this regression plane below



Due to limitations in human visualization, going beyond a 3-dimensional plot (1 response and 2 predictors) is difficult.

Please see the associated video for a little more explanation regarding the contour plot and 3-dimensional plot.

3 Estimating coefficients in MLR

From (5), the predicted response (or fitted values), can be written in matrix form as:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}, \tag{6}$$

where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k)'$.

We use the **method of least squares** to find the estimated coefficients in MLR. This is the same idea when applied in SLR. The method involves minimizing the **sum of squared residuals**, SS_{res}. In SLR, we minimize

$$\sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2$$

with respect to $\hat{\beta}_0, \hat{\beta}_1$. In MLR, the SS_{res} can be expressed in matrix form:

$$Q = \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \right)' \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \right) \tag{7}$$

To minimize the $Q = SS_{res}$ with respect to $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$, We take partial derivatives of Q and set them all to 0, i.e. $\frac{\nabla Q}{\nabla \hat{\beta}} = 0$. Solving for these equations, we get

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} \boldsymbol{X}'\boldsymbol{y}. \tag{8}$$

Residuals are found in the same way in SLR:

$$e_i = y_i - \hat{y}_i,$$

or in matrix form:

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}. \tag{9}$$

3.1 Estimating variance of errors

Similar to SLR, MS_{res} is used to estimate σ^2, the variance of the error terms. MS_{res} if found using

$$MS_{res} = \frac{SS_{res}}{n - p}, \tag{10}$$

where p denotes the **number of regression parameters**. In SLR, $p = 2$, since we have an intercept and one slope. Note: I have seen too many people think p denotes the number of predictors. This is incorrect! As we move forward, we will explore more complicated regression models and we always think in terms of number of regression parameters.

3.2 Distribution of least squares estimators

Following the Gauss Markov theorem, the least squares estimators $\hat{\boldsymbol{\beta}}$ are unbiased, i.e.

$$\boldsymbol{E}\left(\hat{\boldsymbol{\beta}} \right) = \boldsymbol{\beta}, \tag{11}$$

with **variance-covariance matrix** given by

$$\boldsymbol{Var}\left(\hat{\boldsymbol{\beta}} \right) = \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} \tag{12}$$

with σ^2 estimated by MS_{res}. A few notes about the variance-covariance matrix of the least squares estimators:

- it is of dimension $p \times p$,

- the diagonal elements denote the variance of each estimated parameter. For example, the first diagonal element denotes the variance of $\hat{\beta}_0$, the first estimated parameter.
- the off-diagonal elements denote the covariance between respective parameters. For example, the (1,2) entry denotes the covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$.

Please see the associated video for a demonstration on how to read a variance-covariance matrix.

4 ANOVA F Test in MLR

4.1 Sum of squares

As in simple regression, the **analysis of variance (ANOVA) table** for an MLR model displays quantities that measure how much of the variability in the response variable is explained (and not explained) by the regression model. The underlying conceptual idea for the construction of the analysis of variance table is the same:

$$SS_T = SS_R + SS_{res}. \tag{13}$$

What change are the associated degrees of freedom:

- df for SS_R: $df_R = p - 1$
- df for SS_{res}: $df_{res} = n - p$
- df for SS_T: $df_T = n - 1$

Notice the degrees of freedom in SLR has $p = 2$.

4.2 ANOVA table

The ANOVA table is thus

Source of Variation	SS	df	MS	F
Regression	$SS_R = \sum (\hat{y}_i - \bar{y})^2$	$df_R = p - 1$	$MS_R = \frac{SS_R}{df_R}$	$\frac{MS_R}{MS_{res}}$
Error	$SS_{res} = \sum (y_i - \hat{y}_i)^2$	$df_{res} = n - p$	$MS_{res} = \frac{SS_{res}}{df_{res}}$	***
Total	$SS_T = \sum (y_i - \bar{y})^2$	$df_T = n - 1$	***	***

4.3 ANOVA F test

The null and alternative hypotheses associated with the ANOVA F test are:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0, H_a : \text{ at least one of the coefficients is not 0.}$$

So the null hypothesis states the regression coefficients for all predictors are 0. Notice how this statement simplifies in SLR.

There are a few different ways to view these hypothesis statements:

- Is our MLR model **useful**?
- Is our MLR model **preferred** over an intercept-only model?
- Can we drop **all** our predictors from the MLR model?

The test statistic is still

$$F = \frac{MS_R}{MS_{res}} \tag{14}$$

which is compared with an $F_{p-1,n-p}$ distribution.

4.4 Coefficient of determination

The **coefficient of determination,** R^2**,** is still

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{res}}{SS_T}, \tag{15}$$

where R^2 is interpreted as **the proportion of variance in the response variable that is explained by the predictors**.

4.4.1 Caution with R^2

Adding more predictors to a model can only increase R^2, as SS_{res} never becomes larger with more predictors and SS_T remains the same for a given set of responses.

- So even adding predictors that don't make sense will increase R^2.
- R^2 should be used to compare models with the same number of parameters.
- R^2 is a popular measure as it has a nice geometric interpretation.

In response to this caution, we have the **adjusted** R^2, denoted by R_a^2:

$$R_a^2 = 1 - \frac{\frac{SS_{res}}{n-p}}{\frac{SS_T}{n-1}} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SS_{res}}{SS_T}. \tag{16}$$

R_a^2 increases if the added predictors significantly improve the fit of the model, and decreases otherwise.

Let us go back to the `cherry` dataset as an example:

```
summary(result)
```

```
##
## Call:
## lm(formula = volume ~ ., data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4065 -2.6493 -0.2876  2.2003  8.4847
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
## diam          4.7082     0.2643  17.816  < 2e-16 ***
## height        0.3393     0.1302   2.607   0.0145 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.882 on 28 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.9442
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

The ANOVA F statistic is 255‘ with a small p-value. So we reject the null hypothesis and state that our MLR model with diameter and height as predictors is useful.

The R^2 is 0.948. About 94.8% of the variance in volume of cherry trees can be explained by their diameter and height.

The R_a^2 is 0.9442. This value is used in comparison with another model to decide which should be preferred.

5 t Test for Regression Coefficient in MLR

We can assess whether a regression coefficient is significantly different from 0 in an MLR. The null and alternative hypotheses are very much the same as in SLR:

$$H_0 : \beta_j = 0, H_a : \beta_j \neq 0.$$

What these hypotheses mean in words:

- The null hypothesis supports dropping predictor x_j from the MLR model, **in the presence of the other predictors.**
- The alternative hypothesis supports keeping predictor x_j in the MLR model, or that we **cannot drop it in the presence of the other predictors.**

Notice the meaning of the null and alternative hypotheses are a little different than in SLR, where other predictors are not taken into account.

The test statistic is still

$$t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \tag{17}$$

which is compared with a t_{n-p} distribution.

Let us take a look at the `cherry` dataset:

```
summary(result)$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -57.9876589  8.6382259 -6.712913 2.749507e-07
## diam          4.7081605  0.2642646 17.816084 8.223304e-17
## height        0.3392512  0.1301512  2.606594 1.449097e-02
```

Notice the t statistics associated with testing for the coefficient for each predictor is highly significant. So we do not have evidence to drop any of the other predictors to simplify the model.

5.1 Caution in interpretating t test in MLR

An insignificant t test for a coefficient β_j in MLR indicates that predictor x_j can be removed from the model (and leave the other predictors in). It is **not needed in the presence of the other predictors.**

1. A common misstatement many make is that an insignificant t test for a coefficient β_j in MLR implies that predictor x_j has no linear relation with the response variable. This is not necessarily correct!
    - If x_j is highly correlated with at least one of the other predictors, or is a linear combination of a number of other predictors, x_j will probably be insignificant as the addition of x_j doesn't help in improving the model. This concept is called **multicollinearity** which we will explore in more depth in the next module. x_j does not provide independent information from the other predictors, and so will not be needed when the other predictors are in the model.
    - x_j itself may still be linearly related to the response variable, on its own.
    - If your goal is to assess if x_j is linearly related to the response, need to use SLR.
2. Another common misstatement people make is that if they observe more than one t statistic that is insignificant, it means that all of the associated predictors can be dropped from the model. This again is not necessarily correct.
    - An insignificant t test informs us we can drop that particular predictor, while leaving the other predictors in the model. We can only drop one predictor at a time based on t tests.

Notice the limitation of the t test and ANOVA F test in MLR:

- We can **only drop 1 predictor** based on a t test.
- We can **drop all predictors** based on an ANOVA F test.
- What if we wish to drop more than 1 predictor simultaneously, but not all, from the model? We will explore this via another F test in the next module.

6 CIs in MLR

6.1 CI for regression coefficient

The general form for CIs is still the same:

$$\text{estimator} \pm (\text{multiplier} \times \text{s.e of estimator}). \tag{18}$$

The $100(1-\alpha)\%$ CI for β_j is

$$\hat{\beta}_j \pm t_{1-\alpha/2;n-p} se(\hat{\beta}_j) = \hat{\beta}_j \pm t_{1-\alpha/2;n-p} s\sqrt{C_{jj}} \tag{19}$$

where C_{jj} denotes the jth diagonal entry in variance-covariance matrix of the estimated coefficients, $\boldsymbol{Var}\left(\hat{\boldsymbol{\beta}}\right)$.

The multiplier is now based on a t_{n-p} distribution, instead of a t_{n-2} distribution for SLR.

6.2 CI of the mean response

Since we have multiple predictors, we may interested in the CI for the mean of the response, when the predictors are each equal to specific values. Let the vector $\boldsymbol{x_0}$ denote these values on each predictor, specifically

$$\boldsymbol{x_0} = (1, x_{01}, x_{02}, \cdots, x_{0k})',$$

where x_{0j} denotes the value for predictor x_j. The CI for the mean response when $\boldsymbol{x} = \boldsymbol{x_0}$ is

$$\hat{\mu}_{y|\boldsymbol{x_0}} \pm t_{1-\alpha/2,n-p} s\sqrt{\boldsymbol{x_0}'(\boldsymbol{X'X})^{-1}\boldsymbol{x_0}}. \tag{20}$$

6.3 PI of a new response

For a new value of the response when $\boldsymbol{x} = \boldsymbol{x_0}$, the PI is

$$\hat{y}_0 \pm t_{1-\alpha/2,n-p} s\sqrt{1 + \boldsymbol{x_0}'(\boldsymbol{X'X})^{-1}\boldsymbol{x_0}}. \tag{21}$$