Categorical Predictors in MLR

1 Introduction

Thus far, we have only considered predictors that are quantitative in MLR. But what if we need or want to consider predictors that are categorical instead? For example, what if we wish to investigate the earnings of college graduates based on the type of major? The predictor variable, type of major, is clearly categorical and not quantitative. Categorical variables can be incorporated in MLR. In this module, you will learn how to use and interpret the MLR model when categorical predictors are present.

1.1 Quantitative vs categorical variable

First, a quick review on variable types. Variables can be divided into two types: quantitative or categorical. A general way to assess the type of a variable is the answer to the following question: Do arithmetic operations on the variable make sense? If yes, the variable is quantitative.

1.1.1 Quantitative variable

Quantitative variables are measured in terms of numbers, where the number represents an amount. Quantitative variables can be subdivided into either continuous or discrete.

- **Continuous** quantitative variable: takes on any numerical value within a range. E.g.: height can be measured in terms of centimeters or inches. It is continuous as the value can take on any value between the shortest and tallest person.
- **Discrete** quantitative variable: takes on distinct numerical values. E.g. the number of failures among 10 experiments. It is discrete as it can only take on integers between 0 and 10.

A question to determine if the variable is continuous or discrete: Can you potentially list all the plausible values of the variable? If yes, the variable is discrete. For the example above, we can list the plausible values for the number of failures among 10 experiments as: 0, 1, 2, all the way to 10. For the height variable, the list of numerical values for height will be an infinite list as height can take on an infinite number of decimal places.

1.1.2 Categorical variables

Categorical variables express qualitative attributes (often called qualitative variable). The **levels** (or classes) are the various attributes the variable can take on. For example, political affiliation is categorical with three levels: Democrat, Republican, Independent.

A **binary variable** is a categorical variable with two levels. For example, a variable on whether you voted during the 2020 presidential election is binary, as the answer is either yes or no.

The choice of methods used to analyze data are usually driven by whether the variables are quantitative or categorical. You might have noticed this when creating data visualizations: the visualization you use is driven by the type of variable. Discrete variables are interesting since we can use methods meant for quantitative or categorical variables. We will go over how to make this decision in the context of building MLR models later in the module.

We will see how to incorporate categorical predictors in MLR. We will first start with binary predictors before moving on to predictors with more than two levels.

2 Indicator Variables and Dummy Coding

Let us start by considering this simple example from the textbook: Suppose that a mechanical engineer wishes to relate the effective life of a cutting tool (in hours), y, used on a lathe to the lathe speed in revolutions per minute, x_1, and the type of cutting tool (A or B) used.

```
Data<-read.csv("tool.csv", header=TRUE)
head(Data)
```

```
##   hours rpm type
## 1 18.73 610    A
## 2 14.52 950    A
## 3 17.43 720    A
## 4 14.54 840    A
## 5 13.44 980    A
## 6 24.39 530    A
```

So we have one categorical variable, tool type, that is binary, as it has two levels, A or B. We also have a quantitative predictor, lathe speed, x_1, and a quantitative response variable, effective life.

2.1 Indicator variables

Indicator variables that take on the values 0 and 1 are commonly used to represent the levels of categorical predictors. For example, we may have the following two indicator variables to represent the two tool types:

$$x_2 = \begin{cases} 1 & \text{if type A} \\ 0 & \text{otherwise;} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{if type B} \\ 0 & \text{otherwise;} \end{cases}$$

and the MLR model is written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i.$$

However, this will not work in MLR. Recall that the least squares estimators are

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{1}$$

The matrix $(\boldsymbol{X}'\boldsymbol{X})$ is not invertible if we use these indicator variables. Let us consider a toy example where $n = 4$, with the first two observations being tool type A, and the last two observations being tool type B. The design matrix, $\boldsymbol{X}$, becomes

$$\begin{bmatrix} 1 & x_{11} & 1 & 0 \\ 1 & x_{21} & 1 & 0 \\ 1 & x_{31} & 0 & 1 \\ 1 & x_{41} & 0 & 1 \end{bmatrix}$$

Notice that in the design matrix, column 1 equals to the sum of column 3 and column 4. This means we have linear dependence among the columns of the design matrix. And when this happens, $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ cannot be found so unique solutions to the least squares estimators (1) do not exist.

To get around this issue, we use what is known as **dummy coding**.

2.2 Dummy coding

We drop one of the indicator variables. In general, a categorical variable with a levels will be represented by $a-1$ indicator variables, each taking on the values of 0 and 1. This method of coding categorical variables is called **dummy coding**. So for our tool life example, since tool type is binary (two levels), we only need one indicator variable.

We can have the following model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 I_1 + \epsilon, \tag{2}$$

where $x_1 =$ lathe speed and

$$I_1 = \begin{cases} 1 & \text{if type B} \\ 0 & \text{otherwise.} \end{cases}$$

So in this formulation, we only use one indicator variable, I_1. Indicator variables are typically denoted by I. In this formulation, type A is coded 0, and type B is coded 1. The level that is coded 0 is called the **reference** class (sometimes called the baseline class). When the variable is binary, the choice for reference class is not important. The interpretation of the model does not change based on the choice for the reference class.

2.3 Regression coefficient interpretation

To see how we can interpret the regression coefficients with dummy coding, we can substitute the numerical value of the indicator variable in (2) to obtain the regression equation for each tool type:

Type A: $E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2(0) = \beta_0 + \beta_1 x_1$
Type B: $E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2(1) = (\beta_0 + \beta_2) + \beta_1 x_1$

We make the following observations:

- We have two straight lines, with the **same** slope β_1, but the intercepts are **different**, β_0 and $\beta_0 + \beta_2$. This formulation assumes the slope in the scatterplot of y against x_1 is the same for both tool types.
- β_2 indicates the **difference in the mean response for tool type B versus tool type A (type B minus type A), when controlling for lathe speed.**

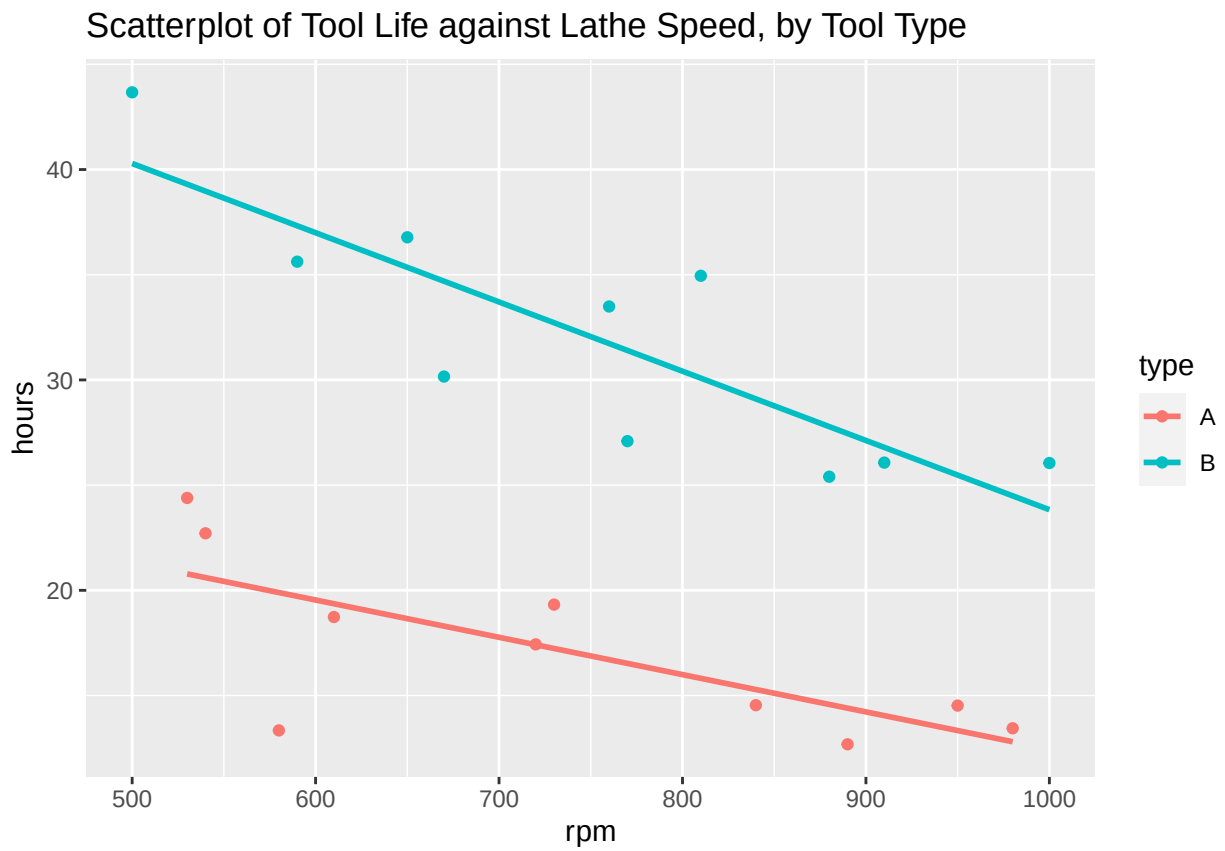
In general, the coefficient of an indicator variable shows how much higher (or lower) the mean response is for the class coded 1 than the reference class, when controlling for x_1.

Let us look at the tool life dataset:

```
##convert type to factor
Data$type<-factor(Data$type)

library(ggplot2)

##scatterplot with separate regression lines
ggplot2::ggplot(Data, aes(x=rpm, y=hours, color=type))+
  geom_point()+
  geom_smooth(method=lm, se=FALSE)+
  labs(title="Scatterplot of Tool Life against Lathe Speed, by Tool Type")
```



We create a scatterplot of the response variable against the quantitative predictor, with separate colors and lines for tool type. Notice that the lines are almost parallel. We noted earlier that the formulation assumes the slopes are parallel.

Assuming that the slopes are not exactly parallel due to random sampling, and that they are truly parallel, β_1 denotes the slopes of both of these lines, and β_2 denotes how much higher the line is for type B than for type A. Let us take a look at the estimated regression coefficients:

```
result<-lm(hours~rpm+type, data=Data)
summary(result)
```

```
##
## Call:
## lm(formula = hours ~ rpm + type, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6255 -1.6308  0.0612  2.2218  5.5044
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.208726   3.738882   9.417 3.71e-08 ***
## rpm         -0.024557   0.004865  -5.048 9.92e-05 ***
## typeB       15.235474   1.501220  10.149 1.25e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3.352 on 17 degrees of freedom
## Multiple R-squared:  0.8787, Adjusted R-squared:  0.8645
## F-statistic:  61.6 on 2 and 17 DF,  p-value: 1.627e-08
```

Note the following from the output:

- We know that since the categorical predictor is binary, we should have only 1 indicator variable representing tool type.
- In the output, note that we have one line for the categorical predictor, called `typeB`. This tells us the name of the predictor, followed by the class that is coded 1. So this tells us that R coded type B as 1, and type A as 0.
- The estimated coefficient for `typeB` is 15.235. This means that the effective life for type B tool is 15.235 hours longer than type A tool, when controlling for lathe speed. The associated p-value when testing this coefficient is small, so the data support claim that there is a significant difference in the mean effective life of the tool based on tool type, when controlling for lathe speed and we do not drop tool type as a term in the model. Since the estimated coefficient is positive, the mean effective life for tool type B is longer than for tool type A, when controlling for lathe speed.
- The estimated coefficient for `rpm` is -0.025. This means the effective life of the tool decreases by 0.025 hours, per unit increase in lathe speed, for given tool type. The associated p-value when testing this coefficient is small, so we do not drop `rpm` as a term in the model.
- The estimated regression equation is $\hat{y} = 35.209 - 0.025x_1 + 15.235I_1$.
- Estimated regression equation for type A: $\hat{y} = 35.209 - 0.025x_1 + 15.235(0) = 35.209 - 0.025x_1$.
- Estimated regression equation for type B: $\hat{y} = 35.209 - 0.025x_1 + 15.235(1) = 50.444 - 0.025x_1$.

Looking at the scatterplot, we noted that the lines may not be exactly parallel. We will need to tweak the regression model as stated in (2) to handle non parallel regressions, if the equations are truly not parallel.

2.4 Thought question

How will the output for this regression change if the dummy coding was changed $I_1 = 1$ if type A and 0 if type B?

Please view the associated video to review this question, as well as for more commentary on dummy coding.

3 Interaction Terms

The regression model as stated in (2) is sometimes called a model with **additive effects**. Additive effects assume that each predictor's effect on the response does not depend on the value of the other predictor. As long as we hold the other predictor constant, changing the value of the predictor is associated with the same change in the mean response. Using the tool life data as an example, this implies that the when looking at the scatterplot of tool life against lathe speed, with separate lines for each tool type, the regression lines are parallel. We noted the lines are not exactly parallel.

When the lines are not parallel, it means the effect of changing lathe speed on the tool life depends on the tool type. When the effect of a predictor on the response depends on the value of the other predictor, we have an **interaction effect** between the predictors on the response variable. To add an interaction effect between the predictors into the model, we have

$$y = \beta_0 + \beta_1 x_1 + \beta_2 I_1 + \beta_3 x_1 I_1 + \epsilon \tag{3}$$

where $I_1 = 1$ if type B and 0 if type A. Substituting the values for I_1 in (3), we have the following regression equations for each tool type:

$$\begin{aligned}\text{Type A: } E(y|x) &= \beta_0 + \beta_2(0) + \beta_1 x_1 + \beta_3 x_1(0) = \beta_0 + \beta_1 x_1.\\ \text{Type B: } E(y|x) &= \beta_0 + \beta_2(1) + \beta_1 x_1 + \beta_3 x_1(1) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1.\end{aligned}$$

We have different intercepts and different slopes for these regressions, whereas in model (2), the slopes are the same.

We can see the effect of changing x_1 on the response depends on the other predictor: for tool type A, increasing x_1 by one unit changes the mean response by β_1; while for tool type B, increasing x_1 by one unit changes the mean response by $\beta_1 + \beta_3$.

Models with interactions are a bit more difficult to interpret than models with no interactions. Therefore, we typical assess if the interaction term is significant or not. This can be tested in a general linear F test framework, since a model with additive effects (2) is the reduced model and a model with interaction effects (3) is the full model. If the coefficient of the interaction term is insignificant, we can drop the interaction term and go with the model with just additive effects.

```
##model with interaction
result.int<-lm(hours~rpm*type, data=Data)
summary(result.int) ##can drop interaction
```

```
##
## Call:
## lm(formula = hours ~ rpm * type, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5534 -1.7088  0.3283  2.0913  4.8652
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.176013   4.724895   6.387 9.01e-06 ***
## rpm         -0.017729   0.006262  -2.831  0.01204 *
## typeB       26.569340   7.115681   3.734  0.00181 **
## rpm:typeB   -0.015186   0.009338  -1.626  0.12345
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.201 on 16 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8764
## F-statistic: 45.92 on 3 and 16 DF,  p-value: 4.37e-08
```

```
##can also do general linear F test. same pvalue as
##t test since only dropping one term
anova(result,result.int)
```

```
## Analysis of Variance Table
##
## Model 1: hours ~ rpm + type
## Model 2: hours ~ rpm * type
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     17 190.98
## 2     16 163.89  1    27.087 2.6443 0.1235
```

We look at the result of the t test for the line `rpm:typeB`, which is the way R denotes the interaction term $x_1 I_1$. This test is insignificant, we do not have evidence that there is an interaction effect between lathe

speed and tool type. So we can drop the interaction and go with the simpler model with just additive effects. The different slopes that we saw in the scatterplot may be attributed to random sampling.

Note that the **hierarchical principle** applies if we have an interaction term: if the interaction term is significant, the lower ordered terms must remain.

3.1 Consideration of interaction terms

Typically, people start with an additive order model. Interactions considered at the start if:

- Exploring interactions is part of your research question.
- An interaction makes sense contextually, or is well-established in the literature.
- We have evidence of interaction based on visualizations.

3.2 Interaction vs correlation

A question that I often get from students have is "Aren't variables that interact also correlated?" The short answer is no, as interaction and correlation are two very different concepts.

A short explanation is that if we say that there is an interaction between two variables, x_1, x_2, it means how each predictor impacts the response variable y depends on the value of the other predictor. Notice that there are three variables, y, x_1, x_2 needed when we talk about an interaction between x_1, x_2.

Correlation only involves two variables.

For a more detailed explanation, including examples, please read this page.

3.3 Dummy coding vs separate regressions

A reasonable question that is often raised is: Why did we not carry out two separate regressions, one for each tool type? When using dummy coding, we are using one regression. The main reason is that it turns out that using one model with dummy coding leads to more precise estimates (smaller standard errors), than creating separate regression for each level. This is true as long as the regression assumptions are met, specifically that the variance of the errors is constant for both levels.

4 Beyond Binary Predictors

If we have a categorical predictor with more than two levels, dummy coding is still used in the same manner: a categorical predictor with a levels will be represented by $a - 1$ indicator variables, each taking on the values of 0 and 1.

Let us use another example. In this example, we consider ratings of wines from California. The response variable is average quality rating, y, with predictors average flavor rating, x_1, and Region indicating which of three regions in California the wine is produced in. The regions are North, Central, and Napa Valley.

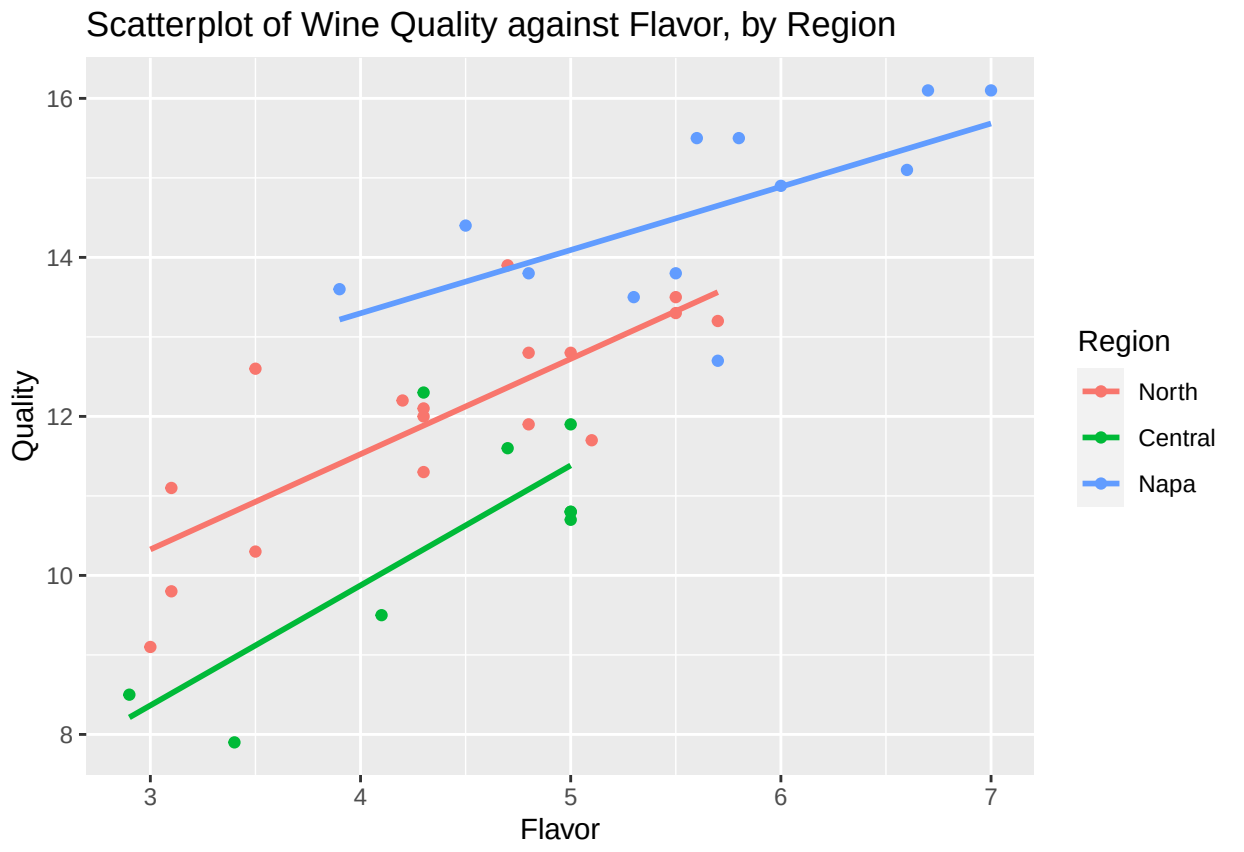
A few notes about using dummy coding in this example:

- Since region has three levels, we will have 2 indicator variables, with one class being the reference class.
- The choice of reference class can be arbitrary, or if there is one class that you are most interested in, make that the reference class.
- The interpretations of the regression model will be consistent regardless of choice for reference class.
- If fitting a model with no interactions, the coefficient of each indicator variable denotes **the difference in the mean response between the level that is coded 1 versus the reference class**, while controlling for the other predictor.

Let us create a scatterplot of quality rating against flavor rating, split by region:

```r
Data<-read.table("wine.txt", header=TRUE, sep="")
##convert Region to factor
Data$Region<-factor(Data$Region)
##assign descriptive labels for each region
levels(Data$Region) <- c("North", "Central", "Napa")

library(ggplot2)
##scatterplot of Quality against Flavor,
##separated by Region
ggplot2::ggplot(Data, aes(x=Flavor, y=Quality, color=Region))+
  geom_point()+
  geom_smooth(method=lm, se=FALSE)+
  labs(title="Scatterplot of Wine Quality against Flavor, by Region")
```



The regression equations are almost parallel so we consider a model with no interactions first. We can use the following indicator variables. Note that there are other ways to define them.

$$I_1 = \begin{cases} 1 & \text{if North} \\ 0 & \text{otherwise} \end{cases}$$

$$I_2 = \begin{cases} 1 & \text{if Central} \\ 0 & \text{otherwise} \end{cases}$$

Since Napa Valley is California's most famous wine region, we would like to make easy comparisons of other regressions with Napa Valley. So it makes sense to make Napa Valley the reference class.

The corresponding model with just additive effects would be

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 I_{i1} + \beta_3 I_{i2} + \epsilon_i$$

So the regression functions are:

North:	$\mathrm{E}\{Y\} = \beta_0 + \beta_1 x_1 + \beta_2(1) + \beta_3(0) = (\beta_0 + \beta_2) + \beta_1 x_1$
Central:	$\mathrm{E}\{Y\} = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(1) = (\beta_0 + \beta_3) + \beta_1 x_1$
Napa Valley:	$\mathrm{E}\{Y\} = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(0) = \beta_0 + \beta_1 x_1$

Recall that this model assumes the slopes are the same for all three regions. β_2, β_3 indicate how different the mean ratings are for the North and Central regions compared with Napa Valley, respectively, when controlling for the other predictor, the average flavor rating.

Let us fit this model and look at the output:

```
##fit regression with no interaction
reduced<-lm(Quality~Flavor+Region, data=Data)
summary(reduced)
```

```
##
## Call:
## lm(formula = Quality ~ Flavor + Region, data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.97630 -0.58844  0.02184  0.51572  1.94232
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     8.3177     1.0100   8.235 1.31e-09 ***
## Flavor          1.1155     0.1738   6.417 2.49e-07 ***
## RegionNorth    -1.2234     0.4003  -3.056  0.00435 **
## RegionCentral  -2.7569     0.4495  -6.134 5.78e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8946 on 34 degrees of freedom
## Multiple R-squared:  0.8242, Adjusted R-squared:  0.8087
## F-statistic: 53.13 on 3 and 34 DF,  p-value: 6.358e-13
```

Some observations from the output:

- $\hat{\beta_2}$ is the estimated coefficient for the indicator of the North region. This value is -1.2234. We interpret this as the average quality rating for wines in the North region is 1.2234 lower than wines in the Napa Valley, when controlling for flavor rating.
- $\hat{\beta_3}$ is the estimated coefficient for the indicator of the Central region.The average quality rating for wines in the South region is 2.7569 lower than wines in the Napa Valley, when controlling for flavor rating.
- As mentioned earlier, we can easily make comparisons for any level with the reference class.

4.1 Difference in mean response between levels excluding the reference class

Estimating the difference in the mean response between levels excluding the reference class can be done by estimating the difference between the regression coefficients of their respective indicator variables. In this example, $\beta_2 - \beta_3$ measures the difference in mean response between the North and Central regions, for given

average flavor rating. The output from our model does not give us this difference immediately. We can either construct a confidence interval (CI) for $\beta_2 - \beta_3$, or perform a hypothesis test for $\beta_2 - \beta_3$.

4.1.1 CI

A $100(1-\alpha)\%$ confidence interval of estimating the differential effects between these regions will be

$$(\hat{\beta}_2 - \hat{\beta}_3) \pm t_{1-\alpha/2,n-p} se\left(\hat{\beta}_2 - \hat{\beta}_3\right), \tag{4}$$

where

$$Var\left(\hat{\beta}_2 - \hat{\beta}_3\right) = Var\left(\hat{\beta}_2\right) + Var\left(\hat{\beta}_3\right) - 2Cov(\hat{\beta}_2, \hat{\beta}_3).$$

Note that in general, the variance for the difference in estimated coefficients is

$$Var\left(\hat{\beta}_j - \hat{\beta}_l\right) = Var\left(\hat{\beta}_j\right) + Var\left(\hat{\beta}_l\right) - 2Cov(\hat{\beta}_j, \hat{\beta}_l). \tag{5}$$

We need to obtain the variance-covariance matrix of the estimated coefficients:

```
##variance covariance matrix of estimated coefficients
round(vcov(reduced),3) ##display 3 dp
```

```
##                (Intercept) Flavor RegionNorth RegionCentral
## (Intercept)          1.020 -0.170      -0.277        -0.277
## Flavor              -0.170  0.030       0.037         0.037
## RegionNorth         -0.277  0.037       0.160         0.113
## RegionCentral       -0.277  0.037       0.113         0.202
```

Let us compute the CI for $\beta_2 - \beta_3$ using (4)

$$\begin{aligned} (\hat{\beta}_2 - \hat{\beta}_3) &\pm t_{1-\alpha/2,n-p}\sqrt{Var\left(\hat{\beta}_2 - \hat{\beta}_3\right)} \\ (-1.2234 + 2.7569) &\pm 2.032245\sqrt{0.160 + 0.202 - 2 \times 0.113} \\ (0.7840453 &, \quad 2.2829547) \end{aligned}$$

Note $t_{1-\alpha/2,n-p}$ found using `qt(0.975, 38-4)`.

The CI excludes 0, so there is a significant difference in the mean quality ratings for wines in the North and Central region, when controlling for flavor rating. The CI consists entirely of positive numbers, so the mean quality rating is higher for wines in the North region than the Central region, when controlling for flavor rating.

View the associated video for more in depth explanation for constructing this CI.

4.1.2 Hypothesis test

We can also compare between the North and Central regions using hypothesis testing. The hypothesis statements will be:

$$H_0 : \beta_2 - \beta_3 = 0, H_a : \beta_2 - \beta_3 \neq 0.$$

The t statistic is

$$
\begin{aligned}
t &= \frac{\hat{\beta}_2 - \hat{\beta}_3}{se\left(\hat{\beta}_2 - \hat{\beta}_3\right)} \\
&= \frac{-1.2234 + 2.7569}{\sqrt{0.160 + 0.202 - 2 \times 0.113}} \\
&= 4.158286,
\end{aligned}
$$

which is larger than the critical value `qt(1-0.05/2, 38 - 4)` $= 2.032245$. So we reject the null hypothesis. Data support the claim that there is a significant difference in the mean quality ratings between wines from the North region and the Central region, when controlling for flavor rating.

4.2 Interactions

Suppose we decide to assess if there is an interaction between flavor rating and region. In other words, the effect of flavor rating on quality rating differs by region. From the scatterplot, the slopes are not exactly parallel, so a significant interaction may exist. The model with interaction would be

$$y = \beta_0 + \beta_1 x_1 + \beta_2 I_1 + \beta_3 I_2 + \beta_4 x_1 I_1 + \beta_5 x_1 I_2 + \epsilon_i$$

So regression functions are:

North:	$\mathrm{E}\{Y\} = \beta_0 + \beta_1 x_1 + \beta_2(1) + \beta_3(0) + \beta_4 x_1(1) + \beta_5 x_1(0) = (\beta_0 + \beta_2) + (\beta_1 + \beta_4)x_1$
Central:	$\mathrm{E}\{Y\} = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(1) + \beta_4 x_1(0) + \beta_5 x_1(1) = (\beta_0 + \beta_3) + (\beta_1 + \beta_5)x_1$
Napa Valley:	$\mathrm{E}\{Y\} = \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3(0) + \beta_4 x_1(0) + \beta_5 x_1(0) = \beta_0 + \beta_1 x_1$

Let us perform a general linear F test to see if we should use the model with no interactions or the model with interactions:

```
##consider model with interactions
##(when slopes are not parallel)
result<-lm(Quality~Flavor*Region, data=Data)

##general linear F test for interaction terms
anova(reduced,result)
```

```
## Analysis of Variance Table
##
## Model 1: Quality ~ Flavor + Region
## Model 2: Quality ~ Flavor * Region
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     34 27.213
## 2     32 25.429  2    1.7845 1.1229 0.3378
```

The general linear F test is insignificant, so we use the reduced model, i.e. the model with no interactions.

5 Pairwise Comparisons

When we have a categorical predictor, we may interested in making comparisons of the mean response between multiple pairs of levels within the categorical predictor. Going back to our wine example, we have a categorical predictor with three levels. This means we can make $\binom{3}{2} = 3$ pairwise comparisons:

- North region vs Napa Valley
- Central region vs Napa Valley
- North region vs Central region

Based on our indicator variables, the regressions equations

North:	$\mathrm{E}\{Y\} = (\beta_0 + \beta_2) + \beta_1 x_1$
Central:	$\mathrm{E}\{Y\} = (\beta_0 + \beta_3) + \beta_1 x_1$
Napa Valley:	$\mathrm{E}\{Y\} = \beta_0 + \beta_1 x_1$

So the parameters denoting the **pairwise comparisons** are

- β_2: North region vs Napa Valley
- β_3: Central region vs Napa Valley
- $\beta_2 - \beta_3$: North region vs Central region

To assess whether there is a significant difference in the mean response between each pair, we can either:

1. Conduct 3 hypothesis tests:
   - $H_0 : \beta_2 = 0, H_a : \beta_2 \neq 0$
   - $H_0 : \beta_3 = 0, H_a : \beta_2 \neq 0$
   - $H_0 : \beta_3 - \beta_2 = 0, H_a : \beta_3 - \beta_2 \neq 0$

or

2. Construct 3 confidence intervals:
   - $\hat{\beta}_2 \pm t_{1-\alpha/2,n-p} se\left(\hat{\beta}_2\right)$
   - $\hat{\beta}_3 \pm t_{1-\alpha/2,n-p} se\left(\hat{\beta}_3\right)$
   - $(\hat{\beta}_2 - \hat{\beta}_3) \pm t_{1-\alpha/2,n-p} se\left(\hat{\beta}_2 - \hat{\beta}_3\right)$

We have to be careful when conducting multiple tests or constructing multiple CIs.

5.1 Significance level

The **significance level**, α, of a hypothesis test is the probability of wrongly rejecting H_0 if it is true. A **Type I** error is defined as wrongly rejecting the null hypothesis if it is true. So an alternate definition of the significance level is the probability of making a Type I error, if the null hypothesis is true.

Suppose we conduct the three hypothesis tests to compare all three pairs of regions, each at significance level α. If the null hypothesis is true for all 3 tests (i.e. there is no significant different in mean response between regions, when controlling for flavor), then the probability of making the right conclusions for all of the tests, assuming the tests are independent, will be $(1-\alpha)^3$. If $\alpha = 0.05$, this probability is $0.95^3 = 0.857375$, not 0.95. A couple of things to consider:

- As we perform more hypothesis tests, the probability of making the right conclusions for all (assuming the null is true for all), decreases.
- Can we do something so that the probability of not making at least one Type I error is still at least $1 - \alpha$?
- For confidence intervals, we want to ensure that the confidence we have that the entire set of intervals capture the true values of the parameters is still at least $1 - \alpha$.

5.2 Multiple pairwise comparisons

To account for the fact that we are making multiple pairwise comparisons, we will need to make our confidence intervals wider, and the critical value larger to ensure the chance of making any Type I error is not more than α. There are a few procedures to do this. We will look two common procedures:

- Bonferroni procedure
- Tukey procedure

With the various procedures in multiple comparison:

1. All the confidence intervals take the form

$$\text{estimate} \ \pm \ \text{multiplier} \ \times \text{se(estimate)} \ . \tag{6}$$

2. For hypothesis tests, we reject the null hypothesis when

$$\text{test statistic} \ > \ \text{critical value.} \tag{7}$$

Only the multiplier and critical values change.

5.3 Bonferroni procedure

Let g denote the number of CIs we wish to construct, or the number of hypothesis tests we need to do to perform our pairwise comparisons.

5.3.1 CIs

The Bonferroni procedure to ensure that we have at least $(1-\alpha)100\%$ confidence that all the g CIs capture the true value is

$$\hat{\beta}_j \pm t_{1-\alpha/(2g);n-p} se(\hat{\beta}_j). \tag{8}$$

The multiplier in (8) is found using $t_{1-\alpha/(2g);n-p}$ instead of $t_{1-\alpha/2;n-p}$.

Going back to the wine example, suppose we wish to make all three pairwise comparisons. The CI to compare the North region with Napa Valley is

$$\begin{aligned} \hat{\beta}_2 &\pm t_{1-\alpha/(2\times 3);38-4} se(\hat{\beta}_2) \\ -1.2234 &\pm 2.518259 \times 0.4003 \\ (-2.2314592 &, \ -0.2153408). \end{aligned}$$

The CI to compare the Central region with Napa Valley is

$$\begin{aligned} \hat{\beta}_3 &\pm t_{1-\alpha/(2\times 3);38-4} se(\hat{\beta}_3) \\ -2.7569 &\pm 2.518259 \times 0.4495 \\ (-3.888858 &, \ -1.624942). \end{aligned}$$

The CI to compare the North region with the Central region is

$$\begin{aligned} (\hat{\beta}_2 - \hat{\beta}_3) &\pm t_{1-\alpha/(2\times 3);38-4} \sqrt{Var\left(\hat{\beta}_2 - \hat{\beta}_3\right)} \\ (-1.2234 + 2.7569) &\pm 2.518259\sqrt{0.160 + 0.202 - 2 \times 0.113} \\ (0.6048119 &, \ 2.4621881) \end{aligned}$$

All three CIs exclude 0, so there is a significant difference in mean quality rating of wines between all pairs of regions, when controlling for flavor rating.

5.3.2 Hypothesis tests

The critical value, based on the Bonferroni procedure, is $t_{1-\alpha/(2g);n-p}$ (instead of $t_{1-\alpha/2;n-p}$). The test statistic still takes the form

$$t = \frac{\text{estimate}}{s.e.\text{ of estimate}}.$$

To compare the North region with Napa Valley, we have $H_0 : \beta_2 = 0, H_a : \beta_2 \neq 0$. The test statistic is

$$\begin{aligned} t &= \frac{\hat{\beta}_2}{se(\hat{\beta}_2)} \\ &= \frac{-1.2234}{0.4003} \\ &= -3.056208. \end{aligned}$$

To compare the Central region with Napa Valley, we have $H_0 : \beta_3 = 0, H_a : \beta_3 \neq 0$. The test statistic is

$$\begin{aligned} t &= \frac{\hat{\beta}_3}{se(\hat{\beta}_3)} \\ &= \frac{-2.7569}{0.4495} \\ &= -6.133259. \end{aligned}$$

To compare the North region with the Central region, we have $H_0 : \beta_2 - \beta_3 = 0, H_a : \beta_2 - \beta_3 \neq 0$. The test statistic is

$$\begin{aligned} t &= \frac{\hat{\beta}_2 - \hat{\beta}_3}{se(\hat{\beta}_2 - \hat{\beta}_3)} \\ &= \frac{-1.2234 + 2.7569}{\sqrt{0.160 + 0.202 - 2 \times 0.113}} \\ &= 4.158286. \end{aligned}$$

The critical value is 2.518259, found using `qt(1-0.05/(2*3), 38-4)`. The magnitudes of all of these test statistics are larger than the critical value, so there is a significant difference in mean quality rating of wines between which pair of regions, when controlling for flavor rating.

View the associated video and the additional set of notes for an explanation of the rationale behind the Bonferroni procedure.

5.4 Tukey procedure

The calculations involved in the Tukey procedure are a little bit more involved so we will not cover the details. We can take a look at some output based on the Tukey procedure:

```
library(multcomp)
pairwise<-multcomp::glht(reduced, linfct = mcp(Region= "Tukey"))
summary(pairwise)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = Quality ~ Flavor + Region, data = Data)
##
## Linear Hypotheses:
##                       Estimate Std. Error t value Pr(>|t|)
## North - Napa == 0     -1.2234     0.4003  -3.056 0.011642 *
## Central - Napa == 0   -2.7569     0.4495  -6.134  < 1e-04 ***
## Central - North == 0  -1.5335     0.3688  -4.158 0.000631 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Each line is showing the result of testing if the mean quality rating differs between the listed pair of regions (when controlling for flavor rating). We note that the results of all three hypothesis tests are significant. So the data support the claim that there is significant difference in the mean quality of wines between all pairs of regions, when controlling for flavor rating.

Given the negative values for the difference in the estimated coefficients, wines from the Napa valley have the highest ratings, followed by wines from the North region, and then wines from the Central region, when flavor rating is controlled.

5.5 Comments about Multiple Comparisons

These procedures to handle multiple pairwise comparisons require there to be no interactions involving the categorical predictor, as we assume the difference in the mean response is the same as long as we control the other predictors.

Some comments about the Bonferroni procedure:

- The probability of making at least one Type I error is $\leq \alpha$. As g increases, this probability becomes a lot less than α.
- A by product is that **power** (the ability to correctly reject the null hypothesis) is sacrificed, especially as g increases.
- Confidence intervals have a higher level of confidence and are wider.
- Bonferroni procedure is **considered conservative** (less powerful, wider intervals).
- Fine to use if g is known prior to looking at the data and is small.
- Easy to implement with simple adjustment to multiplier and critical value.

On the other hand, the Tukey procedure is less conservative and more powerful. Typically used when g is larger.

6 Practical Considerations

6.1 Categorical predictor with many levels

For a categorical predictor with many levels, the output can be daunting to look at, as the number of indicator variables (and hence number of regression coefficients) increases as we have more levels. A few things to consider:

- Are you really interested in exploring the differences in the mean response across all the levels?
- Is there a logical way to **collapse** some levels together that still answers your research question and reduces the number of regression parameters?

- Having more parameters than needed may lead to poorer predictive performance.

6.2 Discrete predictors

Do we treat discrete predictors as quantitative or categorical in MLR? A few things to consider:

- Are we fine with assuming they have a "linear" relationship with the response variable? If yes, more likely we should treat the variable as quantitative.
- How many distinct values are there in the discrete variable? The more distinct values, the more likely we should treat the variable as quantitative.
- Are we concerned about needlessly adding parameters to our model especially if we have a small sample size? If yes, more likely we should treat the variable as quantitative.