General Linear F Test and Multicollinearity

1 Introduction

The purpose of multiple linear regression is to use more than one predictor to predict a response variable. This modules explores an approach to choosing which variables to include in a multiple regression model. For example, many variables can be used to predict someone's systolic blood pressure, such as their age, weight, height, and pulse rate. While all of those predictors are likely to influence the systolic blood pressure, we want to know if we need all of them, or if a subset of those predictors will perform just as well. We will use the general linear F test to do so.

Another issue with having multiple predictors is that the likelihood that at least one of the predictors are linearly dependent, or correlated, with some other predictors increases. This is called multicollinearity. There are some negative consequences if multicollinearity is present. We will learn about these consequences, how to diagnose the presence of multicollinearity, and some solutions if multicollinearity is present.

2 The General Linear F Test

2.1 Motivation

In the previous module, we noted 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 the **general linear F test**. In fact, the t test and ANOVA F test are actually special cases of the general linear F test.

Let us look at a motivating example, using the `nfl.txt` dataset from the textbook. The data are on NFL team performance from the 1976 season. The variables are:

- y: Games won (out of 14 games)
- x_1: Rushing yards (season)
- x_2: Passing yards (season)
- x_3: Punting average (yards/punt)
- x_4: Field goal percentage (FGs made/FGs attempted)
- x_5: Turnover differential (turnovers acquired minus turnovers lost)
- x_6: Penalty yards (season)
- x_7: Percent rushing (rushing plays/total plays)
- x_8: Opponents' rushing yards (season)
- x_9: Opponents' passing yards (season)

We want to assess how the number of games won may be predicted and related to these predictors.

```
Data<-read.table("nfl.txt", header=TRUE)
head(Data)
```

```
##    y   x1   x2   x3   x4 x5  x6   x7   x8   x9
## 1 10 2113 1985 38.9 64.7  4 868 59.7 2205 1917
## 2 11 2003 2855 38.8 61.3  3 615 55.0 2096 1575
## 3 11 2957 1737 40.1 60.0 14 914 65.6 1847 2175
```

```
## 4 13 2285 2905 41.6 45.3 -4 957 61.4 1903 2476
## 5 10 2971 1666 39.2 53.8 15 836 66.1 1457 1866
## 6 11 2309 2927 39.7 74.1  8 786 61.0 1848 2339
```

Let us fit an MLR with all the predictors and take a look at the t tests and ANOVA F test:

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

```
##
## Call:
## lm(formula = y ~ ., data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0408 -0.6802 -0.1131  0.9835  2.9785
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.292e+00  1.281e+01  -0.569 0.576312
## x1           8.124e-04  2.006e-03   0.405 0.690329
## x2           3.631e-03  8.410e-04   4.318 0.000414 ***
## x3           1.222e-01  2.590e-01   0.472 0.642750
## x4           3.189e-02  4.160e-02   0.767 0.453289
## x5           1.511e-05  4.684e-02   0.000 0.999746
## x6           1.590e-03  3.248e-03   0.490 0.630338
## x7           1.544e-01  1.521e-01   1.015 0.323547
## x8          -3.895e-03  2.052e-03  -1.898 0.073793 .
## x9          -1.791e-03  1.417e-03  -1.264 0.222490
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.83 on 18 degrees of freedom
## Multiple R-squared:  0.8156, Adjusted R-squared:  0.7234
## F-statistic: 8.846 on 9 and 18 DF,  p-value: 5.303e-05
```

Notice the t tests are insignificant for a lot of the coefficients. Individually, each t test is informing us that we can drop that specific predictor, while leaving the other predictors in the model.

An erroneous interpretation is to say collectively, these t tests inform us we can drop all of these predictors, except x_2 and maybe x_8, from the model. This is a misconception.

One idea could be to drop the most insignificant predictor, refit the model, and reassess which predictors are insignificant, and continue dropping the most insignificant predictor and refitting the model until all t tests are significant. We will end up conducting multiple hypothesis tests to do so. If possible, we should limit the number of hypothesis tests we conduct: the more tests we do, the likelihood of us wrongly rejecting a null hypothesis increases.

This is where the **general linear F test** (sometimes called a partial F test) is used. We can perform one test to assess if we can simultaneously drop multiple predictors from the model.

Based on this output, we consider dropping $x_1, x_3, x_4, x_5, x_6, x_7, x_9$ since their t tests are insignificant.

2.2 Setting up the general linear F test

The general linear F test allows us to assess if multiple predictors can be dropped simultaneously from the model. The associated F statistic measures the change in the SS_R (or SS_{res}) with the removal of these predictors from the model. The test is based on the following concepts:

- As long as we have the same response variable, SS_T **is constant**, regardless of the model. This is because $SS_T = \sum(y_i - \bar{y})$. It only involves the response variable.
- $SS_T = SS_R + SS_{Res}$.
- Each time predictors are added to the model, the SS_R increases and the SS_{Res} decreases **by the same amount**, since SS_T stays constant.

The general linear F test answers the question: is the change in SS_R (or change in SS_{res}) significant with the removal or addition of predictor(s)?

This question can be answered in a framework that compares two models:

- a **full model**, denoted by F, that uses all predictors under consideration,
- a **reduced model**, denoted by R, that results if some predictors from the full model are dropped.

2.3 Hypothesis statements

Based on this framework, the null and alternative hypotheses for the `nfl.txt` dataset is

$$H_0 : \beta_1 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_9 = 0, H_a : \text{ at least one coeff in } H_0 \text{ is not zero.}$$

In general, the null hypothesis states that the **parameters of the terms that we wish to drop are all 0.** Therefore, the null hypothesis supports the reduced model, R.

The alternative hypothesis states that we cannot drop all the terms that we wish to drop. Therefore, the alternative hypothesis supports the full model, F.

2.4 Test statistic

The associated test statistic for the general linear F test is

$$F_0 = \frac{[SS_R(F) - SS_R(R)]/r}{SS_{res}(F)/(n-p)}, \tag{1}$$

or equivalently

$$F_0 = \frac{[SS_{res}(R) - SS_{res}(F)]/r}{SS_{res}(F)/(n-p)}. \tag{2}$$

The test statistic F_0 is compared with an $F_{r,n-p}$ distribution. The notation is as follows:

- $SS_R(F)$ denotes the SS_R of full model,
- $SS_R(R)$ denotes the SS_R of reduced model,
- r denotes number of parameters being dropped/tested,
- p denotes the number of parameters in the full model,
- $SS_{res}(F)$ denotes the SS_{res} of full model,
- $SS_{res}(R)$ denotes the SS_{res} of reduced model.

Note that the change in SS_R, $SS_R(F) - SS_R(R)$ is always equal to the change in SS_{res}, $SS_{res}(R) - SS_{res}(F)$. Therefore, (1) is always equal to (2).

2.5 Worked example

Let us look at some output for our `nfl.txt` dataset:

```
reduced<-lm(y~x2+x8, data=Data)
anova(reduced, result)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x2 + x8
## Model 2: y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     25 83.938
## 2     18 60.293  7    23.645 1.0084 0.4576
```

- In the output, model 1 has the predictors are x_2, x_8, while model 2 has the predictors $x_1, \cdots, x_9$. So model 1 is the reduced model, and model 2 is the full model.
- We see some information presented in a table. The first line corresponds to model 1, the second line corresponds to model 2.
- Under the column `RSS`, we have the values for SS_{res}. Admittedly, this can be a bit confusing, but it is not SS_R. So we can see that $SS_{res}(R) = 83.938$ and $SS_{res}(F) = 60.293$. Note that SS_{res} is always smaller for the full model, and SS_R is always larger for the full model.
- Under the column `Res.DF`, we have the degrees of freedom for the SS_{res} of that model. In the calculation of the F statistic, we want the value associated with the full model, 18.
- Under the column `Df`, we have the number of parameters that we are testing to drop, which is 7.
- Under the column `Sum of Sq`, we have the difference in SS_{res} between both models, $SS_{res}(R) - SS_{res}(F) = 83.938 - 60.293 = 23.645$
- Under the column `F`, we have the F statistic, $F_0 = 1.0084$. We can verify this calculation using (2), $F_0 = \frac{(83.938-60.293)/7}{60.293/18}$.
- The p-value of this general linear F test is reported in the last column. This can be found using:

```
1-pf(1.0084, 7, 18)
```

```
## [1] 0.4575954
```

The critical value can be found using:

```
qf(1-0.05, 7, 18)
```

```
## [1] 2.576722
```

So we fail to reject the null hypothesis. The data do not support the alternative hypothesis (i.e. the full model). So we go with the reduced model.

2.6 Comparison of general linear F test with other hypothesis tests in MLR

The t test and ANOVA F test in MLR are special cases of the general linear F test, when $r = 1$ and $r = p - 1$ respectively.

- For the t test, the reduced model has 1 less term than the full model. The F_0 statistic is compared with an $F_{1,n-p}$ distribution. It turns out that an $F_{1,n-p}$ distribution is directly related with a t_{n-p} distribution, and so the general linear F test is exactly the same as the t test when dropping 1 term.
- For the ANOVA F test. The reduced model drops all the terms and has only the intercept. Some call this the intercept-only model.

2.7 Alternative approach to general linear F test

There is another way in which the information needed to perform a general linear F test. This approach is called the **sequential sums of squares** (sometimes called extra sums of squares). It works on the same principle that every time a predictor is added to the model, the SS_R of the model increases, and the SS_{res}

decreases by the same amount, since SS_T is constant. The information is displayed as we add one predictor at a time. Let us define some notation:

- $SS_R(x_1)$ denotes SS_R when x_1 is the only predictor in the model.
- $SS_R(x_1, x_2)$ denotes SS_R when x_1, x_2 are in the model.
- $SS_R(x_2|x_1)$ denotes the increase in SS_R when x_2 is added to the model with x_1 already in it. It is read as SS_R of x_2 given x_1.

Based on this example, $SS_R(x_2|x_1) = SS_R(x_1, x_2) - SS_R(x_1)$, and so $SS_R(x_1, x_2) = SS_R(x_1) + SS_R(x_2|x_1)$. Note that $SS_R(x_2|x_1)$ is not equal to $SS_R(x_2)$, as the latter denotes SS_R when x_2 is the only predictor in the model.

Let us see how we can use the sequential sums of squares with the `nfl.txt` dataset:

```
result.seq<-lm(y~x2+x8+x1+x3+x4+x5+x6+x7+x9, data=Data)
anova(result.seq)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x2         1  76.193  76.193 22.7469 0.0001533 ***
## x8         1 166.833 166.833 49.8064 1.392e-06 ***
## x1         1  11.190  11.190  3.3408 0.0842060 .
## x3         1   1.806   1.806  0.5390 0.4722931
## x4         1   0.836   0.836  0.2497 0.6233301
## x5         1   0.652   0.652  0.1946 0.6643619
## x6         1   0.905   0.905  0.2702 0.6095609
## x7         1   2.907   2.907  0.8680 0.3638490
## x9         1   5.348   5.348  1.5967 0.2224904
## Residuals 18  60.293   3.350
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The values under the column "Sum Sq" give the sequential SS_Rs. So,

- for the first line, we have $SS_R(x_2) = 76.193$,
- the second line, we have $SS_R(x_8|x_2) = 166.833$,
- then $SS_R(x_1|x_2, x_8) = 11.190$,
- and so on for each term,
- finally, $SS_R(x_9|x_1, x_2, \cdots, x_8) = 5.348$.
- The very last line refers to $SS_{Res}(x_1, x_2, \cdots, x_9) = 60.293$.

Essentially, the output for each term informs us the increase in SS_R when that term is added to the model, given that the previously listed terms are already in the model.

Notice the order the sequential sums of squares are displayed is the same order used when entering the predictors in `lm()`.

We are still testing

$$H_0 : \beta_1 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_9 = 0, H_a : \text{ at least one coeff in } H_0 \text{ is not zero.}$$

Using (1), the F statistic for this test is

$$F_0 = \frac{[SS_R(F) - SS_R(R)]/r}{SS_{res}(F)/(n-p)}$$
$$= \frac{(11.190 + 1.806 + 0.836 + 0.652 + 0.905 + 2.907 + 5.348)/7}{60.293/18}$$
$$= 1.00839.$$

Compare this F_0 statistic using this approach with the example shown in Section 2.5. We have the exact same result (other than rounding).

Please view the associated video for more explanation on the extra sums of squares approach.

2.7.1 Practice questions

We will use the sequential sums of squares for the `nfl.txt` dataset:

```
anova(result.seq)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x2         1  76.193  76.193 22.7469 0.0001533 ***
## x8         1 166.833 166.833 49.8064 1.392e-06 ***
## x1         1  11.190  11.190  3.3408 0.0842060 .
## x3         1   1.806   1.806  0.5390 0.4722931
## x4         1   0.836   0.836  0.2497 0.6233301
## x5         1   0.652   0.652  0.1946 0.6643619
## x6         1   0.905   0.905  0.2702 0.6095609
## x7         1   2.907   2.907  0.8680 0.3638490
## x9         1   5.348   5.348  1.5967 0.2224904
## Residuals 18  60.293   3.350
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. Carry out a general linear F test to assess if we can drop x_6, x_7, x_9 from the model with all predictors.
2. What is the value of $SS_R(x_2, x_8, x_1)$?
3. What is the value of $SS_{res}(x_2, x_8, x_1)$?
4. What is the value of $SS_{res}(x_1, x_2, \cdots, x_8)$?

Please view the associated video for a review of these practice questions.

3 Multicollinearity

What happens if at least one predictor is almost a linear combination of other predictors? This is called multicollinearity, and there are negative consequences on our MLR model. We will learn what these negative consequences are, how to detect multicollinearity, and some solutions. As we consider more and more predictors for our model, multicollinearity is more likely to exist.

3.1 Linear dependency & multicollinearity

Before we define multicollinearity, we have to define linear dependency. Recall that we can write the MLR model in matrix form as

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

where $\boldsymbol{X}$ is the design matrix and is

$$\begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & & & \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}$$

Note that each column of the design matrix (other than column 1) represents each predictor variable.

The columns of a matrix are **linearly dependent** if at least one column can be expressed as a linear combination of the other columns (there exist nonzero constants c_i such that $c_1x_1 + c_2x_2 + ... + c_kx_k = 0$).

As an example, suppose we have three predictors: x_1 denoting SAT verbal score, x_2 denoting SAT math score, and x_3 denoting SAT score. Since SAT score is the sum of the SAT verbal and math scores, $x_3 = x_1 + x_2$. So if we were to create the design matrix for these three predictors, we have linear dependency. With linear dependency, we can predict x_3 from x_1, x_2 with no error. Recall that the least squares estimators are found using

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}. \tag{4}$$

If there is a linear dependence among the columns of $\boldsymbol{X}$, then $\mathbf{(X'X)^{-1}}$ does not exist. This means that unique estimates of β_j's cannot be determined.

Multicollinearity exists in our model when at least one predictor is **almost linearly dependent**, or can be predicted with a high degree of accuracy, from the other predictors.

An example of multicollinearity would be if we have predictors x_1 denoting right arm length, x_2 denoting right thigh length, and x_3 denoting right calf length. If we know someone's right arm and right thigh lengths, we can probably predict their right calf length with a high degree of accuracy. In this example, we are likely to have multicollinearity.

With multiple predictors, we will always find some degree of **collinearity**. The question is whether this degree is high enough warrant our concern.

When predictors are linearly dependent on each other, they do not provide independent information in their association to the response variable. It becomes difficult to **separate** their effects on the response variable.

3.2 Sources of multicollinearity

There are a few reasons for the presence of multicollinearity.

3.2.1 Study design

The design of the study might lead to multicollinearity, and so a solution will be to change its design. Let us consider this example:

Suppose the Virginia Department of Motor Vehicles (DMV) wants to study the waiting time customers spend waiting in line, based on the number of people ahead in line and number of counters open.

- The number of people ahead in line and number of counters open could be highly positively correlated; the more people in line, the more counters will be staffed by the DMV. So the nature of the study leads to multicollinearity.
- To break the multicollinearity between the number of people ahead in line and number of counters open, we could collect data on instances where the number of people in line is high, yet the number of counters opened is low, and vice versa. This will allow us to isolate the effect of each predictor on waiting times.

3.2.2 Nature of the data

Sometimes, the very nature of the variables lead to multicollinearity and we cannot do much to remedy this.

Suppose we wish to investigate electric consumption in households based on income and size of the home in a city.

- Income and size of the home are likely to be highly correlated, due to high income earners wanting to buy bigger homes, and low income earners being unable to buy bigger homes.
- We cannot force high income earners to live in small homes, or have low income earners buy bigger homes to break the multicollinearity. In this setting, we have to choose one of the predictors.

3.2.3 Too many predictors

As we collect data on more and more variables, we are more likely to encounter multicollinearity. We have to ask if some predictors are provide the same, or similar information, as other predictors.

3.3 Consequences of multicollinearity

The main consequence with multicollinearity is that we have **high variance with the estimated coefficients**. This means the value of the estimated coefficient may be very different from the true value. The consequences from this are:

- Estimated coefficients can be difficult to interpret, as the estimated value may be different from the true parameter. Also, if 2 predictors are correlated, then holding one constant while increasing the other one may not make much sense.
- Algebraic sign of coefficients can be different from what is known theoretically. If the true coefficient is positive, but because the estimated coefficient is different, it could be negative. So we may think the direction of the association is opposite.
- Predictors that we know should impact the response variable are found to be insignificant, as the standard error of the estimated coefficient is large, and hence the t statistic is small. We may erroneously think that predictor is not related to the response variable.

Interestingly, predictions may still be unbiased if the regression assumptions are met.

So depending on what you are using your regression model for, multicollinearity may or may not be a huge problem. Recall the two main uses of regression models:

1. **Prediction**: Predict a future value of a response variable, using information from predictor variables.
2. **Association**: Quantify the relationship between variables. How does a change in the predictor variable change the value of the response variable?

- If the goal of your regression analysis is to interpret the coefficients and understand the effects of each the predictors on the response variable, multicollinearity is a big issue.
- If the goal of your regression analysis is predict future values of the response, then multicollinearity may be less of an issue as long as you do not extrapolate.

4 Detecting Multicollinearity

The following are indicators of the presence of multicollinearity:

1. **Insignificant** results in individual tests on the regression coefficients for important predictor variables. A significant ANOVA F test provides more evidence of multicollinearity.
2. The presence of estimated coefficients with **large standard errors**.
3. Estimated regression coefficients with an algebraic sign that is the **opposite** of that expected from theoretical considerations or prior experience.

4. **High correlation** between pairs of predictor variables.
5. **High variance inflation factors** (VIFs).

We have touched upon the first three ways earlier, and using correlation makes intuitive sense. Next, we will look at VIFs in a bit more detail.

4.1 Variance inflation factors (VIFs)

Variance inflation factors (VIFs) are associated with the coefficients of the predictor variables in MLR. VIFs measure **how much the variance of the corresponding coefficient is multiplied by due to the presence of collinearity versus the lack of collinearity being present**. Mathematically, VIFs are defined as:

$$(VIF)_j = \frac{1}{1 - R_j^2}, \quad j = 1, 2, \ldots, k, \tag{5}$$

where R_j^2 is the **coefficient of determination when x_j is regressed on the other $k-1$ predictors in the model**.

Larger VIFs indicate stronger evidence of multicollinearity. Generally, VIFs greater than 5 indicate some degree of multicollinearity, and VIFs greater than 10 indicate a high level of multicollinearity.

Let us look at the VIFs for the `nfl.txt` dataset:

```
library(faraway)
round(faraway::vif(result),3)
```

```
##    x1    x2    x3    x4    x5    x6    x7    x8    x9
## 4.828 1.420 2.127 1.566 1.924 1.276 5.415 4.536 1.423
```

The VIFs for the coefficients for x_7 is above 5, indicating some degree of multicollinearity in our data.

4.2 Handling multicollinearity

Depends on the source of multicollinearity, as discussed in Section 3.2.

- If due to study design, we can collect data on observations to break the collinearity.
- If due to the nature of the data where some predictors are linearly dependent on others, drop predictor(s). Choose a subset of these predictors (maybe even just one) and remove the rest from the model.
- Abandon least squares regression and use other methods. Other methods such as shrinkage methods and principal components regression help improve predictions, but may not aid in helping explore the relationship between the predictors and response variable. So it depends on what you want to use your regression for.