Paper 3277-2019

Conducting Tests in Multivariate Regression

Chii-Dean Lin, San Diego State University

ABSTRACT

Linear regression models are used to predict a response variable based on a set of independent variables (predictors). Multivariate regression is an extension of a linear regression model with more than one response variable in the model. In a linear regression model, a linear relationship between the response variable and the one or more predictors is assumed. In addition, the random errors are assumed to follow a normal distribution with a constant variance and are also assumed to be independent. In conducting a multivariate regression analysis, the assumptions are similar to the assumptions of a linear regression model but in a multivariate domain. In this paper, we first review the concepts of multivariate regression models and tests that can be performed. In correspondence with the tests under multivariate regression analyses, we provide SAS® code for testing relationships among regression coefficients using the REG procedure. The *mtest* statement in PROC REG is the key statement for conducting related tests. To correctly specify the necessary syntax, we first re-write the hypothesis test we want to test into a form of $LBM = 0$, where L and M are matrices determined by the hypothesis and B is the parameter matrix. The matrices L and M help us to correctly specify the syntax in the *mtest* statement. Various hypothesis tests from an example are used to demonstrate how the L and M are decided and how the *mtest* statement in PROC REG is written.

INTRODUCTION

Regression analysis was first developed in 19th century and is one of the most used statistical methods (Kutner et al, 2004). It can be used for prediction or used for assessing an association between two variables. The purpose of this paper is to review multivariate regression models and to discuss how one can use the PROC REG procedure to test hypotheses in multivariate regression. Multivariate regression is a statistical method that is useful in many fields including medical industry and psychology among others. It is an extension of a univariate regression model (single dependent variable) to a model with multiple response variables. An example of fitting a multivariate regression model is to predict a subject's systolic blood pressure and diastolic blood pressure based on BMI, age, and alcohol consumption. In this example, there are two response variables (systolic blood pressure and diastolic blood pressure) and three predictor variables (BMI, age, and alcohol). In this example, the three predictor variables (BMI, age, and alcohol consumption) are used to predict both response variables: systolic blood pressure and diastolic blood pressure. An approach for the analysis is to fit two univariate multiple linear regression models (one model for each response variable) for the two response variables and interpret the results independently. While the parameter estimates are the same by using either univariate or multivariate regression model, univariate approaches may not be able to address important scientific questions. In addition, using a univariate approach will be less efficient in constructing simultaneous confidence intervals for regression coefficients when the correlations among the response variables are high.

In any statistical data analysis, checking assumptions are always the necessary steps. Assumptions needed for the regression analysis are that random errors follow a normal distribution with a constant variance and they are uncorrelated. Assumptions for a multivariate regression analysis are similar to the assumptions under a univariate regression analysis but extended to a multivariate domain. The assumptions include that the random

error vector follows a multivariate normal distribution and the variance-covariance matrix of the random error vector is homogeneous (Johnson and Wichern, 2007). To test multivariate normality, an introduction and a related SAS code can be found in SAS Customer Support website (SAS Institute online website). A macro (*multnorm*) that is used to test multivariate normality can also be found under the same SAS Customer Support website (SAS Institute online website). To test homogeneity of variance covariance matrix, the Box's M test can be applied. In doing so, one can partition the data into several groups based on *X* values and apply the Box's M test to test homogeneity of a variance-covariance matrix among the partitioned groups. The Box's M test can be produced using the PROC DISCRIM procedure. More information for the Box's M test can be found in SAS STAT manual (SAS Institute (2013)). This paper emphasizes on providing SAS codes for hypothesis tests in multivariate regression analyses through an example.

Note this paper is an extension of Lin (Lin, 2015). In this paper, a quick overview of multiple linear regression and multivariate regression is given. An example is used to test interesting scientific questions and how the corresponding SAS codes are written. Various tests related to multivariate regression are provided. Finally, a conclusion that summarizes this paper is provided.

MULTIPLE LINEAR REGRESSION VS. MULTIVARIATE REGRESSION

For a linear regression model with one predictor variable (simple linear regression), we can state the model as:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where Y_i is the i^{th} response, X_i is the i^{th} observed independent variable, β_0 and β_1 are unknown parameters, and ε_i is a random error following a normal distribution with *0* mean and a constant variance σ^2. The random errors ε_i and ε_j are assumed to be uncorrelated. In addition, to fit a linear regression model, *Y* and *X* should be linearly associated. In this model, the slope β_1 represents the expected change of the outcome variable *Y* when the value of *x* is changed by one unit. The intercept β_0 represents the expected value of *Y* when *X* is *0*. Depending on the range of the collected data, it is possible that the intercept is meaningless (in the case that the observed *x* values does not cover *0*, which will run into an extrapolation issue for interpreting the meaning of the intercept).

If there is more than one predictor variable in a regression model, it is called a multiple linear regression model. We can use a matrix format to present the multiple linear regression model:

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon},$$

where $\underline{Y}$ is an *n x 1* response vector, *X* is an *n x (p+1)* matrix, $\underline{\beta}$ is a *(p+1) x 1* parameter vector, and $\underline{\varepsilon}$ is an *n x 1* random vector. In this model, we assume that there are *p* predictor variables. The least square estimator of $\underline{\beta}$ is $(X^tX)^{-1}X^t\underline{Y}$ and the variance of the least square estimator is $\sigma^2(X^tX)^{-1}$. An extension of a multiple linear regression model is to consider the model with more than one response variable. In this case, it is called multivariate regression analysis. A multivariate regression model with *k* response variables can be expressed as

$$Y = XB + \varepsilon,$$

where *Y* is an *n x k* response matrix, *X* is an *n x (p+1)* matrix, B is a *(p+1) x k* parameter matrix, and ε is an *n x k* random error matrix. For the two models described above, the design matrix *X* is identical for both models. That is, we use the same set of independent variables to predict different response variables. The four matrices in the model can be expressed as follow:

$$Y = \begin{bmatrix} \underline{y}_1' \\ \underline{y}_2' \\ \vdots \\ \vdots \\ \underline{y}_n' \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \vdots & x_{1p} \\ 1 & x_{21} & \vdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \vdots & x_{np} \end{bmatrix}, B = \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0k} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ \beta_{p1} & \beta_{p2} & \cdots & \beta_{pk} \end{bmatrix}, \epsilon = \begin{bmatrix} \underline{\varepsilon}_1' \\ \underline{\varepsilon}_2' \\ \vdots \\ \vdots \\ \underline{\varepsilon}_n' \end{bmatrix}.$$

Note $\underline{y}_i^t$ represents the *k* outcomes from the i^{th} subject. Note also if *k* = 1, the above multivariate regression model is the same as the usual univariate multiple linear regression model. The least square estimator of B is $(X^tX)^{-1}X^tY$. To conduct parameter tests in multivariate regression, most of the *mtest* statement in PROC REG is straightforward. For complicated multivariate tests, we suggest to rewrite the hypothesis test into a multivariate general linear hypothesis $H_0: LBM = 0$, where *L* and *M* are matrices to be decided so that $H_0: LBM = 0$ and the desired hypothesis are identical where the matrix *B* is the parameter matrix defined above. The elements in *L* and *M* are used to decide the linear functions of the *mtest* statement.

USING PROC REG FOR MULTIVARIATE REGRESSION

The SAS procedure, PROC REG, provides tools for fitting regression models, model selections, and diagnostic analyses, etc. Diagnostic plots such as residual plot, studentized residual plot, histogram of the residual, quantile-quantile plot (QQ plot), and Cook's distance are automatically produced for a newer version of SAS.

When a multivariate regression model is considered, normally the *k* outcome measures are correlated. If the correlations among the response variables are small, individual univariate regression approaches can be applied since there is not much difference if we compare the results from univariate regression approaches to the results from a multivariate approach. When the correlations among the response variables are high, it is more efficient for applying a multivariate approach. As mentioned in the Introduction section, the SAS macro, *multnorm*, can be used to test multivariate normal assumption and the Box's M test under *PROC DISCRIM* can be used to test homogeneity of variance-covariance matrix assumption.

The *mtest* statement in *PROC REG* is used for analyses related to multivariate regression models. If there is no expression in the *mtest* statement, the *mtest* will test the hypothesis that all parameters (coefficients of predictors) except the intercept are zero. That is, it will test if there is no linear association between the predictors and the set of response variables. The *mtest* statement for different tests is introduced through an example in next section.

AN EXAMPLE

In this section, we use an example to state several scenarios and to demonstrate the use of the *mtest* statement in PROC REG. The example we use is weightlifting data collected from the International Weightlifting Federation (IWF) (IWF(2015)). Weightlifting competition is an Olympic event that is categorized by an athlete's weight and gender. There are eight categories (from 56 kg to 105+ kg) for men and seven categories (from 48 kg to 75+ kg) for women. Two lift styles (the snatch style and the clean and jerk style) are required for each competition. At most three attempts are allowed for each lift style. Three champions are awarded for each category (the snatch style, the clean and jerk style, and the sum of the snatch and the clean and jerk (TOTAL)). While there is no age category at the Olympic game, the IWF does maintain world records categorized by age group as well. This example was chosen to demonstrate tests in multivariate regression analyses. We use this example to test hypotheses under a multivariate regression model. In this example, we consider two

predictors (AGE and bodyweight (WT)) and three outcome variables (the snatch style, the clean and jerk style, and the total). Note this example can be analyzed using two-factor factorial multivariate analysis of variance (MANOVA) with AGE and WT as the two factors due to the characteristics of the two predictors.

The records maintained by the IWF show that the peak performance age for this sport is around 40 and it is older than the peak performance age of most other sports. For demonstration purposes, we use six age groups (age from 42 to 72) and seven WT categories (range from 56 kg to 105 kg) from men's records only (linear trends are observed within the selected ranges). We want to assess if AGE and WT variables are linearly associated with the three outcome variables (the snatch style, the clean and jerk style, and the total). We want to assess if the predictor variables AGE and WT are good predicting variables for the response variables. Can we build a model with good predicting power to predict the three response variables based on AGE and WT? If a linear association is identified, we can also test if the coefficients of AGE are the same for the snatch style and the clean and jerk style responses. That is, under same WT category, we want to test if the "effect" of AGE on the snatch style and on the clean and jerk style responses is identical. Similarly, we can evaluate the impact of WT predictor variable on the snatch style and the clean and jerk style outcomes as well. Since the response variable TOTAL is the sum of the snatch lift style and the clean and jerk lift style from an athlete, we can also test if the AGE coefficient associated with the TOTAL response variable equals the sum of the AGE coefficients related to the snatch style and the clean and jerk style responses. The tests mentioned above can be performed using the *mtest* statement. Depending on the desired tests, some SAS codes are straightforward to generate. For a more complicated test, we suggest rewriting the null hypothesis into a form of $H_0: LBM = 0$, where B is the parameter matrix and L and M are matrices to be decided so that $LBM = 0$ and the desired test are equivalent. Note the matrix L is used to assess the impact of predictors within the same response variables while the matrix M is used to evaluate the impact of predictors among response variables. The elements of the matrices L and M are used to code the *mtest* statement.

In this example, the dimensions in the multivariate regression model are $n = 42$ (42 observations), $p = 2$ (two predictor variables, WT and AGE), and $k = 3$ (three response variables, SNATCH, CLEAN, and TOTAL). The parameter matrix B for this example is

$$B = \begin{bmatrix} \beta_{01} & \beta_{02} & \beta_{03} \\ \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{21} & \beta_{22} & \beta_{23} \end{bmatrix}.$$

Note β_{0i} is the intercept for the i^{th} outcome variable; β_{1i} (β_{2i}) is the slope associated with the WT(AGE) variable for the i^{th} response variable ($i = 1, 2, or\ 3$), respectively.

Before we conduct an overall test for the multivariate regression analysis, scatterplots are generated to check linearity assumption. The scatterplots (Figure 1) show that negative linear trends are observed between the outcome variables and the AGE variable (left panel). That is, from age 40 to age 70, the older an athlete is, the lighter an athlete can lift. This is consistent with all three response variables (SNATCH, CLEAN, and TOTAL). Similar findings are detected for the WT variable as well. The linearity trends are also suitable for the associations between the WT variable and the response variables. This shows within each age group, the heavier an athlete's body weight is, the more weight an athlete can lift. From the scatterplots, we can fit linear regression models for the response and the predictor variables we considered here.

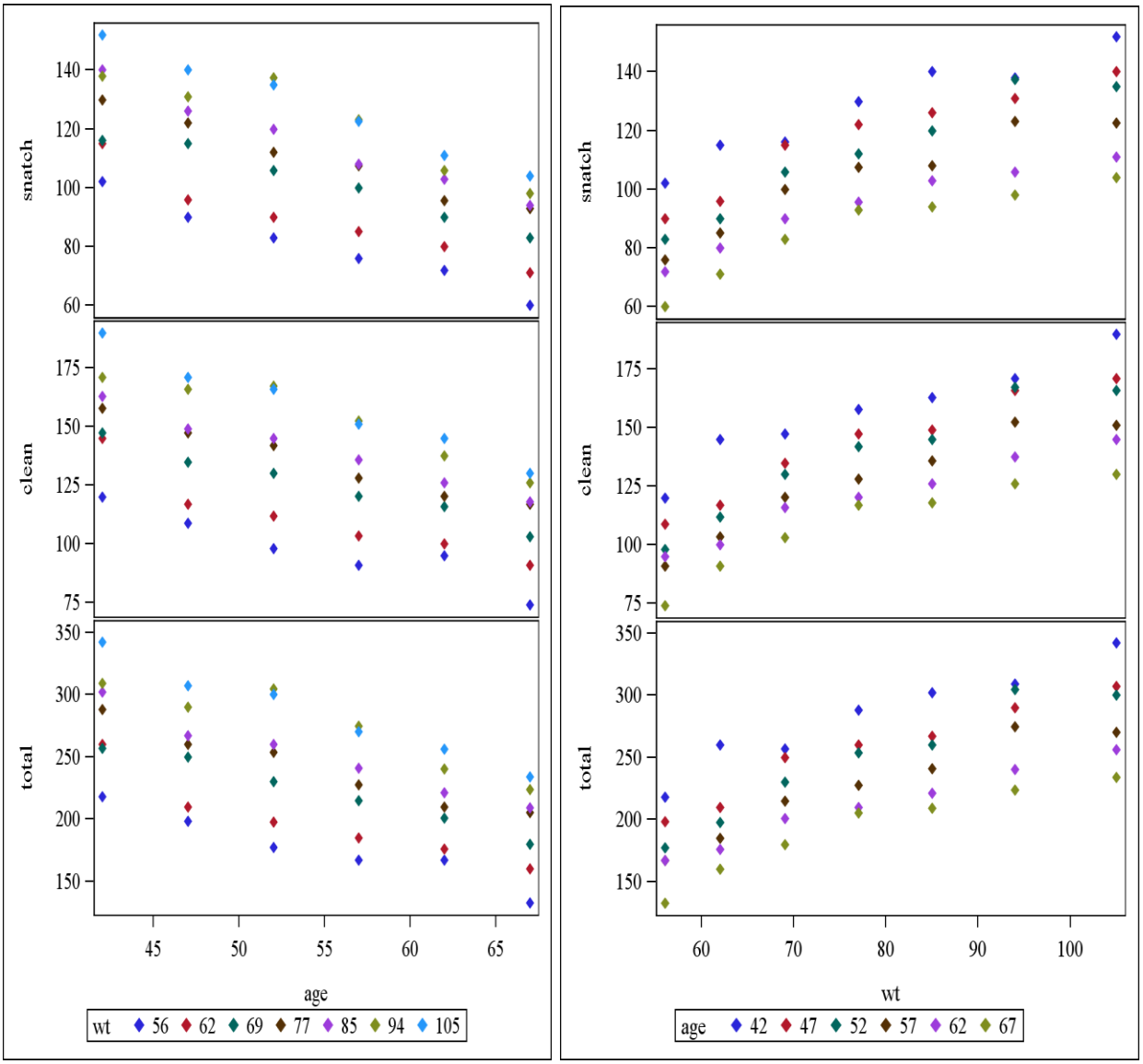

Figure 1. Scatterplots of Snatch (top), Clean (middle), and Total (bottom) variables categorized by age (left panel) and categorized by wt (right panel).

The left panel in Figure 1 was generated by the PROC SGSCATTER procedure. The SAS code is shown below:

```
proc sgscatter data = wtlift;
   compare y=(snatch clean total) x=age /group = wt
   markerattrs=(symbol="diamondfilled");
 run;
```

Similarly, the SAS code for the right panel plots is:

```
proc sgscatter data = wtlift ;
   compare y=(snatch clean total) x=wt /group = age
   markerattrs=(symbol="diamondfilled");
run;
```

To evaluate the dependency among the three response variables and the two predictor variables, a Pearson correlation coefficients matrix is generated and shown below. As expected, the three response variables are highly correlated. The correlations between the response variables and the AGE variable are negative (from -0.6 to -0.65) while the correlations between the response variables and the WT variable are positive (from 0.72 to 0.77). These findings are consistent with what we observed from previous scatterplots (Figure 1).

Pearson Correlation Coefficients, N = 42 Prob > |r| under H0: Rho=0					
	snatch	**clean**	**total**	**wt**	**Age**
snatch	1.00000	0.98928 <.0001	0.99420 <.0001	0.72269 <.0001	-0.65166 <.0001
clean	0.98928 <.0001	1.00000	0.99680 <.0001	0.76608 <.0001	-0.59813 <.0001
total	0.99420 <.0001	0.99680 <.0001	1.00000	0.75071 <.0001	-0.62231 <.0001
wt	0.72269 <.0001	0.76608 <.0001	0.75071 <.0001	1.00000	0.00000 1.0000
age	-0.65166 <.0001	-0.59813 <.0001	-0.62231 <.0001	0.00000 1.0000	1.00000

With the understanding of the associations among the variables, we conduct the following tests.

- Is there a linear association between the outcome variables and the set of independent variables (an overall effect test)? (H_0: $\underline{\beta_1} = \underline{\beta_2} = \underline{0}$)
- Are the fitted regression lines associated with the SNATCH and the CLEAN outcomes identical? (H_0: $\beta_{01} = \beta_{02}$, $\beta_{11} = \beta_{12}$, $\beta_{21} = \beta_{22}$)
- Are the AGE coefficients for the SNATCH and for the CLEAN responses identical? H_0: $\beta_{21} = \beta_{22}$
- Test H_0: $\beta_{11} + \beta_{12} = \beta_{13}$
- Test H_0: $\beta_{21} + \beta_{22} = \beta_{23}$
- Test H_0: $\beta_{13} - \beta_{11} = \beta_{23} - \beta_{21}$, $\beta_{13} - \beta_{12} = \beta_{23} - \beta_{22}$

The SAS codes, the matrices *L* and *M*, and the SAS outputs are discussed in next sections.

Is there a linear association between the outcome variables and the set of independent variables (an overall test)?

This is the first hypothesis test to be performed in regression analyses. The null hypothesis for this test is H_0: $\underline{\beta_1} = \underline{\beta_2} = \underline{0}$. To conduct this test, we can add *mtest* statement to the PROC REG procedure. The SAS code is provided below:

```
proc reg data = wtlift;
 model snatch clean total = wt age;
 mtest;
quit;
```

Note in the MODEL statement, there are three variables (SNATCH, CLEAN, and TOTAL) on the left-hand side of the equal sign. These are the three response variables we considered in the model. The independent variables WT and AGE are listed on the right-hand side of the equal sign. The WT and AGE variables are used to predict the SNATCH, the CLEAN, and the TOTAL response variables in this analysis. With the *mtest* statement, a multivariate test will be performed in addition to the univariate regression analysis outputs from the PROC REG procedure. If there is no expression used after the *mtest* statement, the test that there is no linear association between the outcome variables and the set of independent variables (H_0: $\underline{\beta_1} = \underline{\beta_2} = \underline{0}$ in this example) will be assumed. We expect the estimated coefficients of WT and AGE are different for different response variables.

Partial outputs including parameter estimates tables for the three univariate regression analyses and a multivariate test table are shown below (generated by the *mtest* statement). The three parameter estimates tables are for SNATCH, CLEAN, and TOTAL response variables, respectively. Both WT and AGE predictor variables are linearly associated with the three response variables. The three fitted regression lines are

$$\widehat{SNATCH} = 121.27 + 0.949*WT - 1.632*AGE$$
$$\widehat{CLEAN} = 135.15 + 1.217*WT - 1.811*AGE$$
$$\widehat{TOTAL} = 252.35 + 2.187*WT - 3.456*AGE$$

As we expected, the WT variable is positively correlated with all three response variables while the AGE variable is negatively correlated with these three response variables. For the SNATCH response, we can predict that the SNATCH record will be 0.949 kg more if the body weight is increased by 1 kg (AGE predictor is held constant). Similarly, the record will be 1.632 kg lighter if the AGE is increased by 1 year (WT predictor is held constant). If we compare the SNATCH and the CLEAN outcomes, we can see that the record change using the CLEAN lift style is larger than the record change using SNATCH lift style (1.217kg to 0.949kg) if an athlete changes one kg in weight (AGE is held constant). The R^2 (not shown) for the three fitted models are all very high (range from .94 to .96). The associated small *p*-values and the high R^2 indicate that both WT and AGE are good predictor variables to predict weightlifting records in the three considered categories (SNATCH, CLEAN, and TOTAL).

Parameter Estimates (SNATCH)					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	121.27092	6.35127	19.09	<.0001
Wt	1	0.94918	0.04844	19.59	<.0001
Age	1	-1.63184	0.09236	-17.67	<.0001

Parameter Estimates (CLEAN)					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	135.15105	7.84591	17.23	<.0001
Wt	1	1.21685	0.05984	20.34	<.0001
Age	1	-1.81143	0.11409	-15.88	<.0001

Parameter Estimates (TOTAL)					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	252.34745	13.55889	18.61	<.0001
Wt	1	2.18659	0.10341	21.14	<.0001
Age	1	-3.45592	0.19717	-17.53	<.0001

With the added *mtest* statement to the Proc REG procedure, a multivariate statistics table shows test statistics and *p* values is provided. In the table, four test statistics (Wilks' Lambda, Pillai's Trace, Hotelling-Lawley Trace, and Roy's Greatest Root) are presented. All four *p* values in this table are very small. This indicates that there are linear associations between the two predictor variables and the three response variables. Note an R^2 type measure for the model can be calculated from the Wilk's Lambda. The subtraction of the Wilk's Lambda from 1 can be used as an R^2 type measure. In this case, the value is *1 – 0.0376 = 0.9624*. This indicates that about *96%* of the variation of the 3 response variables can be explained by the two predictor variables in the model. This value is similar to the R^2s calculated from the three univariate regression approaches (range from .94 to .96).

Multivariate Statistics and F Approximations					
S=2 M=0 N=17.5					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.03755975	51.31	6	74	<.0001
Pillai's Trace	1.17464085	18.03	6	76	<.0001
Hotelling-Lawley Trace	19.97456401	121.63	6	47.596	<.0001
Roy's Greatest Root	19.68759753	249.38	3	38	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					
NOTE: F Statistic for Wilks' Lambda is exact.					

The overall model test shows that there are strong linear associations between the three considered response variables and the two considered predictor variables. We further perform some subject-specific multivariate tests from this example.

Are the fitted regression lines for the SNATCH and for the CLEAN outcomes identical?

Since the world records using the SNATCH and the CLEAN lift styles are similar, we may want to test if the two regression lines are identical. The hypothesis for this test is H_0: $\beta_{01} = \beta_{02}$, $\beta_{11} = \beta_{12}$, $\beta_{21} = \beta_{22}$. The null hypothesis assumes that the intercepts, the coefficients of AGE, and the coefficients of WT are all identical for the fitted lines associated with the SNATCH and the CLEAN responses. To be able to specify the correct *mtest* statement, we can determine the matrices *L* and *M* so that *LBM = 0* and $\beta_{01} = \beta_{02}$, $\beta_{11} = \beta_{12}$, $\beta_{21} = \beta_{22}$ are equivalent. The matrix *L* for this test is I_3 (an identity matrix with a dimension of *3*) and matrix *M* is $[1, -1, 0]^t$. The first column of *L* represents the intercept. The second column is for WT while the third column is related to AGE. For the matrix *M*, the element of the first row is associated with the first response variable (SNATCH in our model). The second row is related to the second response variable, etc. In our example, CLEAN and TOTAL are associated with the second and third rows. Based on the elements of *L* and *M*, we can generate the following SAS code for this test:

```
proc reg data = wtlift;
 model snatch clean total = wt age;
 mtest intercept, wt, age, snatch-clean;
quit;
```

The inclusion of `intercept, wt, age` is decided from *L* and `snatch-clean` is based on the elements of *M*. The output table shown below suggests that the two fitted lines are not identical (*p value < 0.0001)*.

Multivariate Statistics and Exact F Statistics					
S=1 M=0.5 N=18.5					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.01601561	798.71	3	39	<.0001
Pillai's Trace	0.98398439	798.71	3	39	<.0001
Hotelling-Lawley Trace	61.43909100	798.71	3	39	<.0001
Roy's Greatest Root	61.43909100	798.71	3	39	<.0001

Are the AGE coefficients equal for the SNATCH and the CLEAN responses?

In this test, we are interested in evaluating if the AGE impact on the SNATCH and the CLEAN outcomes are equal under the same WT. That is, we want to test if $\beta_{21} = \beta_{22}$. Note β_{21} and β_{22} are the coefficients of AGE associated with the SNATCH and the CLEAN responses, respectively. From previous tables, we see the estimated AGE coefficients associated with SNATCH and CLEAN outcomes are -1.63 and -1.81, respectively. We want to test if the two coefficients are significantly different. This can be done by adding a linear function into the *mtest* statement. In this test, we have *mtest* `snatch-clean, age;`. The SNATCH-CLEAN expression describes the coefficient difference for the SNATCH and the CLEAN responses while the AGE expression restricts the coefficient difference we are testing to the AGE predictor only. The associated *L* and *M* are [0, 0, 1] and $[1, -1, 0]^t$, respectively. The complete code for this test is shown below.

```
proc reg data = wtlift;
 model snatch clean total = wt age;
```

```
 mtest snatch-clean, age;
quit;
```

To be able to identify the output table easily, we can also add a label into the *mtest* statement. This can be done by specifying any label followed by the "**:**" symbol before the *mtest* statement. The following output table shows the result for testing the equivalence of the two coefficients associated with the AGE variable.

Multivariate Statistics and Exact F Statistics					
S=1 M=-0.5 N=18.5					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.81784061	8.69	1	39	0.0054
Pillai's Trace	0.18215939	8.69	1	39	0.0054
Hotelling-Lawley Trace	0.22273214	8.69	1	39	0.0054
Roy's Greatest Root	0.22273214	8.69	1	39	0.0054

The result shows that the coefficients of AGE for the SNATCH and the CLEAN are significantly difference. That means, the weightlifting records for the SNATCH and the CLEAN AND JERK styles is significantly different when the AGE is changed by one year while the WT predictor is held constant. Similarly, we can test the slopes for the WT variable on the SNATCH and the CLEAN response variables are equal as well (H_0: $\beta_{21} = \beta_{22}$). The derived *L* and *M* are [0, 1, 0] and [1, -1, 0]^t, respectively. The corresponding SAS code is to replace AGE with WT in the *mtest* statement.

```
proc reg data = wtlift;
 model snatch clean total = wt age;
 mtest snatch-clean, wt;
quit;
```

The result table is shown below.

Multivariate Statistics and Exact F Statistics					
S=1 M=-0.5 N=18.5					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.35731346	70.15	1	39	<.0001
Pillai's Trace	0.64268654	70.15	1	39	<.0001
Hotelling-Lawley Trace	1.79866314	70.15	1	39	<.0001
Roy's Greatest Root	1.79866314	70.15	1	39	<.0001

A highly significant result is observed for this test as well.

Test H_0: $\beta_{21} + \beta_{22} = \beta_{23}$

Since the outcome variable TOTAL is the sum of the snatch and the clean and jerk lifting weights from an athlete, it is interesting to test if the coefficients are additive for either AGE

or WT variables. The stated hypothesis is for the AGE variable only. The *L* and *M* matrices are [0, 0, 1] and [1, 1, -1]^t, respectively. The corresponding SAS code is

```
proc reg data = wtlift;
 model snatch clean total = wt age;
 mtest snatch+clean-total, age;
quit;
```

The *p*-value from the multivariate test is 0.8085 and thus we can conclude that there is no statistically significant difference between the sum of the AGE coefficients associated with the SNATCH and the CLEAN and the coefficient of AGE for the TOTAL response.

Multivariate Statistics and Exact F Statistics					
S=1 M=-0.5 N=18.5					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.99847513	0.06	1	39	0.8085
Pillai's Trace	0.00152487	0.06	1	39	0.8085
Hotelling-Lawley Trace	0.00152720	0.06	1	39	0.8085
Roy's Greatest Root	0.00152720	0.06	1	39	0.8085

Test H_0: $\beta_{11} + \beta_{12} = \beta_{13}$

Similarly, the argument of the *mtest* statement for testing if the sum of the coefficients of WT for the SNATCH and for the CLEAN responses is equal to the coefficient of WT for the TOTAL response is shown below. The *L* and *M* matrices are [0, 1, 0] and [1, 1, -1]^t, respectively.

```
proc reg data = wtlift;
 model snatch clean total = wt age;
 mtest snatch+clean-total, wt;
quit;
```

As similar to the previous test, a non-significant result has been observed. That is, the test fails to reject that the sum of the coefficients of WT for the SNATCH and the CLEAN responses is equal to the coefficient of WT for the TOTAL response.

Multivariate Statistics and Exact F Statistics					
S=1 M=-0.5 N=18.5					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.98555744	0.57	1	39	0.4542
Pillai's Trace	0.01444256	0.57	1	39	0.4542
Hotelling-Lawley Trace	0.01465420	0.57	1	39	0.4542
Roy's Greatest Root	0.01465420	0.57	1	39	0.4542

Test H_0: $\beta_{13} - \beta_{11} = \beta_{23} - \beta_{21}$, $\beta_{13} - \beta_{12} = \beta_{23} - \beta_{22}$

This test is used to demonstrate a more complicated test in multivariate regression models. The matrices *L* and *M* will make it easier to specify the *mtest* statement. After some algebraic work, the corresponding *L* and *M* are:

$L = [0, 1, -1],\ M = \begin{bmatrix} -1 & 0 \\ 0 & -1 \\ 1 & 1 \end{bmatrix}$. The associated SAS code is:

```
proc reg data = wtlift;
   model snatch clean total = wt age;
   mtest total-snatch, total-clean, wt-age;
quit;
```

Note there is one expression from *L* `(wt-age)` and two expressions `(total-snatch, total-clean)` from *M*. This is due to that there is only one row in *L* and 2 columns in *M* matrix. A significant result is detected in this test.

Multivariate Statistics and Exact F Statistics					
S=1 M=0 N=18					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.06181288	288.38	2	38	<.0001
Pillai's Trace	0.93818712	288.38	2	38	<.0001
Hotelling-Lawley Trace	15.17785851	288.38	2	38	<.0001
Roy's Greatest Root	15.17785851	288.38	2	38	<.0001

CONCLUSION

The SAS procedure, PROC REG, accompanying with the *mtest* statement, is the key procedure for multivariate regression related analyses. This paper reviews the concepts of multivariate regression. We use a weight lifting example to demonstrate the use of the *mtest* statement in variety of tests related to multivariate regression models. The matrices *L* and *M* that are used to set up the expressions in the *mtest* statement are discussed. The SAS macro, *multnorm*, can be used to check the assumption of a multivariate normal

distribution. In checking homogeneity of variance-covariance matrix for the residuals, we can apply Box's M test that can be generated from PROC DISCRIM procedure.

REFERENCES

IWF (2015). *IWF Masters Weightlifting Records*, Retrieved in May 2015 from http://www.mastersweightlifting.org/.

Johnson and Wichern (2007). *Applied Multivariate Statistical Analysis(6th ed).*Upper Sallde River, NJ: Pearson Prentice Hall.

Kutner, Nachtsheim, and Neter. 2004. *Applied Linear Regression Models*- **4th** Edition. McGraw-Hill Education.

Lin, J (2015). "*Data Analyses in Multivariate Regression*". SAS Conference Proceedings: Western Users of SAS Software 2015, San Diego, California.

SAS Institute Inc, Macro to Test Multivariate Normality, http://support.sas.com/kb/24/983.html.

SAS Institute Inc. 2013. SAS® STAT® 9.4 User's Guide. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Chii-Dean Joey Lin
Department of Mathematics & Statistics, San Diego State University
San Diego, CA 92182
E-mail: cdlin@sdsu.edu