**STAT 3501/8109 Regression, Time Series, and Forecasting for Business Applications**

# Group Project on Multiple Regression Analysis

Instructor: Dr. Boris Iglewicz

Group 5:

Yin Jiang, Chen Yan Wu, Kelly Yarusso, Daobin Ye,

Kan Zhang, Tianchi Zhang, Yoel Zuman

---

October 22, 2013

Temple University

**Section 1: Introduction of Data Set and Purpose of Project**

The art of predicting a person's weight based on height alone has long been a practice. However, the quality of this model has been found to be very poor. Therefore, a few Professors in *the Journal of Statistics Education* have shown that by using specific body measurements along with age, height and gender, an excellent model to predict weight can be obtained.

The study provided data of 507 physically active (a few hours of exercise a week) individuals; 260 women and 247 men, most being young with an average age of 30 years. The original data contains the individuals' measurements of 12 body girths and 9 skeletal diameters, plus their respective weight, height, age, and gender. In this data set, if we predict weight using only height (Table 1), the coefficient of determination ($R^2$) – which measures the fit quality of the regression line, is only 51.5% – which is very lousy. Therefore, we will start by using all of the above mentioned measurements and then conduct a series of multiple regression analyses that will eventually narrow down the best model predictor of weight to only 11 variables; 5 body girths, 3 skeletal diameters, height, age and gender. Gender is a dummy variable, while the other variables are numerical variables. Moreover, since we will find that two of the body girths, chest and shoulder girth, are highly correlated to each other, we will end up with one model, in which we can essentially interchange those two girths into the one model and observe only minor differences.

**Section 2: Data Analysis**

**Part 1: To Determine the Best Predictive Model for Weight**

Step 1: Regression Analysis of the 1st predictive model with all 24 predictors for weight

First, we used the regression to test the predictive model with all 24 predictors for weight.

$$weight=\beta_0+\beta_1(biacromial)+\beta_2(pelvicbreadth)+\beta_3(bitrochanteric)+\beta_4(chestdepth)+\beta_5(chestdiam)+\beta_6(elbowdiam)+\beta_7(wristdiam)+\beta_8(kneediam)+\beta_9(anklediam)+\beta_{10}(shouldergirth)+\beta_{11}(chestgirth)+\beta_{12}(waistgirth)+\beta_{13}(navelgirth)+\beta_{14}(hipgirth)+\beta_{15}(thighgirth)+\beta_{16}(bicepgirth)+\beta_{17}(forearmgirth)+\beta_{18}(kneegirth)+\beta_{19}(calfgirth)+\beta_{20}(anklegirth)+\beta_{21}(wristgirth)+\beta_{22}(age)+\beta_{23}(height)+\beta_{24}(gender)+\varepsilon_i$$

However, we found the Variance Inflation Factor (VIF) of six variables-shouldergirth, chestgirth, waistgirth, hipgirth, bicepgirth and forearmgirth were larger than the rest, which indicates that multicollinearity exists in the model. However, the multiple coefficient of determination $R^2$ (97.6%) is very large. So the data fit is very good for the predictive model for weight. Subsequently our target is to eliminate multicollinearity.

Step 2: Best Subsets

Using the Best Subsets Regression, we found that given the variable number of 14 (K=15), the $C_p$ is 15.7. As a result, we excluded 10 variables and concluded the first new predictive model for weight as follow:

$$weight=\beta_0+\beta_1(prelvicbreadth)+\beta_2(chestdepth)+\beta_3(kneediam)+\beta_4(shouldergirth)+\beta_5(chestfirth)+\beta_6(waistgrith)+\beta_7(hipgirth)+\beta_8(thighgirht)+\beta_9(forearmgirth)+\beta_{10}(kneegirth)+\beta_{11}(calfgirth)+\beta_{12}(age)+\beta_{13}(height)+\beta_{14}(gender)+\varepsilon_i$$

Step 3: Regression Analysis of the 2nd predictive model with 14 predictors for weight

After finding the best subsets regression, we used Minitab to analyze the second predictive model after 10 predictor variables were excluded. Given this regression equation by Minitab, we still found that the R square is 97.6%, which looks fine, however, the VIF value of chestgirth (13.208) is much higher than the other variables. So there was still a multicollinearity issue existing for this new predictive model. We needed to continue to improve the predictive model.

Step 4: Stepwise

We undertook Stepwise Regression 3 times in Minitab in order to figure out a better predictive model for weight with low VIF values and little change in $R^2$. After the three separating Stepwise tests, we found that the variable thighgirth had the largest insignificant P-value of 0.401, given by the first Stepwise test. We also found that the variable calfgirth had the largest insifnificant P-value of 0.194 given by the second Stepwise test. The third stepwise test displayed no variables with insignificant p-value. As a result, we had our third predictive model as follows after excluding the variable thighgirth and calfgirth from the second predictive model.

$$weight=\beta_0+\beta_1(prelvicbreadth)+\beta_2(chestdepth)+\beta_3(kneediam)+\beta_4(shouldergirth)+\beta_5(chestgirth)+\beta_6(waistgirth)+\beta_7(hipgirth)+\beta_8(forearmgirth)+\beta_9(kneegirth)+\beta_{10}(age)+\beta_{11}(height)+\beta_{12}(gender)+\varepsilon_i$$

Step 5: Regression Analysis of the 3rd predictive model with 12 predictors for weight

We tested the third predictive model with 12 predictors for weight by regression in Minitab. However, the given VIF of shouldergirth is 13.006, which indicated multicollinearity

still existing in this predictive model. Thus, we needed to continue to fix the multicollinearity issue and find a better predictive model.

Step 6: Pearson Correlation Test

Furthermore, in order to understand the multicollinearity from the third predictive model for weight, we used the Pearson Correlation Test in Minitab. The test concluded that the correlation coefficient between shouldergirth and chestgirth is 0.927. Thus, chestgirth and shouldergirth are highly correlated.

**Part 2. Analysis of the Best Regression Equations**

After analyzing the Pearson Correlation test and finding the chest girth and shoulder girth to be highly correlated, we considered leaving out one of the two variables. We used the regression analysis to figure out which one to exclude from our equation.

During our first regression analysis, we left out chest girth and used a dummy variable for gender; 1 represents male and 0 represents female. We found the best regression equation for weight from those 11 selected variables to be:

*weight = - 119 + 0.0998 pelvicbreadth + 0.410 chestdepth + 0.616 kneediam*

*+ 0.171 shouldergirth + 0.404 waistgirth + 0.361 hipgirth*

*+ 0.908 forearmgirth + 0.368 kneegirth - 0.0794 age + 0.278 height*

*- 2.29 gender.*

The R squared is 97.1%, which did not drop significantly from the original model with all the 24 predictors. Since the multiple coefficient of determination (R squared) is close to 1, it presents a very good fit. The Variance Inflation Factors (VIF) values looked reasonable since they are all under 10, this means there is no multicollinearity between the 11 predictors. In addition, we tested the assumption of multiple regression for the selected equation above. The first test of normal distribution for the error, showed by Figure 1 and Figure 2, indicates the equation above is in fact a normal distribution with minimal outliers. The second assumption is to test the independence of the error. We used the Durbin-Watson (DW) Statistic to prove that there is no serial correlation, thus, to prove the independence of the error. Using Minitab we found the DW is equal to 2.00286, which is close to 2 proving that the null hypothesis ($\rho_{\epsilon,\varepsilon\text{-}1}=0$) is reasonable. Therefore, the two assumptions were met.

Using the given Analysis of Variance (ANOVA) Table by Minitab, we used the F test to evaluate the best regression equation for weight from those 11 selected variables. The hypothesis test is:

$$H_0: \beta_1=\beta_2=\beta_3=\beta_4=\beta_5=\beta_6=\beta_7=\beta_8=\beta_9=\beta_{10}=\beta_{11}=0$$

$$H_1: \textit{At least one of the } \beta \textit{ is not equal to zero.}$$

The F value is determined by the mean squared of regression divided by the mean squared of error. Since the computed value of F=1518.51 is greater than 1.8, we reject the null hypothesis and conclude that the regression is significant at a significance level of 5%.

Furthermore, we tested the Graphical Analysis of Residuals (residuals against fitted values) shown by Figure 3. We found that residual model against the fitted values is constant, with one influential point. This indicates the model is reasonable. Given the unusual observation

output there were only 6 observations within the entire model where the X value gives it large leverage.

Moreover, we computed a similar regression analysis for the second model, leaving out the shoulder girth instead of the chest girth, and found very similar results. The difference between the two predictive models for weight is that the VIF of the chest girth was close to 10, however the R squared is 0.1% higher. The two assumption hypotheses (shown by Figure 4 and Figure 5) are met as well. The F test (Table 3) is also significant. The model of residuals against the fitted values (Figure 6) is reasonable. In this analysis, we found only 4 unusual observations whose X variable gives it large leverage.

**Section 3: Conclusion and recommendation**

From our analysis utilizing multiple methods of data processing technique, we have determined that two acceptable models are applicable to the data. We simplified the original 12 body girths and 9 skeletal diameters factors to 8 dimensions, plus their weight, height, age and gender. The first model includes factors including prelvicbreadth, chestdepth, kneediam, *chestgirth*, waistgirth, hipgirth, forearmgirth, kneegirth, age, height, and dummy variable gender. The second model includes factors including prelvicbreadth, kneediam, *shouldergirth*, chestgirth, waistgirth, hipgirth, forearmgirth, kneegirth, age, height, and dummy variable gender. Since there exists multicollinearity between chestgirth and shouldergirth, we have to pick one of them in the prediction model.

Applying regression analysis, best subsets, stepwise, Durbin-Watson (DW) statistic and Pearson correlation test, we obtained the optimal linear regression prediction functions. For

further analysis of the data, we recommend using multivariate multiple linear regression analysis that includes response surface models.

## APPENDIX

### Table 1: Regression (weight vs. height)

weight = - 105 + 1.02 height

| Predictor | Coef | SE Coef | T | P | VIF |
|---|---|---|---|---|---|
| Constant | -105.011 | 7.539 | -13.93 | 0.000 | |
| height | 1.01762 | 0.04399 | 23.13 | 0.000 | 1.000 |

S = 9.30804  R-Sq = 51.5%  R-Sq(adj) = 51.4%

### Table 2: ANOVA (for the 1st best regression equation)

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Regression | 11 | 87529.5 | 7957.2 | 1518.51 | 0.000 |
| Residual Error | 495 | 2593.9 | 5.2 | | |
| Total | 506 | 90123.3 | | | |

### Table 3: ANOVA (for the 2nd best regression equation)

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Regression | 11 | 87561.4 | 7960.1 | 1538.01 | 0.000 |
| Residual Error | 495 | 2561.9 | 5.2 | | |
| Total | 506 | 90123.3 | | | |

### Figure 1: Assumption Test for the 1st best regression equation

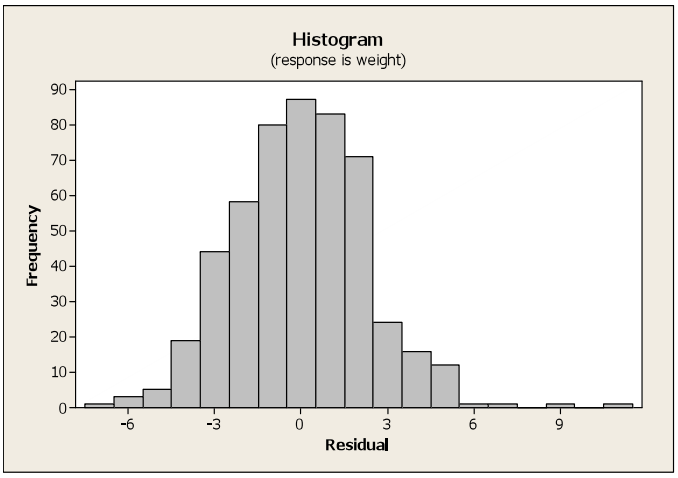

**Figure 2: Assumption Test for the 1st best regression equation**

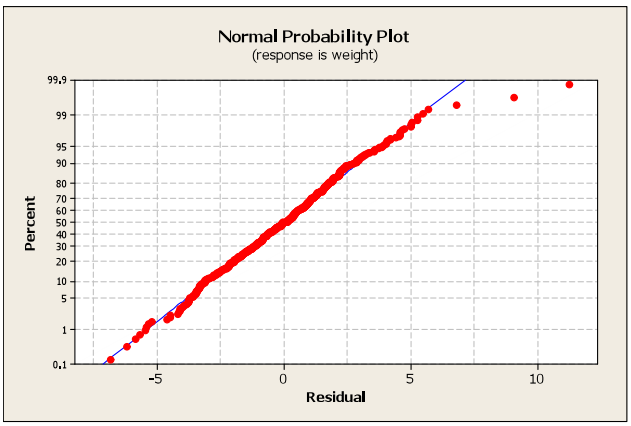


**Figure 3: Graphical Analysis of Residuals for the 1st best regression equation**

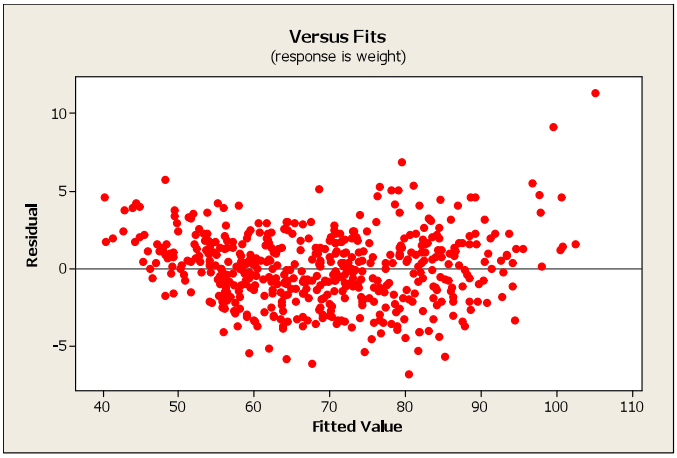


**Figure 4: Assumption Test for the 2nd best regression equation**

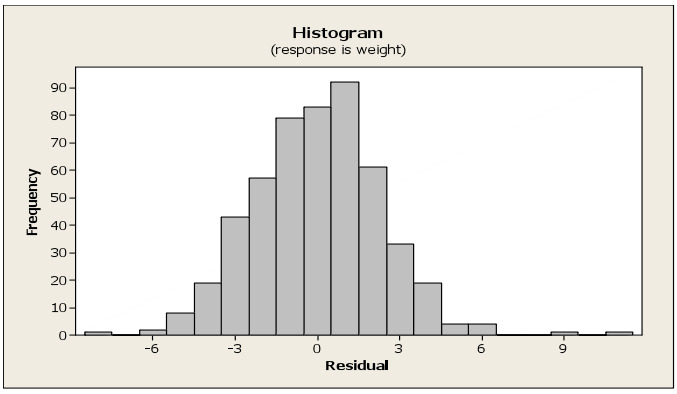

**Figure 5: Assumption Test for the 1st best regression equation**

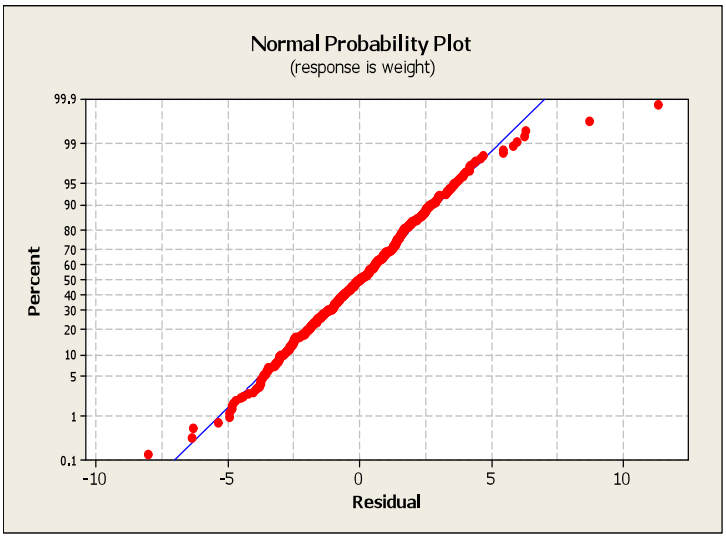


**Figure 3: Graphical Analysis of Residuals for the 2nd best regression equation**