Group #7

Jiaqi Li, Yifei Deng, Changqing Su, Yanxiang Zhu, Sichen Feng

Professor Cremona

November 10th, 2017

**Research Questions and Exploratory Data Analysis**

**Introduction for the Data set:**

* Reference:

<http://ww2.amstat.org/publications/jse/jse_data_archive.htm>

* Description:

This dataset contains 21 body dimension measurements as well as age, weight, height, and gender on 507 individuals. The 247 men and 260 women were primarily individuals in their twenties and thirties, with a scattering of older men and women, all exercising several hours a week.

* Source:

Measurements were initially taken by the first two authors - Grete Heinz and Louis J. Peterson - at San Jose State University and at the U.S. Naval Postgraduate School in Monterey, California. Later, measurements were taken at dozens of California health and fitness clubs by technicians under the supervision of one of these authors.

* Variable Descriptions:

The dataset is about the body dimensions of 507 individual with 247 men and 260 women and contains the following variables:

* biacromialD: biacromial diameter (cm)
* biiliacD: biiliac diameter, or "pelvic breadth" (cm)
* bitrochantericD: bitrochanteric diameter (cm)
* chest\_depth: chest depth between spine and sternum at nipple level, mid-expiration (cm)
* chestD: chest diameter at nipple level, mid-expiration (cm)
* elbowD: elbow diameter, sum of two elbows (cm)
* wristD: wrist diameter, sum of two wrists (cm)
* kneeD: knee diameter, sum of two knees (cm)
* ankleD: ankle diameter, sum of two ankles (cm)
* shoulderG: shoulder girth over deltoid muscles (cm)
* chestG: chest girth, nipple line in males and just above breast tissue in females, mid-expiration (cm)
* waistG: waist girth, narrowest part of torso below the rib cage, average of contracted and relaxed position (cm)
* navelG: navel (or "Abdominal") girth at umbilicus and iliac crest,  iliac crest as a landmark (cm)
* hipG: hip girth at level of bitrochanteric diameter (cm)
* highG: thigh girth below gluteal fold, average of right and left girths (cm)
* bicepG: bicep girth, flexed, average of right and left girths (cm)
* forearmG: forearm girth, extended, palm up, average of right and left girths (cm)
* kneeG: knee girth over patella, slightly flexed position, average of right and left girths (cm)
* calfMaxG: calf maximum girth, average of right and left girths (cm)
* ankleMinG: ankle minimum girth, average of right and left girths (cm)
* wristMinG: wrist minimum girth, average of right and left girths (cm)
* age: age (years)
* weight: weight (kg)
* height: height (cm)
* gender: gender (1 - male, 0 - female)

**Research Questions:**

* Objective:

Setting up models for how one’s weight is related to some factors of one’s body features.

* Questions:
* Do any of the predictors need to be removed from the model?
* Do all assumptions of the error term of our model are satisfied?
* Do there exist any errors or unusual data?
* Does collinearity problem exist in our data?
* Does gender have significant effect on the response weight?

**Exploratory data analysis:**

First, we need to set up the full model including all the predictors from the data set. Then using backward elimination with a specific threshold alpha-to-remove to eliminate predictors that do not have significant impacts on the response and check the correlation between the response and all the predictors with assistance of scatter plots.

When we get a simpler model based on the threshold we set up with compare to the full model, we will use diagnostic graphs to check if all assumptions of error terms (constant variance, linearity, normality) are satisfied. If assumptions are not all satisfied, we need to do several steps to adjust the model. First, we need to check whether outliers exist; if outliers exist, we need to remove the observations where those outliers come from. Second, we need to check whether collinearity problems exist between the predictors in the model; if collinearity problems exist, we need to calculate the variance inflation factor (VIF) and drop the predictor with the highest VIF, and re-compute to see if the collinearity problems still exist. We do this step because we want to produce a useful and simple model. We also want to consider adding interaction terms to the model if we believe that the effect of one predictor in the response varies with the value of the other predictors. Third, we need to check if any transformations or weighted least squares (WLS) are necessary to satisfy linearity and constant variance.

In addition, since the data set contains one categorical predictor (gender), we want to check whether this categorical predictor has significant impact on the response of the model by comparing models with and without the categorical predictor (model selection). If this categorical predictor does not have significant impact, we may want to consider removing this predictor to get a simpler model.

Finally, we also want to check R square and adjusted R square as well as Mallow’s Criterion, Akaike’s Information Criterion (AIC), and Bayesian Information Criterion (BIC). This step (model selection) gives us a better view about which model we want to choose as our “best” model with predictive ability and minimal complexity if we come up with several models containing different predictors.

**Statistical Summary:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min | 1st Qu. | Median | Mean | 3rd Qu. | Max | Sd |
| biacromialD | 32.4 | 36.2 | 38.7 | 38.81 | 41.15 | 47.4 | 3.059132 |
| biiliacD | 18.7 | 26.5 | 28 | 27.83 | 29.25 | 34.7 | 2.206308 |
| bitrochantericD | 24.7 | 30.6 | 32 | 31.98 | 33.35 | 38 | 2.030916 |
| chest\_depth | 14.3 | 17.3 | 19 | 19.23 | 20.9 | 27.5 | 2.515877 |
| chestD | 22.2 | 25.65 | 27.8 | 27.97 | 29.95 | 35.6 | 2.74165 |
| elbowD | 9.9 | 12.4 | 13.3 | 13.39 | 14.4 | 16.7 | 1.352906 |
| wristD | 8.1 | 9.8 | 10.5 | 10.54 | 11.2 | 13.3 | 0.944361 |
| kneeD | 15.7 | 17.9 | 18.7 | 18.81 | 19.6 | 24.3 | 1.347595 |
| ankleD | 9.9 | 13 | 13.8 | 13.86 | 14.8 | 17.2 | 1.247351 |
| shoulderG | 85.9 | 99.45 | 108.2 | 108.2 | 116.55 | 134.8 | 10.37483 |
| chestG | 72.6 | 85.3 | 91.6 | 93.33 | 101.15 | 118.7 | 10.02762 |
| waistG | 57.9 | 68 | 75.8 | 76.98 | 84.5 | 113.2 | 11.01269 |
| navelG | 64 | 78.85 | 84.6 | 85.65 | 91.6 | 121.1 | 9.424128 |
| hipG | 78.8 | 92 | 96 | 96.68 | 101 | 128.3 | 6.680623 |
| highG | 46.3 | 53.7 | 56.3 | 56.86 | 59.5 | 75.7 | 4.459889 |
| bicepG | 22.4 | 27.6 | 31 | 31.17 | 34.45 | 42.4 | 4.246941 |
| forearmG | 19.6 | 23.6 | 25.8 | 25.94 | 28.4 | 32.5 | 2.830579 |
| kneeG | 29 | 34.4 | 36 | 36.2 | 37.95 | 49 | 2.61757 |
| calfMaxG | 28.4 | 34.1 | 36 | 36.08 | 38 | 47.7 | 2.847661 |
| ankleMinG | 16.4 | 21 | 22 | 22.16 | 23.3 | 29.3 | 1.862337 |
| wristMinG | 13 | 15 | 16.1 | 16.1 | 17.1 | 19.6 | 1.380931 |
| age | 18 | 23 | 27 | 30.18 | 36 | 67 | 9.608472 |
| weight | 42 | 58.4 | 68.2 | 69.15 | 78.85 | 116.4 | 13.34576 |
| height | 147.2 | 163.8 | 170.3 | 171.1 | 177.8 | 198.1 | 9.407205 |
| gender | 0 | 0 | 0 | 0.4872 | 1 | 1 | 0.5003293 |