Bodyfat Dataset Analysis

Thomas Grunauer, Jonathan Gorman, Seamus McCrave

December 2020

Introduction

Body fat percentage – like other measures of body composition – seek to assess the health of an individual's body. Despite the importance of such measures, techniques to accomplish this are expensive and time consuming. The proposed data set, acquired from the DASL - Data and Story Library website, aims to find an alternative method of measuring body fat percentage using body dimensions. [5]

Figure 1: Normal plot of percentage body fat response variable

The data set contains observations of 250 male subjects and measurements of the 16 variables outlined in table 1.[2] All except age is a continuous variable, and the division of age into broader categories is explained in the analysis of hypothesis 1. The primary goal of this analysis is to determine whether or not one can build a linear model based solely on easily measurable values. This excludes the body density as a predictive variable because it requires the subject to access a water displacement chamber. Furthermore, as explained in the hypothesis 2 section, body density is related to body fat percentage through Siri's Equation for body fat percentage[6].

Variable Index	Variable Name
1	Age
2	Weight
3	Height
4	Neck
5	Chest
6	Abdomen
7	Hip
8	Waist
9	Thigh
10	Knee
11	Ankle
12	Bicep
13	Forearm
14	Wrist
15	Density
16	% Body Fat

Table 1: Variables measured for the subjects. Each variable is a measurement of the specified body dimension, except the age variable which is simply the participant's age.

Each body dimension variable is measured as a circumference in centimeters except weight and height. To check that the percentage body fat response variable is normally distributed, one can use a normal plot and confirm the results with the Shapiro-Wilk test for normality.

In reference to figure 1, a majority of the data follows a normal trend, with some deviations at the tails of the distribution. These deviations are deemed unimportant by the Shapiro-Wilk test, which outputs a p-value of 0.1697. The Shapiro-Wilk test's null hypothesis is that the data is normally distributed, thus, we cannot reject the null hypothesis that the data is normal even at a significance level of $\alpha = 0.1697$. The results of these tests imply that a transformation is not required to normalize the data for the response variable.

Hypothesis 1

Upon visual inspection, there was a noticeable difference between the body fat percentages of individuals under the age of 60 against those who were 60 years old or older. Originally there were 51 distinct categories of age in the data set. To consolidate the number of categories, the data was broken down into two subsets. Setting age to be either less than 60 or greater than/equal to 60, age is able to be considered a categorical

variable which allows age to be used as a parameter for analysis. For the purpose of simplicity, the latter group of individuals 60 years old or older will be referred to as senior citizens [3]. A side-by-side box plot is displayed below to aid as a visual in helping understand the distribution of each of the groups.

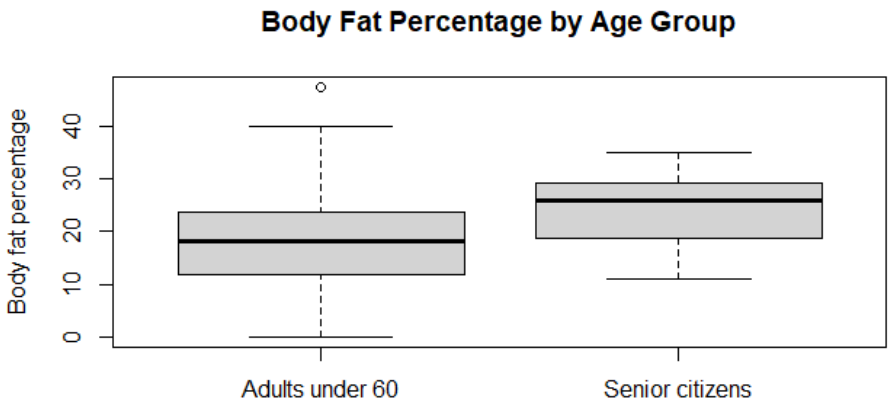


Figure 2: side-by-side box plot of body fat percentage by age group

The above figure tells us the sample mean of the senior citizen group is larger than the sample median for the younger adults. We can also see that the whiskers of the senior citizen's box plot are much smaller, suggesting a smaller sample variance than the younger adults. This plot gave provided a good idea of the underlying distributions, but a formal hypothesis test was required to make a definite decision.

The younger group had a sample size of $n_1 = 214$ compare to the sample size of the older group of $n_2 = 36$. Sample mean for the body fat percentage of the younger crowd was observed to be $\mu_{younger} = 18.13598$ compare to the sample mean of the seniors whose observed sample mean was measured to be $\mu_{older} = 24.36$. The younger sample standard deviation was measured to be $s_{younger} = 8.25$ compared to $s_{older} = 6.40$ of the seniors. To formally test if the population means of the body fat percentage of the two groups were equivalent, or if they showed a significant difference from one another, the following null and alternate hypotheses are observed:

$$H_0 : \mu_{younger} = \mu_{older}, H_\alpha : \mu_{younger} \neq \mu_{older}.$$

Given that both sample sizes are above 30, the central limit theorem could be applied to assume that both populations are normally distributed. This meant a Z-test could be performed even though the variance of the two samples is currently unknown. The z-test statistic for such an inference is as follows:

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_{younger} - \mu_{older})}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}. \tag{1}$$

In the z-test statistic it is noted that $(\mu_{younger} - \mu_{older}) = 0$ under the observed null hypothesis. Plugging in the values above results in a z score of $Z = -5.1563$. This is a very low Z-value which tells us, based on the knowledge of the area under the standard normal curve, that the associated p-value will be very close to zero; this value can be approximated as $p = 2.518 * 10^{-7}$. Given that the p-value is less than $\alpha = 0.05$, null hypothesis can be rejected and the conclusion can be made that there is a significant difference in the mean body fat percentage of adults under 60 and senior citizens.

Now that we know there is a significant difference between $\mu_{younger}$ and μ_{older}, we can test to see which mean is larger. We initially suspected that $\mu_{younger} < \mu_{older}$ which would correspond to the hypotheses:

$$H_0 : \mu_{younger} = \mu_{older}, H_\alpha : \mu_{younger} < \mu_{older}$$

The test statistic for this test would be the same as for the previous test. Hence, $Z = -5.1563$ with a corresponding p-value of $p = 1.259 * 10^{-7}$ which is half of the p-value of our first test. Since $1.259 * 10^{-7} < 0.05$ we can reject the null hypothesis and conclude that the mean body fat percentage of seniors is higher than younger adults.

Hypothesis 2

The body fat data set was originally collected with the purpose of devising a better way to evaluate body fat percentage. Due to practical difficulties of existing methods, it was hypothesized that the variables measured in the data set could predict body fat, providing a more cost and time efficient alternative. To build an *inferential* model and test whether or not this hypothesis holds, one can construct a multiple linear regression model, which accounts for multiple predictor variables influencing a response variable. The process of building such a model involves determining which predictor variables are statistically significant, and refining the model according to the assumptions of linear regression. The exact hypotheses take the form:

$$H_0 : \beta_i = 0, H_A : \beta_i \neq 0 \tag{2}$$

Where, in reference to table 1, i ranges from 0 to 14, where β_0 is the intercept. The response variable is body fat percentage. The original data set contained a body density variable, however, this will not be used as a predictor variable. In fact, percentage body and body density are related through Siri's equation [6]:

$$\%BF = (\frac{4.95}{Density} - 4.5) \times 100 \tag{3}$$

If the predictor variable has a statistically significant slope, then it should be included in the linear model, mathematically represented as:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4}$$

Where $\mathbf{Y}$ is the response variable vector, $\mathbf{X}$ is the predictor variable matrix, β is the slope vector, and ϵ is the error. Inference on each β_i is performed under a t-distribution with a t statistic:

$$|t_i| = \frac{|\hat{\beta}_i|}{SE(\hat{\beta}_i)} > t_{n-(k+1),\alpha/2} \tag{5}$$

Where $\hat{\beta}_i$ is the estimated slope of the ith variable, $SE(\hat{\beta}_i)$ is the standard error of each variable's slope, n is the number of observations, k is the number of variables, and α is the significance level. The inequality indicates the case of rejection, or that the variable has a statistically significant slope.

The F-test for the overall regression model assesses whether or not at least one predictor variable is related to the model. This test has hypotheses of the form:

$$H_0 : \beta_0 = ... = \beta_{14} = 0, H_A : \beta_i \neq 0 \quad \text{for some i} \tag{6}$$

Thus, if the multiple linear regression model has at least one accurate predictor variable, the F-test will reject H_0. Such as scenario is given by the following criterion:

$$F = \frac{MSR}{MSE} > f_{k,n-(k+1),\alpha} \tag{7}$$

Where MSR is the mean square regression and is defined as $MSR = \frac{SSR}{1}$, and MSE is the mean square error, defined by $MSE = \frac{SSE}{n-2}$. The last important parameter is the coefficient of multiple determination, which represents the percentage of variation in the response variable accounted for by regression on the predictor variables:

$$r^2 = 1 - \frac{SSE}{SST} \tag{8}$$

Where SSE and SST, as referenced in the previous equation for the F-statistic, represent the error sum of squares and total sum of squares respectively. In simple linear regression, the square root of the coefficient of determination yields the correlation coefficient, which measures the strength of the relation between the response variable and predictor variable. However, when many predictor variables are considered, the correlation coefficient becomes the *multiple correlation coefficient*, and measures the same strength but between the response variable and the *set* of predictor variables.

Now that the essential parameters and test statistics are briefed, the normality assumption must be checked to proceed with multiple linear regression. This is accomplished in the introduction.

The linear model command was used in R to output a multiple regression model. This command automatically performs inference on the β_i values, and assigns a coefficient of multiple determination along with the overall F-statistic. Initially, the following model was used:

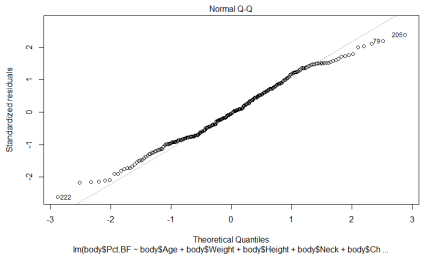
$$y_{BF\%} = \beta_0 + \beta_1 x_1 + ... + \beta_{14} x_{14} \tag{9}$$

Where, referring to the variable table, the integer subscript denotes to the corresponding variable. The regression output indicates which variables have a statistically significant slope, and is pictured in figure 3.

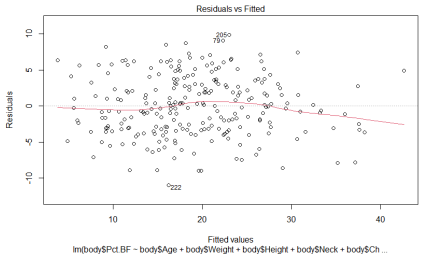
```
Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.68516   23.37412   0.072 0.942587
body$Age      0.07189    0.03217   2.234 0.026389 *
body$Height  -0.24675    0.19114  -1.291 0.197989
body$Neck    -0.38682    0.23486  -1.647 0.100887
body$Abdomen  0.90452    0.09140   9.897  < 2e-16 ***
body$Waist         NA         NA      NA       NA
body$Hip     -0.15878    0.14586  -1.089 0.277446
body$Chest   -0.11919    0.10825  -1.101 0.272004
body$Weight  -0.01762    0.06714  -0.263 0.793153
body$Thigh    0.17299    0.14683   1.178 0.239926
body$Knee    -0.04580    0.24560  -0.186 0.852230
body$Ankle    0.18502    0.21985   0.842 0.400862
body$Bicep    0.17968    0.17039   1.054 0.292732
body$Forearm  0.27605    0.20692   1.334 0.183454
body$Wrist   -1.80162    0.53304  -3.380 0.000848 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.255 on 236 degrees of freedom
Multiple R-squared:  0.7505,    Adjusted R-squared:  0.7368
F-statistic: 54.61 on 13 and 236 DF,  p-value: < 2.2e-16
```

Figure 3: Initial model's output in R Studio.

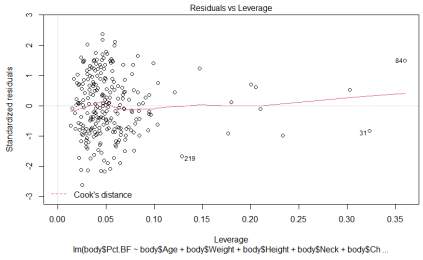
We can check the assumptions of multiple linear regression with several plots, as seen in figure 4. The normal plot for residuals shows that the residuals are normally distributed. The residuals vs. fitted values plot checks for constant variance, and spread should be random around 0. The residuals vs. leverage plot should be random around



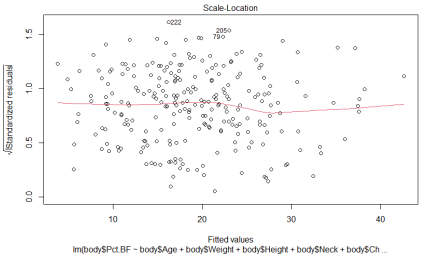
(a) Normal plot for residuals



(b) Residuals vs. fitted values



(c) Residuals vs. leverage



(d) Standardized residuals vs fitted values (scale location)

Figure 4: Plots to check assumptions for initial model

0 and leverage should be low, where leverage is how the fitted response behaves when the actual response changes. The standardized residuals vs fitted values plot is a standardized version of the residual plot. The output in figure 3 shows that approximately 75% of the variation in y is accounted for by the multiple regression model. Additionally, the p-value related to the F-statistic implies that there is a statistically significant linear relation with at least on the variables in the model. However, only three of of the fourteen variables are statistically significant at varying significance. This suggests that the model can be refined to be simpler, while retaining accuracy in the form of r^2. Typically, more variables will increase the multiple coefficient of determination, but convolutes the model and possibly introduce *multicollinearity*, which invalidates the model results at high levels. Multicollinearity is a linear relation among several of the predictor variables. In other words, the correlation coefficient is non-trivial among mixed pairs of the predictor variables. This raises issues of linear dependence among the columns of $\mathbf{X}$ in equation (3), which must be invertible to perform inference on β. One technique of assessing this issue is calculating *variance inflation factors*, which are the diagonal elements of the inverted correlation matrix $\mathbf{R}$. Variance inflation factors greater than 10 usually give rise to problems in the model [5].

It should be noted that for all additional analysis, the waist variable was removed from the model. This is due to its strong linear dependence with at least one the other predictor variables, leading to errors during regression. This behavior is confirmed with the alias function in R.

After removing the waist variable, variance inflation factors were computed for each predictor variable. The variables with VIFs greater than 10 are: Weight, Chest, Hip, and Abdomen. Unfortunately, Abdomen is the strongest predictor of percentage body fat according to the initial model output, so it would be unfavorable to remove this variable form the model. Perhaps, at this point, it would be best to remove all statistically insignificant variables, and address multicollinearity afterward. It is possible that abdomen is linearly dependent with one of the insignificant variables, so removing them would cause no issues with the model's accuracy and improve abdomen's VIF.

The stepwise regression variable choosing algorithm was used to select statistically significant variables and refine the model. This algorithm runs partial F-tests on successive subsets of the variables, until the whole list of variables has been iterated through. It eliminates variables if their partial F-statistics do not exceed some criterion [5]. The result is a refined model, which will have a similar coefficient of determination but fewer variables, hopefully mitigating multicollinearity and simplifying the practical components of using the model. In a clinical setting, fewer body dimensions would be required to get the same prediction of percentage body fat. The results of stepwise linear regression can be seen in figure 5.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.04038    8.35881   0.603 0.547074
body$Age      0.07258    0.03030   2.396 0.017361 *
body$Height  -0.26807    0.12612  -2.125 0.034567 *
body$Neck    -0.45133    0.21774  -2.073 0.039252 *
body$Abdomen  0.82271    0.06880  11.958  < 2e-16 ***
body$Hip     -0.19488    0.12984  -1.501 0.134689
body$Thigh    0.22387    0.12900   1.735 0.083943 .
body$Forearm  0.29550    0.19166   1.542 0.124440
body$Wrist   -1.73072    0.49360  -3.506 0.000542 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.241 on 241 degrees of freedom
Multiple R-squared:  0.7469,    Adjusted R-squared:  0.7385
F-statistic:  88.9 on 8 and 241 DF,  p-value: < 2.2e-16
```

Figure 5: Stepwise linear regression algorithm applied to the original model.

This algorithm was applied using the stepAIC function in R. Backward stepwise regression was used, which eliminates variables from the original set. This function is slightly different than the one explained above, in that it eliminates variables based on the Aikake Information Criterion as opposed to the partial F-statistic. As displayed, the algorithm eliminated the chest, weight, and bicep variables. The waist variable was not included in the algorithm input.

The output shows that approximately 74.7% of the variation in y is accounted for by the model. This is a drop of merely 0.4% when compared to the original model. By reducing the number of variables, not only is the model simplified, but multicollinearity risk has been reduced. Re-running VIF tests yields only one value close to 10, that of the Hip variable. Additionally, if a significance level of 0.05 is used, the forearm and

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.90033    8.08402   0.359   0.7201
body$Age      0.05602    0.02382   2.351   0.0195 *
body$Height  -0.32299    0.12155  -2.657   0.0084 **
body$Abdomen  0.77097    0.03362  22.932  < 2e-16 ***
body$Wrist   -1.91138    0.40953  -4.667 5.03e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.277 on 245 degrees of freedom
Multiple R-squared:  0.7383,    Adjusted R-squared:  0.7341
F-statistic: 172.8 on 4 and 245 DF,  p-value: < 2.2e-16
```

Figure 6: Final multiple linear regression model for percentage body fat response variable.

thigh variables can be eliminated as well. Removing these outputs a slightly lower r^2 of 0.7404. This model now indicates that the neck variable is insignificant, so it is eliminated too. The age, height, abdomen, and wrist variables remain, with a coefficient of multiple determination r^2 of 0.7383.

This result, in a clinical setting, allows for the measurement of only three body dimensions with a reduction of only 1.22% in the amount of variation in y accounted for by the model. The results can be seen in figure 6.

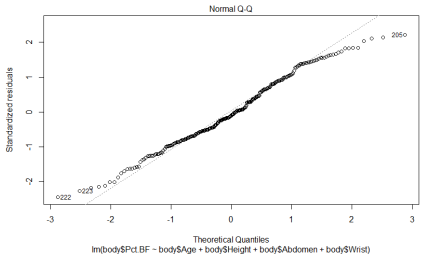
All variables have statistically significant slope estimations, and the VIF for each variable is below 2, which suggests minimal multicollinearity. The assumptions for this model can be checked with a sequence of plots as seen in figure 7. It should be noted that the goal of this particular regression was to determine if bodily dimensions and age could solely predict a man's percentage body fat.

Random Removal of Data

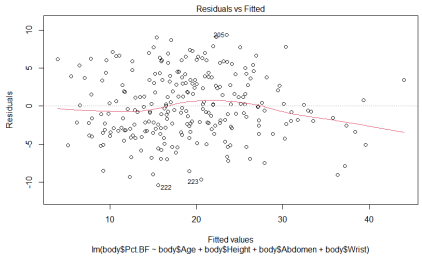
Random removal of data from the data set simulates having a smaller data set. Theoretically, random omission of data will not have a systematic pattern, and will not favor any phenomena or trends in the data. However, fewer observations will affect the statistical outcome of the experiment, and coincidentally, it may remove some outliers from the set, having a significant impact on results.

The data was reduced by 20%. This was done by generating an array of 50 random numbers and subtracting that array from the data table stored in R. This removes the corresponding rows, and changes the total data set to 200 observations as opposed to 250.

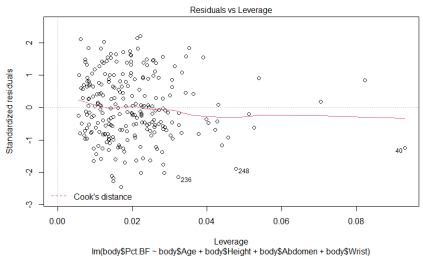
After performing a regression analysis identical to the full data set, the results in the corresponding figure were obtained.



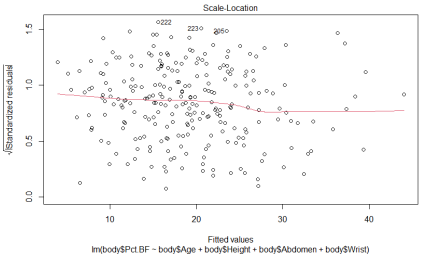
(a) Normal plot for residuals



(b) Residuals vs. fitted values



(c) Residuals vs. leverage



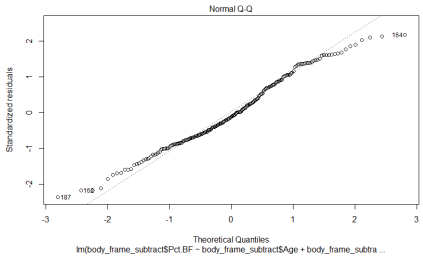
(d) Standardized residuals vs fitted values (scale location)

Figure 7: Plots to check assumptions for refined model

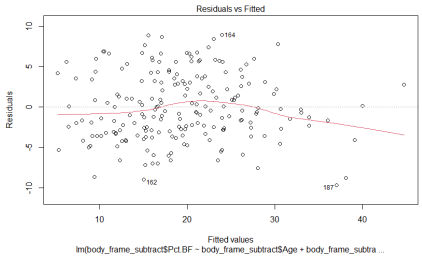
```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                5.18554    9.11583   0.569  0.57011
body_frame_subtract$Age    0.07326    0.02638   2.777  0.00603 **
body_frame_subtract$Height -0.36050    0.13586  -2.654  0.00862 **
body_frame_subtract$Abdomen  0.77595    0.03684  21.063  < 2e-16 ***
body_frame_subtract$Wrist  -1.94863    0.45126  -4.318  2.5e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.178 on 195 degrees of freedom
Multiple R-squared:  0.7533,    Adjusted R-squared:  0.7483
F-statistic: 148.9 on 4 and 195 DF,  p-value: < 2.2e-16
```

Figure 8: Multiple linear regression model on a randomly omitted data set.

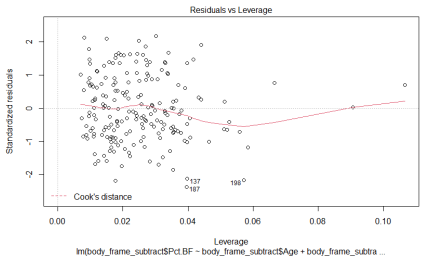
Figure 9 depicts the assumption plots. Each factor has a statistically significant slope at a 0.05 significance level. Additionally, the VIF for each variable is below two. Even after removing the same amount of variables, the coefficient of multiple determination in this case is $r^2 = 0.7533$, approximately 0.015 higher than the original model's r^2 value. The plots for this results can be seen in the corresponding figure. This increase in the coefficient of multiple determination can be explained by the possible removal of outliers. Equation (8) shows that decreasing SSE, the error sum of squares, will increase r^2. SSE measures the overall goodness of fit, or how much the response values differ from the model. So if outliers from the model were randomly removed, this can increase the accuracy of the model[5].



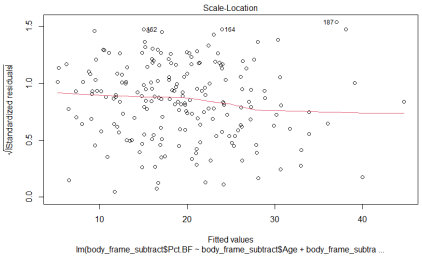
(a) Normal plot for residuals



(b) Residuals vs. fitted values



(c) Residuals vs. leverage



(d) Standardized residuals vs fitted values (scale location)

Figure 9: Plots to check assumptions for randomly removed data model

Non-Ignorable Missing Data

Dealing with data that is missing for a specific reason requires more care than if the data was missing completely at random. If the data was missing because someone just forgot to put a few values into the dataset, we could simply delete this observations from the data without worrying about bias. Our option for dealing with non-ignorable missing data would be to impute the missing values using any of a number of techniques.

Suppose younger men with a large weight and high body fat percentage were too self conscious to complete some of the body measurements. Deleting these observations from the data would make it seem as though seniors had a higher body fat percentage and skew the data. Our other option would be to use imputation to replace these values. Some less rigorous methods of imputation include hot-deck and cold-deck. For the cold-deck method, we would just look at similar datasets, take the necessary data where needed, and plug it into our original dataset where the values are missing. The hot-deck method would include ordering our data in some way, possibly from smallest value to largest, and then filling in the missing value with the previous index's data. More concretely, if our jth variable had a missing value in index ith observation, we would fill it with the data from the $(i-1)th$ observation. Neither of these methods are very rigorous and introduce a large amount of bias. There are other methods which yield much better results. [4]

We can also use linear regression to predict these missing values. There are two methods, deterministic and stochastic regression imputation. The problem with deterministic regression is that the replacement values fit the data too well, the value would sit directly on the hyperplane. The real value would have some variation. Stochastic regression imputation overcomes this problem by adding a normally distributed random error. This helps to further reduce the bias.[1]

There are many methods for imputation but stochastic imputation would have been our go-to method.

References

[1] Columbia. *Missing-data imputation*. URL: http://www.stat.columbia.edu/~gelman/arm/missing.pdf. accessed: 12.1.2020.

[2] Roger W. Johnson. *Body measurements to predict percentage of body fat in males*. URL: http://math.uprag.edu/regresion/fat.html. accessed: 12.1.2020.

[3] Nithyashri. *Classification of human age based on Neural Network using FG-NET Aging database and Wavelets*. URL: https://ieeexplore.ieee.org/document/6416855. accessed: 12.1.2020.

[4] Gabriella Schoier. *On partial nonresponse situations:the hot deck imputation method*. URL: http://www.stat.fi/isi99/proceedings/arkisto/varasto/scho0502.pdf. accessed: 12.1.2020.

[5] Tamhane. *Statistics and Data Analysis from Elementary to Intermediate*. Prentice Hall, 2000. ISBN: 9780137444267.

[6] BabetteS. Zemel. *Classification of human age based on Neural Network using FG-NET Aging database and Wavelets*. URL: https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/body-density. accessed: 12.1.2020.