Summary of Body Fat Study - Group 9

Nowadays fighting obesity and weight problems remains a serious health concern. We hope through this study, people can have clearer understandings of body fat and lead healthier lives.

Background Knowledge:

Based on existing research evidence, we know that men's bodies are more apt to store excess fat in the upper body, especially in the abdominal region[1]. While broad shoulders, knees, and wrists were significantly associated with high fat-free mass, e.g. bone mineral density[2]. So a high-level idea is: the factors related to body frame size should be inversely proportional to body fat and the factors reflecting fat accumulation should be in direct proportion.

Data Cleaning:

~Recalculate Bodyfat: According to the background information, we know that the BODYFAT is calculated from DENSITY. So to begin with, we build a simple linear model of BODYFAT ~ DENSITY and replaced the biased data with fitted values. We also spot a sample whose BODYFAT is negative after recalculation. On further examination, the sample is seen to be more than an observational error but an outlier thus can be removed.

~Analyze Outliers: To detect and analyze other possible outliers, we started by fitting a full model. By looking at the influence plot, we screen out a few samples with large student residual, cook's distance or hat-values, and check whether the unusual combination of variables makes sense. Once the value goes against with the background knowledge or common sense, we will adopt these methods accordingly: **Firstly,** if related variables exist, recalculate the value. *(Recalculate HEIGHT(29.9) with bmi and WEIGHT for No.42.)* **Secondly,** if we can find a group of similar people, replace the value with group mean. *(Replace No.96's BODYFAT(1.42) with the mean of a group of people whose ABDOMEN in (98,100), ADIPOSITY in (25,27), and AGE in (40,70))* **Finally,** if it is an influential point and outlier, delete the sample. *(Delete No.39 since this person is unusually strong and probably muscly with a very low body fat percentage.)*

Rule of Choosing Model:

Our basic idea is to use stepwise regression to eliminate the variables with low contribution or high multicollinearity. We first introduce all the variables into the model. Then, an F-test is conducted to check whether all coefficients of the variables are zero. At the same time, T-test is conducted for the variables previously selected in the model. Our goal is to ensure that every time introducing a new variable step by step, all the coefficients still remain significant. Stepwise regression can ensure the set of explanatory variables in the final regression model is optimal.

Model Selection:

Firstly, Density was excluded because it is used for the outliers of BodyFat. Since adiposity is calculated from Weight and Height, we deleted these two variables to avoid multicollinearity. Secondly, in backward regression, the coefficients of Age, Adiposity, Neck, Chest, Abdomen, Hip, Wrist, these seven variables are significant. The Adjusted R Square and MSE of the model with seven variables was similar with that of the full models with all variables. We decided to choose the model with fewer parameters to reduce the complexity of the models. Since the significance level of Age and Neck was found to be weak, the T-test was carried out for forward regression methods respectively. ANOVA Test was conducted between the model with and without Age. The null hypothesis that adding Age variable would increase the fitting effect of the model was rejected. Therefore, Age variable was excluded to reduce the model complexity. Similarly, the Neck variable is eliminated. Since the Adjusted R Square of the reduced model without Age and Neck is quite similar with that full model, after trading off model complexity and accuracy, we finally choose Adiposity, Chest, Abdomen, Hip, Wrist these five variables.

Model Diagnostics

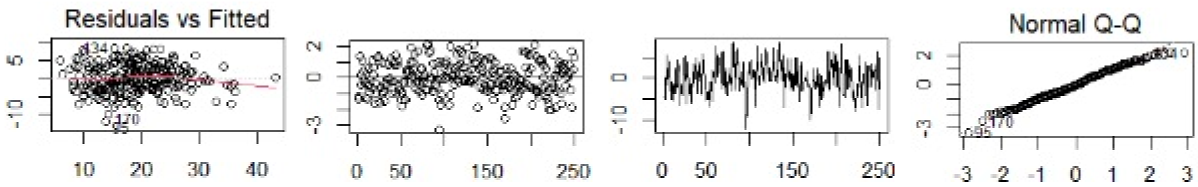

Figure 1: Linearity, Homoscedasticity, Independence, and Normality

After reaching our final model, which uses the Bodyfat as response variable and uses Adiposity, Chest, Abdomen, Hip, and Wrist as predictors, we check the four assumptions associated with a linear regression model: Linearity, Homoscedasticity, Independence, and Normality.

The linearity assumption can be checked by inspecting the Residuals vs Fitted plot (**Figure 1** – First). In our example, there is no pattern in the residual plot and the red line is approximately horizontal at zero. This suggests that we can assume a linear relationship between the predictors and the outcome variables. What is important to know is that the red line is the Locally Estimated Scatterplot Smoothing (LOESS) regression line, which creates a smooth line through scatter plot to help to see relationships between variables and foresee trends..

For Homoscedasticity of variance, this assumption can be checked by plotting the standardized residuals. (**Figure 1** – Second). This plot shows the residuals are spread equally. Hence, there is no violation of the Homoscedasticity assumption.

Next, we check Independence by making a line plot of residuals (**Figure 1** – Third). The plot shows no obvious trends, so the independence assumption is satisfied.

Then, we check the Normality with a qq-plot of the standardized residuals (**Figure 1** – Fourth). The points on the qq-plot fall roughly on a straight line, so the normality assumption is satisfied.

Model Strengths/Weaknesses

There are some strengths of our model. First, our model satisfies the four linear regression assumptions, Linearity, Homoscedasticity, Independence, and Normality, making our results more credible. Second, we have a moderate adjusted R square at 0.722, which indicates that 72.2% of the variation in the output variables are explained by the input variables.

There are also some weaknesses of our model. In particular, there still might be some correlations between our predictors. Besides, the amount of training data is quite small.

Conclusion/Discussion:

By using multiple linear regression combined with VIF, F-Test, T-Test, and ANOVA Test, we got our final model as:

$$\begin{aligned} Bodyfat = 13.70 + 0.73 \times Adiposity - 0.29 \times Chest + 0.84 \times Abdomen \\ - 0.30 \times Hip - 1.78 \times Wrist\,. \end{aligned}$$

After checking the linear assumptions, all the four assumptions are satisfied. Therefore, we can interpret that the Bodyfat can be influenced by Adiposity, Chest, Abdomen, Hip, and Wrist. To explain our final model, taking Adiposity as an example, considering other variables as fixed, as Adiposity increases one unit, the Bodyfat increases 0.73 unit. The coefficients of Adiposity and Abdomen are positive because it is commonly acknowledged that a person with larger adiposity and abdominal circumference has a higher body fat percentage. The coefficients of Chest and Hip are negative because all the observations in our dataset are from men. For most men, their chest and hip have more muscle or lean mass, which means they have lower body fat percentage. The coefficient of wrist is negative because most human's wrist mostly consists of bone or skeleton and hence has little fat.

Contributions:

JiachengYu wrote the rule of choosing model and model selection for report and model selection for presentation.

Kaiyan Ma reviewed relative articles, conducted data cleaning and wroted corresponding parts in the reports.

Xiyu Hao figured the plots of linear regression assumptions, and wrote the model diagnostics, model strength/weakness, model explanation, and presentation slides: model diagnostics and model strength/weakness.

Jiahao (James) Wan edited the conclusion/discussion part of the summary, and worked on the part of the slides. JW also developed, maintained, and deployed the shiny app codes. Ultimately, JW is responsible for the whole web app. Overall, we met four times, spent about ten hours.

References:

[1]Erchonia. (2022, September 24). *How men and women store fat differently*. Erchonia. Retrieved October 16, 2022, from https://www.erchonia.com/how-men-and-women-store-fat-differently/#:~:text=Men's%20bodies%20are%20more%20apt,fewer%20stored%20in%20fat%20reserves.

[2]William Cameron Chumlea, Wayne Wisemandle, Shumei S Guo, Roger M Siervogel, Relations between frame size and body composition and bone mineral status, The American Journal of Clinical Nutrition, Volume 75, Issue 6, June 2002, Pages 1012–1016, https://doi.org/10.1093/ajcn/75.6.1012