**1. Introduction**

In the quest for improved health management, accurately predicting body fat percentage has become crucial. This project aims to develop a simple, accurate and robust model with clinical indicators, and provide users with a friendly interface that enables people to track their body fat index.

**2 Data Processing**

We first check the summary statistics for variables. The target variable, body fat, follows a normal distribution ; its extreme values are 0% and 45.1%, both of which are abnormal in reality. For these two outliers, we replace them with the average body fat within ±5 years of age of the corresponding samples. Similarly, we use z-score with a threshold of 3 to detect outliers for other variables and impute them in the same way. For dataset segmentation, we split the population at an 8:2 ratio and construct the test set by randomly selecting an equal number of samples from three age groups: [20-40), [40-60) and [60, 80]; then the remaining data automatically becomes the training set.

**3 Final Model Statement**

**3.1 Usage Example**

A man aged 40 with abdomen, chest, hip and thigh circumferences(cm) of 86.6, 97, 92.6 and 55.9 respectively is expected to have a body fat of 16.06% based on our model; His 95% prediction interval is 11.71% and 20.41%.

**3.2 Interpretation**

Estimated coefficients are shown as () and in the units of centimeter for all circumferences and year for age. This means that for each year increase in age while all the other measurements remaining constant, the model predicts that body fat will increase, on average, by 0.0345%.

**4 Relevant Statistical Analysis**

**4.1 Model Selection**

**4.1.1 Selection Criteria**

In order to obtain the model that best meets the goal, we develop the selection criteria in terms of accuracy, simplicity and robustness.

To evaluate model accuracy, we consider the root mean square error(RMSE) and the coefficient of determination() on the training set. A lower RMSE indicates that the model fits training data well; An closer to 1 indicates a better degree of fit. For model simplicity, we count how many predictors are included, and evaluate whether the rationale is easy to interpret. For robustness, we consider RMSE and  on the test set. Since the test set is unseen during the training, a good performance suggests that the model is able to produce reliable predictions for new inputs, and thus is generalizable and adaptive.

**4.1.2 Candidate Models**

We have proposed four candidate models in total:

1. Baseline. A basic linear regression model involving only height and weight.
2. Ridge regression. A linear regression model with ridge penalty and all the 14 predictors1.
3. Lasso regression with 5-fold cross validation. We first input all the predictors. Among features with largest absolute coefficients, we picked five the most common ones to fit again from scratch. The final model involves age, chest, abdomen, hip and thigh circumstances.
4. An SVM regression model involving the same five predictors as candidate model (3).

**4.1.3 Comparison**

From Table 1, although (1) is the simplest, it performs the worst. We do not want to sacrifice accuracy too much for simplicity. (2), (3) and (4) result in similar accuracy and robustness. Despite a slightly better accuracy, (2) includes 14 predictors, which is too many to be further generalized. (4) applies a complex methodology which is not friendly for comprehension; it performs worse than (3). Therefore, we choose (3) as our final model. (3) contains five common predictors and achieves fairly good results; its rationale is also intuitive. It should be noted that we made this decision only based on our criteria and did not consider the process of model construction.

Since coefficients returned by regression are biased, we can not conduct standard significance

testing. But cross-validation results show that (3) performs similarly on training and test sets, which confirms its robustness and implies that it can explain about 63% of the body fat variance.

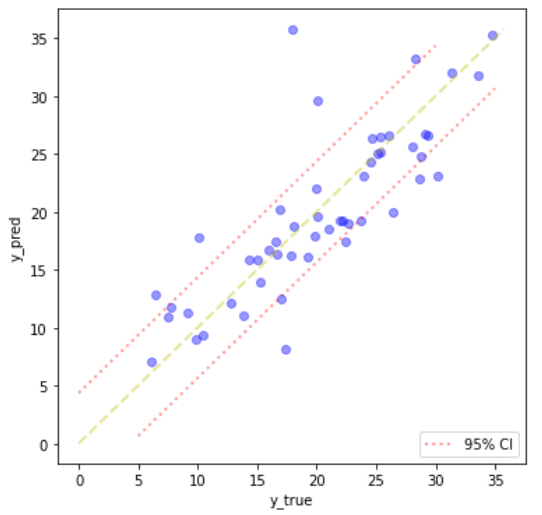
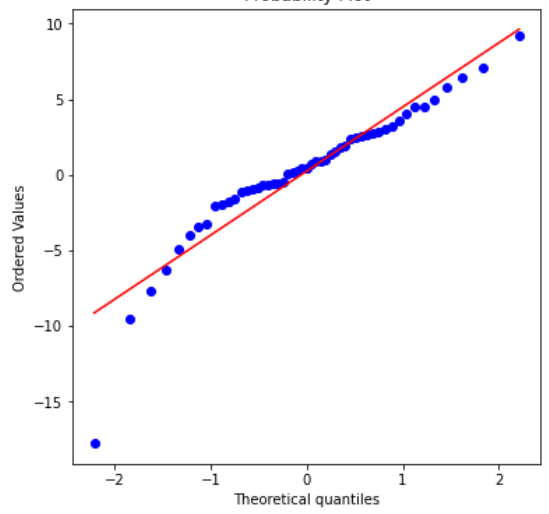
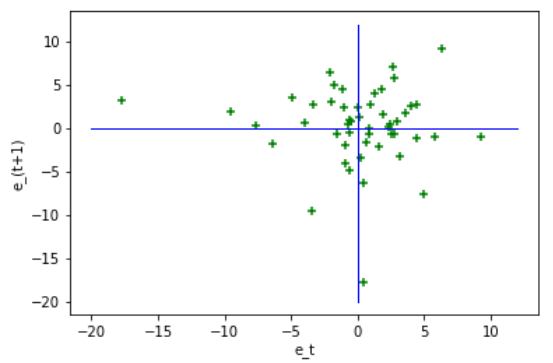
Table 1.Comparison among Candidate Models

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Criteria  Candidate Model | Accuracy(Training set) | | Simplicity | | Robustness(Test set) | |
| RMSE |  | #Parameters | Methodology | RMSE |  |
| 1. Baseline: Linear regression | 5.4799 | 0.4445 | 2 | Linear regression | 5.0462 | 0.5154 |
| 1. Ridge regression | 4.2354 | 0.6682 | 14 | Linear regression | 4.3720 | 0.6363 |
| 1. LassoCV regression | 4.4722 | 0.6300 | 5 | Linear regression | 4.3568 | 0.6388 |
| 1. SVM regression | 4.5940 | 0.6096 | 5 | SVM regression | 4.4429 | 0.6244 |

**4.2 Model Diagnostics**

**4.2.1 Residual Analysis**

Results are shown in Figure 2. There is no obvious distribution in Figure 2(a), meaning that the residuals are irrelevant. Except for a few outliers, most residuals are normally distributed(Figure 2(b)); And most true values are located in the 95% confidence interval of predicted ones.



(a) (b) (c)

Figure 2. Residual Analysis.

(a): Residual Plot. (b). Q-Q Plot of the Residuals. (c) Scatter Plot of predicted and true values on the test set.

**4.2.1 Outlier Analysis**

We check the outlier with the largest residual value(Table 2). It turns out that the original body fat is a outlier and the imputed one used as true value is very far from the original. However, the prediction provided by our model falls between the two, balancing this difference to some extent. This shows that our model is able to adjust for aberrant values and again confirms its robustness.

Table 2.Outlier Analysis for the Largest Residual

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Body Fat - Original | Body Fat - imputed | Body Fat - predicted | Age | Chest | Abdomen | Hip | Thigh |
| 45.1 | 17.98 | 35.74 | 51 | 119.8 | 122.1 | 112.8 | 62.5 |

**5 Discussion**

**5.1 Model Strengths and Weaknesses**

Generally, our final model meets the requirements of this project. Compared with other candidates, our final model reduces the number of predictors from 14 to 5 and achieves the best overall performance, which also helps enhance the robustness of model. However, the on the test set is 0.6388, which is acceptable but not convincing enough for a mature product. We believe this could be improved with a larger dataset and more powerful methodology. Due to limited dataset capacity and time constraint, we are not able to build a more advanced model while keeping it simple and understandable.

**5.2 Conclusions**

To find the best model for body fat, we have compared four regression models and chosen the one with lasso regularization and cross-validation because a returns a lower RMSE and only requires five predictors. This model shows good accuracy and robustness on both training and test sets, although limitations still exist. If time permits, we believe it can be better polished with superior techniques.

**References**

[1] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

**Contribution Table**

|  |  |  |  |
| --- | --- | --- | --- |
| Contributions | Shan Leng | Siyan Wang | Ruotong Zhang |
| Presentation | Reviewed, edited and provided feedback on all slides. | Responsible for slides 5-10 (results).  Reviewed/edited slides 1-4. | Reviewed/edited and provided feedback on all slides. |
| Summary | Responsible for final model statement, relevant statistical analysis and model diagnostics;  Reviewed and edited all sections. | Responsible for discussion and provided feedback on whole document. | Responsible for introduction and data processing; Provided feedback on whole document. |
| Code | Responsible for model construction and diagnostics code. | Reviewed code. | Responsible for data cleaning code. |
| Shiny App | Provided feedback on Shiny app. | Responsible for Shiny app. | Reviewed Shiny app. |