**1. Introduction**

(a brief introduction about motivation/what the report is about/what you plan to describe) 咱们三个每人都写一段；最后整合。

**2 Data Processing/Cleaning** 若彤

1. Discuss any relevant/key aspects about the data (e.g. mean/sd of Y, some Xs)
2. Discuss any outliers removed/fixed/imputed **even before building a model** (e.g. body fat % looks suspicious, some predictors look suspicious, etc.)
   1. Example 1: I removed individual with body fat % BLANK because of BLANK.
   2. Example 2: I fixed/imputed individuals with body fat % BLANK by BLANK

**3 Final Model Statement** 若彤和偲妍

1. State your final model/easy-to-use rule of thumb.
2. State an example usage of the final model
   1. Example 1: a man with BLANK is expected to have a body fat % of BLANK based on our model). His 95% prediction interval is between BLANK and BLANK.
3. Interpret your model (in *laymen’s terms*\*\*)
   1. Example 1: Our estimated coefficients are BLANK and BLANK, which are in the units of BLANK and BLANK. This means that for every BLANK increase in BLANK, the model predicts that body fat % will increase, on average, by BLANK.

**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.

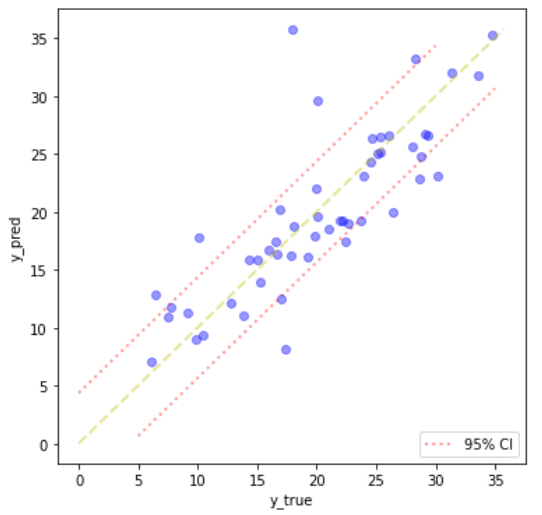
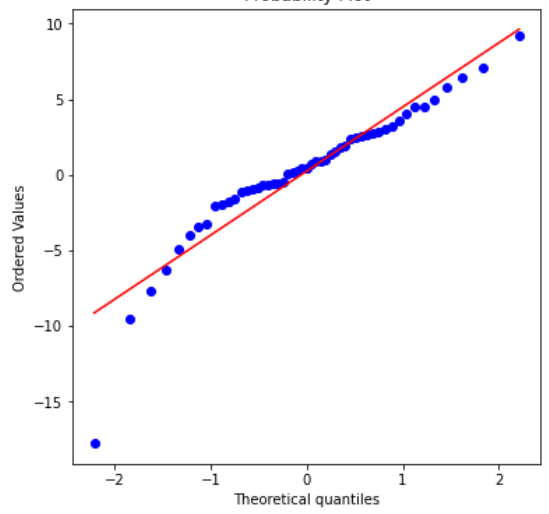
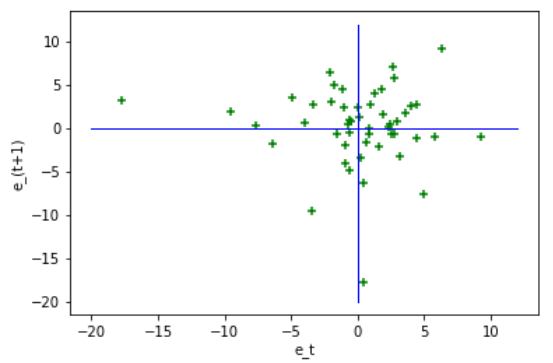
Table 1.Comparison among Candidate Models

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 Model Strengths/Weaknesses** 偲妍

**5.2.1 Model Strengths**

High accuracy and High Robustness: The final model we choose, the LassoCV regression, has the best overall performance in accuracy and robustness comparing with other models.

Simplicity: Comparing with other model in our experiment, we reduce the feature of model from 14 to 5, while keeping the level of accuracy. The reduction of feature number help enhancing the robustness of model obviously.

Object alignment: The model’s performance meets the specific requirements and goals of the project, as we aim to calculate people’s bodyfat with related features. The data set capacity is not enough for us to use more complicated model while it remains to be simple and understandable.

**5.2.2 Model Weaknesses**

Underfitting: The R2 is only 0.6388 in test set, which means the model is still not convincing enough for a mature product. We conclude that is because the model remains to be too simple.

Data quality: The capacity of data set is too small, which not allows us to split it into training set and testing set big enough. Besides, there are some outliers inside the data set, we transformed these data in data processing while maybe these data still have impact on the analysis.

**6 Conclusion/Discussion** (summarize what you wrote above; final thoughts/discussions)

咱们三个每人都编一段；最后整合。

**References:** must be only on the 3rd page and may not exceed more than one page.(glmnet + 数据)

+ contribution table