**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** 冷杉

Rationale for your Final Model and or Model Diagnostics

**4.1 Model Selection**

**4.1.1 Selection Criteria**

In order to obtain the model that best meets the goal, we develop the following 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. RMSE measures the average difference between estimated values and actual values; A lower RMSE indicates that the model fits training data well and has potential ability to make accurate predictions. represents the proportion of response variation explained by the model; An closer to 1 indicates a better degree of fit on the training set.

As for model simplicity, we count how many predictors are included, and evaluate whether the rationale of the model is easy to understand and interpret. A model that requires fewer predictors and applies more basic methodology is considered more simple.

For robustness, similarly, we consider RMSE and on the testing set. But since the testing set is unseen during the training stage, a good performance on the testing set 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 predictors[[1]](#footnote-0).
3. Lasso regression with cross validation. A linear regression model with lasso penalty and 5-fold cross validation. We First feed all the predictor into the model. Among the top six features with largest absolute coefficients, we pick five the most common ones and use them to fit another model of the same type from scratch. We end up with a lasso regression model involving 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**

1. Discuss, in some length, why you chose the final model. Include relevant tables/figures/etc (e.g. scatterplot of final model, table of key statistics, etc.)
   1. Example 1: We chose this model because of the following reasons. First, BLANK (e.g. background research). Second, BLANK (e.g. statistical analysis/figures/tables). Third, other models using BLANK had BLANK (e.g. some key statistics), which was BLANK in comparison to our final model.

Table 1 summarizes the results for all the candidate models. As we can see, although (1) is the simplest, it performs the worst on both training and testing sets. We do not want to sacrifice accuracy to much for simplicity. (2), (3) and (4) result in similar results in terms of accuracy and robustness. Despite a slightly better accuracy, (2) includes 14 predictors, which is too many to be further generalized. (4) applies a more complex methodology and is not friendly for comprehension.

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

1. Discuss any key estimates or statistical tests you have conducted to support your model. Make sure to interpret your results carefully.
   1. Example 1: We conducted the following test to see whether the predictor(s) we have chosen are significant in predicting the outcome. (formally state hypothesis; formally state which test statistic you used; state the p-value/test stat value; state the Type I error you’re willing to tolerate). From our statistical test/p-value, we can conclude that BLANK (interpret your statistical tests/p-values in a laymen’s term based on this data’s predictors; see lecture notes for details).
   2. Example 2: We found our R^2 to be BLANK, which implies (interpret it in layman’s terms)
   3. Example 3: The estimated slope and intercept are BLANK and BLANK, with 95% CIs BLANK and BLANK. This implies that BLANK (interpret in laymen’s terms). Also, based on the 95% CI, we can reject/retain the null hypothesis of BLANK and BLANK. In other words (interpret them in laymen’s terms)

**4.2 Model Diagnostics**

Include relevant model diagnostics you did (e.g. plots/tables/etc.) and explain your model diagnostics (e.g. why you did it, what assumption is this diagnostic checking, what is your conclusion from the diagnostic checks, how did you resolve any violations of model assumptions)

* 1. Example 1: We checked the following four assumptions for SLR/MLR. First, we checked BLANK using BLANK (see Figure BLANK). Because BLANK, we believed BLANK is plausible, even though there is slight violations of BLANK. Second, we checked BLANK.

**5 Model Strengths/Weaknesses** 偲妍

1. Discuss strengths and weaknesses of your model
   1. Example 1: Some strengths of our model include BLANK, BLANK, and BLANK. In particular, our model satisfies the linear regression assumptions of BLANK and BLANK, brining credence to our results/interpretation in BLANK.
   2. Example 2: Some weaknesses of our model include BLANK, BLANK, and BLANK. In particular, for certain groups of males, the model may provide BLANK and BLANK because of BLANK.

**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

1. All the predictors: Age, height, wight, Adioposity, and ten kinds of circumferences(neck, chest, abdomen, hip, thigh, knee, ankle, biceps, forearm and wrist). [↑](#footnote-ref-0)