CS4049 Assesment 1 Report

Dimitris Dinis-Gridian

November 13, 2022

1 Introduction

The given task is to analyse the relationship between body fat and different types of body measurements based on a data set containing 252 samples of different body measurements and body fat of men. The analysis can be broken down into

- prediction of body fat based on features which can be measured using a scale and a ruler
- inference of the body measurements with the most positive correlation concerning body fat

The data set contains the dependent variable y, 'Percent body fat', and the independent features X, which are ten different body measurements, age, weight, height, adiposity and body density.

The Ridge Regression model was chosen for this task. It is a type of multiple regression model which introduces a new parameter, λ, called the shrinkage parameter, in the Least Square estimate (LSE) closed-form equation defining linear regression [2]:

$$\vec{b} = (X^T X)^{-1} X^T y$$

Where $\vec{b}$ is the trainable coefficients associated with the independent features. The ridge regression solution adds the scalar λ to the LSE equation by multiplying it with the identity matrix:

$$\vec{b_{ridge}} = (X^T X + \lambda I)^{-1} X^T y$$

2 Model Selection

The reason for choosing Ridge Regressions is that, as described previously, it introduces the parameter λ in the LSE solution, which is used to place a constraint over $\vec{b_{ridge}}$ so that the coefficients are penalized—penalizing means that the more significant the increase, the bigger the shrinkage over $\vec{b_{ridge}}$. This is used to solve multicollinearity, a phenomenon where independent features have similar or the same slopes, which causes inaccurate estimations of coefficients and large prediction variance[4].

2.1 Multicolinearity, correlation

Multicollinearity, defined as a metric, is the inflation of a coefficient's variance due to its linear dependencies on other coefficients [3]. Plotting the correlation matrix can help understand the colinearity of the data, which can be seen in figure 1. As can be seen, the data has high multicollinearity. Features like the abdomen and chest circumferences have an almost perfect positive correlation of 0.92, whereas Density and abdomen have an excellent negative correlation of -0.80.

Plots against the target variable can also give insight into the correlation between each feature and the target variable. For example, on first look, it can be observed that weight, hips and abdomen circumferences are positively correlated with body fat, which can be observed in figure 2.

On the other hand, Density gm/cm^3 is highly negatively correlated to body fat, as shown in figure 3.

The correlation between each feature can be measured using Pearson's correlation coefficient formula, confirming the conclusions reached from analysing the distributions [1].

Using NumPy's corrcoef function, each feature's correlation was measured against the target variable body fat. As shown in figure 4, Density has an almost perfect negative correlation with body fat, whereas abdomen circumference and adiposity index have a relatively good positive correlation.

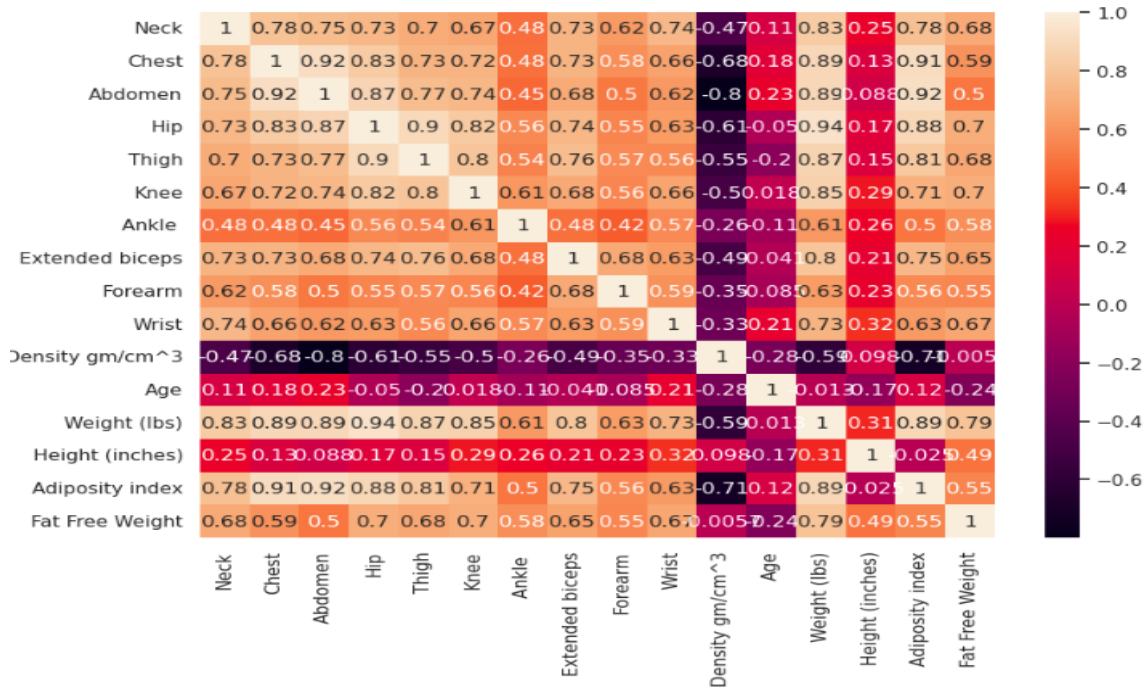

Figure 1: Correlation matrix plotted for body fat data, excluding body fat. Column names have been shortened for readability and page fit.

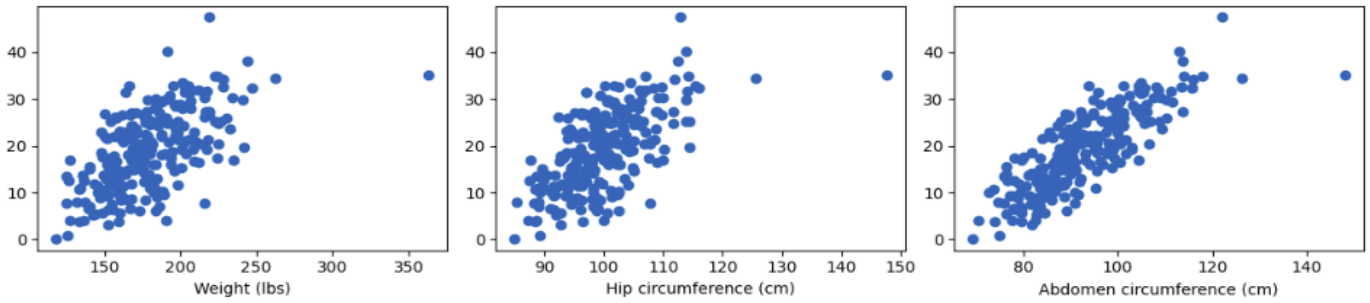

Figure 2: Hip circumference, Abdomen circumference and weight plotted against body fat

3 Data preprocessing

The data was loaded into the environment using pandas.

```
dataset = pd.read_csv("body_fat_data.csv", index_col=0)

dataset.describe()
```

It has been found that weight has the highest mean, where the average mean is roughly 65; it has a mean of 178.

Before fitting the data, it was split into training and test data using a $0.67 - 0.33$ split. Standardization was used so that the coefficients of the independent variables would be shrunk fairly by λ. Standardization ensures all features are on the same scale so that shrinking would not affect their coefficients unfairly, as different scales contribute differently to the penalty. Although the features were roughly on the same scale (shared the same magnitude), standardization was still applied to respect the methodology of standardizing data before training a model [5].

There also needs to be mentioned that the test data was standardized using the mean and standard deviation of the training data. If the test data had been standardized using its own mean and standard deviation, it would have a different scale from what the model was trained on, resulting in biased predictions.

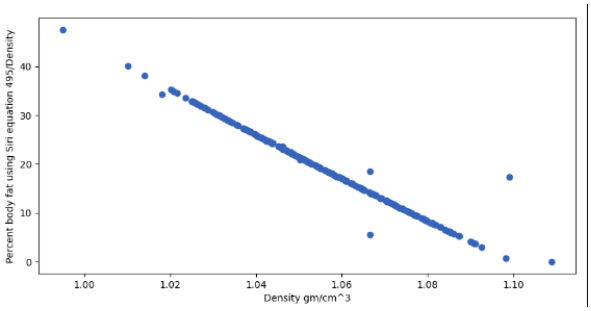

Figure 3: Negative correlation between Density gm/cm^3 and body fat

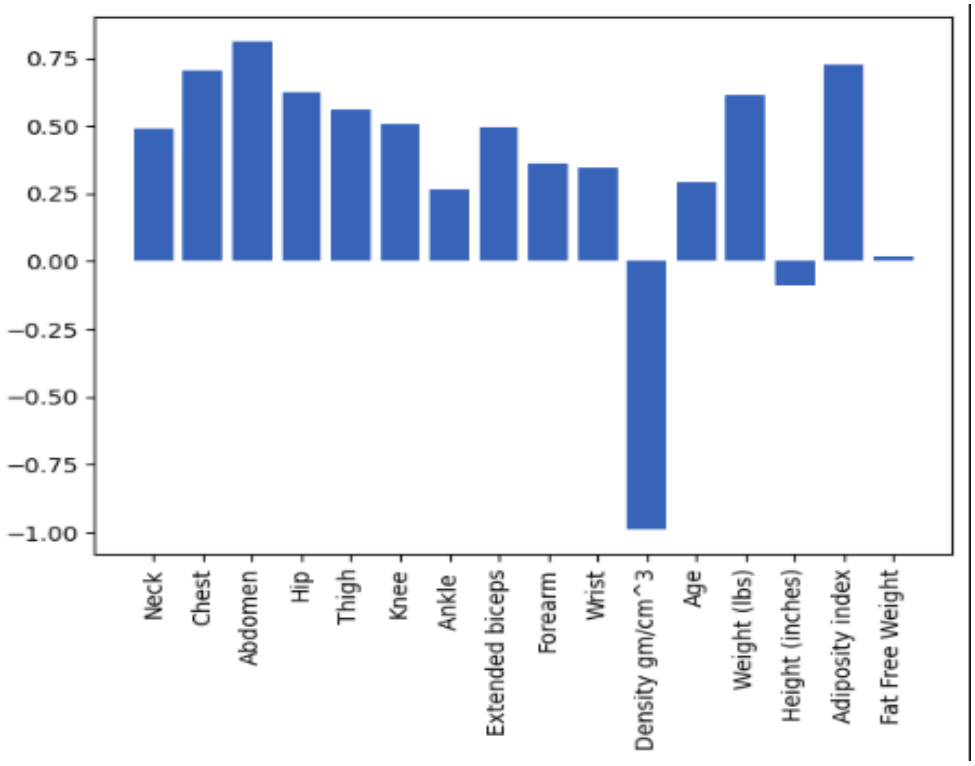

Figure 4: Pearson's correlation coefficient for each feature concerning body fat

On the other hand, min-max normalization has been considered using the following function signature:

```
def normalize(data, maxv,minv):
    return (data- minv) / (maxv + minv)
```

It has been found that coefficients are on a different scale than standardized ones. However, the loss remained the same.

4 Cross-Validation

The assignment requires estimating body fat using only a scale and a measuring tape. Body density cannot be determined using a ruler or a measuring tape. Furthermore, body fat is derived using Density; thus, having it as a feature would mean we are using a linearly dependent feature of y to predict y. Using Density to fit the model would result in a minor loss; however, since correlation to the target is almost perfect, the cross-validation result would be a λ term close to, or even 0. This would defeat the purpose of this assignment as if the lambda term is 0; then the model becomes linear

regression. For these reasons, the density column was dropped. Fat-free weight is also derived using a fraction of body fat; therefore, it was dropped. Finally, the adiposity index represents the weight over height, which makes it a duplicate coefficient. Consequently, it was dropped as well.

K-fold cross-validation was used to perform cross-validation for finding the best value for λ. Per the assignment's requirements, λ values of [0, 0.5... 50] have been trialled. For each λ, the training data was split into k partitions, leaving one out for testing. The ridge model was fit with each partition, and the mean squared error of the prediction was computed using the validation partition. An implementation of the cross-validation algorithm has been provided in the RidgeFromScratch jupyter file. The results were compared with the sklearn implementation, which can also be found in the SkLearnRidge jupyter file.

The optimal λ found after cross-validation was 0.5:

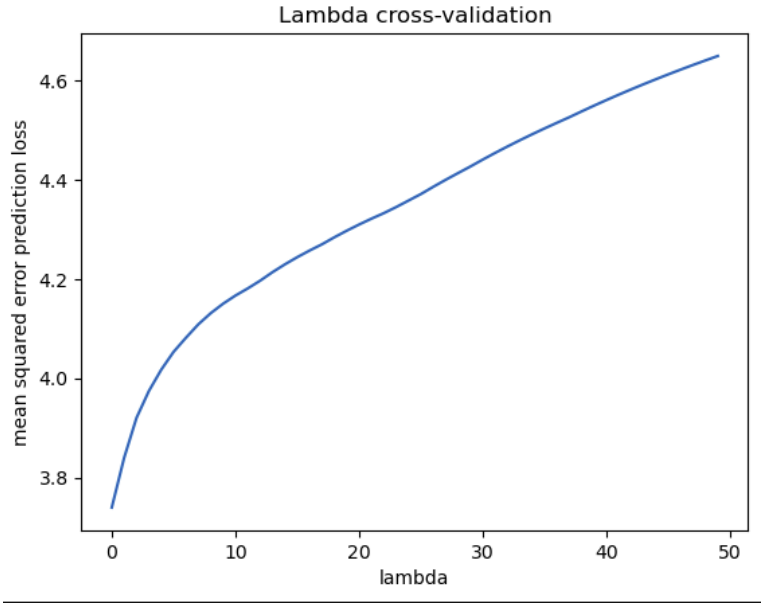

Figure 5: Lambda and their respective loss during cross-validation

5 Predictions, measurements

The model was fit using the λ value at 0.5. Ridge has also been implemented from scratch. The coefficients can be computed using the ridge formula:

```
ridge = np.linalg.inv(X.T @ X + penalty) @ X.T @ y
```

It can then be used for prediction by applying it to the test data:

```
predictions = test_X @ ridge
return predictions
```

As shown in figure 6, the abdomen circumference is the optimal measurement for estimating body fat. It is followed by thigh and wrist circumferences and weight and hip circumference. The coefficients can be confirmed by looking at the respective feature's correlation in Pearson's correlation coefficient figure 4. Using these coefficients, estimations of body fat can be made. The performance can be seen in figure 7.

5.1 Conclusions, Limitations

The loss of the prediction was computed using mean absolute error, which resulted in a loss of roughly 3.48. When checking the losses manually, there are five test points with a difference more significant than five from the predicted values: 1,14,18,24,26. These test points share a familiar pattern: they

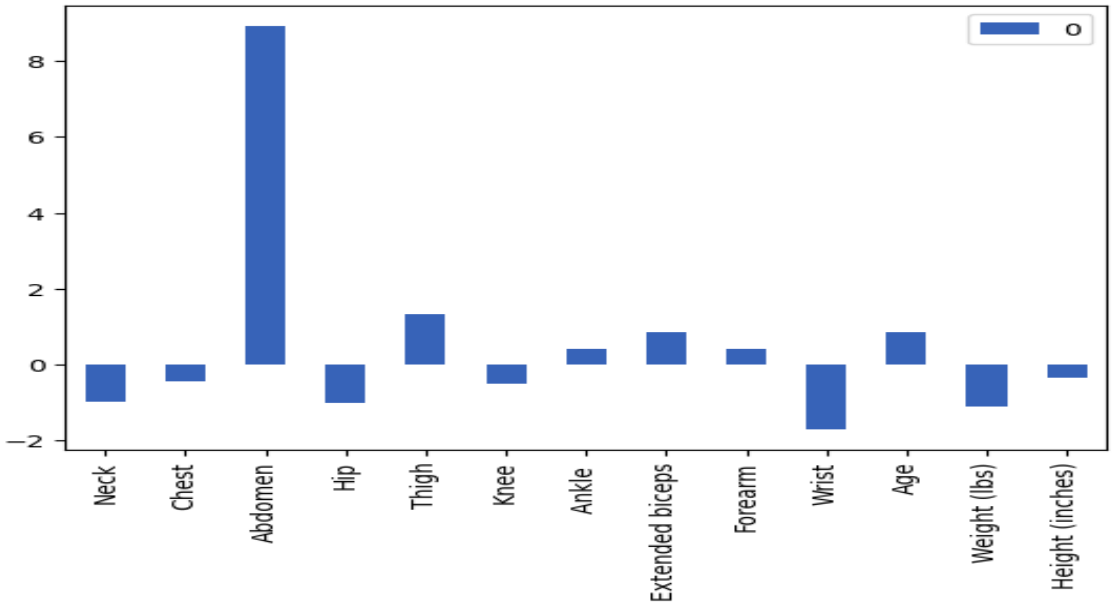

Figure 6: Coefficients for ridge model with lambda = 0.5

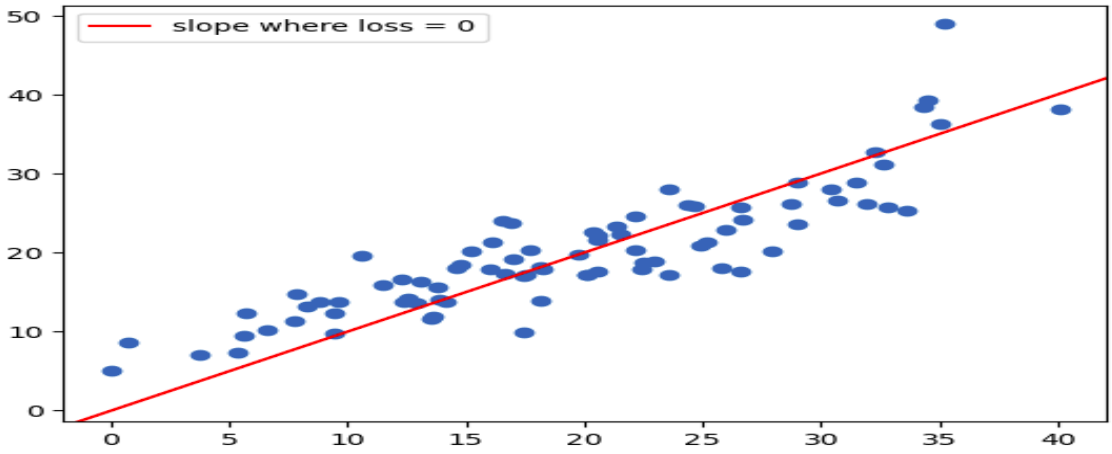

Figure 7: Model scatter distribution, compared to the true values of y

all have abnormal features such as high lower body circumference and low higher body circumference (14) or simply a strange element such as a tall wrist circumference (1). Given this, it can be said that the model performs fairly well in predicting body fat, using the abdomen circumference's coefficient as the main predictor.

References

[1] S. L. Crawford. Correlation and regression. *Circulation*, 114(19):2083–2088, 2006.

[2] T. Daniya, M. Geetha, B. S. Kumar, and R. Cristin. Least square estimation of parameters for linear regression. *International Journal of Control and Automation*, 13(2):447–452, 2020.

[3] J. I. Daoud. Multicollinearity and regression analysis. *Journal of Physics: Conference Series*, 949(1):012009, dec 2017.

[4] H. Yu, S. Jiang, and K. C. Land. Multicollinearity in hierarchical linear models. *Social science research*, 53:118–136, 2015.

[5] K.-H. Yuan and W. Chan. Biases and standard errors of standardized regression coefficients. *Psychometrika*, 76(4):670–690, 2011.