ISYE 7406: Homework # 2 Report

Introduction:

The goal of this assignment is to analyze a health dataset to predict body fat percentage (calculated using Brozek's equation) based on various other predictors. These predictors include another body fat measure (calculated using Siri's equation), age, weight, height, body part measurements, and more. The ultimate goal is to develop accurate predictive models for body fat percentage and compare their performance to find the optimal model.

The dataset used in this analysis comes from Johnson R. *Journal of Statistics Education* (1996). It consists of data obtained from 252 male participants. The target variable is body fat percentage calculated using Bzoek's equation (brozek). There are 17 additional variables, including physical measurements like body circumferences, weight and height. For the analysis I split the data into a training set used for exploratory data analysis and model building, and a test set for model evaluation. I implemented seven different regression algorithms and evaluated each model's performance. Cross Validation was used to ensure robustness of the models to identify the optimal model.

Exploratory Data Analysis (EDA):

The dataset is well-structured, with no missing values in any of the variables. The target variable, body fat percentage (Brozek), showed an approximately normal distribution. The histogram showed a bell shaped, symmetric curve with minimal skewness, suggesting a balanced target variable.

To explore the relationships between the variables, I created a correlation matrix which revealed key insights about the variable relationships. Siri's equation, another method for calculating body fat percentage, showed a near perfect positive correlation (0.999) with Brozek's equation. This was expected since both variables represent body fat percentage calculated using slightly different formulas. Similarly, density also had a near perfect negative correlation (-0.996) with Brozek's equation. Since density is directly used in the Brozek's formula (brozek=457/Density- 414.2), this strong correlation was also expected. These findings suggest Siri and density are highly informative in predicting body fat percentage, however they appear to be nearly identical to the target variable, and including both could be redundant.

Other predictors with strong correlations to body fat percentage included abdomen circumference (0.816), chest circumference (0.705), and adiposity index (0.729). Moderate correlations were observed with hip circumference (0.633) and weight (0.631). The correlations suggest that body measurements, particularly those related to the chest and abdomen area, are strong predictors of body fat percentage. This aligns with the role of adipose tissue, which stores fat in the chest and abdomen to protect vital internal organs. This explains the high correlation between adiposity, chest, and abdomen.

Variables like height (-0.094) and fat free weight (0.038) showed very weak correlations with body fat percentage. These variables may contribute very little to predictive the model's value and could potentially be excluded from simpler models. Their predictive value will be further assessed during variable selection and modeling.

It's also important to note, the EDA revealed signs of multicollinearity among predictors. When predictors are highly correlated with each other, this can lead to overfitting and unreliable predictions. So it's essential that multicollinearity is addressed. Adiposity index is strongly correlated with chest circumference (0.915), abdomen circumference (0.926), and weight (0.893), which is consistent with their shared relationship to body fat storage. Similarly, weight was highly correlated with several body part measurements including neck (0.84), chest (0.89), abdomen(0.898), hip(0.94), thigh(0.88), and knee (0.87). These overlapping relationships suggest some variables may provide redundant information, requiring further investigation.

Multicollinearity must be assessed and confirmed through additional steps. Variance inflation factors (VIF) scores were calculated to confirm multicollinearity. A high VIF indicates multicollinear variables. Several variables had high VIF scores that exceeded the threshold. Siri, density, weight, and fat free weight significantly exceeded the threshold, indicating high multicollinearity in these variables. While variables like chest, abdomen, hip, and adipose showed mild multicollinearity. Later in my analysis, LASSO regularization and dimensionality reduction via Principal component analysis (PCA) will be used to address multicollinearity.

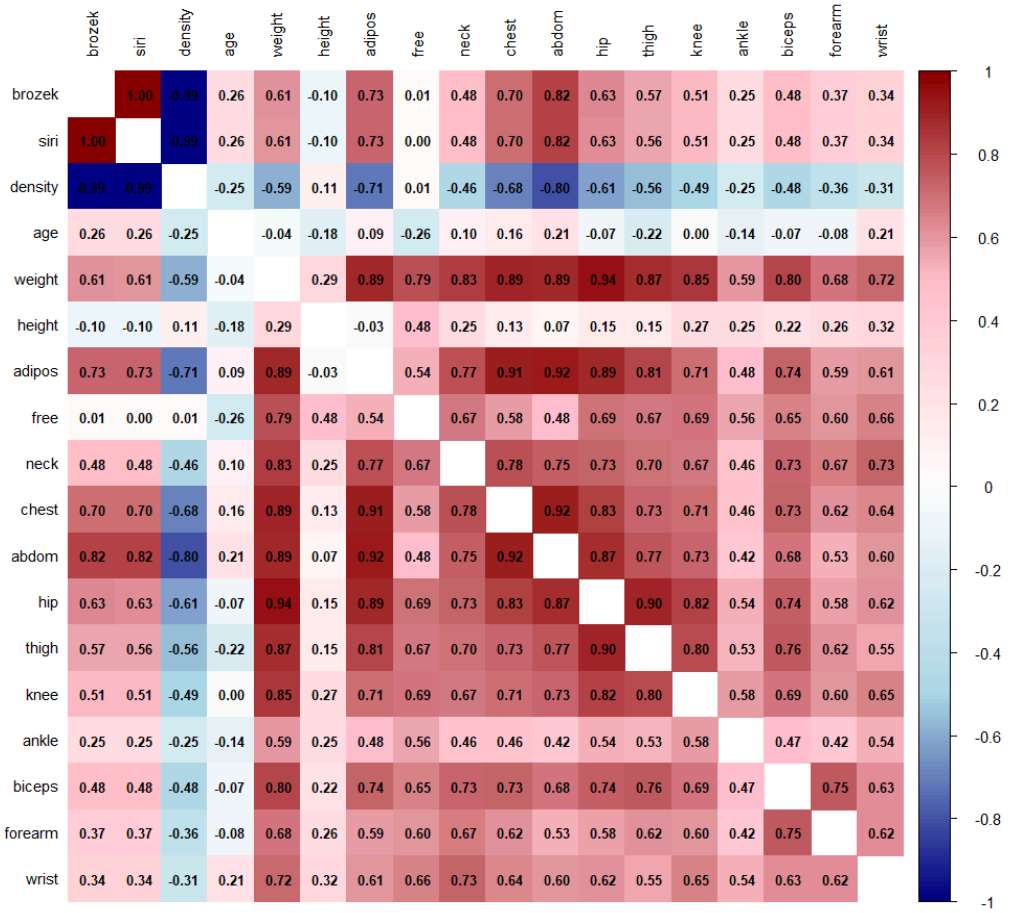
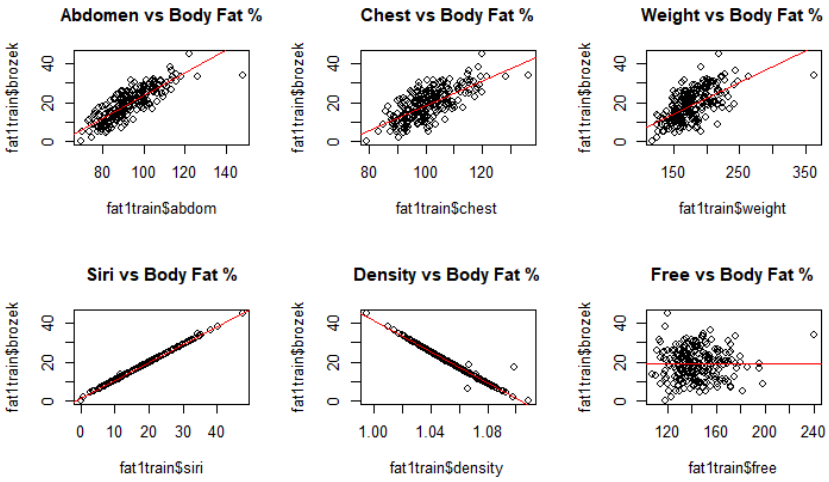


Figure 1: Correlation Matrix with self correlations (diagonal) removed. A coefficient of 1 indicates a strong positive relationship, 0 indicates no relationship, and -1 indicates a strong negative relationship. Blue represents negative relationships and red represents positive relationships.

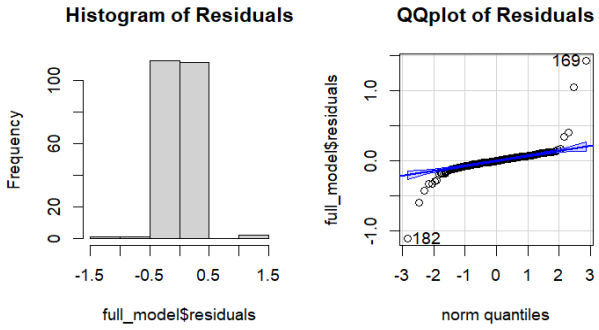
Scatterplots were used to examine linearity between body fat percentage and some of the key predictors with high, moderate, and low correlations. For siri and density, scatter plots showed nearly perfect linear

relationships (straight line) with body fat percentage. Siri showed a positive trend, while density showed a negative trend, consistent with their respective correlation coefficients of 0.999 and -0.996. Scatterplots for chest, abdomen, and weight appeared to show moderate to strong positive linear relationships with body fat percentage, indicating that as these measurements increase, body fat percentage also increases. For variables like fat free weight, the scatter plot showed weak positive relationships with body fat percentage, as seen in the correlation matrix. The points are more scattered and a less distinctive trend.

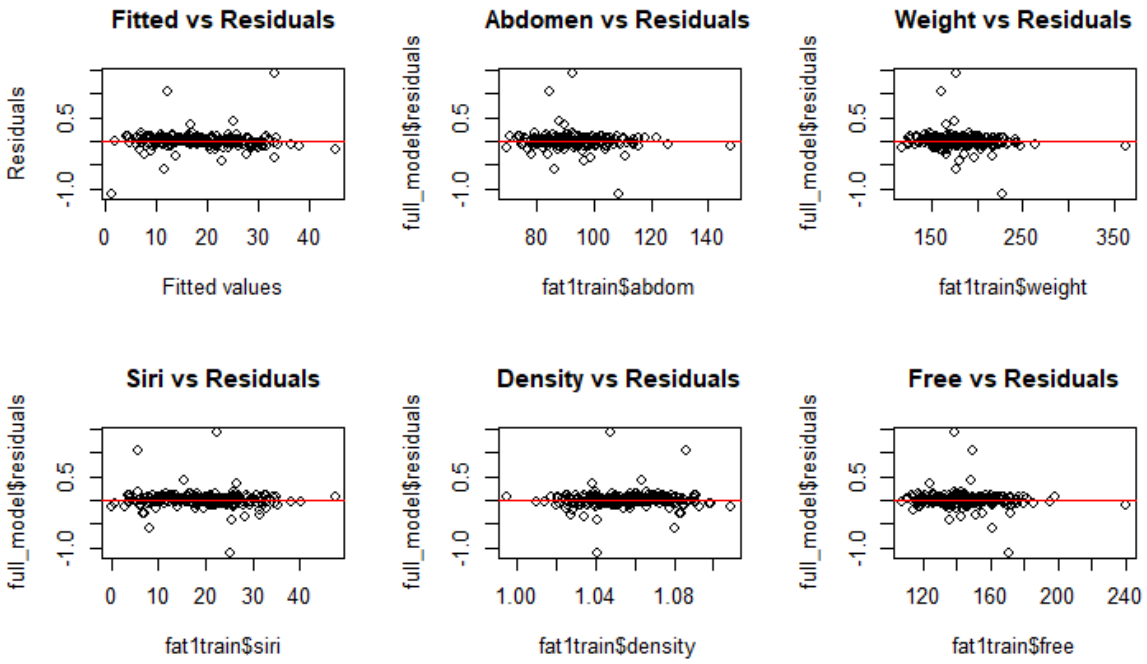


2A. Assumptions: Scatterplots of key predictors and the target variable body fat percentage, displaying linear relationship.

In order for Linear regression results to be valid and reliable, all of the Linear regression assumptions must hold. Linearity assumption was examined using scatterplots, which showed clear linear relationships between body fat percentage and key predictors. It was further supported by scatterplots of the residuals vs predictors. Constant variance assumption was checked through the scatterplot of the residuals vs fitted values. The residuals appeared to be randomly scattered (no clear trend) around the zero line and showed consistent variance, with minor deviations, supporting linearity and homoscedasticity assumption. Few outlier points are observed and can be investigated further. The normality assumption was assessed through the histogram and QQ plot of the residuals. The histogram suggests that the residuals are approximately normally distributed, while the QQ plot shows most of the points follow the reference line, with slight deviations at the tails. The normality assumption is reasonably met. However, the Independence assumption appears to be violated due to multicollinearity, which was identified through correlation analysis and high VIF scores. This issue will be addressed using regularization and dimensionality reduction techniques. Overall, the diagnostic plots confirm the key linear regression assumptions of linearity, homoscedasticity, and normality are reasonably met. However, addressing multicollinearity is essential to ensure model validity.



2B. Assumptions: Normality Assumption Plots showing a histogram and the QQplot of the residuals.



2C. Assumptions: Constant Variance & Linearity Assumption Plots showing fitted vs residuals and predictors vs residuals.

The EDA identified key predictors for body fat percentage and also uncovered multicollinearity concerns among several predictors.These findings will help with model selection and it highlights how crucial variable selection and addressing multicollinearity will be in achieving good performance and reliable results.

Methodology:

In this analysis, I built seven different regression models involving different approaches to analyze and predict body fat percentage (Brozek's equation). These models were evaluated and compared using several evaluation metrics to ultimately determine which model was the best fit.

I started with a full Linear regression model that included all of the predictors. I'll refer to this model as the LR Full model. Linear Regression is an algorithm that models the relationship between a dependent variable and independent variables by finding the best fitting line. It is used to predict continuous outcomes based on the features. I used the lm() function in R to implement this model and evaluation metrics that included r-squared, adjusted r-squared (found in the summary statistics), and Mean Squared Error (MSE) to evaluate its performance.

The next model implemented was a Linear Regression model with the best (k=5) subset. To determine which subset of predictors was the best choice, I performed the best subset regression using the regsubsets() function in the leaps library. I used r-squared, adjusted r-squared, and Mean Squared Error (MSE) to evaluate performance.

Stepwise regression was the third model implemented in this analysis. Stepwise regression is an algorithm used for variable selection. It automates the process of variable selection, by adding or removing predictors that are unimportant for predicting the dependent variable.There are three types of Stepwise Regression,

backwards, forward, and both. Backwards Stepwise, begins with all the predictors in the model and removes predictors one by one until it gets the optimal Akaike Information Criterion (AIC). Forward Stepwise Regression on the other hand begins with no predictors (just the intercept) and adds one predictor at a time until the AIC no longer significantly improves. Both Stepwise is a combination of the two. I opted for backwards stepwise direction using the step() function to identify the optimal set of predictors for the model that has the best tradeoff for model complexity and goodness of fit. I used the coef() function on the stepwise model to view the selected features that resulted in the optimal AIC..

Ridge regression, which is a regularization technique, was implended next. It adds a penalty term and shrinks the coefficients of predictors, shrinking the least important ones close to zero. It's great for addressing multicollinearity and identifying the most important predictors for variable selection. In earlier steps I identified multicollinearity in the data. The correlation matrix in my EDA suggested several predictors were highly correlated and I confirmed this by performing VIF analysis. Several predictors presented VIF scores significantly above the threshold indicating multicollinearity. Ridge regression helps with multicollinearity by shrinking coefficients (the impact) of predictors that are highly correlated with others. The lm.ridge() function from the MASS package was used to fit ridge regression models. The optimal regularization parameter (λ) values were selected using Generalized Cross-Validation (GCV) and k-fold cross-validation (via cv.glmnet() function from the glmnet package). By removing highly correlated variables, ridge regression also helps with reducing model complexity for models with many predictors compared to observations (or $p > n$), which helps prevent overfitting.

The Least Absolute Shrinkage and Selection Operator (LASSO) regression was implemented next. LASSO is also a regularization technique and is very similar to Ridge regression. However, unlike Ridge, LASSO shrinks the coefficients of less important variables all the way to zero, essentially removing their contribution to the model altogether. Basically it performs variable selection directly by removing irrelevant predictors . It's good with high dimensional data and data with multicollinearity present. I implemented LASSO using lars().

Principal Component Regression (PCR) was performed next. PCR is used to reduce dimensionality in data while also retaining most of the variance in the data. PCR works by transforming correlated variables into uncorrelated principal components which addresses multicollinearity as well. It selects the optimal number of principal components that keep as much useful information as possible, while also removing information that is not helpful. PCR was implemented by using the pcr() function and cross validation was used to find the optimal number of principal components.

Partial Least Squares (PLS) Regression was the last model implemented. This model is very similar to PCA regression, except it takes into account the relationships between the predictors and the response variable. This is different from PCA, because PCA only considers relationships between the predictors. PLS makes sure the selected components are relevant to predicting the target. It is essentially a combination of PCA and linear regression. PLS handles multicollinearity and high dimensional data and reduces overfitting. It simplifies the model while still retaining the ability to predict the response variable. This model was implemented using the plsr() function.

To ensure my results were reliable for model comparison, and not just dependent on one specific split of the data, I used Monte Carlo cross validation with 100 iterations for each model. For each iteration, the data was randomly split into training and test subsets repeatedly. By randomly reshuffling the data each time, this ensures the same data isn't always used for the training set vs the test set. Training and testing errors (measured by MSE) were calculated and stored for comparison.

Results:

The performance of the seven regression models for predicting body fat percentage was evaluated and compared using metrics like Mean Squared Error (MSE), variance, and R-squared. The results include the initial training and testing errors and the Monte Carlo cross-validation mean testing errors and variance. A summary of the results is provided below.

Initial Models

Model	Training Error	Testing Error
Linear Regression (Full)	0.02930823	0.008755981
LR Best Subset (k=5)	0.03146801	0.002786218
Stepwise Regression	0.02945827	0.008955971
Ridge Regression	0.02930890	0.008859234
LASSO Regression	0.03085618	0.003158102
PCR Regression	0.03032433	0.008846963
PLS Regression	0.03303619	0.020043438

Figure 3A.Training and Testing error for each model.

Monte-Carlo Cross Validation

Model	Variance	Testing Error
Linear Regression (Full)	0.006435608	0.05561845
LR Best Subset (k=5)	0.006209426	0.04075667
Stepwise Regression	0.007733174	0.05823229
Ridge Regression	0.006635656	0.05639683
LASSO Regression	0.005425468	0.05216016
PCR Regression	0.006435608	0.05561845
PLS Regression	0.006431716	0.05552176

3B. Mean Test error and Variance using cross-validation B=100

The Linear Regression model (full model) included all of the predictors, producing an R2 and adj R2 value of 0.9996. The statistically significant predictors included siri (body fat percentage by Siri's equation), density, weight, free fat weight, thigh, and knee circumference. In the monte-carlo cross validation it had a mean testing error of 0.0556 and a variance of 0.0064. The relatively higher mean test error suggests potential overfitting due to including all of the predictors (many of which were multicollinear). This resulted in its reduced ability to generalize on new data compared to other models.

The Best Subset (k=5) Linear Regression model was one of the best performing models. It resulted in an r2 of 0.9996 and an adjusted r2 of 0.9995 and produced a mean error of 0.04 and variance of 0.0062 during cross-validation. It achieved the lowest mean testing error among the other models. This shows the model generalized on the test data better and achieved a good balance between model simplicity and accuracy. The best k=5 subset model mitigated overfitting while retaining the most relevant and strong predictors for body fat percentage. The best k=5 subset included siri, density, thigh, knee, and wrist. The model also has high explanatory power for body fat percentage (99%).

Stepwise Regression model with the optimal AIC retained 10 predictors. It removed neck, chest, abdomen, hip, weight, ankle, and age in the initial split. As uncovered in my EDA some of these predictors showed high multicollinearity, however the stepwise had the highest testing error and variance in cross-validation out of all the models. It also had higher initial training and testing errors compared to other models. Stepwise has some limitations and the results show it's not recommended for small datasets because of overfitting or the inclusion of unimportant predictors.

```
> round(coef(stepwise_model),3)
(Intercept)        siri     density      height      adipos        free       thigh        knee      biceps
      8.587       0.888      -9.619       0.045       0.047      -0.010       0.016      -0.027      -0.015
   forearm       wrist
      0.017       0.035
```

Ridge Regression showed the predictors siri and density remained consistently dominant through both cross validation methods, similarly to the Stepwise model. Ridge regression shrinks coefficients without completely zeroing them out, and although it regularized the coefficients effectively, the testing error is higher than some of the other models. Its training and testing errors were 0.02930889 and 0.008859234 respectively. While it performed reasonably well, its testing error and variance were higher than the top models.

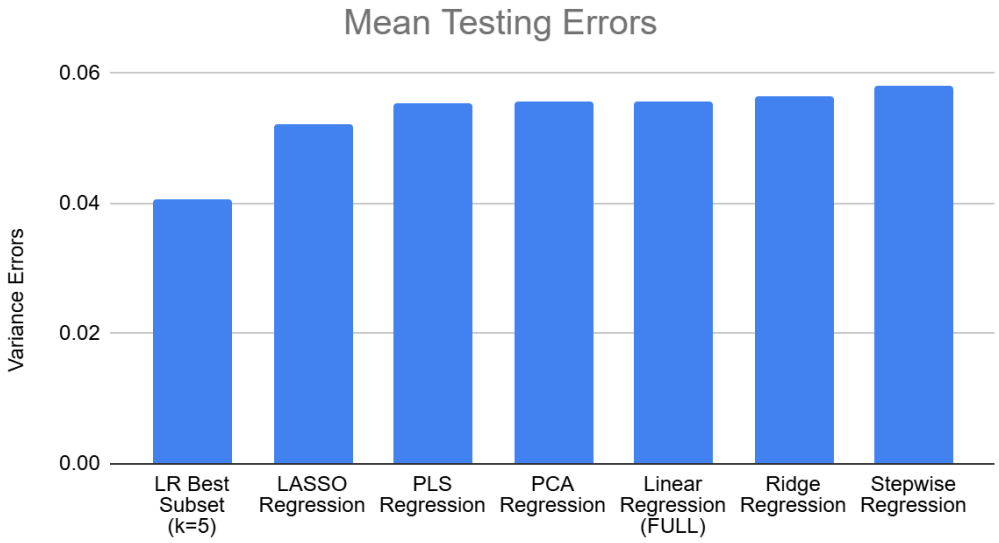
```
> print(round(ridge_coefficients_final, 5))
Intercept      siri   density       age    weight    height    adipos      free      neck     chest     abdom       hip
  12.60951   0.88255 -10.20168  -0.00066   0.01168  -0.00081  -0.01912  -0.01348  -0.00061   0.00259   0.00076  -0.00356
     thigh      knee     ankle    biceps   forearm     wrist
   0.01464  -0.02613   0.00325  -0.01720   0.02388   0.03276
```

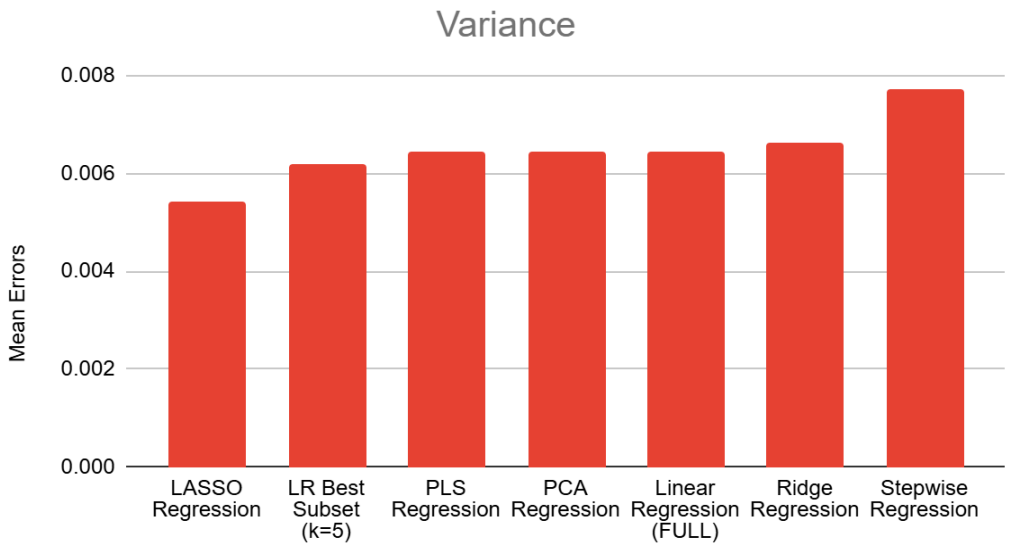
The LASSO Regression model was one of the strongest performers. It produced a mean testing error of 0.0522 and the lowest variance among all models at 0.0054 in cross-validation. It also showed siri and density were consistently dominant, showing the importance of these predictors. LASSO performed variable selection by eliminating (zeroing coefficients) predictors of neck, chest, abdomen, hips, height, adiposity, free, and weight. The LASSO model performs very well which demonstrates its strength in achieving a simpler model while maintaining strong predictive power.

```
> print(round(LASSO_coeff, 5))
intercept      siri   density       age    weight    height    adipos      free      neck     chest     abdom       hip
  11.27371   0.90402  -9.40594  -0.00037   0.00000   0.00000   0.00000   0.00000   0.00000   0.00000   0.00000   0.00000
     thigh      knee     ankle    biceps   forearm     wrist
   0.00697  -0.01356   0.00000  -0.00893   0.01334   0.01719
```

Principal Component Regression (PCR) model achieved moderate performance while reducing dimensionality of the data. It transforms correlated predictors into uncorrelated components. The testing error was comparable to some of the other models, however, it did not outperform the Best Subset (k=5) or LASSO models.

Partial Least Squares Regression (PLS) model performed similarly to the PCR model. It focused on maximizing the covariance between predictors and the response variable while reducing dimensionality. Its testing error of 0.0555 and variance of 0.0064 were comparable to those of PCR. Like PCR, the PLS model also did not surpass the Best Subset or LASSO models in terms of overall performance.





Conclusion:

This analysis showed that choosing the right variables and avoiding unnecessary complexity leads to better model performance and reliability. The Best Subset Regression (k=5) linear regression model was the best model, with the lowest mean test error and variance. It found the right balance between bias and variance by selecting only five key predictors (siri, density, thigh, knee, and wrist), proving that a simpler model with relevant features performs better. Variable selection played a critical role in improving model accuracy.

LASSO regression was the next best performer, following closely behind the best subset model. It did well in variable selection, completely eliminating the impact of unnecessary predictors by shrinking their coefficients to zero. It maintained low error and variance, and retained just 8 predictors, removing more than half of the original predictors.

PCR and PLS regression showed moderate results, but they did not outperform some of the simpler models like Best subset or LASSO. Ridge Regression and the Full Linear Model performed less favorably, with higher testing errors and variances, likely due to including too many predictors, some of which were highly correlated. This led to overfitting and reduced generalization. This also confirmed that keeping all predictors does not always improve performance, especially with multicollinearity. Stepwise regression was the weakest model, showing the highest testing error and variance.

Overall, this study highlights that choosing the right predictors is crucial for building an accurate and reliable model. The success of the Best Subset and LASSO regression models showed that removing redundant or unimportant variables improves accuracy. Simpler models with carefully selected predictors consistently generalize the best and perform better than complex models.