Body Fat Prediction Dataset

Professor Yu Yue
STA 9705 (Spring 2025)
May 19, 2025

Saisrijith Reddy Maramreddy

Peter Wong

Bryan Yeung

Zahidul Zahin

Introduction

The goal of our project is to investigate the relationship between body structure measurements and body composition, specifically, body fat percentage and body mass index (BMI), using multivariate statistical techniques. Accurate measurement of body fat, such as through underwater weighing, is often expensive, time-consuming, and impractical for everyday clinical or fitness use. If there are simpler, more accessible physical measurements like age, waist, or height that can be used to reliably estimate body fat, then these less invasive measurements could be useful proxies to determine body composition in the real world.

To investigate this, the statistical methods that were applied are multivariate regression, MANOVA, principal component analysis (PCA), and canonical correlation analysis (CCA) to a dataset of 252 men, sourced from the Body Fat Prediction dataset on Kaggle. This dataset includes 15 different physical attributes, which range from standard measurements like height and weight to more detailed circumferential data such as wrist, forearm, thigh, and abdomen size. Though this dataset came pre-cleaned, we checked min/max's of the different features to ensure there were no nonsensical values, and dropped any columns with missing or 0 data for further accuracy.

Each statistical tool we use plays a distinct role in uncovering the structure and the predictive power in our data. Multivariate regression allowed us to model multiple outcomes simultaneously, accounting for potential correlations between them. MANOVA helped us test whether certain groups, based on BMI category, differed across multiple dependent variables at once. PCA offered dimensionality reduction, which allowed us to find combinations of measurements that account for most of the variation in our dataset. Finally, canonical correlation allows us to study the shared structure between sets of predictor and response variables, which helps us isolate the most relevant features for prediction.

These techniques allow us to build accurate predictive models, but also help us identify the most statistically significant and practically used body measurements. These findings can inform more efficient health assessments by helping us reduce the number of required measurements needed from a person while maintaining a strong predictive understanding of body composition.

Data Collection and Description

The dataset, “Body Fat Prediction Dataset,” was retrieved from Kaggle, an online platform for data science studies. Kaggle, owned by Google, has many datasets that are often used for educational, analytical, and modeling purposes.

This particular dataset consists of 15 features collected from 252 men, aimed at predicting body fat percentage through physical measurements. The variables included are:

1. Density determined from underwater weighing 2. Percent body fat from Siri's (1956) equation 3. Age (years) 4. Weight (lbs) 5. Height (inches) 6. Neck circumference (cm) 7. Chest circumference (cm) 8. Abdomen 2 circumference (cm) 9. Hip circumference (cm) 10. Thigh circumference (cm) 11. Knee circumference (cm) 12. Ankle circumference (cm) 13. Biceps (extended) circumference (cm) 14. Forearm circumference (cm) 15. Wrist circumference (cm)

Although the dataset came pre-cleaned according to the source, we conducted additional reviews to ensure quality data. We examined the minimum and maximum values of each variable to check for outliers and removed any rows that had missing or zero values for further accuracy.

To support our analysis, we created a new variable called Body Mass Index (BMI), which is commonly used to classify weight. It was calculated using the standard formula:

$$\text{BMI} = \frac{\text{weight in pounds} \times 703}{(\text{height in inches})^2}$$

This dataset can be used to illustrate multivariate analysis techniques. Accurate measurement of body fat is inconvenient and costly. It is desirable to have easy methods of estimating body fat that are not inconvenient or costly.

Multivariate Regression

Multivariate regression is a statistical modeling approach used when there are multiple continuous outcomes that we want to predict using a shared set of explanatory variables. Unlike univariate regression, which models just one response variable at a time, multivariate regression allows us to model multiple related outcomes. This is especially useful when the dependent variables may have some potential correlation, since the method takes into account these correlations and the modeling process. For our project, we applied a multivariate regression to understand how various body measurements could predict 2 important markers of body composition. The model assumes the following form -

$$\begin{bmatrix} \text{BodyFat} \\ \text{BMI} \end{bmatrix} = \mathbf{X} \begin{bmatrix} \beta_{\text{BodyFat}} \\ \beta_{\text{BMI}} \end{bmatrix} + \varepsilon$$

The 1st is body fat percentage and body mass index, or BMI. These two outcomes are usually studied in health and fitness research and are known to be related, making them ideal candidates for joint modeling.

From our predictor set, we excluded height and weight because BMI is a ratio between them- including them would introduce redundancy and potential multicollinearity. As part of our exploratory data analysis, we generated a correlation matrix (Appendix Figure MR1) to understand the relationships between all predictors and both response variables. This helped us identify which variables were mostly strongly associated with body fat and BMI. We found that abdomen, hip, and chest had strong positive correlations with both outcomes, while neck was negatively associated. Wrist and forearm also showed relative correlations. (MR1)

Using SAS's PROC GLM procedure, with body fat and BMI as the joint responses and the selected body measurements as predictors, we were able to fit our multivariate regression model. The model for body fat had an R^2 of approximately 0.739, indicating strong predictive performance (Appendix MR2). Significant predictors in this model included abdomen, forearm, age, neck, hip, and thigh. The model for BMI did not have as much predictive power, with an R^2 of 0.257, identifying hip and age as significant predictors. (Appendix MR3) These results were reasonable because of physiological differences in how fat percentage and BMI respond to different parts of the body.

To refine our model, we were able to perform a stepwise regression. For body fat, this gave us a final model that included abdomen, forearm, and wrist as predictors. This model had a high R-squared of 0.7308, and all predictors were statistically significant with variance inflation factors below 10, meaning that there was no serious multicollinearity (Appendix MR4) . Below is our final model:

$$\hat{\text{BodyFat}} = -34.03 + 0.99(\text{Abdomen}) - 0.14(\text{Weight}) + 0.46(\text{Forearm}) - 1.52(\text{Wrist})$$

The BMI model produced a final model with hip, wrist, and age as predictors and an R^2 of about 0.24. Below is the stepwise model for BMI:

$$\hat{\text{BMI}} = -18.46 + 0.81(\text{Hip}) - 2.19(\text{Wrist}) + 0.088(\text{Age})$$

To try and improve our model for BMI, we test the transformations of the response variables. Using a Box-Cox transformation, we found the optimal Lambda to be approximately -1.75, meaning that a log or inverse transformation could help normalize our residuals and help stabilize the variance. Once this transformation was applied, our models' R^2 increased to

about 0.5, which is a substantial improvement and suggests that the transformed model better captures the underlying structure of the data.(Appendix MR5) Below is the corresponding model:

$$\widehat{\log(\text{BMI})} = 1.44219 + 0.02015 \cdot \text{Hip} - 0.01746 \cdot \text{Wrist} + 0.00212 \cdot \text{Age}$$

Overall, this process allowed us to better understand which body measurements contribute most strongly to body composition metrics. It also showed how appropriate transformations and model selection techniques can significantly improve model performance.

MANOVA

Multivariate Analysis of Variance (MANOVA) is a statistical method used to examine if there are significant differences between groups using multiple dependent variables. Unlike ANOVA, which analyzes one dependent variable, MANOVA looks at a combination of dependent variables and determines whether the mean differences among groups on these variables are likely to have occurred by chance. This method is useful when the dependent variables are correlated, as it reduces the risk of Type I error when conducting multiple ANOVAs separately. MANOVA checks whether different groups show different trends across several outcome measures. This gives us a bigger picture of how groups are different.

To explore whether there are significant differences in body measurements based on body fat percentage classification and Body Mass Index classification, we applied the Multivariate Analysis of Variance (MANOVA) technique. MANOVA is appropriate when there are multiple dependent variables and one or more categorical independent variables. In our analysis, the dependent variables include various body measurements such as age, weight, height, neck, chest, abdomen, hip, thigh, knee, ankle, bicep, forearm, and wrist. The independent variables were the BMI group, categorized into four levels (underweight, normal, overweight, and obese), and the body fat percentage group, categorized into four levels (low, normal, high, and very high). We excluded height and weight from the list of dependent variables for the BMI group because BMI is a derived variable based on those two measurements. Including all three would introduce multicollinearity and redundancy into the analysis, which could distort the results and violate MANOVA assumptions.

We also used contrasts in MANOVA to test specific group comparisons within the BMI and body fat percentage classifications. We examined whether combining the Underweight and Obese categories versus the Normal and Overweight categories in BMI, and combining Low and Very High versus Normal and High in body fat percentage, would reveal significant multivariate differences. We also tested whether the body measurements of individuals classified as Normal were significantly different from those classified as Overweight in the BMI group, and from those classified as High in the body fat percentage group. These contrasts helped us look more closely at specific patterns or differences that might be missed when just comparing all the groups as a whole.

The first hypothesis we tested is:
Null Hypothesis (H0): The means of the groups are equal across the four BMI groups and the four body fat percentage groups ($\mu1 = \mu2 = \mu3 = \mu4$).
Alternative Hypothesis (Ha): At least one group differs in its mean ($\mu1 \neq \mu2 \neq \mu3 \neq \mu4$).

MANOVA Tests for the Hypothesis of No Overall BMI_GROUP Effect
H = Type III SSCP Matrix for BMI_GROUP
E = Error SSCP Matrix

S=3 M=3.5 N=118

Statistic	Value	P-Value
Wilks' Lambda	0.23906523	<.0001
Pillai's Trace	0.86607531	<.0001
Hotelling-Lawley Trace	2.75031711	<.0001
Roy's Greatest Root	2.58691049	<.0001

MANOVA Tests for the Hypothesis of No Overall BODYFAT_GROUP Effect
H = Type III SSCP Matrix for BODYFAT_GROUP
E = Error SSCP Matrix

S=3 M=4.5 N=117

Statistic	Value	P-Value
Wilks' Lambda	0.32161183	<.0001
Pillai's Trace	0.73876014	<.0001
Hotelling-Lawley Trace	1.92519348	<.0001
Roy's Greatest Root	1.82745185	<.0001

Based on the MANOVA results, both the BMI group and the Body Fat Percentage group show statistically significant overall multivariate effects on the dependent variables. This conclusion is supported by all four multivariate test statistics (Wilks' Lambda, Pillai's Trace, Hotelling-Lawley Trace, and Roy's Greatest Root), each having a p-value below 0.0001, which is well below the 0.05 significance. Therefore, we reject the null hypothesis that there are no differences between groups. The results show that differences between groups in BMI and Body Fat Percentage lead to variations that are statistically significant across all of the dependent variables.

The second hypothesis we tested is:
Null Hypothesis (H0): There is no difference in the means between the combined groups of Underweight and Obese vs. Normal and Overweight for BMI, and Low and Very High vs. Normal and High for Body Fat Percentage ($\frac{\mu 2 + \mu 4}{2} = \frac{\mu 1 + \mu 3}{2}$).
Alternative Hypothesis (Ha): There is a significant difference in the means between these combined groups ($\frac{\mu 2 + \mu 4}{2} \neq \frac{\mu 1 + \mu 3}{2}$).

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall Underweight and Obese vs Normal and Overweight Effect
H = Contrast SSCP Matrix for Underweight and Obese vs Normal and Overweight
E = Error SSCP Matrix

S=1 M=4.5 N=118

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.97216494	0.62	11	238	0.8115
Pillai's Trace	0.02783506	0.62	11	238	0.8115
Hotelling-Lawley Trace	0.02863203	0.62	11	238	0.8115
Roy's Greatest Root	0.02863203	0.62	11	238	0.8115

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall Low and Very High vs Normal and High Effect
H = Contrast SSCP Matrix for Low and Very High vs Normal and High
E = Error SSCP Matrix

S=1 M=5.5 N=117

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.94036864	1.15	13	236	0.3171
Pillai's Trace	0.05963136	1.15	13	236	0.3171
Hotelling-Lawley Trace	0.06341275	1.15	13	236	0.3171
Roy's Greatest Root	0.06341275	1.15	13	236	0.3171

Based on the MANOVA results, neither the BMI contrast between Underweight and Obese vs. Normal and Overweight, nor the Body Fat Percentage contrast between Low and Very High vs. Normal and High, shows statistically significant overall multivariate effects on the dependent variables. This is supported by all four multivariate test statistics (Wilks' Lambda, Pillai's Trace, Hotelling-Lawley Trace, and Roy's Greatest Root), each having p-values above the 0.05 significance ($p = 0.8115$ and $p = 0.3171$). Therefore, we fail to reject the null hypothesis in both cases. The results suggest that differences in the means between these combined groups in BMI and Body Fat Percentage do not differ.

The last and final hypothesis that we tested is:
Null Hypothesis (H0): There is no significant difference in the means of body measurements ($\mu1 = \mu3$).
Alternative Hypothesis (Ha): There is a significant difference in the means of body measurements ($\mu1 \neq \mu3$).

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall Normal vs Overweight Effect
H = Contrast SSCP Matrix for Normal vs Overweight
E = Error SSCP Matrix

S=1 M=4.5 N=118

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.43868226	27.68	11	238	<.0001
Pillai's Trace	0.56131774	27.68	11	238	<.0001
Hotelling-Lawley Trace	1.27955423	27.68	11	238	<.0001
Roy's Greatest Root	1.27955423	27.68	11	238	<.0001

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall Normal vs High Effect
H = Contrast SSCP Matrix for Normal vs High
E = Error SSCP Matrix

S=1 M=5.5 N=117

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.71355803	7.29	13	236	<.0001
Pillai's Trace	0.28644197	7.29	13	236	<.0001
Hotelling-Lawley Trace	0.40142771	7.29	13	236	<.0001
Roy's Greatest Root	0.40142771	7.29	13	236	<.0001

Based on the MANOVA results, both the BMI group contrast (Normal vs. Overweight) and the Body Fat Percentage group contrast (Normal vs. High) show statistically significant overall multivariate effects on the dependent variables. This is supported by all four multivariate test statistics (Wilks' Lambda, Pillai's Trace, Hotelling-Lawley Trace, and Roy's Greatest Root), each having p-values less than 0.0001, which is well below the 0.05 significance. Therefore, we reject the null hypothesis that there are no differences between the groups. The results indicate that differences between BMI categories (Normal vs. Overweight) and Body Fat Percentage levels (Normal vs. High) lead to statistically significant variations across the dependent variables.

This study used MANOVA to investigate whether body measurements differ based on BMI and Body Fat Percentage (BFP) classifications. The results showed significant differences across BMI and BFP groups overall, supporting the first hypothesis. While the second hypothesis, comparing combined extreme categories, was not supported, the third hypothesis, comparing Normal vs. Overweight (BMI) and Normal vs. High (BFP), showed significant differences in body measurements. These findings suggest that even moderate changes in BMI and BFP are associated with meaningful differences in body composition. Some limitations of this study include possible issues with MANOVA assumptions, leaving out certain variables to avoid overlap, and combining groups in ways that may have missed important differences. Future research should look at more variables and use other methods to get a clearer understanding.

PCA

Principal Component Analysis(PCA) is a dimensionality reduction technique that transforms a set of correlated variables into a smaller set of uncorrelated components, called principal components, which capture the maximum variance in the data. In this analysis, we apply PCA using both the correlation and covariance matrices to explore the dominant dimensions of body composition and identify meaningful groupings based on variable loadings and explained variance.

As we can see from the table below, the variables exhibit a wide range of standard deviations. For example, Density has the smallest standard deviation(0.0190), while Age(12.60) and Abdomen(10.78) have much larger spreads. Because of this imbalance, using the correlation matrix(R) for PCA is more appropriate, as it standardizes the variables and ensures no single variable disproportionately influences the principal components.

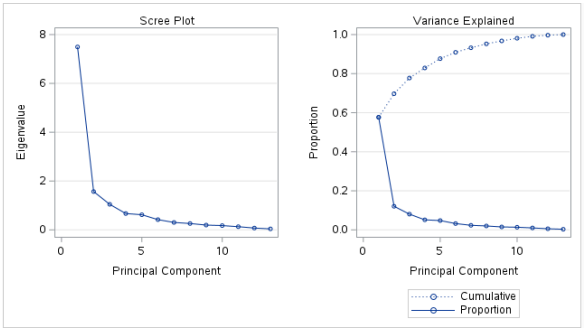
Simple Statistics

	Density	Age	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
Mean	1.055573810	44.88492063	70.14880952	37.99206349	100.8242063	92.55595238	99.90476190	59.40595238	38.59047619	23.10238095	32.27341270	28.66388889	18.22976190
StD	0.019031434	12.60203972	3.66285579	2.43091323	8.4304755	10.78307680	7.16405767	5.24995203	2.41180459	1.69489340	3.02127375	2.02069117	0.93358493

Looking at the eigenvalues of the R matrix, the first four components account for 82.91% of the variance (PC1: 57.64%, PC2: 12.07%, PC3: 8.05%, PC4: 82.91%). In the scree plot, the slope flattens from PC5 onwards, implying only the first 4 PCs should be retained.

Eigenvalues of the Correlation Matrix

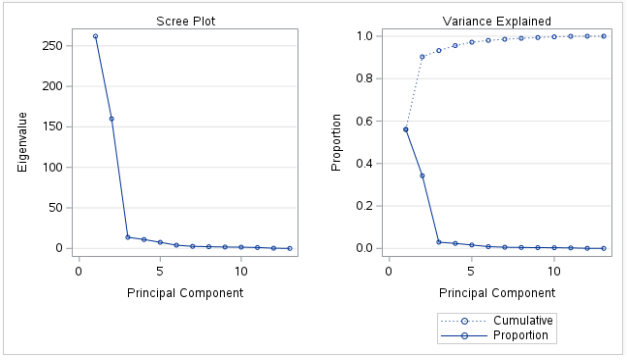
	Eigenvalue	Difference	Proportion	Cumulative
1	7.49341024	5.92417817	0.5764	0.5764
2	1.56923207	0.52239783	0.1207	0.6971
3	1.04683424	0.37812597	0.0805	0.7777
4	0.66870827	0.04877798	0.0514	0.8291
5	0.61993029	0.20035324	0.0477	0.8768
6	0.41957705	0.11578131	0.0323	0.9091
7	0.30379574	0.04329177	0.0234	0.9324
8	0.26050397	0.06420590	0.0200	0.9525
9	0.19629806	0.02089122	0.0151	0.9676
10	0.17540684	0.04497860	0.0135	0.9811
11	0.13042824	0.05546233	0.0100	0.9911
12	0.07496591	0.03405683	0.0058	0.9969
13	0.04090908		0.0031	1.0000



Looking at the eigenvalues of the S matrix, the first two components account for 90.30% of the variance (PC1: 56.06%, PC2: 34.24%). In the scree plot, the slope flattens from PC3 onwards, implying only the first 2 PCs should be retained.

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	261.882295	101.915817	0.5606	0.5606
2	159.966479	146.255682	0.3424	0.9030
3	13.710796	2.757493	0.0294	0.9324
4	10.953304	3.454186	0.0234	0.9558
5	7.499117	3.562938	0.0161	0.9719
6	3.936180	1.394004	0.0084	0.9803
7	2.542175	0.460743	0.0054	0.9858
8	2.081432	0.386975	0.0045	0.9902
9	1.694458	0.166331	0.0036	0.9938
10	1.528126	0.430202	0.0033	0.9971
11	1.097925	0.846975	0.0024	0.9995
12	0.250949	0.250853	0.0005	1.0000
13	0.000096		0.0000	1.0000



Based on the scree plot for R matrix, we retained the first four principal components, which together explain a substantial portion of the total variance in the dataset. The first principal component (PC1) is dominated by positive loadings from variables such as chest, abdomen, hip, thigh, neck, biceps, and forearm, while body density has a strong negative loading. This suggests that PC1 primarily captures overall body size and fat distribution, with individuals scoring higher on this component likely having larger body measurements and lower body density-consistent with higher body fat levels. The second component (PC2) shows strong negative loadings for age and height, and moderate negative contributions from abdomen and chest, while biceps and forearm load positively. This component appears to reflect a contrast between age/height and upper-body muscularity, potentially distinguishing older, taller individuals from younger individuals with more upper-arm mass. The third component (PC3) is mostly influenced by a large positive loading for age, followed by moderate positive loadings from height and neck, and slight negative loadings for hip and abdomen. This component appears to reflect an age-related body structure, contrasting older individuals with relatively leaner lower-body profiles. Lastly, the fourth component (PC4) has its strongest loadings from forearm and biceps, with moderate influence from wrist and neck, possibly separating individuals with more muscular arms from those with leaner upper limbs.

Eigenvectors													
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12	Prin13
Density	-.241999	0.424896	0.196224	0.137898	0.271062	0.541922	0.092274	-.019810	0.409854	0.298371	0.145745	0.013756	0.221094
Age	0.022754	-.593661	0.602329	-.047815	0.092334	-.011464	0.271564	0.201123	0.226326	-.053475	0.306890	0.073622	-.073645
Height	0.092186	0.502024	0.475489	-.234080	-.634452	-.186001	-.013218	0.073974	0.017538	0.006824	0.116762	0.040019	-.015178
Neck	0.314768	0.005188	0.166882	0.218095	-.024949	0.312422	-.456665	-.119783	0.267005	-.640181	-.106109	-.091314	-.060305
Chest	0.330860	-.166477	0.002199	0.010870	-.109365	0.027334	-.275568	-.139133	0.310579	0.554359	-.287096	0.409786	-.317127
Abdomen	0.330345	-.252530	-.058859	-.129704	-.177834	0.008989	-.131241	-.141941	0.057927	0.192543	0.073034	-.173708	0.813511
Hip	0.336420	-.001858	-.198887	-.165191	-.068991	0.219818	0.088020	-.108659	0.035944	0.196917	0.411509	-.604484	-.415369
Thigh	0.320215	0.092966	-.328513	-.068963	-.035891	0.211959	0.192442	0.062212	-.089912	-.207307	0.486547	0.632003	0.042515
Knee	0.317077	0.086341	-.002963	-.245259	0.040885	0.115911	0.668742	-.097662	0.154254	-.172058	-.551280	-.044132	0.027038
Ankle	0.234135	0.255361	0.069232	-.474760	0.617742	-.396130	-.247303	0.120431	0.161762	-.039318	0.085264	0.002191	0.024641
Biceps	0.309522	0.081800	-.066416	0.319485	-.004267	0.038693	-.004575	0.856552	-.047726	0.122815	-.139305	-.127472	0.052806
Forearm	0.257947	0.174411	0.045265	0.657421	0.106201	-.499234	0.246813	-.302716	0.147360	0.031807	0.174798	-.038255	0.031337
Wrist	0.285189	0.071199	0.427258	0.086791	0.265784	0.242538	-.042010	-.186531	-.725294	0.157649	-.068691	0.043114	-.003152

Based on the scree plot for the S matrix, we retained the first two principal components, which together explain a substantial portion of the total variance in the dataset. The first component (PC1) is dominated by large positive loadings from abdomen, chest, and hip, with additional contributions from thigh and biceps. This component reflects overall body size and fat accumulation. The second component (PC2) shows a strong negative loading for age, along with moderate negative contributions from hip, thigh, and ankle. This suggests that PC2 captures an age-related contrast in lower-body composition, distinguishing younger individuals with leaner legs and hips from older individuals. The third component (PC3) is mostly influenced by height, followed by moderate contributions from biceps and neck, indicating that it is related to stature and upper-body muscularity. Lastly, the fourth component (PC4) is shaped primarily by the forearm and biceps, with moderate influence from wrist and neck, suggesting that it reflects upper-limb strength and muscular development, helping distinguish individuals with more muscular arms from those with leaner upper limbs.

Eigenvectors													
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12	Prin13
Density	-.000894	-.000090	0.001182	0.000317	0.001236	-.001684	0.000572	-.000101	0.001263	0.001643	0.000915	0.003994	0.999986
Age	0.237533	0.947395	0.079856	0.167440	0.084045	-.026430	-.032865	-.025340	0.003895	-.018082	-.038114	-.018141	0.000150
Height	0.024280	-.067773	0.921658	0.008125	-.327573	0.003174	-.104155	-.132601	-.015387	-.048411	-.079349	-.025986	-.000345
Neck	0.119381	-.027260	0.139447	-.009784	0.170091	0.203077	0.166592	0.062949	0.629440	0.630189	0.207901	-.164378	-.001379
Chest	0.493329	-.073774	0.095867	-.643388	0.407329	-.333711	-.208120	-.071267	-.026590	-.039180	-.010567	-.004259	-.000304
Abdomen	0.653157	-.055664	-.215071	-.120893	-.600851	0.364432	0.055978	0.097201	-.050876	-.005239	0.002767	0.019370	0.002198
Hip	0.392994	-.197599	-.011221	0.505659	0.004221	-.572731	0.423447	-.180271	0.066206	-.033705	-.068090	-.026328	-.000904
Thigh	0.252193	-.197876	-.040953	0.486779	0.264337	0.238374	-.683479	-.101306	0.111414	-.036764	-.200814	0.020489	0.000578
Knee	0.115540	-.047298	0.127480	0.194593	0.083254	-.065967	-.122158	0.449727	-.155035	-.186150	0.800150	-.053921	-.000228
Ankle	0.050936	-.037590	0.103433	0.075159	0.089263	-.090739	0.028687	0.712443	-.339697	0.348957	-.450119	-.123586	0.000450
Biceps	0.136226	-.069334	0.132166	0.076692	0.403485	0.469609	0.358762	-.322967	-.555307	0.139925	0.093418	-.016226	0.000440
Forearm	0.068047	-.043679	0.125079	-.024320	0.277466	0.309112	0.342445	0.298018	0.361473	-.641050	-.228734	-.068389	0.001007
Wrist	0.039644	-.000768	0.086087	0.024599	0.072000	0.019045	0.075474	0.132561	0.070721	0.095875	0.007223	0.973310	-.004302

Both the correlation-based and covariance-based PCA analyses reveal meaningful insights into the underlying structure of the dataset, though they differ in the number of components retained and the emphasis of interpretation. The correlation matrix PCA, which adjusts for differences in variable scale, suggests retaining the first four principal components, capturing approximately 83% of the total variance. These components reflect key dimensions related to body size and fat distribution, age-related variation, and upper-limb muscularity, offering a more balanced representation across all variables. In contrast, the covariance matrix PCA, which preserves the raw scale of variables, recommends retaining only two components, which account for over 90% of the total variance, heavily driven by high-variance variables such as age and abdomen. While both approaches highlight similar body composition trends, the correlation-based PCA provides a more interpretable and equitable summary, especially when variables differ significantly in their measurement scales. Therefore, for comprehensive pattern recognition and interpretation, the correlation-based analysis may be more appropriate in this context.

Canonical Correlation

Canonical Correlation Analysis (CCA) examines the linear relationship between multiple response and predictor variables. The goal of CCA is to simplify complex datasets while retaining significant relationships. In our study of the Body Fat dataset, the dataset contained 15 data features. However, many of these features may not add statistical significance to the model.

Using the Wilk's Lambda test, there are 2 significant canonical correlations between X and Y variables. By examining the standardized canonical vectors, it was revealed that response vectors Density and Body Fat have a high correlation with predictors chest, abdomen, and hip sizes. In a second canonical variate pair, it shows that there is a relationship between Body Fat and BMI using the predictors height, abdomen, and hip sizes. The third canonical variate pair showed weaker correlations between the response and predictor variables.

These standardized canonical vectors can therefore be used to build a more compact/efficient multivariate model for further examination. The first set of canonical vectors showed that we can reduce the model from 3 response variables to 2 variables and from 12 predictor variables to 3 variables. The second set of canonical vectors revealed a relationship between two response variables with 3 predictor variables.

The SaS output and code for this analysis are located in the Appendix, under this topic.

Significance Testing for Canonical Correlation ρ1, ρ2, ρ3

(n = 252, p = 3, q = 12, V(H) = q = 12, V(E) = n – q – 1 = 252 – 12 – 1= 239)

Ho: ρ1 = 0 vs. Ha: ρ1 ≠ 0

Wilks Λ1 = Π (1-ri^2) = (1- 0.841119) (1- 0.610597) (1- 0.052539)

= (0.15888)(0.3894)(.94746) = **0.0586**

Λα (p, Vh, Ve) = Λ(.05) (p, Vh, Ve) = Λ(.05) = (3, 12, 239) = ~ Λ(.05) (3, 12, 240) = 0.811

Λ1= **0.0586** < Λα = 0.811 à conclude Reject Ho. There is a significant linear relationship between the Y and X variables on at least one dimension using the first canonical correlation.

Ho: ρ2 = 0 vs. Ha: ρ2 ≠ 0

Wilks Λ2 = Π (1-ri^2), i=2 = (1- 0.610597) (1- 0.052539) = (0.3894)(.94746) = 0.3689

Λα (p-k+1, q-k+1, n-k-q) = @k=2 à Λ(.05) (2, 11, 238) = ~ Λ(.05) (2, 11, 240) = 0.87

Λ2= 0.3689< Λα = 0.87à conclude Reject Ho. There is a significant linear relationship between the Y and X variables on at least one dimension using the second canonical correlation.

Ho: ρ3 = 0 vs. Ha: ρ3 ≠ 0

Wilks Λ3 = Π (1-ri^2), i=3 = (1- 0.052539) = (.94746) = 0.94746

Λα (p-k+1, q-k+1, n-k-q) = @k=3 Λ(.05) (1, 10, 237) = ~ Λ(.05) (1, 10, 240) = 0.928

Λ2= 0.94746 > Λα = 0.928 Do Not Reject Ho. The third canonical correlation is not significant.

Canonical Vectors Response

μ1=-0.188 Density + 0.325 BodyFat + 0.690 BMI

μ2=-0.0117 Density - 0.959 BodyFat + 0.827 BMI

μ3=6.415 Density + 6.356 BodyFat + 0.0038 BMI

Canonical Vectors Predictors

V1 = -.0117 Age - .6675 Height – .0695 Neck - .0271 Chest + .5555 Abdomen + .1770 Hip

-.0178 Thigh + .2237 Knee + .0364 Ankle + .0373 Biceps + .0807 Forearm - .06660 Wrist

V2 = -.2340 Age - .8138 Height + .2428 Neck + .1566 Chest – 1.6194 Abdomen + .8417 Hip

-.3098 Thigh + .3658 Knee + .0028 Ankle - .0487 Biceps - .1027 Forearm + .3013 Wrist

V3 = .2672 Age + .1247 Height + .2355 Neck + 1.238 Chest – 1.7777 Abdomen + 1.5131 Hip

-.5945 Thigh + .1200 Knee - .5201 Ankle - .6376 Biceps - .0105 Forearm + .5010 Wrist

References

Rencher, Alvin C. Methods of Multivariate Analysis. New York, J. Wiley, 2002.

Yu Yue PhD. Chapter 6: Multivariate Analysis of Variance (MANOVA), New York, 2025.

Yu Yue PhD. Chapter 10: Multivariate Multiple Regression, New York, 2025.

Yu Yue PhD. Chapter 11: Canonical Correlation, New York, 2025.

Yu Yue PhD. Chapter 12: Principal Component Analysis, New York, 2025.

STA -9705 – Homework's 1 through 6 solutions

SAS Institute Inc. (2024). SAS/STAT 15.3 User's Guide. Cary, NC: SAS Institute Inc.

Body Fat Prediction Dataset (Kaggle)
https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset?resource=download

Appendix

SAS Code and Output:

Multivariate Regression - Correlation Matrix - MR1

Pearson Correlation Coefficients, N = 251
Prob > |r| under H0: Rho=0

	BodyFat	BMI	Age	Weight	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
BodyFat	1.00000	0.36818 <.0001	0.29106 <.0001	0.60505 <.0001	0.48274 <.0001	0.69561 <.0001	0.80973 <.0001	0.61798 <.0001	0.55003 <.0001	0.49902 <.0001	0.25446 <.0001	0.48675 <.0001	0.34947 <.0001	0.33543 <.0001
BMI	0.36818 <.0001	1.00000	0.03868 0.5419	0.38762 <.0001	0.26250 <.0001	0.38041 <.0001	0.41221 <.0001	0.45974 <.0001	0.43030 <.0001	0.36083 <.0001	0.20615 0.0010	0.30772 <.0001	0.21053 0.0008	0.18558 0.0032
Age	0.29106 <.0001	0.03868 0.5419	1.00000	-0.01608 0.7999	0.11153 0.0778	0.17484 0.0055	0.22927 0.0002	-0.05404 0.3939	-0.20600 0.0010	0.01438 0.8207	-0.10838 0.0866	-0.04372 0.4905	-0.08892 0.1602	0.21218 0.0007
Weight	0.60505 <.0001	0.38762 <.0001	-0.01608 0.7999	1.00000	0.82847 <.0001	0.89238 <.0001	0.88600 <.0001	0.93986 <.0001	0.86626 <.0001	0.85058 <.0001	0.60832 <.0001	0.79838 <.0001	0.62409 <.0001	0.72566 <.0001
Neck	0.48274 <.0001	0.26250 <.0001	0.11153 0.0778	0.82847 <.0001	1.00000	0.78209 <.0001	0.75066 <.0001	0.73129 <.0001	0.69125 <.0001	0.66777 <.0001	0.47165 <.0001	0.72838 <.0001	0.61847 <.0001	0.74155 <.0001
Chest	0.69561 <.0001	0.38041 <.0001	0.17484 0.0055	0.89238 <.0001	0.78209 <.0001	1.00000	0.91426 <.0001	0.82610 <.0001	0.72337 <.0001	0.71361 <.0001	0.47448 <.0001	0.72525 <.0001	0.57172 <.0001	0.65428 <.0001
Abdomen	0.80973 <.0001	0.41221 <.0001	0.22927 0.0002	0.88600 <.0001	0.75066 <.0001	0.91426 <.0001	1.00000	0.87178 <.0001	0.76192 <.0001	0.73233 <.0001	0.44523 <.0001	0.68139 <.0001	0.49460 <.0001	0.61379 <.0001
Hip	0.61798 <.0001	0.45974 <.0001	-0.05404 0.3939	0.93986 <.0001	0.73129 <.0001	0.82610 <.0001	0.87178 <.0001	1.00000	0.89448 <.0001	0.82032 <.0001	0.55216 <.0001	0.73643 <.0001	0.53728 <.0001	0.62436 <.0001
Thigh	0.55003 <.0001	0.43030 <.0001	-0.20600 0.0010	0.86626 <.0001	0.69125 <.0001	0.72337 <.0001	0.76192 <.0001	0.89448 <.0001	1.00000	0.79523 <.0001	0.53276 <.0001	0.75910 <.0001	0.55870 <.0001	0.55120 <.0001
Knee	0.49902 <.0001	0.36083 <.0001	0.01438 0.8207	0.85058 <.0001	0.66777 <.0001	0.71361 <.0001	0.73233 <.0001	0.82032 <.0001	0.79523 <.0001	1.00000	0.60611 <.0001	0.67504 <.0001	0.54824 <.0001	0.65927 <.0001
Ankle	0.25446 <.0001	0.20615 0.0010	-0.10838 0.0866	0.60832 <.0001	0.47165 <.0001	0.47448 <.0001	0.44523 <.0001	0.55216 <.0001	0.53276 <.0001	0.60611 <.0001	1.00000	0.47950 <.0001	0.41101 <.0001	0.56064 <.0001
Biceps	0.48675 <.0001	0.30772 <.0001	-0.04372 0.4905	0.79838 <.0001	0.72838 <.0001	0.72525 <.0001	0.68139 <.0001	0.73643 <.0001	0.75910 <.0001	0.67504 <.0001	0.47950 <.0001	1.00000	0.67463 <.0001	0.62810 <.0001
Forearm	0.34947 <.0001	0.21053 0.0008	-0.08892 0.1602	0.62409 <.0001	0.61847 <.0001	0.57172 <.0001	0.49460 <.0001	0.53728 <.0001	0.55870 <.0001	0.54824 <.0001	0.41101 <.0001	0.67463 <.0001	1.00000	0.57935 <.0001
Wrist	0.33543 <.0001	0.18558 0.0032	0.21218 0.0007	0.72566 <.0001	0.74155 <.0001	0.65428 <.0001	0.61379 <.0001	0.62436 <.0001	0.55120 <.0001	0.65927 <.0001	0.56064 <.0001	0.62810 <.0001	0.57935 <.0001	1.00000

Multivariate Regression - SAS Output BodyFat Model Summary - MR2

The GLM Procedure

Dependent Variable: BodyFat

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	12723.40070	1156.67279	61.61	<.0001
Error	239	4487.37508	18.77563		
Corrected Total	250	17210.77578			

R-Square	Coeff Var	Root MSE	BodyFat Mean
0.739269	22.53635	4.333085	19.22709

Multivariate Regression - SAS Output BMI Model Summary - MR3

The GLM Procedure

Dependent Variable: BMI

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	5866.87649	533.35241	7.50	<.0001
Error	239	17002.43626	71.13990		
Corrected Total	250	22869.31275			

R-Square	Coeff Var	Root MSE	BMI Mean
0.256539	32.47294	8.434447	25.97377

Multivariate Regression - Stepwise Regression for BodyFat - MR4

Stepwise Selection for BodyFat

The REG Procedure
Model: MODEL1
Dependent Variable: BodyFat

Number of Observations Read	251
Number of Observations Used	251

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	12578	3144.48812	166.97	<.0001
Error	246	4632.82329	18.83262		
Corrected Total	250	17211			

Root MSE	4.33966	R-Square	0.7308
Dependent Mean	19.22709	Adj R-Sq	0.7264
Coeff Var	22.57053		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-34.02579	7.27491	-4.68	<.0001	0
Weight	1	-0.13505	0.02474	-5.46	<.0001	6.92454
Abdomen	1	0.99211	0.05611	17.68	<.0001	4.78959
Forearm	1	0.46027	0.18186	2.53	0.0120	1.77072
Wrist	1	-1.51718	0.44247	-3.43	0.0007	2.24298

Multivariate Regression - Transformed BMI Output - MR5

The REG Procedure
Model: MODEL1
Dependent Variable: logBMI

Number of Observations Read	251
Number of Observations Used	251

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4.51060	1.50353	105.86	<.0001
Error	247	3.50816	0.01420		
Corrected Total	250	8.01876			

Root MSE	0.11918	R-Square	0.5625
Dependent Mean	3.23356	Adj R-Sq	0.5572
Coeff Var	3.68562		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.44219	0.14916	9.67	<.0001
Hip	1	0.02015	0.00140	14.41	<.0001
Wrist	1	-0.01746	0.01094	-1.59	0.1120
Age	1	0.00212	0.00063011	3.36	0.0009

Multivariate Regression - SAS Code

```
/* Step 4: Correlation matrix + scatterplot matrix */
ods graphics on;
proc corr data=bodyfat_clean plots=matrix(histogram);
  var BodyFat BMI Age Weight Neck Chest Abdomen Hip Thigh Knee Ankle Biceps Forearm Wrist;
run;

/* Step 5: Multivariate regression (GLM) */
proc glm data=bodyfat_clean plots=all;
  model BodyFat BMI = Abdomen Wrist Forearm Age Neck Chest Hip Thigh Knee Ankle Biceps / solution;
run;

/* Stepwise selection for BodyFat */
proc reg data=bodyfat_clean;
  model BodyFat = Age Weight Neck Chest Abdomen Hip Thigh Knee Ankle Biceps Forearm Wrist
        / selection=stepwise slentry=0.05 slstay=0.05 vif;
  title "Stepwise Selection for BodyFat";
run;
quit;

proc reg data=bodyfat_clean;
  model BMI = Age Neck Chest Abdomen Hip Thigh Knee Ankle Biceps Forearm Wrist
        / selection=stepwise slentry=0.05 slstay=0.05 vif;
run;


data bodyfat_clean;
  set bodyfat_clean;
  logBMI = log(BMI);
run;

proc reg data=bodyfat_clean;
  model logBMI = Hip Wrist Age;
run;
```

MANOVA

MANOVA Tests for the Hypothesis of No Overall BMI_GROUP Effect
H = Type III SSCP Matrix for BMI_GROUP
E = Error SSCP Matrix

S=3 M=3.5 N=118

Statistic	Value	P-Value
Wilks' Lambda	0.23906523	<.0001
Pillai's Trace	0.86607531	<.0001
Hotelling-Lawley Trace	2.75031711	<.0001
Roy's Greatest Root	2.58691049	<.0001

MANOVA Tests for the Hypothesis of No Overall BODYFAT_GROUP Effect
H = Type III SSCP Matrix for BODYFAT_GROUP
E = Error SSCP Matrix

S=3 M=4.5 N=117

Statistic	Value	P-Value
Wilks' Lambda	0.32161183	<.0001
Pillai's Trace	0.73876014	<.0001
Hotelling-Lawley Trace	1.92519348	<.0001
Roy's Greatest Root	1.82745185	<.0001

MANOVA CONTRAST

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall Underweight and Obese vs Normal and Overweight Effect
H = Contrast SSCP Matrix for Underweight and Obese vs Normal and Overweight
E = Error SSCP Matrix

S=1 M=4.5 N=118

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.97216494	0.62	11	238	0.8115
Pillai's Trace	0.02783506	0.62	11	238	0.8115
Hotelling-Lawley Trace	0.02863203	0.62	11	238	0.8115
Roy's Greatest Root	0.02863203	0.62	11	238	0.8115

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall Normal vs Overweight Effect
H = Contrast SSCP Matrix for Normal vs Overweight
E = Error SSCP Matrix

S=1 M=4.5 N=118

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.43868226	27.68	11	238	<.0001
Pillai's Trace	0.56131774	27.68	11	238	<.0001
Hotelling-Lawley Trace	1.27955423	27.68	11	238	<.0001
Roy's Greatest Root	1.27955423	27.68	11	238	<.0001

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall Low and Very High vs Normal and High Effect
H = Contrast SSCP Matrix for Low and Very High vs Normal and High
E = Error SSCP Matrix

S=1 M=5.5 N=117

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.94036864	1.15	13	236	0.3171
Pillai's Trace	0.05963136	1.15	13	236	0.3171
Hotelling-Lawley Trace	0.06341275	1.15	13	236	0.3171
Roy's Greatest Root	0.06341275	1.15	13	236	0.3171

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall Normal vs High Effect
H = Contrast SSCP Matrix for Normal vs High
E = Error SSCP Matrix

S=1 M=5.5 N=117

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.71355803	7.29	13	236	<.0001
Pillai's Trace	0.28644197	7.29	13	236	<.0001
Hotelling-Lawley Trace	0.40142771	7.29	13	236	<.0001
Roy's Greatest Root	0.40142771	7.29	13	236	<.0001

MANOVA SAS CODE

```
DATA BODYFATDATA;
    INFILE '/home/u64147743/HW/bodyfat.csv' DSD FIRSTOBS=2 DELIMITER=',';
    INPUT DENSITY BODYFAT AGE WEIGHT HEIGHT NECK CHEST ABDOMEN HIP THIGH KNEE ANKLE BICEPS FOREARM WRIST;
    BMI = (WEIGHT * 703) / (HEIGHT**2);
RUN;


DATA BODYFAT_GROUPED;
  SET BODYFATDATA;
  LENGTH BMI_GROUP $12.;
  IF BMI < 18.5 THEN BMI_GROUP = 'Underweight';
  ELSE IF BMI < 25 THEN BMI_GROUP = 'Normal';
  ELSE IF BMI < 30 THEN BMI_GROUP = 'Overweight';
  ELSE BMI_GROUP = 'Obese';
RUN;


PROC GLM DATA=BODYFAT_GROUPED;
  CLASS BMI_GROUP;
  MODEL AGE NECK CHEST ABDOMEN HIP THIGH KNEE ANKLE BICEPS FOREARM WRIST = BMI_GROUP;
  MANOVA H=BMI_GROUP / PRINTH PRINTE MSTAT=EXACT;
RUN;


PROC GLM DATA=BODYFAT_GROUPED;
  CLASS BMI_GROUP;
  MODEL AGE NECK CHEST ABDOMEN HIP THIGH KNEE ANKLE BICEPS FOREARM WRIST = BMI_GROUP;
  CONTRAST 'Underweight and Obese vs Normal and Overweight'
    BMI_GROUP -1 1 -1 1;
  CONTRAST 'Normal vs Overweight'
    BMI_GROUP 1 0 -1 0;
  MANOVA H= BMI_GROUP / PRINTE PRINTH;
RUN;
```

```sas
/* BODY FAT % */
DATA BODYFAT_GROUPED_FAT;
  SET BODYFATDATA;
  LENGTH BODYFAT_GROUP $12.;

  /* Body fat % groupings for men */
  IF BODYFAT < 6 THEN BODYFAT_GROUP = 'Low';
  ELSE IF BODYFAT < 18 THEN BODYFAT_GROUP = 'Normal';
  ELSE IF BODYFAT < 25 THEN BODYFAT_GROUP = 'High';
  ELSE BODYFAT_GROUP = 'Very High';
RUN;

PROC GLM DATA=BODYFAT_GROUPED_FAT;
  CLASS BODYFAT_GROUP;
  MODEL AGE WEIGHT HEIGHT NECK CHEST ABDOMEN HIP THIGH KNEE ANKLE BICEPS FOREARM WRIST = BODYFAT_GROUP;
  MANOVA H=BODYFAT_GROUP / PRINTH PRINTE MSTAT=EXACT;
RUN;


PROC GLM DATA=BODYFAT_GROUPED_FAT;
  CLASS BODYFAT_GROUP;
  MODEL AGE WEIGHT HEIGHT NECK CHEST ABDOMEN HIP THIGH KNEE ANKLE BICEPS FOREARM WRIST = BODYFAT_GROUP;
  CONTRAST 'Low and Very High vs Normal and High'
    BODYFAT_GROUP -1 1 -1 1;
  CONTRAST 'Normal vs High'
    BODYFAT_GROUP -1 0 1 0;
  MANOVA H=BODYFAT_GROUP / PRINTH PRINTE;
RUN;
```

Canonical Correlation:

Body Fat Analysis - Canonical Correlation

The CANCORR Procedure

Canonical Correlation Analysis

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)				Test of H0: The canonical correlations in the current row and all that follow are zero				
					Eigenvalue	Difference	Proportion	Cumulative	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.917125	0.912631	0.010029	0.841119	5.2940	3.7260	0.7653	0.7653	0.05861841	31.39	36	700.97	<.0001
2	0.781407	0.771856	0.024579	0.610597	1.5680	1.5126	0.2267	0.9920	0.38894437	13.98	22	476	<.0001
3	0.229214	0.156318	0.059803	0.052539	0.0555		0.0080	1.0000	0.94746091	1.33	10	239	0.2174

Multivariate Statistics

S=3 M=4 N=117.5

Statistic	Value	P-Value
Wilks' Lambda	0.05861841	<.0001
Pillai's Trace	1.50425438	<.0001
Hotelling-Lawley Trace	6.91748557	<.0001
Roy's Greatest Root	5.29400119	<.0001

Standardized Canonical Coefficients for the VAR Variables

	V1	V2	V3
Age	-0.0117	-0.2340	0.2672
Height	-0.6675	-0.8138	0.1247
Neck	-0.0695	0.2428	0.2355
Chest	-0.0271	0.1566	1.2380
Abdomen	0.5555	-1.6194	-1.7777
Hip	0.1770	0.8417	1.5131
Thigh	-0.0178	-0.3098	-0.5945
Knee	0.2237	0.3658	0.1200
Ankle	0.0364	0.0028	-0.5201
Biceps	0.0373	-0.0487	-0.6376
Forearm	0.0807	-0.1027	0.0105
Wrist	-0.0660	0.3013	0.5010

Standardized Canonical Coefficients for the WITH Variables

	W1	W2	W3
Density	-0.1883	-0.0117	6.4150
BodyFat	0.3254	-0.9593	6.3566
BMI	0.6901	0.8271	0.0038

Canonical Correlation SAS Code:

```
/* Canonical Correlation */

DATA BODYFAT;
  INFILE '/home/u64138820/PW-STA9705/kaggle_bodyfat.csv' dsd FIRSTOBS=2;
  INPUT Density BodyFat Age Weight Height Neck Chest Abdomen Hip Thigh Knee Ankle Biceps Forearm Wrist;
  BMI_num = Weight*703;
  BMI_den = ((Height/12)**2);
  BMI = BMI_num/BMI_den;


PROC CANCORR ALL MSTAT=exact;
  WITH Density BodyFat BMI;  /*Y-variables */
  VAR Age Height Neck Chest Abdomen Hip Thigh Knee Ankle Biceps Forearm Wrist;  /* predictor X variables */
  TITLE 'Body Fat Analysis - Canonical Correlation';
RUN;
```

PCA Additional Output(Correlation & Covariance Matrices):

Correlation Matrix

	Density	Age	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
Density	1.0000	-.2776	0.0979	-.4730	-.6826	-.7990	-.6093	-.5531	-.4950	-.2649	-.4871	-.3516	-.3257
Age	-.2776	1.0000	-.1716	0.1135	0.1764	0.2304	-.0503	-.2001	0.0175	-.1051	-.0412	-.0851	0.2135
Height	0.0979	-.1716	1.0000	0.2537	0.1349	0.0878	0.1704	0.1484	0.2861	0.2647	0.2078	0.2286	0.3221
Neck	-.4730	0.1135	0.2537	1.0000	0.7848	0.7541	0.7350	0.6957	0.6724	0.4779	0.7311	0.6237	0.7448
Chest	-.6826	0.1764	0.1349	0.7848	1.0000	0.9158	0.8294	0.7299	0.7195	0.4830	0.7279	0.5802	0.6602
Abdomen	-.7990	0.2304	0.0878	0.7541	0.9158	1.0000	0.8741	0.7666	0.7372	0.4532	0.6850	0.5033	0.6198
Hip	-.6093	-.0503	0.1704	0.7350	0.8294	0.8741	1.0000	0.8964	0.8235	0.5584	0.7393	0.5450	0.6301
Thigh	-.5531	-.2001	0.1484	0.6957	0.7299	0.7666	0.8964	1.0000	0.7992	0.5398	0.7615	0.5668	0.5587
Knee	-.4950	0.0175	0.2861	0.6724	0.7195	0.7372	0.8235	0.7992	1.0000	0.6116	0.6787	0.5559	0.6645
Ankle	-.2649	-.1051	0.2647	0.4779	0.4830	0.4532	0.5584	0.5398	0.6116	1.0000	0.4849	0.4190	0.5662
Biceps	-.4871	-.0412	0.2078	0.7311	0.7279	0.6850	0.7393	0.7615	0.6787	0.4849	1.0000	0.6783	0.6321
Forearm	-.3516	-.0851	0.2286	0.6237	0.5802	0.5033	0.5450	0.5668	0.5559	0.4190	0.6783	1.0000	0.5856
Wrist	-.3257	0.2135	0.3221	0.7448	0.6602	0.6198	0.6301	0.5587	0.6645	0.5662	0.6321	0.5856	1.0000

Covariance Matrix

	Density	Age	Height	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
Density	0.0003622	-0.0665871	0.0068232	-0.0218812	-0.1095189	-0.1639594	-0.0830776	-0.0552616	-0.0227224	-0.0085444	-0.0280083	-0.0135232	-0.0057871
Age	-0.0665871	158.8114052	-7.9230459	3.4771707	18.7462230	31.3100503	-4.5440713	-13.2383561	0.5323658	-2.2439480	-1.5672153	-2.1659252	2.5122036
Height	0.0068232	-7.9230459	13.4165125	2.2590543	4.1654074	3.4683338	4.4713005	2.8543896	2.5270205	1.6435686	2.2997889	1.6923473	1.1013304
Neck	-0.0218812	3.4771707	2.2590543	5.9093392	16.0842168	19.7664219	12.7994403	8.8786132	3.9422349	1.9689831	5.3698678	3.0634971	1.6903567
Chest	-0.1095189	18.7462230	4.1654074	16.0842168	71.0729177	83.2546561	50.0939879	32.3032418	14.6292753	6.9012967	18.5403673	9.8834672	5.1958504
Abdomen	-0.1639594	31.3100503	3.4683338	19.7664219	83.2546561	116.2747453	67.5221229	43.3990680	19.1715709	8.2831730	22.3157963	10.9668891	6.2398022
Hip	-0.0830776	-4.5440713	4.4713005	12.7994403	50.0939879	67.5221229	51.3237223	33.7148321	14.2282129	6.7801081	16.0012426	7.8898141	4.2142003
Thigh	-0.0552616	-13.2383561	2.8543896	8.8786132	32.3032418	43.3990680	33.7148321	27.5619963	10.1189812	4.8031730	12.0782067	6.0133632	2.7382684
Knee	-0.0227224	0.5323658	2.5270205	3.9422349	14.6292753	19.1715709	14.2282129	10.1189812	5.8168014	2.5001024	4.9455625	2.7091766	1.4962208
Ankle	-0.0085444	-2.2439480	1.6435686	1.9689831	6.9012967	8.2831730	6.7801081	4.8031730	2.5001024	2.8726636	2.4828126	1.4351859	0.8959050
Biceps	-0.0280083	-1.5672153	2.2997889	5.3698678	18.5403673	22.3157963	16.0012426	12.0782067	4.9455625	2.4828126	9.1280951	4.1407891	1.7829857
Forearm	-0.0135232	-2.1659252	1.6923473	3.0634971	9.8834672	10.9668891	7.8898141	6.0133632	2.7091766	1.4351859	4.1407891	4.0831928	1.1047045
Wrist	-0.0057871	2.5122036	1.1013304	1.6903567	5.1958504	6.2398022	4.2142003	2.7382684	1.4962208	0.8959050	1.7829857	1.1047045	0.8715808

PCA SAS Code:

```
/* Load CSV Data */
proc import datafile="/home/u64154370/HW assignments/Homework/bodyfat_cleaned_numeric_fixed.dat"
    out=bodyfat
    dbms=dlm
    replace;
    delimiter=' ';
    getnames=yes; /* or yes, depending on whether first row has variable names */
run;

/* PCA using covariance matrix on predictors only */
proc princomp data=bodyfat cov;
    var Density Age Height Neck Chest Abdomen Hip Thigh Knee Ankle Biceps Forearm Wrist;
run;
/* PCA using correlation matrix with output and 2D score plot */
proc princomp data=bodyfat out=results2 plots(ncomp=2)=score(ellipse);
    var Density Age Height Neck Chest Abdomen Hip Thigh Knee Ankle Biceps Forearm Wrist;
run;
/* Print first 3 principal components */
proc print data=results2;
    var Prin1 Prin2 Prin3;
run;
proc contents data=bodyfat;
run;
```