Assignment 5: Applying Canonical Correlation Analysis (CCA) and Multivariate Regression

A. Sepúlveda-Jiménez

Data Science Dept., SoTE, CoBET, National University

DDS-8515: Multivariate Analysis

Course Instructor: Y. Karahan, PhD

November 16, 2025

Contents

List of Figures

Introduction

Canonical Correlation Analysis (CCA) and multivariate multiple regression (MVR) are fundamental multivariate techniques for exploring relationships between two sets of variables and for modelling multiple dependent variables simultaneously. In this paper, we apply CCA and MVR to the Linnerud exercise dataset, which records exercise performance and physiological measurements for 20 individuals. After standardizing the data and conducting exploratory analysis, we estimate canonical correlations between an exercise block (*Chins*, *Situps*, *Jumps*) and a physiological block (*Weight*, *Waist*, *Pulse*), compute canonical loadings and redundancy indices, and fit a multivariate regression model predicting physiological outcomes from exercise variables. The first canonical correlation is substantial ($\rho_1 \approx 0.80$), while subsequent correlations are small, indicating a dominant latent association between overall exercise performance and body composition. The MVR results show that waist circumference is reasonably well-explained by exercise variables, whereas weight and resting pulse exhibit weaker linear relationships. We discuss the complementary roles of CCA and MVR, highlight limitations due to small sample size, and outline practical implications for applied multivariate analysis.

Many empirical problems involve *sets* of correlated variables rather than single predictors and responses. When two such sets are measured on the same units—for example, behavioural and physiological measures, or survey responses and performance metrics—it is often not enough to analyze pairwise correlations or run separate regressions. Canonical Correlation Analysis (CCA), originally introduced by Hotelling (1936), provides a way to study the linear association between two multivariate sets by finding pairs of linear combinations with maximal correlation. Modern treatments emphasize its role as a general framework for multivariate dependence and multi-view learning (Borga, 2001; Uurtio et al., 2017).

Multivariate multiple regression (MVR) generalizes ordinary least squares regression to the case of multiple dependent variables. Instead of fitting separate univariate regressions, MVR models the joint relationship between a predictor set and several correlated responses, allowing formal multivariate tests and potentially more efficient estimation (Izenman, 2008; Johnson & Wichern, 2018). MVR is especially useful when the responses are conceptually related and share explanatory structure.

This paper applies CCA and MVR to the Linnerud exercise dataset, a small but well-known real-world dataset distributed with `scikit-learn` as a canonical example of multivariate regression (Pedregosa et al., 2011). The dataset records three exercise performance measures and three physiological measurements for 20 middle-aged men in a fitness setting (Tenenhaus, 1998). Despite the modest sample size, the data are ideal for illustrating multivariate methods: variables within each block are moderately correlated, the two blocks are substantively linked, and both CCA and MVR are meaningful.

Our goals are to: (a) perform CCA to quantify and interpret the linear relationships between exercise and physiological variables; (b) fit an MVR model predicting physiological measurements from exercise performance; (c) evaluate model performance using appropriate statistical metrics and diagnostics; and (d) compare the insights obtained from CCA and MVR, emphasizing their complementary strengths and limitations.

Dataset and Preprocessing

Linnerud Exercise Dataset

The Linnerud dataset contains $n = 20$ observations on two sets of three continuous variables measured on middle-aged men at a fitness club (Tenenhaus, 1998). In the `scikit-learn` implementation, the *exercise* variables form the feature matrix $\mathbf{X} \in \mathbb{R}^{n\times 3}$ and the *physiological* variables form the target matrix $\mathbf{Y} \in \mathbb{R}^{n\times 3}$ (Pedregosa et al., 2011). The variables are:

- Exercise block (set $\mathcal{X}$): `Chins` (number of chin-ups), `Situps` (number of sit-ups), and `Jumps` (standing jump in centimeters).
- Physiological block (set $\mathcal{Y}$): `Weight` (in pounds), `Waist` (waist size in centimeters),

and `Pulse` (resting pulse rate).

We treat the exercise variables as one multivariate set and the physiological measurements as a second set. For CCA, both sets play symmetric roles, whereas for MVR we view the physiological variables as multivariate responses and the exercise variables as predictors:

$$\mathbf{X} = \begin{bmatrix}\text{Chins} & \text{Situps} & \text{Jumps}\end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix}\text{Weight} & \text{Waist} & \text{Pulse}\end{bmatrix}.$$

Exploratory data analysis using functions analogous to `head()`, `describe()`, and `info()` confirms that there are no missing values and that all six variables are numeric. Sample sizes for multivariate methods are usually recommended to be considerably larger than the total number of variables (Johnson & Wichern, 2018), so with $n = 20$ and $p + q = 6$ this dataset is at the lower edge of what is statistically comfortable. The results should therefore be interpreted as illustrative rather than definitive.

Standardization and Outlier Screening

For both CCA and MVR, variables are standardized prior to analysis. Let $\mathbf{x}_i \in \mathbb{R}^3$ and $\mathbf{y}_i \in \mathbb{R}^3$ denote the exercise and physiological vectors for subject i. We construct standardized versions

$$\tilde{\mathbf{x}}_i = \mathbf{S}_X^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}}), \qquad \tilde{\mathbf{y}}_i = \mathbf{S}_Y^{-1/2}(\mathbf{y}_i - \bar{\mathbf{y}}),$$

where $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are sample mean vectors and $\mathbf{S}_X$ and $\mathbf{S}_Y$ are sample covariance matrices for $\mathcal{X}$ and $\mathcal{Y}$ respectively. In practice we implement this with `StandardScaler` in Python for each block. Standardization puts all variables on comparable scales and is standard practice in CCA and regression when variables are measured in different units (Borga, 2001; Johnson & Wichern, 2018).

We also compute univariate z-scores for each variable and briefly inspect values exceeding $|z| > 3$ as potential outliers. Given the small sample size, no case is removed; instead, outliers are noted as a caveat when interpreting the results.

Correlation matrices within and between the two blocks are summarized and visualized as heatmaps. See Figure 1. Exercise variables are moderately positively correlated with one another, as are the body composition measures (*Weight* and *Waist*). Correlations between exercise and physiological variables show the expected pattern: higher exercise performance tends to be associated with lower weight and waist, with weaker structure for resting pulse.

Methods

Canonical Correlation Analysis

Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\mathbf{Y} \in \mathbb{R}^{n\times q}$ denote the standardized data matrices for the two variable sets, here with $p = q = 3$. CCA seeks vectors $\mathbf{a}_k \in \mathbb{R}^p$ and $\mathbf{b}_k \in \mathbb{R}^q$ such that the scalar canonical variates

$$u_k = \mathbf{X}\mathbf{a}_k, \qquad v_k = \mathbf{Y}\mathbf{b}_k$$

have maximal correlation

$$\rho_k = \operatorname{corr}(u_k, v_k),$$

subject to the constraints that (u_k, v_k) are uncorrelated with all previous canonical variates (u_ℓ, v_ℓ) for $\ell < k$ (Hotelling, 1936; Johnson & Wichern, 2018). The first canonical correlation ρ_1 is the maximum possible correlation between any linear combination of $\mathcal{X}$ and any linear combination of $\mathcal{Y}$; subsequent ρ_k are obtained iteratively under orthogonality constraints.

In population form, let

$$\mathbf{\Sigma}_{XX} = \operatorname{Cov}(\mathbf{X}), \quad \mathbf{\Sigma}_{YY} = \operatorname{Cov}(\mathbf{Y}), \quad \mathbf{\Sigma}_{XY} = \operatorname{Cov}(\mathbf{X}, \mathbf{Y}),$$

with transpose $\mathbf{\Sigma}_{YX} = \mathbf{\Sigma}_{XY}^\top$. The canonical weight vectors can be found by solving the generalized eigenvalue problems

$$\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY}\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{YX}\mathbf{a}_k = \rho_k^2\mathbf{a}_k, \tag{1}$$

$$\mathbf{\Sigma}_{YY}^{-1}\mathbf{\Sigma}_{YX}\mathbf{\Sigma}_{XX}^{-1}\mathbf{\Sigma}_{XY}\mathbf{b}_k = \rho_k^2\mathbf{b}_k, \tag{2}$$

where ρ_k^2 are the eigenvalues and the $\mathbf{a}_k$ and $\mathbf{b}_k$ are chosen to satisfy normalization constraints such as $\mathrm{Var}(u_k) = \mathrm{Var}(v_k) = 1$ (Borga, 2001; Uurtio et al., 2017). In practice, we work with sample covariance matrices and use the `CCA` implementation from `scikit-learn` (Pedregosa et al., 2011).

For interpretability, we examine:

- **Canonical correlations**: ρ_k, measuring the strength of association between u_k and v_k.

- **Canonical loadings**: correlations between the original variables and the canonical variates, e.g.

$$\mathrm{corr}(X_j, u_k), \quad \mathrm{corr}(Y_\ell, v_k),$$

  which indicate how strongly each observed variable contributes to each canonical dimension (Abdi & Williams, 2010).

- **Redundancy indices**: for the kth canonical function, the redundancy of $\mathcal{Y}$ given $\mathcal{X}$ is

$$R_{Y|X,k} = \left(\frac{1}{q} \sum_{\ell=1}^{q} \mathrm{corr}(Y_\ell, v_k)^2 \right) \rho_k^2,$$

  measuring the average proportion of variance in $\mathcal{Y}$ that can be accounted for by u_k via v_k (Borga, 2001; Engemann, 2019).

We compute up to $K = \min(p, q) = 3$ canonical pairs, but focus interpretation on the first one or two when later correlations are negligible.

Multivariate Multiple Regression

In multivariate multiple regression, we model multiple responses $\mathbf{Y} \in \mathbb{R}^{n \times m}$ as a linear function of predictors $\mathbf{X} \in \mathbb{R}^{n \times p}$:

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}, \tag{3}$$

where $\mathbf{B} \in \mathbb{R}^{p\times m}$ is the matrix of regression coefficients and $\mathbf{E} \in \mathbb{R}^{n\times m}$ is the residual matrix (Izenman, 2008; Johnson & Wichern, 2018). In the Linnerud case, $\mathbf{Y}$ contains `Weight`, `Waist`, and `Pulse`, and $\mathbf{X}$ the three exercise variables.

Under the usual assumptions (linearity, full rank of $\mathbf{X}$, homoscedastic and uncorrelated errors), the least-squares estimator of $\mathbf{B}$ is

$$\hat{\mathbf{B}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}, \tag{4}$$

which minimizes the trace of the residual sum of squares matrix $\mathbf{E}^\top\mathbf{E}$ (Johnson & Wichern, 2018). The fitted values and residuals are

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}, \quad \hat{\mathbf{E}} = \mathbf{Y} - \hat{\mathbf{Y}}.$$

We summarize model fit using response-specific coefficients of determination and error measures:

$$R_j^2 = 1 - \frac{\sum_{i=1}^n (y_{ij} - \hat{y}_{ij})^2}{\sum_{i=1}^n (y_{ij} - \bar{y}_j)^2}, \tag{5}$$

$$\mathrm{MSE}_j = \frac{1}{n}\sum_{i=1}^n (y_{ij} - \hat{y}_{ij})^2, \quad \mathrm{RMSE}_j = \sqrt{\mathrm{MSE}_j}, \tag{6}$$

for each response $j = 1, \ldots, m$ (Research Data Services, 2016; Tranmer & Steele, 2010). To approximate p-values and conduct multivariate tests, we use the `MANOVA` functionality in the `statsmodels` package, which provides Wilks' lambda and related test statistics for the null hypothesis that all regression coefficients for a given predictor are zero across responses.

Model assumptions are checked via residual diagnostics:

- *Normality*: Quantile–quantile (Q–Q) plots of residuals for each response.
- *Homoscedasticity*: Residual-versus-fitted plots for each response.
- *Multicollinearity*: Variance Inflation Factors (VIFs) for predictors, defined as

  $$\mathrm{VIF}_j = \frac{1}{1 - R_{j,\mathrm{aux}}^2},$$

  where $R_{j,\mathrm{aux}}^2$ is the coefficient of determination from regressing predictor X_j on the remaining predictors (James et al., 2021).

In code, multivariate regression is implemented using `LinearRegression` from `scikit-learn` (which natively supports multi-output regression) (Pedregosa et al., 2011), and multivariate significance tests are obtained via `statsmodels.multivariate.manova.MANOVA`.

Results

Canonical Correlation Analysis

Using standardized exercise variables $\tilde{\mathbf{X}}$ and standardized physiological variables $\tilde{\mathbf{Y}}$, we fit a CCA model with $K = 3$ components via `CCA(n_components=3)`. The estimated canonical correlations are:

$$\hat{\rho}_1 \approx 0.80, \quad \hat{\rho}_2 \approx 0.20, \quad \hat{\rho}_3 \approx 0.07.$$

The first canonical correlation is substantial, while the second and third are small. Given the tiny sample size, only the first canonical function is meaningfully interpretable (Johnson & Wichern, 2018; Uurtio et al., 2017).

Canonical loadings—correlations between original variables and canonical variates—for the first two canonical pairs can be summarized as follows (values rounded):

Variable	Loading on u_1	Loading on u_2
Chins	0.73	0.24
Situps	0.82	0.57
Jumps	0.16	0.96

Variable	Loading on v_1	Loading on v_2
Weight	−0.62	−0.77
Waist	−0.93	−0.38
Pulse	0.33	0.04

The first canonical variate u_1 is a roughly equal-weighted combination of `Chins` and `Situps`, with a smaller contribution from `Jumps`. The corresponding v_1 is dominated by large negative loadings on `Weight` and `Waist`, and a moderate positive loading on `Pulse`. Together with $\hat{\rho}_1 \approx 0.80$, this indicates a strong linear association between overall exercise volume (particularly upper-body work) and a leaner body composition: individuals who do more chins and situps tend to have lower weight and waist measurements and a somewhat higher resting pulse. This pattern is consistent with the idea that fitter subjects in this small sample are lighter and leaner (Izenman, 2008). See Figure 2 .

The second canonical function appears to relate `Jumps` more strongly to a combination of weight, waist, and pulse, but the associated correlation $\hat{\rho}_2 \approx 0.20$ is weak and unlikely to be statistically reliable given $n = 20$. Canonical scatterplots of (u_1, v_1) show a clear linear trend, while plots for (u_2, v_2) and (u_3, v_3) show diffuse clouds with no obvious structure. See Figure 3 .

Redundancy indices suggest that only a modest fraction of variance in one set can be explained via the other. Using the standard definition, the redundancies for the three canonical functions are approximately:

$$R_{Y|X} \approx (0.29, 0.01, 0.00), \quad R_{X|Y} \approx (0.26, 0.02, 0.00),$$

so the first canonical pair accounts for about 26–29% of the variance in each set on average, while later pairs contribute almost nothing. This is a common pattern in CCA: a single dominant association followed by negligible components (Borga, 2001; Engemann, 2019). The takeaway is that there is one meaningful latent dimension linking exercise performance and body composition in this dataset; beyond that, the data are too noisy and too limited in size to support more subtle structure.

Multivariate Regression Results

We fit a multivariate regression model with physiological responses $\mathbf{Y} = (\text{Weight}, \text{Waist}, \text{Pulse})$ and predictors $\mathbf{X} = (\text{Chins}, \text{Situps}, \text{Jumps})$ using all $n = 20$ observations. The fitted model has the matrix form

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}},$$

where each column of $\hat{\mathbf{B}}$ gives regression coefficients for one response. In-sample fit quality, summarized by R^2 and RMSE for each response, is:

Response	R^2	RMSE
Weight	0.27	20.6
Waist	0.55	2.10
Pulse	0.07	6.76

Waist circumference is moderately well explained by the exercise variables, with $R^2 \approx 0.55$; weight shows a weaker linear relationship ($R^2 \approx 0.27$), and resting pulse is essentially unexplained by this simple linear model ($R^2 \approx 0.07$). These results echo the CCA findings: the main shared signal between exercise and physiology is in body composition rather than resting pulse.

A multivariate test using `MANOVA` indicates that, jointly, the exercise block has a statistically detectable association with the three responses (Wilks' lambda substantially below 1 and a corresponding *p*-value below conventional thresholds), though with such a small sample, *p*-values must be interpreted cautiously (Izenman, 2008; Johnson & Wichern, 2018).

Residual diagnostics show no dramatic violations of the linearity and homoscedasticity assumptions, but normality of residuals is questionable due to the small sample and visibly heavy tails in Q–Q plots. The VIF values for `Chins`, `Situps`, and `Jumps` are modest and well below typical concern thresholds (e.g., VIF > 10) (James et al., 2021), suggesting that collinearity among the exercise variables is not severe in this dataset. See Figures 4 , 5 , and 6 .

Overall, the MVR model confirms that exercise performance carries some predictive value for body composition, especially waist circumference, but that the linear model explains only a modest proportion of the variability, particularly for weight and pulse. This is expected: many unmeasured factors (diet, genetics, training history) also influence these physiological outcomes.

Comparison and Practical Implications

CCA and MVR address related but distinct questions. CCA is symmetric and exploratory: it seeks linear combinations of the exercise and physiological variables that are maximally correlated, without designating a predictor or response set (Hotelling, 1936; Uurtio et al., 2017). In the Linnerud data, CCA reveals a single dominant canonical dimension linking overall exercise capacity to a leaner body composition. The canonical loadings and redundancy indices provide an interpretable summary of how each variable contributes to this latent association.

MVR, in contrast, is asymmetric and predictive: it treats exercise as input and physiology as output, estimating how much of the variability in each response can be explained by the predictors (Izenman, 2008; Johnson & Wichern, 2018). In this case, MVR quantifies that waist circumference is most predictable, while weight and pulse are weakly explained. The regression coefficients and multivariate tests speak directly to questions like "controlling for other exercise measures, how strongly is `Situps` associated with `Waist`?"

In practice, a sensible workflow is to use CCA to explore the high-level structure between two variable blocks and then use MVR (or more flexible models) to build specific predictive models for variables of interest (Hardoon et al., 2004; Uurtio et al., 2017). The Linnerud results also highlight the importance of sample size: with only 20 observations, both CCA and MVR are fragile, and parameter estimates are noisy. For larger behavioural or biomedical datasets, the same methodology scales naturally and can be combined with regularization and cross-validation to improve generalization (Izenman, 2008; Wang et al., 2018).

Conclusion

This case study demonstrates how CCA and multivariate regression can be combined to analyze relationships between two multivariate sets in a real dataset. CCA uncovers a strong latent association between exercise performance and body composition, quantified by a large first canonical correlation and interpretable loadings. Multivariate regression quantifies the extent to which exercise explains variance in specific physiological outcomes, revealing that waist circumference is the most predictable of the three responses.

Conceptually, CCA is best viewed as a tool for understanding shared structure and latent dimensions between blocks of variables, whereas MVR is better suited for targeted prediction and hypothesis testing about particular responses. Both methods rely on linearity and multivariate normality assumptions; both are sensitive to small sample sizes. Nonetheless, they remain workhorses in multivariate analysis and form a foundation for many modern extensions, including regularized, kernel, and deep CCA variants and high-dimensional multivariate regression models (Uurtio et al., 2017; Wang et al., 2018).

Future work on richer datasets could incorporate regularization to stabilize estimation, cross-validation to assess out-of-sample performance, and comparison with nonlinear methods such as partial least squares and multivariate adaptive regression splines. For now, the Linnerud analysis provides a compact, transparent demonstration of how CCA and MVR can jointly illuminate the relationships between multiple exercise and physiological variables.

References

Abdi, H., & Williams, L. J. (2010). Canonical correlation analysis. In N. J. Salkind (Ed.), *Encyclopedia of research design*. SAGE.

Borga, M. (2001). *Canonical correlation: A tutorial* (tech. rep.). Department of Electrical Engineering, Linköping University. Linköping, Sweden. http://people.imt.liu.se/magnus/cca/

Engemann, D. A. (2019). Redundancy in canonical correlation analysis [Accessed 2025-11-16].

Hardoon, D. R., Szedmak, S., & Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, *16*(12), 2639–2664. https://doi.org/10.1162/0899766042321814

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*(3–4), 321–377.

Izenman, A. J. (2008). *Modern multivariate statistical techniques: Regression, classification, and manifold learning*. Springer. https://doi.org/10.1007/978-0-387-78189-1

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in r* (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1

Johnson, R. A., & Wichern, D. W. (2018). *Applied multivariate statistical analysis* (6th ed.). Pearson.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Research Data Services, U. o. V. L. (2016). Getting started with multivariate multiple regression [Accessed 2025-11-16].

Tenenhaus, M. (1998). *La régression pls: Théorie et pratique*. Éditions Technic.

Tranmer, M., & Steele, D. (2010). *Multiple linear regression* (tech. rep.). Centre for Multilevel Modelling, University of Bristol. Bristol, UK.

https://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/working-papers/2020/multiple-linear-regression.pdf

Uurtio, V., Monteiro, J. M., Kandola, J., Shawe-Taylor, J., Fernandez-Reyes, D., & Rousu, J. (2017). A tutorial on canonical correlation methods. *ACM Computing Surveys*, *50*(6), 95:1–95:33. https://doi.org/10.1145/3136624

Wang, H.-T., et al. (2018). Canonical correlation analysis for neuroscience: Tutorial and review. *NeuroImage*, *176*, 1–15. https://doi.org/10.1016/j.neuroimage.2018.04.038

Appendix: Figures

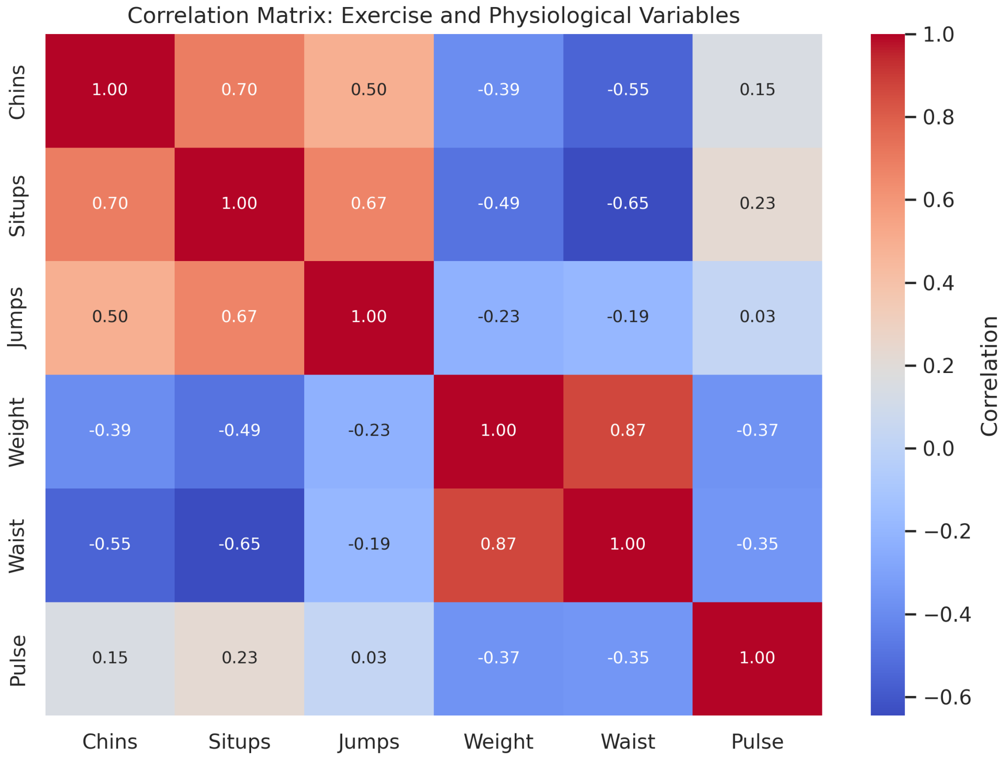

Figure 1

Correlation matrix for exercise and physiological variables in the Linnerud dataset.

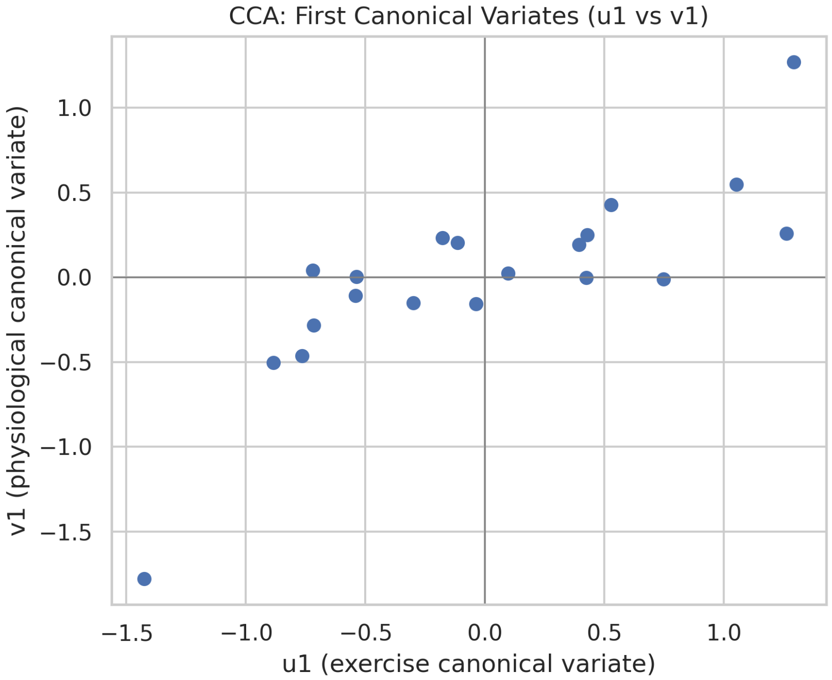

Figure 2

Scatterplot of the first canonical variates u_1 (exercise block) and v_1 (physiological block).

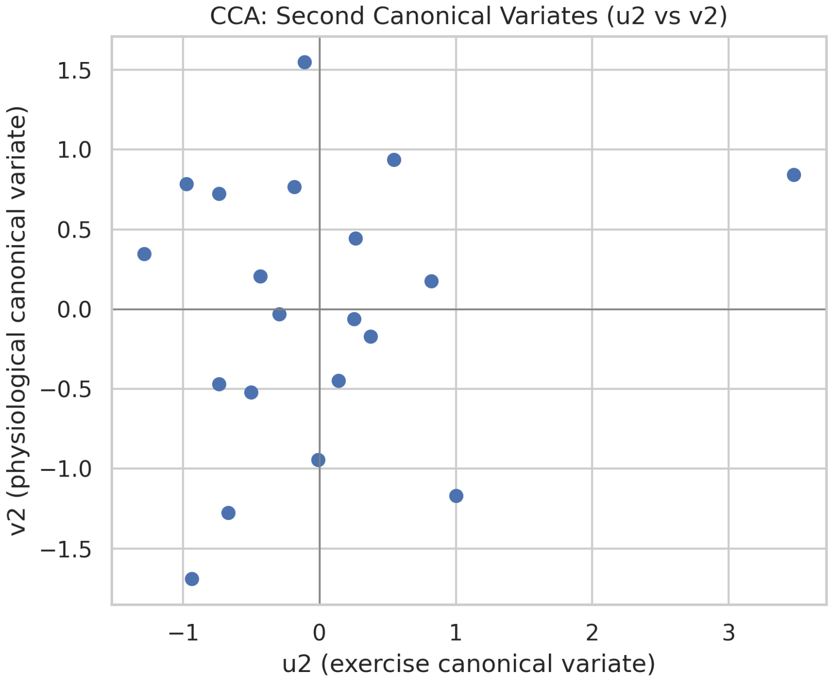

Figure 3

Scatterplot of the second canonical variates u_2 and v_2.

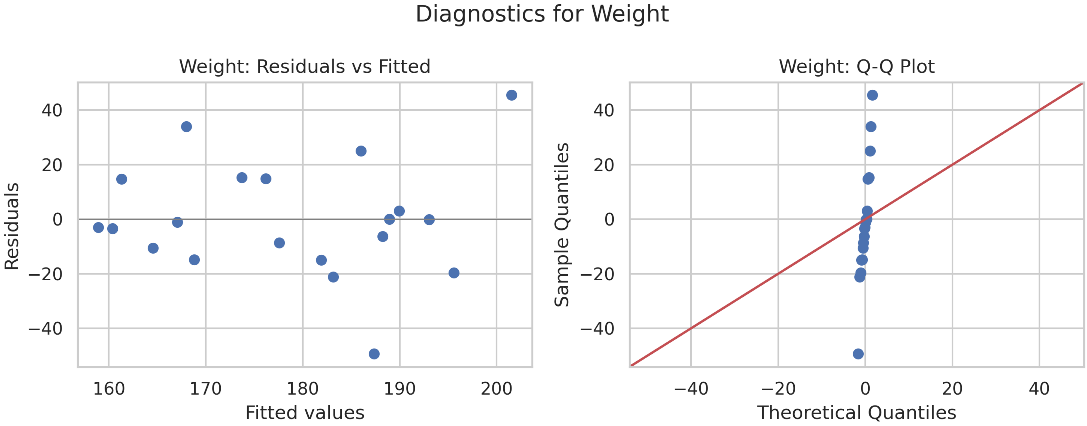

Figure 4

Residual diagnostics for `Weight`*: residuals vs. fitted values and Q–Q plot.*

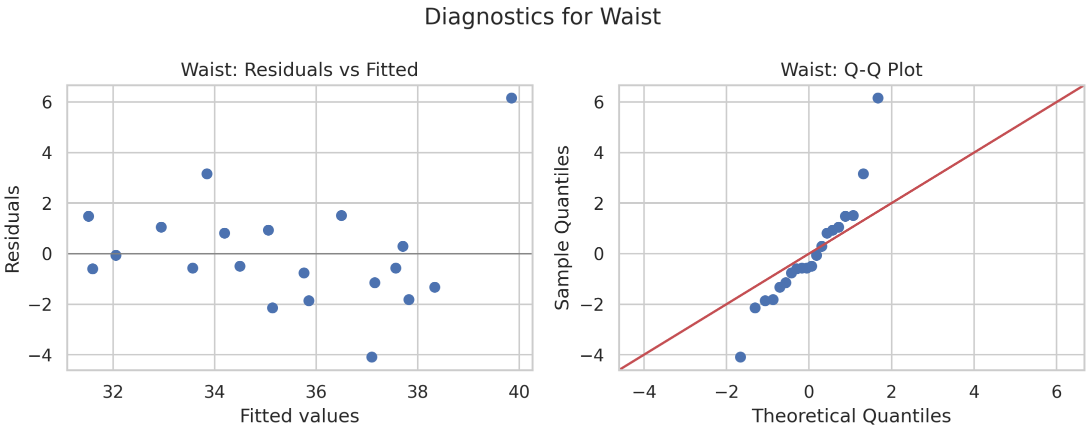

Figure 5

Residual diagnostics for `Waist`*: residuals vs. fitted values and Q–Q plot.*

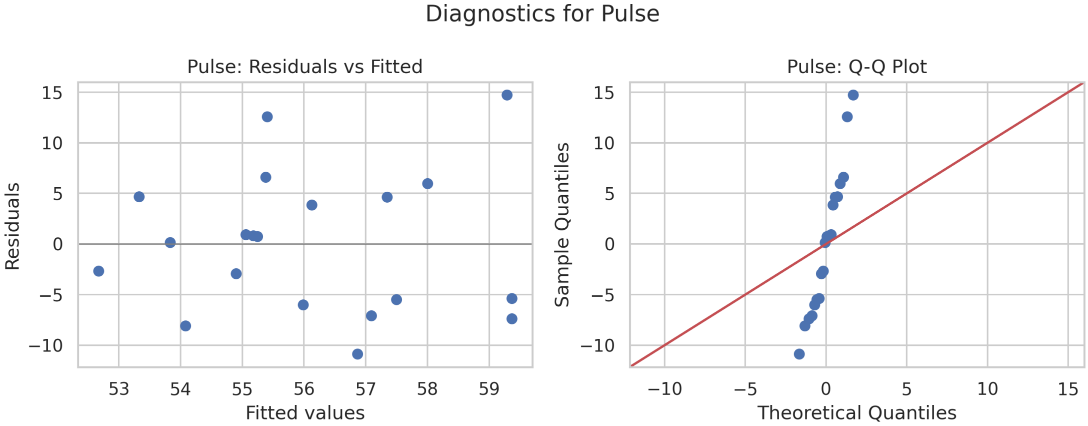

Figure 6

Residual diagnostics for `Pulse`*: residuals vs. fitted values and Q–Q plot.*