1. Diabetes Data Analysis

1.1 Data loading and association structure

The diabetes dataset contains observations from 442 patients and includes ten explanatory variables (age, sex, body mass index, average blood pressure, and six blood serum measurements) along with one outcome variable that reflects disease progression after one year. After importing the data into Python, pairwise correlations among the explanatory variables were computed and summarized in a correlation matrix. A heat-map visualization was originally used to display these associations; here, the emphasis is on interpretation rather than graphics.

Key relationships identified: - The serum measure **S5** exhibits the strongest overall associations, particularly with **S4** (correlation ≈ 0.62). - **Body mass index (BMI)** and **blood pressure (BP)** show a moderate positive relationship (correlation ≈ 0.39). - **S1** is moderately negatively related to **sex** (correlation ≈ −0.38). - Most remaining variable pairs demonstrate weak to moderate dependence, though several serum variables show patterns that suggest possible redundancy.

Overall, the structure of the correlation matrix suggests that some predictors may convey overlapping information.

1.2 Meaning and consequences of collinearity

Collinearity occurs in a multiple regression setting when two or more predictor variables are strongly linearly related, such that one predictor can be reasonably approximated by a combination of others.

Implications for regression estimates: 1. Estimated coefficients tend to have inflated standard errors. 2. Coefficient values become unstable, meaning small data changes can lead to large swings in estimates. 3. The estimated signs of coefficients may be misleading. 4. Isolating the individual effect of a single predictor becomes difficult. 5. Overall statistical power is reduced, making it harder to detect truly important predictors.

1.3 Full multivariate linear regression model

A linear regression model was fitted using all ten predictors along with an intercept term.

Model performance: - Mean Squared Error (MSE): **2859.70** - Adjusted R^2: relatively high, indicating reasonable explanatory power

Statistical significance: - Predictors showing statistical significance at the 5% level include **sex**, **BMI**, **blood pressure**, and **S5**. - Variables such as **age**, **S1**, **S2**, **S3**, **S4**, and **S6** do not reach conventional significance thresholds.

Interpretation: The coexistence of a high adjusted R^2 with several non-significant coefficients strongly suggests multicollinearity. In particular, the strong correlation between **S4** and **S5** contributes to this issue.

1.4 Forward versus backward variable selection

Forward selection begins with an empty model and introduces predictors one at a time. At each step, the variable that most improves a chosen criterion (such as AIC, BIC, or an F-test) is added, and the process stops when no remaining variable satisfies the inclusion rule. - *Strength:* Efficient when handling many candidate predictors. - *Limitation:* May miss variables that only become important in combination with others.

Backward selection starts with the full model containing all predictors. Variables are removed sequentially based on minimal contribution until all remaining predictors satisfy a retention criterion. - *Strength:* Accounts for joint effects among variables at the outset. - *Limitation:* Computationally demanding with large feature sets.

1.5 Stepwise selection using a forward approach

Stepwise selection blends automated decision rules with iterative model building. Using a forward strategy, the procedure operates as follows: 1. Start with no predictors. 2. Evaluate all candidate variables not yet in the model. 3. Add the variable that yields the greatest improvement in adjusted R^2 while meeting a significance threshold ($p < 0.05$). 4. Repeat until no further improvement is possible.

Variables selected: - **BMI** - **S5**

Model performance: - Mean Squared Error: **2876.68** - Adjusted R^2: comparable to the full model

Although the MSE is slightly higher than that of the full model, the reduced number of predictors provides a more parsimonious model with similar explanatory ability.

---

2. Analysis of the Titanic Dataset

2.1 Linear regression versus logistic regression

Linear regression is designed for continuous outcomes. It assumes normally distributed errors, estimates parameters via ordinary least squares, and produces predictions on the entire real line.

Logistic regression, in contrast, is used for binary or categorical outcomes. It models the probability of an event through a logistic (sigmoid) function, constraining predictions to the interval [0, 1]. Parameters are estimated using maximum likelihood methods, and normality of errors is not required.

2.2 Estimating survival probability

The Titanic dataset records survival outcomes for passengers (excluding crew). By fitting a logistic regression model, the probability of survival for any given passenger can be computed as a function of their characteristics.

2.3 Survival probabilities by class, gender, and age

When survival rates are grouped by passenger class, sex, and age category, several clear patterns emerge: - Females consistently show higher survival probabilities than males. - First-class passengers fare better than those in second or third class. - Children have higher survival rates than adults.

These findings align with historical accounts of evacuation priorities.

2.4 Logistic regression model for survival

A logistic regression model was estimated using **passenger class**, **sex**, and **age** as predictors.

Estimated coefficients (all statistically significant, p < 0.001): - Intercept: **2.092** - Passenger class: **−1.133** - Sex (female = 1): **2.497** - Age: **−0.034**

Interpretation: - Lower passenger class is associated with reduced survival chances. - Being female substantially increases the probability of survival. - Older passengers face lower survival probabilities.

2.5 Classification performance

Using a confusion matrix, the model’s predictive accuracy was evaluated.

Confusion matrix components: - True negatives: **523** - False positives: **96** - False negatives: **126** - True positives: **301**

Derived metrics: - Sensitivity (true positive rate): **70.4%** - Specificity (true negative rate): **84.5%** - Overall classification accuracy: approximately **78%**

These results indicate solid predictive performance, particularly in identifying non-survivors.

---

3. Principal Component Analysis (PCA)

3.1 Conceptual overview and applications

Principal Component Analysis is a technique that transforms a set of potentially correlated variables into a new set of uncorrelated components. Each component captures as much variance as possible while remaining orthogonal to the others.

Why PCA is useful: - Reduces dimensionality while retaining most of the information. - Mitigates multicollinearity by constructing independent components. - Enhances visualization of high-dimensional data. - Improves computational efficiency in downstream models.

3.2 Mathematical formulation

Let **X** be an (n × p) data matrix. 1. Subtract column means to obtain a centered matrix. 2. Compute the covariance matrix from the centered data. 3. Solve the eigenvalue problem for this covariance matrix. 4. Form a transformation matrix from the eigenvectors. 5. Project the centered data onto these eigenvectors to obtain principal components.

Eigenvalues quantify the variance explained by each component, while eigenvectors define their directions.

3.3 Correlation structure of Dow Jones stocks

For the 30 constituents of the Dow Jones Industrial Average, the correlation matrix can be derived from the covariance matrix by standardizing variances. The resulting matrix summarizes pairwise relationships among stock returns, with values bounded between −1 and 1.

3.4 PCA loadings for the first two components

Applying PCA to the correlation matrix yields component loadings that represent the contribution of each stock to the principal components. Bar charts of the first and second components reveal which stocks exert the strongest influence on overall market variation.

3.5 Comparison with an equally weighted market portfolio

Similarity between a principal component and the overall market can be assessed by correlating the component with an equally weighted index of all stocks. Typically, the first principal component closely tracks the general market movement.

3.6 Loading the Dow Jones data

The Dow Jones dataset can be imported directly into the programming environment using standard financial data libraries, after which returns and correlations can be computed.

3.7 Variance explained and scree analysis

A scree plot of cumulative explained variance shows how many components are needed to capture most of the information. In this case, **11 principal components** are sufficient to explain **95%** of the total variance.

3.8 Two-dimensional PCA representation

By plotting observations using the first two principal components, clusters and outliers among stocks become visible. The three most distant stocks are those with the largest deviations from the mean position in this reduced space, indicating behavior that differs most from the general market pattern.