TU Dortmund

Introductory Case Studies

Project III: Regression Analysis

Lecturers:
Prof. Dr. Jörg Rahnenführer
Dr. Franziska Kappenberg
M. Sc. Marieke Stolte

Author: Bushra Tariq Kiyani

Group number: 5

Group members: Prerana Rajeev Chandratre, Sathish Ravindranth, Shivam Shukla, Janani Veeraraghavan

January 27, 2023

Contents

1 Introduction

Height inference is a crucial part of individual identification work in many cases, for example in elderly people, inpatients and bedridden patients, people with skeletal deformities, and in forensic anthropology. In the actual detection process, height prediction plays a key role in finding the dead source as well as detecting the broken corpse and the unknown body. Height is always an interesting factor not only in the medical field but also in fashion, sports, armed services etc. Height, estimated from linear body measurements proved to be a useful proxy for stature. However, the relationship between body measurements and height implies the need to develop a predictive equation. The regression analysis method is widely used in statistical analysis for height predictions. It is commonly used in economic, scientific and technological fields to create empirical formulae for predictions (Alemayehu et al., 2018).

This project aims to predict the height of a person using regression analysis. The given dataset bodymeasurements.csv contains 11 different types of body measurements (variables) of 424 persons. The aim of this project is to develop a formula that predicts height from chest circumference, belly circumference, biceps circumference, knee circumference, ankle circumference, wrist circumference, thigh circumference, calf circumference, body weight measurements and age and sex information. We also find the optimal combination of explanatory variables.

To do this, we first fit a linear regression model that explains the body height based on all other given variables. To find the best set of explanatory variables for the body height we use the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) as the selection criteria. Then we compare the included variables of the two resulting best models. Then using BIC criteria we estimate the best linear model for the height. We look into the coefficients of the model, interpret them and their statistical significance. We calculate confidence intervals for the regression parameters and evaluate the goodness of fit.

The second section describes the structure and quality of the data set in more detail. Additionally, we state the goals of the project in the second section. The third section explains the Multivariate linear regression, Best Subset Selection, with the criteria AIC and BIC, Goodness-of-fit (adjusted $R-$squared) and Residual plot. The fourth section focuses on the application of these methods and the interpretation of the graphs and results. Finally, the fifth section summarizes the most important findings.

2 Problem statement

2.1 Data set and data quality

This report deals with the regression analysis of a sample dataset bodymeasurements.csv, given by the lecturers of the course. It contains body measurements of 424 persons aged between 18 to 40 years old, which were collected in a study. The aim of the study was to determine a model that can be used to explain body height. The data set consists of 11 variables. A description of the variables is given below:

Variable	Type	Description
Height	Numeric	Body height (in cm)
Chest	Numeric	Chest circumference (in cm)
Belly	Numeric	Belly circumference (in cm)
Biceps	Numeric	Biceps circumference (in cm)
Knee	Numeric	Knee circumference (in cm)
Ankle	Numeric	Ankle circumference (in cm)
Wrist	Numeric	Wrist circumference (in cm)
Ankle	Numeric	Ankle circumference (in cm)
Thigh	Numeric	Thigh circumference (in cm)
Calf	Numeric	Calf circumference (in cm)
Weight	Numeric	Body weight (in kg)
Age	Numeric	Age of the person at the time of the survey (in years)
Sex	Nominal	Sex ('m' for males, 'f' for females)

Table 1: Data Description

The data set does not contain missing values and the measurements are taken with high precision. The data set is collected for study purposes and provided by the TU Dortmund University lecturers, so we expect high data quality.

2.2 Project objectives

In this project, we aim at finding a model that can be used to explain body height and determining the 'best' set of explanatory variables for the body height using the Best Subset Selection. We interpret the model's coefficients and their statistical significance, provide confidence intervals for the regression parameters and evaluate the goodness of fit. To do so, first, we fit a full linear regression model using all the explanatory variables that explain the body height. Evaluate the statistical significance of the parameters and develop a linear regression equation. Then we find the best set of explanatory variables using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) as the selection criteria and we compare the included explanatory variables of the two resulting best models. Then we estimate the 'best' linear model for the dependent variable w.r.t. the BIC and we check model assumptions using the Q-Q plot and residual plot for the estimated model. Then We interpret the model's coefficients and their statistical significance, provide confidence intervals for the regression parameters and evaluate the goodness of fit i.e. adjusted $R-$squared.

3 Statistical methods

This section discusses statistical methods used to determine the best model to explain the body height, e.g. Multivariate linear regression, Best Subset Selection, with the criteria AIC and BIC, Goodness-of-fit and Residual plot.

3.1 Multiple linear regression

The linear regression model is a statistical modelling method used to model a relationship between a continuous variable of interest y called the response variable and a set of explanatory variables $x_1, ..., x_k$. Explanatory variables can be continuous or categorical. In general, we model the relationship between y and $x_1, ..., x_k$ with a function $f(x_1, ..., x_k)$. This relationship is affected by random noise ϵ. In practice, we assume additive errors and thus obtain (Fahrmeir et al., 2022, p. 74-75):

$$Y = f(x_1, ..., x_k) + \epsilon$$

Where ϵ is a random variable and the function f is a linear combination of covariates, such that:

$$f(x_1, ..., x_k) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

The parameters $\beta_0, \beta_1, ..., \beta_k$ are unknown and need to be estimated. The parameter β_0 represents the intercept. If we combine the covariates and the unknown parameters into $p = k + 1$ dimensional vectors, $\mathbf{x_i} = (1, x_{i1}, ..., x_{ik})'$ and $\boldsymbol{\beta} = (\beta_0, ..., \beta_k)'$ then:

$$\hat{y}_i = \mathbf{x_i}'\boldsymbol{\beta} + \epsilon_i$$

Considering n observations, the model can be represented in vector notation. If we define the vectors $\mathbf{y}, \boldsymbol{\epsilon}$ and the design matrix $\mathbf{X}$ as follows (Fahrmeir et al., 2022, p. 75):

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}$$

Then we can write the model as (Fahrmeir et al., 2022, p. 75):

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

3.1.1 Model Assumptions

The following assumptions regarding the noise ϵ are made (Fahrmeir et al., 2022, p. 75-76):

Expectation of the errors: The errors have mean or expectation zero, i.e. $E[\epsilon_i] = 0$ or in matrix notation $E[\boldsymbol{\epsilon}] = \mathbf{0}$

Homoscedastic Error Variances: We assume a constant error variance σ^2 across observations, that is homoscedastic errors with $\text{Var}(\epsilon_i) = \sigma^2$ The errors are called heteroscedastic when the variances vary among observations i.e. $\text{Var}(\epsilon_i) = \sigma_i^2$.

Uncorrelated Errors: In addition to homoscedastic variances, we assume that errors are uncorrelated, which means $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$. The assumption of homoscedastic and uncorrelated errors lead to the covariance matrix $\text{Cov}(\boldsymbol{\epsilon}) = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon'}) = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. This shows ϵ is stochastically independent of covariates.

The design matrix X has full column rank: We assume that the design matrix has full column rank i.e. $\mathbf{rk}(\mathbf{X}) = k + 1 = p$ which ensures linear independence of the columns in the design matrix $\mathbf{X}$. It is necessary in order to obtain unique estimators of the regression coefficients $\boldsymbol{\beta}$. The assumption of linear independence is violated when at least one of the explanatory variables can be represented as a linear combination of the other covariates, implying redundancy of information.

Gaussian errors: We often assume normally distributed errors to construct confidence intervals and hypothesis tests for the regression coefficients. Together with assumptions of Homoscedastic Error Variance and uncorrelated errors, we get $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ or in matrix notation $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, which follows $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$

3.1.2 Dummy Coding

A dummy variable is a numeric variable that represents categorical data, such as gender and region etc. In regression, we observe the linear effect of continuous covariates on the response variable. Dummy coding is a way of incorporating categorical variables into regression analysis. For modelling the effect of a covariate $x \in (1, ..., c)$ with c categories using dummy coding, we define the $c - 1$ dummy variables as follows:

$$x_{i1} = \begin{cases} 1, & \text{if } x_i = 1 \\ 0, & \text{otherwise} \end{cases} \quad ... \quad x_{i,c-1} = \begin{cases} 1, & \text{if } x_i = c - 1 \\ 0, & \text{otherwise} \end{cases}$$

For $i = 1, ..., n$. We get a modified linear function as follows:

$$f(x_1, ..., x_k) = \beta_0 + \beta_1 x_{i1} + ... + \beta_{i,c-1} x_{i,c-1} + ... + \beta_k x_k$$

We omit one of the dummy variables for reasons of identifiability, in this case, the dummy variable for category c. This category is called the reference category. The estimated effects can be interpreted by direct comparison with the reference category (Fahrmeir et al., 2022, p. 95-97).

3.1.3 Estimation

We differentiate the model parameters from their estimates by a 'hat', which means estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^2$ are denoted by $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\sigma}}^2$ respectively.

Regression parameters Estimation: Regression parameters are usually estimated using **the method of least squares**. According to the principle of least squares, the unknown regression coefficients $\boldsymbol{\beta}$ are estimated by minimizing the sum of the squared deviations from the true response value y_i and the predicted $\hat{y_i} = \mathbf{x}_i' \beta$.

$$\text{LS}(\beta) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i' \beta)^2 \text{ or in Matrix notation} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^2$$

To estimate the $\boldsymbol{\beta}$, we minimize LS($\boldsymbol{\beta}$) by taking the first derivative and equating it to zero and solving it for $\boldsymbol{\beta}$. We get least square estimator $\hat{\boldsymbol{\beta}}$ (Fahrmeir et al., 2022, p. 105):

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{y}$$

Predicted Values and Residuals: Based on an estimator $\hat{\beta}$ for β, a straightforward estimator of the mean $E[y_i]$ of y_i is given by:

$$\widehat{E[y_i]} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_k x_{ik} = \mathbf{x}_i'\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$\widehat{E[y_i]}$ is usually referred as $\hat{y}_i$. The deviation $y_i - \mathbf{x}_i'\hat{\beta}$ between true value y_i and estimated value $\hat{y}_i$ is called **residual** and denoted by $\hat{\epsilon}$.

$$\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

Residuals $\hat{\boldsymbol{\epsilon}}$ are not identical to the errors $\boldsymbol{\epsilon}$. The residuals $\hat{\boldsymbol{\epsilon}}$ can be seen as estimates of errors $\boldsymbol{\epsilon}$ (Fahrmeir et al., 2022, p. 107).

Estimation of the Error Variance: Maximum likelihood estimator for error variance is (Fahrmeir et al., 2022, p. 108):

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}'}{n}$$

which is biased, so Restricted Maximum Likelihood Estimator (REML) is a commonly used estimator for σ^2. Which is given as:

$$\hat{\sigma}^2 = \frac{1}{n-p}\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}'$$

3.1.4 Hypothesis Testing and Confidence Intervals for $\hat{\beta}_j$

Assuming independently and identically normally distributed errors i.e. $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ to evaluate the relationship between the independent and dependent variable, we perform hypothesis tests of regression coefficients $\hat{\beta}_j$ for $j = 1, ..., k$. The null hypothesis H_0 states that j^{th} covariate has no significant effect on the response variable, while the alternative hypothesis H_1 states that j^{th} covariate has a significant effect on the response variable.

$$H_0 : \hat{\beta}_j = 0 \text{ and } H_1 : \hat{\beta}_j \neq 0, \qquad j = 1, ..., k$$

t−test: A $t-$test is used to check the significance of individual regression coefficients in multiple linear regression. The test statistic t_j for this test is based on the $t-$distribution with $n - p$ degrees of freedom.

$$t_j = \frac{\hat{\beta}_j}{se_j}$$

Where $se_j = \sqrt{\widehat{Var(\hat{\beta}_j)}}$ is the estimated standard deviation or standard error of $\hat{\beta}_j$. The critical value for the rejection region of the null hypothesis H_0 is obtained as the $(1-\alpha/2)$-quantile of the $t-$distribution with $n-p$ degrees of freedom. Thus, we reject the null hypothesis, if

$$|t_j| > t_{1-1/\alpha}(n-p)$$

Rejecting the null hypothesis H_0 implies that the covariate x_j has a significant effect on the response variable y.

Confidence Intervals for $\hat{\beta}_j$: Assuming normally distributed errors, using $t_{1-1/\alpha}(n-p)$ and standard errors se_j, we obtain the following $(1-\alpha)$-confidence intervals for β_j:

$$\left[\hat{\beta}_j - t_{n-p}\left(1-\frac{\alpha}{2}\right) . \ se_j, \hat{\beta}_j + t_{n-p}\left(1-\frac{\alpha}{2}\right) . \ se_j\right]$$

(Fahrmeir et al., 2022, p. 125-139).

3.2 Best Subset Selection

We need model choice criteria for the comparison of different models and the selection of the best model. The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are widely used criteria for model choice in linear models.

3.2.1 The Akaike information criterion (AIC)

The Akaike information criterion (AIC) is one of the most widely used criteria for model choice within the scope of likelihood-based inference. AIC is defined as:

$$\text{AIC} = -2.l(\hat{\beta}_M, \hat{\sigma}^2) + 2(|M|+1)$$

where $l(\hat{\beta}_M, \hat{\sigma}^2)$ is the maximum value of the log-likelihood. Smaller values of AIC corresponds to a better model fit. The total number of parameters is $|M|+1$ because the error variance σ^2 is also counted as a parameter. In a linear model with Gaussian errors, we get:

$$\text{AIC} = n.log(\hat{\sigma}^2) + 2(|M|+1)$$

In AIC, ML estimator $\hat{\sigma}^2 = \frac{\hat{\epsilon}\hat{\epsilon}'}{n}$ is considered instead of unbiased variance estimator.

(Fahrmeir et al., 2022, p. 148).

3.2.2 The Bayesian information criterion (BIC)

The form of the BIC is similar to that of the AIC, and again smaller value indicates a better model fit. The Bayesian information criterion (BIC) is defined as:

$$\mathrm{BIC} = -2.l(\hat{\beta}_M, \hat{\sigma}^2) + log(n)(|M| + 1)$$

Assuming Gaussian errors, we obtain

$$\mathrm{BIC} = n.log(\hat{\sigma}^2) + log(n)(|M| + 1)$$

The main difference is that BIC penalizes complex models more than AIC. Therefore, the resulting "best" models are generally more parsimonious when using BIC rather than AIC. (Fahrmeir et al., 2022, p. 149).

3.3 The Coefficient of determination (adjusted $R-$squared)

The Coefficient of determination is the proportion of total variance that is explained by the regression model.

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}\hat{\epsilon}^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

The closer the coefficient of determination is to 1, the smaller the sum of the residual squares and the better the fit to the data. Adjusted R^2 measures the proportion of variation explained only by the explanatory variables that really help in explaining the dependent variable. It penalizes adding independent variables that don't help in predicting the dependent variable.

$$\tilde{R}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

The main difference between $R-$squared (R^2) and adjusted $R-$squared ($\tilde{R}^2$) is when we add an independent variable to a model, the $R-$squared increases, even if the independent variable is insignificant. While the adjusted $R-$squared increases only when the independent variable is significant and affects the dependent variable. (Fahrmeir et al., 2022, p. 112-115).

3.4 Interpretation of Parameter Estimates

The sign of a linear regression coefficient tells whether there is a positive or negative relationship between each explanatory variable and the dependent variable. A positive coefficient indicates that as the value of the explanatory variable increases, the mean of the dependent variable also increases. A negative coefficient indicates that the dependent variable decreases as the explanatory variable increases. The coefficient value indicates how much the average of the dependent variable changes for a one-unit shift in the independent variable while holding other variables in the model constant. Holding the other variables constant is important because it allows for estimating the effect of each variable in isolation from the others.(Fahrmeir et al., 2022, p. 107).

3.5 Residual plot

A residual plot plots residuals $\hat{\epsilon}$ on the y-axis against the predicted values $\hat{y_i}$ on the x-axis. The ideal residual plot shows a random scatter of points around zero with a constant variance. If the points exhibit an increasing, decreasing or non-constant fluctuation, then the variance is not constant (heteroscedastic variance). The linearity of a model function can also be checked through residual plots. If the points form a pattern then the model function is incorrect. If points are scattered equally above and below the average line then $E[\boldsymbol{\epsilon}] = \mathbf{0}$. (Fahrmeir et al., 2022, p. 79-80).

4 Statistical analysis

In this section, a detailed description of the application of the linear regression model is discussed. The R software (R Core Team, 2022) Version: 4.2.1 with additional packages **ggpubr** (Kassambara, 2022), **dplyr** (Wickham et al., 2022), **ggplot2** (Wickham, 2016), **reshape2** (Wickham, 2007), **GGally** (Schloerke et al., 2021),**olsrr**(Hebbali, 2020) is used for fitting the linear regression model, finding the best set of explanatory variables for the body height using AIC and BIC and for visualizing graphical representations.

4.1 Descriptive Analysis

First, we perform a descriptive analysis of the data set. Table 3 on page 17 of the appendix summarizes the descriptive statistical information about the data. The average height is 170.8. We observe the age of persons is between $18 - 40$ years with a standard deviation of 5.85 years and the median age is 25 years. Weight measurements show the highest standard deviation of 12.74 while the lowest standard deviation of 1.35 is in the wrist measurements. But a small or large standard deviation is due to small or large values for wrist and weight measurements. The mean and the median of all measurements are almost the same or very close to each other. Observed minimum and maximum heights are 147.20 cm and 198.10 cm respectively. Figure 1 shows scatter plots of Height vs all other covariates which show that age and thigh have a very weak correlation with height. Figure 2(b) shows a correlation heatmap and we observe that the correlation coefficient value for age is 0.06 and for thigh is 0.14 which are the lowest values. The highest correlation value is for weight 0.73, it indicates a strong correlation between height and weight.

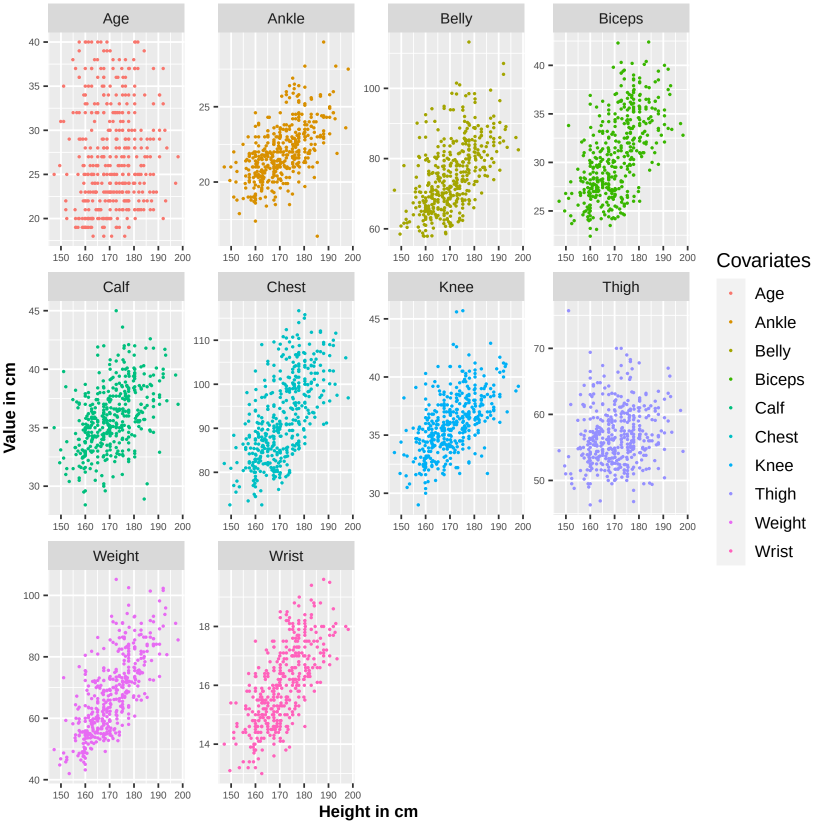
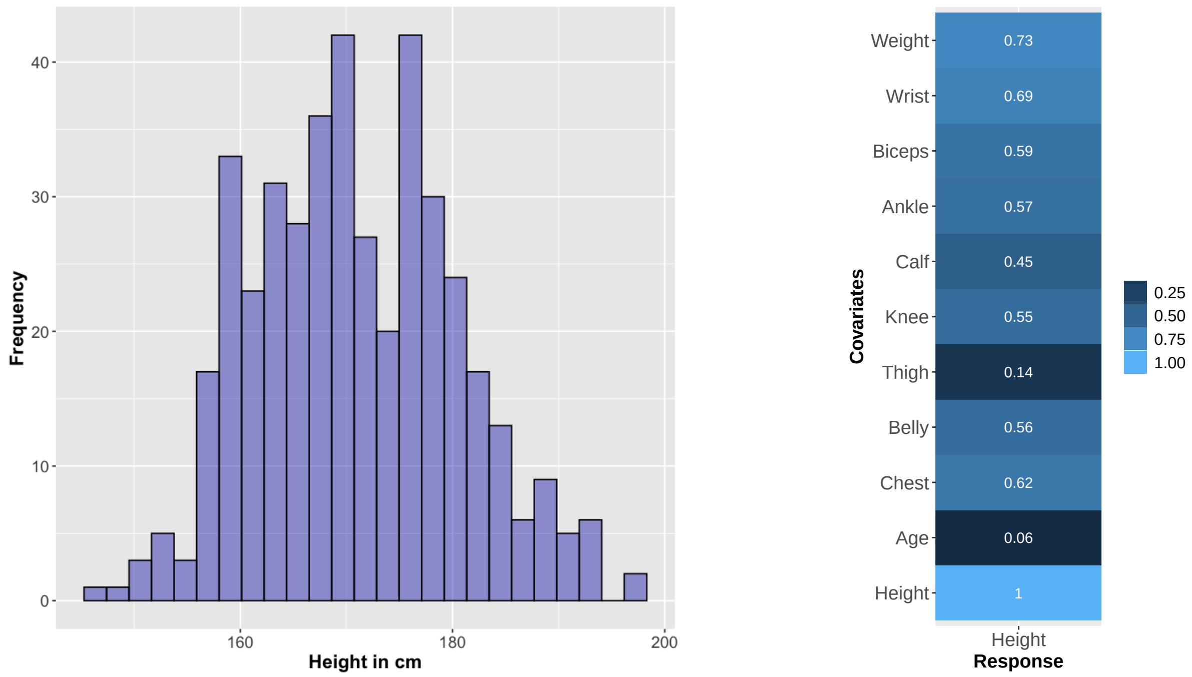


Figure 1: Scatter plots of Height vs all other covariates

Frequency Distribution of response variable (body height): Figure 2(a) shows the frequency distribution of body height.The graph shows a normal distribution. Which shows that the response is normally distributed. We observe some extreme left and right values. There are three persons whose height is above 190 cm and three with less than 150 cm. Maximum people have a height between $170 - 180$ cm.



(a) Frequency Distribution of Body Height

(b) Correlation Heatmap

Figure 2: Frequency Distribution Histogram and Correlation Heatmap of the response

4.2 Full Linear Regression Model

A linear regression model that explains the body height based on all other 11 covariates is fit to the given data set. Since Sex is a categorical variable, a dummy variable is created for male as x_{sexm} and female is taken as the reference category. The following model equation is formulated:

$$\begin{aligned}\hat{y}_{Height} = \hat{\beta}_{Intercept} + \hat{\beta}_{Age}.x_{Age} + \hat{\beta}_{Chest}.x_{Chest} + \hat{\beta}_{Belly}.x_{Belly} + \hat{\beta}_{Biceps}.x_{Biceps} \\ +\hat{\beta}_{Knee}.x_{Knee} + \hat{\beta}_{Ankle}.x_{Ankel} + \hat{\beta}_{Wrist}.x_{Wrist} + \hat{\beta}_{Thigh}.x_{Thigh} + \hat{\beta}_{Calf}.x_{Calf} \\ +\hat{\beta}_{Weight}.x_{Weight} + \hat{\beta}_{Sexm}.x_{Sexm}\end{aligned}$$

The table 4 on page 17 of the appendix shows the output of the model. With significance level $\alpha = 0.05$. The $p-$values of coefficients related to Sexm, Belly, Thigh, Calf, Biceps and Weight are less than α , showing these variables are statistically significant and a worthwhile addition to the regression model. Adjusted $R-$squared is 0.79 which indicates a good fit for the model.

4.3 Best Subset Selection

Now to find the best set of explanatory variables for the body height the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) as the selection criteria discussed in section 3.2 are used. This starts by fitting the model with one covariate and choosing the model with the lowest AIC and BIC values. Then the model is fitted with two covariates and the best models are chosen, the process continues until we include all the covariates. The table 5 on page 18 of the appendix shows the resulting combinations of best models. According to the AIC criterion, the minimum AIC value is 2438.82, and eight covariates Sexm, Chest, Belly, Thigh, Calf, Biceps, Wrist, and Weight are selected with adjusted $R-$squared 0.796 which shows a good fit. Using the BIC criterion the lowest BIC value is 2474.51 and the Model selects six covariates Sexm, Belly, Thigh, Calf, Biceps and Weight with adjusted $R-$squared 0.794 which also shows a good fit. We see BIC criterion is more restrictive than AIC. When BIC is used Chest and Wrist become insignificant.

4.4 Fitting the best model using BIC

Now we fit the linear model for the dependent variable w.r.t. the BIC from section 4.2. After fitting the linear model we calculate the confidence intervals for the regression parameters. The table 2 shows the summary.

Parameters	Estimated coefficients	Confidence intervals	$p-$values
(Intercept)	213.72	[205.194 , 222.247]	< 0.01
Sexm	6.61	[4.746 , 8.472]	< 0.01
Belly	-0.73	[-0.832 , -0.623]	< 0.01
Thigh	-0.60	[-0.78 , -0.416]	< 0.01
Calf	-0.53	[-0.778 , -0.278]	< 0.01
Biceps	-1.11	[-1.34 , -0.877]	< 0.01
Weight	1.42	[1.302 , 1.53]	< 0.01

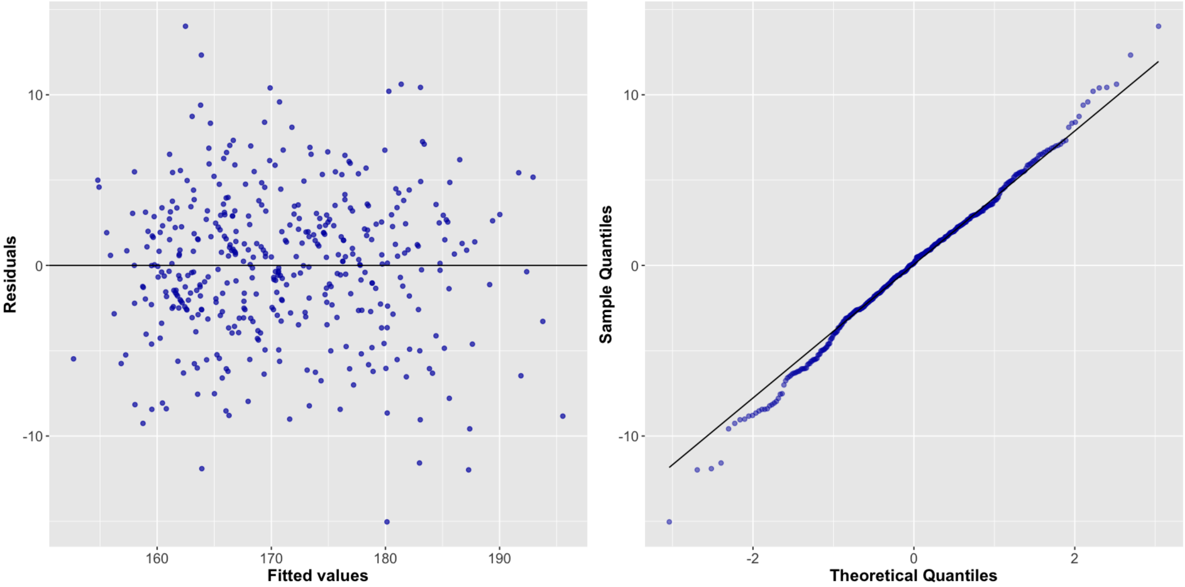
Table 2: Summary of a fitted linear model with the best subset using BIC

We get the following model equation:

$$\hat{y}_{Height} = 213.72 + 6.61.x_{Sexm} - 0.73.x_{Belly} - 0.6.x_{Thigh} - 0.53.x_{Calf} \\ -1.11.x_{Biceps} + 1.42.x_{Weight}$$

4.4.1 Verifying the Assumptions

As discussed in section 3.1.1, here the assumptions of the classical linear model are analyzed for our data set. **Expectation of Errors:** To verify $E[\boldsymbol{\epsilon}] = \mathbf{0}$, we observe the Figure 3(a) which shows the residual plot. We plot the selected best model's fitted values against residuals. We see data points are scattered almost equally above and below the average line. Which shows $E[\boldsymbol{\epsilon}] = \mathbf{0}$. **Homoscedastic Error Variances:** Also from Figure 3 we observe constant variation around the average line, showing variance stays constant with an increasing $\hat{y}$. **Gaussian Errors:** Figure 3(b) shows a Q-Q plot, we see the points fall almost along the straight line indicating that residuals are normally distributed. **Absence of Multicollinearity:** The variable measurements are random with high precision so we don't expect multicollinearity here.



(a) Residual Plot

(b) Q-Q Plot

Figure 3: Residual and Q-Q Plots of fitted linear regression model using BIC

4.4.2 Parameters Interpretation

Coefficients of Belly, Thigh, Calf and Biceps are negative, which shows a negative relationship between these variables and the response variable. While Sexm and Weight are positively related to the response. The weight coefficient represents the average increase of height in centimetres for every additional kilogram in weight. So If the weight of a person increases by 1 kg, the predicted height increases by 1.42 cm. Similarly, if the Belly circumference of a person increases 1 cm, the predicted body height decreases 0.73 cm and by increasing the Thigh size of a person 1 cm, the predicted body height decreases 0.6 cm. When the Calf size increases by 1 cm the predicted body height decreases by 0.53 cm. 1 cm increase in the Biceps circumference of a person decreases the predicted body height by 1.11 cm. Sexm shows, by keeping all other covariates constant, on average the predicted height for males is larger than for females. Males are on average predicted 6.61 cm taller than females. All $p-$values for the coefficients are significant which indicates that these covariates are statistically significant. Adjusted $R-$squared is 0.794 which shows a good model fit.

5 Summary

Height is always a subject of interest especially in the medical field, fashion, sports and fitness. In this project, we aimed at finding the best model that can explain body height. We were given a data set of body measurements, consisting of 424 observations and 12 body measurements including one response variable (body height) and 11 explanatory variables. We determined a linear regression model to explain the body height using chest circumference, belly circumference, biceps circumference, knee circumference, ankle circumference, wrist circumference, thigh circumference, calf circumference, body weight measurements and age and sex information as explanatory variables.

we were interested in determining the best linear regression model to predict the body height of a person. To find the optimal subset of explanatory variables, we used the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) as the selection criteria. We measured the goodness of fit using adjusted $R-$squared.

We started with descriptive data analysis and looked at the summary statistics of the data set which showed that the average observed body height was 170.8, body weight showed the highest standard deviation of 12.74 and wrist measurement had the lowest standard deviation of 1.35. But this could be the effect of large weight measurements and small wrist measurements. The median age was 25. The mean and median of each measurement were very close. The frequency distribution graph of body height showed an overall normal distribution.

First, we fitted a full linear regression model based on all the given covariates. We observed six covariates Sexm, Belly, Thigh, Calf, Biceps and Weight were statistically significant. Adjusted $R-$squared was 0.79 depicting a good fit. Then we used AIC and BIC criteria to select the best subset of covariates which resulted in two model choices. Using AIC as a criterion, 8 explanatory variables were selected with adjusted $R-$squared 0.796. While using BIC as a selection criterion, 6 covariates were selected with adjusted $R-$squared 0.794.

Then we estimated the linear model for the body height using covariates chosen by the BIC criterion. Based on the signs of parameter estimates we concluded that Sexm, and weight were positively, while Belly, Thigh, Calf and Biceps were negatively related to body height. Confidence intervals for the parameter estimates were not too wide. We observed all $p-$values were less than 0.05, showing that these variables were statistically significant. Adjusted $R-$squared was 0.794 which indicated that the model was a good fit.

Although the linear regression model proved to be a good model in this project, it does not guarantee to be the best model. We took into account only the physical features of a person as covariates, while genetics (Parents' height, genes variations/DNA structure etc) can also influence the predicted height. So next study can be done by adding other influencing factors.

Bibliography

Digssie Alemayehu, Argaw Alemayehu, and Belachew Tefera. Developing an equation for estimating body height from linear body measurements of ethiopian adults. *JPhysiol Anthropol 37, 26*, 2018. doi: https://doi.org/10.1186/s40101-018-0185-7.

Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, and Brian D. Marx. *Regression.* Springer-Verlag GmbH, 2022. 2nd Edition.

Aravind Hebbali. *olsrr: Tools for Building OLS Regression Models*, 2020. URL `https://olsrr.rsquaredacademy.com/`. Version 0.5.3, R package version 3.3.

Alboukadel Kassambara. *ggpubr: ggplot2' Based Publication Ready Plots*, 2022. URL `https://rpkgs.datanovia.com/ggpubr/`. R package version 3.4.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2022. URL `https://www.R-project.org/`.

Barret Schloerke, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Jason Crowley. *GGally: Extension to 'ggplot2'*, 2021. https://ggobi.github.io/ggally/, https://github.com/ggobi/ggally.

Hadley Wickham. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12):1–20, 2007. URL `http://www.jstatsoft.org/v21/i12/`.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL `https://ggplot2.tidyverse.org`.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2022. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

Appendix

A Additional tables

	Min	Q1	Median	Mean	Q3	Max	SD
Age	18.00	22.00	25.00	26.90	31.25	40.00	5.85
Height	147.20	163.20	170.20	170.88	177.80	198.10	9.40
Chest	72.60	84.78	90.95	92.22	99.83	116.70	9.60
Belly	57.90	67.50	74.10	75.28	82.00	113.20	9.90
Thigh	46.30	53.70	56.30	56.82	59.50	75.70	4.43
Knee	29.00	34.30	35.90	36.01	37.70	45.70	2.54
Calf	28.40	34.00	35.80	35.88	37.70	45.00	2.76
Ankle	16.40	20.90	21.90	22.02	23.10	29.30	1.85
Biceps	22.40	27.27	30.35	30.83	34.12	42.40	4.27
Wrist	13.00	14.88	15.90	15.97	17.00	19.60	1.35
Weight	42.00	57.30	66.80	67.82	75.62	105.20	12.74

Table 3: Summary Statistics

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	212.3588	6.9685	30.47	< 0.01
Age	0.0416	0.0401	1.04	0.3001
Sexm	4.6545	1.3623	3.42	0.0007
Chest	-0.1285	0.0701	-1.83	0.0675
Belly	-0.6792	0.0594	-11.44	< 0.01
Thigh	-0.6007	0.1000	-6.01	< 0.01
Knee	0.1062	0.1704	0.62	0.5334
Calf	-0.7251	0.1481	-4.90	< 0.01
Ankle	0.1964	0.2138	0.92	0.3587
Biceps	-1.0035	0.1497	-6.70	< 0.01
Wrist	0.4546	0.3939	1.15	0.2491
Weight	1.4067	0.0684	20.56	< 0.01

Table 4: Full Linear Regression Model

Model	Covariates	Adjusted.R^2	AIC	BIC
1	Weight	0.53	2789.50	2801.65
2	Thigh Weight	0.66	2646.98	2663.18
3	Belly Thigh Weight	0.74	2534.27	2554.52
4	Belly Thigh Biceps Weight	0.77	2495.71	2520.00
5	Sex Belly Thigh Biceps Weight	0.79	2457.25	2485.60
6	Sex Belly Thigh Calf Biceps Weight	0.79	2442.11	2474.51
7	Sex Chest Belly Thigh Calf Biceps Weight	0.80	2439.86	2476.31
8	Sex Chest Belly Thigh Calf Biceps Wrist Weight	0.80	2438.82	2479.32
9	Sex Chest Belly Thigh Knee Calf Biceps Wrist Weight	0.80	2440.37	2484.92
10	Sex Chest Belly Thigh Knee Calf Ankle Biceps Wrist Weight	0.80	2442.12	2490.72
11	Age Sex Chest Belly Thigh Knee Calf Ankle Biceps Wrist Weight	0.80	2443.90	2496.54

Table 5: Best subset selection with adjusted R^2, AIC and BIC