TU Dortmund

Introductory Case Studies

Project 3: Regression Analysis

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

Author: Prerana Rajeev Chandratre

Group number: 5

Group members: Janani Veeraghavan, Bushra Kiyani, Sathish Ravindranath Kabatkar, Shivam Shukla

January 27, 2023

Contents

1 Introduction

Developmental psychologists and other health care services use growth charts to monitor the growth of humans at regular intervals of time. These growth charts are constructed using the data from a large number of physically fit humans. The height, weight, age, sex, and measurements of body such as the circumference of the head, biceps, and chest can be compared to the expected parameter of other healthy human with same age and sex for further research on human growth. These measurements and growth charts can be used to predict the height of a human, and further research can be done to study human anatomy (Wikipedia, 2022). The data set includes human body measurements of 424 persons along with their weight, age and gender. This data set is provided by course instructors of TU Dortmund University. It provides additional body measurements in centimeters and weight in kilograms for each observation.

The main objective of this project report is to perform linear regression to determine how the height of people can be predicted by other variables such as body measurements, weight, age, and sex. The four assumptions are validated for the linear regression model: no plausible multicollinearity between the covariates, homoscedastic errors, independent observations, and normally distributed errors with constant variance and zero means. Further, the best subset of covariates is identified using the best subset selection method with the following selection criteria: the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). From the two models, the best subset of covariates used in the models are compared, and the best model is found by using the BIC criterion to fit the data. Estimated model coefficients are interpreted, and goodness-of-fit is evaluated.

This report has five sections in total including the introduction section. Section 2 explains the data set, the quality of the data, and provides an overview of the objective of the project. Section 3 talks about the methods used for statistical analysis. Here the concepts of multiple linear regression, assumptions, estimation of coefficients, residuals, adjusted coefficient of determination, residual plots, dummy coding for categorical covariates, hypothesis testing and confidence intervals, best subset selection with Akaike information criterion (AIC) and Bayesian information criterion (BIC) with formulations, and graphical representations are explained. In the subsequent Section 4, a linear regression model is fitted to the given data set and the results are compared and interpreted. The last section contains a review of the findings and the valuable takeaways from the analysis.

2 Problem Statement

2.1 Data Set and Data Quality

This project uses a data set "bodymeasurement.csv" provided by the course instructors. The data set contains information on human body measurements such as the chest, belly, biceps, knee, ankle, wrist, thigh, and calf circumferences, as well as the body height, weight, age, and sex for an individual aged between 18 and 40 years. It has a total of 424 observations and 13 variables. The variable *ID* has been removed as it is not relevant for regression analysis. The measurements of each variable are precise and accurate, with no missing values across all the variables, ensuring the quality of the data.

The variables *Height*, *Chest*, *Belly*, *Biceps*, *Knee*, *Ankle*, *Wrist*, *Thigh*, and *Calf* are numeric continuous variables, and these body measurement values are given in centimeters (cm). Another numeric continuous variable *Weight* whose values are given in kilograms (kg). Moreover, *Sex* is a nominal variable that gives information about the sex of individuals.

2.2 Project Objectives

The goal of this project is to fit a multiple linear regression model on the provided data set. Initially, a brief descriptive analysis is performed on the given data set, and the results are summarized. Next, a linear regression model with all the covariates is fit to the model, and the goodness-of-fit is interpreted. To find the best subset of covariates, using the AIC and BIC criteria of the best subset selection procedure, two subsets of covariates with the lowest AIC and BIC values are selected. The selected variables in these two models are compared. The model with the lowest BIC is selected, and the assumptions of the linear model are verified using residual and Q-Q plots. The regression coefficients of the covariates are estimated and interpreted. The estimated confidence intervals and the statistical significance of these covariates using t-tests are interpreted and summarised. Finally, the goodness-of-fit of the fitted model is evaluated using the adjusted coefficient of determination value.

3 Statistical Methods

In this section, all the statistical methods which are used for the analysis of this project are discussed in detail. For this regression analysis task, software Jupyter Notebook (Fernando Perez, 2015) Version 6.3.0, R (R Core Team, 2020) Version 4.1.0 is used. The graphs are created using R packages namely, ggplot2 (Wickham, 2016), making use of the packages dplyr (Wickham, 2022), olsrr (Aravind, 2020) and gridExtra (Auguie, 2017).

3.1 Multiple Linear Regression Model

Multiple linear regression analysis is a statistical method to model the linear relationship between one or more covariates and a dependent, continuous response. Let $(x_1, \ldots, x_p)$ represents the set of covariates (continuous or categorical) and Y represents the response variable (continuous) with the function $f(x_1, \ldots, x_p)$ modelling their relationship. The presence of random errors ε makes the relationship non-deterministic. As a result, it follows that the response variable Y is a random variable (Fahrmeir, 2013, p.73-75). The model is represented as:

$$Y = f(x_1, \ldots, x_p) + \varepsilon \ .$$

This function f is the linear combination of covariates and represented as follows:

$$f(x_1, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \ ,$$

where β_0 is the intercept and $\beta_1, \ldots, \beta_p$ are parameters of the model that are unknown and need to be estimated. Using the vector of unknown parameters $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)'$ and the vector of covariates $\boldsymbol{x} = (1, x_1, \ldots, x_p)'$ the model function is defined as:

$$f(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}.$$

The vector of responses and error terms are represented as,

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \ ,$$

and the design matrix of the covariates $\boldsymbol{X}$,

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix},$$

then the matrix notation is given as:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

(Fahrmeir, 2013, p. 73-75).

3.1.1 Assumptions on Error Terms

Fitting a multiple linear regression model requires the following assumptions to be valid. First, the covariates and the response variable are assumed to have a linear relationship. Second, the expected value of each error term is equal to zero, i.e., $E(\varepsilon_i) = 0, \quad \forall\ i = 1, \dots, n$. Third, the variance of the error terms are constant for each observation i.e., $Var(\varepsilon_i) = \sigma^2$. Furthermore, there is no correlation between error terms, and homoscedastic errors with covariance matrix, $\text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix. There exists no linear dependency between the covariates (no multi-collinearity), i.e., the design matrix X has a full column rank, i.e., $rk(X) = p+1$. The final and important assumption for significance testing is that the error terms ϵ_i are considered to be identical, independent and normally distributed. From the two assumptions mentioned above, the distribution has a mean value of 0 and a constant variance i.e., $\varepsilon_i \sim N(0, \sigma^2)$. Finally, the covariates and the error terms are assumed stochastically independent (Fahrmeir, 2013, p. 75-76).

3.1.2 Dummy Coding for Categorical Variables

Covariates can be either continuous or categorical. In the case of categorical variables, the interpretation is not possible as a linear additive effect to the response. Consider a covariate with k categories, dummy coding is one method where it introduces dummy variables for $k-1$ categories. These dummy variables are assigned to have binary values

(0 or 1). The introduced dummy variables are represented as:

$$x_{i1} = \begin{cases} 1 & x_i = 1, \\ 0 & \text{otherwise}, \end{cases} \qquad \dots \qquad x_{i,k-1} = \begin{cases} 1 & x_i = k-1, \\ 0 & \text{otherwise}, \end{cases}$$

for $i = 1, ..., n$. The dummy coded variables are then included in the regression model as:

$$Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{i,k-1} x_{i,k-1} + ... + \varepsilon_i \ .$$

To maintain identifiability, the dummy variable for one category is excluded and used as the reference category to interpret the coefficients of other dummy variables as a direct comparison (Fahrmeir, 2013, p. 97).

3.1.3 Estimation of Regression Coefficients and Interpretation

The most commonly used approach for estimating the coefficients of multiple linear regression model is the least squares estimation (LS), as the estimator function is differentiable and can be solved analytically. The goal is to find the regression coefficients which minimises the sum of squared deviations between the true response and the predicted response values. The least squares estimator is given by the formula:

$$LS(\beta) = \sum_{i=1}^{n} (y_i - x_i'\beta)^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon \ .$$

The vector notation of the estimator is given as:

$$LS(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^2.$$

Since the estimator function is differentiable, taking the first partial derivative of the LS estimator function with respect to the $\boldsymbol{\beta}$ and solving for $\boldsymbol{\beta}$ by equating to 0 gives the least squares estimates $\hat{\boldsymbol{\beta}}$ for the regression coefficient.

$$\frac{\partial LS(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0 \ , \qquad \hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \ .$$

The coefficient $\hat{\beta}_j$ is interpreted as a linear additive value to the predicted response. The j^{th} covariate is interpreted while fixing all other covariates to a constant value. If the $\hat{\beta}_j$ is positive, increase in one unit of the covariate will increase the predicted response

with estimated coefficient value, and if negative, it decreases the predicted response by factor of coefficient value. If the value is 0, the predicted response is not affected by the respective covariate. In case of the dummy coded covariate, the estimated coefficient is added to the predicted response when compared to the reference category. The intercept $\hat{\beta}_0$ can be interpreted as the average response (Fahrmeir, 2013, p. 104-107).

3.1.4 Residuals

The residuals are used to validate the assumptions of the multiple regression linear model. The residuals ε_i is defined as the deviation between the true response y_i and the predicted response $\hat{y}_i$, given by the formula:

$$\varepsilon_i = y_i - \hat{y}_i, \quad \forall\, i = 1, ..., n,$$

where the $\hat{y}_i$ is the predicted response calculated by the estimated coefficiencts $\hat{\boldsymbol{\beta}}$ and the vector of predicted responses are given by:

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

where $\boldsymbol{H}$ is a $n \times n$ hat matrix. The hat matrix $\boldsymbol{H}$ is idempotent and symmetric. For n observations, the vector of residuals is given by:

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}.$$

From the assumption of normally distributed error terms i.e., $\varepsilon_i \sim N(0, \sigma^2)$, it follows that the predicted response is normally distributed, $\hat{y}_i \sim N(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2)$. Using the maximum likelihood estimation, the unbiased estimate for the variance σ^2 with the calculated residuals is given by the formula:

$$\hat{\sigma}^2 = \frac{1}{n - (p+1)}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}\ ,$$

(Fahrmeir, 2013, p.124).

3.1.5 Residual Plots

In the multiple linear regression model, a residual plot is a simple scatter plot that shows the residuals, which is the difference between the true response values and the predicted response values on the y-axis, and the fitted values or the predicted responses on the x-axis. These plots are used to check for patterns in the residuals, which can indicate a good or poor fit of the model. The mean value of the residuals is taken as the intercept of the reference line. If the residuals are randomly scattered around the reference line with no pattern, it implies that the residuals are homoscedastic (Fahrmeir, 2013, p. 183).

3.1.6 Adjusted Coefficient of Determination

The coefficient of determination is the percentage of the total variance in the response variable that is explained by the fitted covariates in the linear regression model. The coeffecient of determination and the empirical correlation coefficient are closely related and it can be used to measure the goodness of fit. The coefficient of determination R^2 is given by the formula:

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

where $\bar{y}$ is the mean value of the true response variable. $\sum_{i=1}^{n}(\hat{y}_i - \hat{y})^2$ is the sum of squared errors of predicted values and the response mean. The model with R^2 value close to 1 indicates that the model has smaller residuals and hence the model fit is better. If the R^2 value is close to 0 then the model is not fitting the data better as it gives larger residuals. If R^2 is exactly equals to 0 it means that the predicted response values $\hat{y}_i$ are same as the response mean value $\bar{y}$ (Fahrmeir, 2013, p. 112). The R^2 value limits comparing models with different number of covariates and selecting the model with the better fit is difficult as the R^2 value, is either constant or increasing as a covariate is added to the model. Including the number of covariates, an adjusted coefficient of determination is calculated, given by the formula:

$$Adj\ R^2 = 1 - \frac{n-1}{n-|M|}(1 - R^2) \ ,$$

where n and $|M|$ is the total number of observations and the number of covariates respectively and the interpretation is same as the R^2.

3.1.7 Hypothesis Testing and Confidence Intervals

For the linear regression model, hypothesis testing for a j^{th} covariate is used to determine whether the coefficient of the j^{th} covariate is statistically significant or not under the assumption of normality of error terms, i.e., $\varepsilon_i \sim N(0, \sigma^2)$. The null hypothesis H_0 is that the coefficent of the j^{th} covariate is equal to zero, which means that the covariate has no effect on the response variable. The alternative hypothesis H_1 is that the j^{th} coefficient is not equal to zero, which implies that the covariate has an effect on the response variable:

$$H_0 : \beta_j = 0 \qquad \text{and} \qquad H_1 : \beta_j \neq 0, \quad j = 1, ..., p.$$

The most commonly used testing method to test the above mentioned null hypothesis is t-test. The test statistic t_j for the j^{th} covariate is a t-distributed random variable with $n-(p+1)$ degrees of freedom with a chosen significance level α is given by the formula:

$$t_j = \frac{\hat{\beta}_j}{\hat{se}_j} \ ,$$

where $\hat{se}_j = \widehat{Var(\hat{\beta}_j)}^{1/2}$ indicates the estimated standard deviation or standard error of the regression coefficients $\hat{\beta}_j$. From the t-distribution the $(1-\alpha/2)$-quantile with $n-(p+1)$ degrees of freedom is used to determine the critical value of the rejection region. Therefore, if $|t_j| > t_{n-(p+1)}(1-\alpha/2)$ the null hypothesis is rejected. The significance is also decided by p-value of the test statistic, if it is less than the significance level α and it is found in the t-distribution table with $n-(p+1)$ degrees of freedom (Fahrmeir, 2013, p.125-131).

For the least squares estimated coefficients of linear regression, the $100(1-\alpha)\%$ confidence interval describes the interval range corresponding to β_j with a $100(1-\alpha)\%$ probability for $j = 0, 1, ..., p$. The t_j-test statistic and estimated standard deviation $\hat{se}_j$ can be used to construct the confidence intervals for a single coefficient β_j under the assumption of normality of errors is constructed as:

$$[\hat{\beta}_j - t_{n-(p+1)}(1-\alpha/2) \cdot \hat{se}_j, \ \hat{\beta}_j + t_{n-(p+1)}(1-\alpha/2) \cdot \hat{se}_j] \ ,$$

(Fahrmeir, 2013, p. 136).

3.2 Best Subset Selection

The best subset selection is a method for selecting the best subset of covariates in a linear regression model. The method involves fitting a linear model for all possible combination of subsets of covariates. The best subset can be selected based on two criteria: AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion). The first step of the procedure is to fit a model with no covariates, and in the second step to fit a model with one covariate at a time and select the covariates which have the lowest AIC or BIC value. Repeat the second step for all possible combinations of subsets, which results in fitting 2^p models, where p is the total number of covariates (Fahrmeir, 2013, p. 146).

3.2.1 Akaike Information Criterion (AIC)

As discussed above, the AIC is one of the goodness-of-fit measure for selecting the best subset of covariates for a model. It determines how much information is lost by a model. It is mostly applied in likelihood-based inference, where the lower the AIC value, the better the model. The log likelihood of the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ with a penalizing component based on the number of parameters employed in the model, the AIC for a model with $p+1$ parameters is given by the formula:

$$\text{AIC} = -2 \cdot l(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) + 2(p+1) \ .$$

Assuming that the response are normally distributed as the error terms are normally distributed i.e., $Y \sim N(X\beta, \sigma^2 I)$), the least squares estimates are equal to the maximum likelihood estimation, the AIC is:

$$\text{AIC} = n \log(\hat{\sigma}^2) + 2(p+1) \ ,$$

where $\hat{\sigma}^2$ is the biased variance estimator $\hat{\sigma}^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/n$. The model with smallest AIC value is considered the best-fit model and the p parameters are best subset of covariates (Fahrmeir, 2013, p. 148).

3.2.2 Bayesian Information Criterion (BIC)

Similar to the AIC, the Bayesian information criterion (BIC) is also a goodness-of-fit measure for the selection of the best subset of covariates. The model is better fit when the BIC value is lower. The model will not be a better fit if additional parameters are added to increase the value of a likelihood function during the model fitting. By adding a penalty term for the total number of parameters, and the total number of observations, the BIC is calculated by the formula:

$$\text{BIC} = -2.l\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2\right) + \log(n)(|p| + 1) \ .$$

When Gaussian errors are taken into account, BIC is defined as:

$$\text{BIC} = n \cdot \log\left(\hat{\sigma}^2\right) + \log(n)(|p| + 1) \ .$$

Like AIC, the model with smallest BIC value is considered the best-fit model and the p parameters are best subset of covariates (Fahrmeir, 2013, p. 149).

4 Statistical Analysis

In this section, all the tests that are performed and the underlying assumptions required for these models are discussed. The interpretation of the results with regard to the problem statement is presented in this section.

4.1 Descriptive Analysis of the Data set

A descriptive analysis is performed on the given data set. The value of the variable age ranges from 18 to 40 years, and the median value of age in years is 25. The average height of a person is 170.88 cm with a standard deviation of 9.40 cm and the tallest and shortest individuals in the data set are measured at 198.10 cm and 147.20 cm respectively. In Figure 1, the histogram representing the distribution, which follows a bell shaped structure, implying that the variable height is normally distributed and symmetric around its mean. Table 3 on page 18 in the appendix provides values of central tendency, variability of the variables (continuous) and height (response variable).

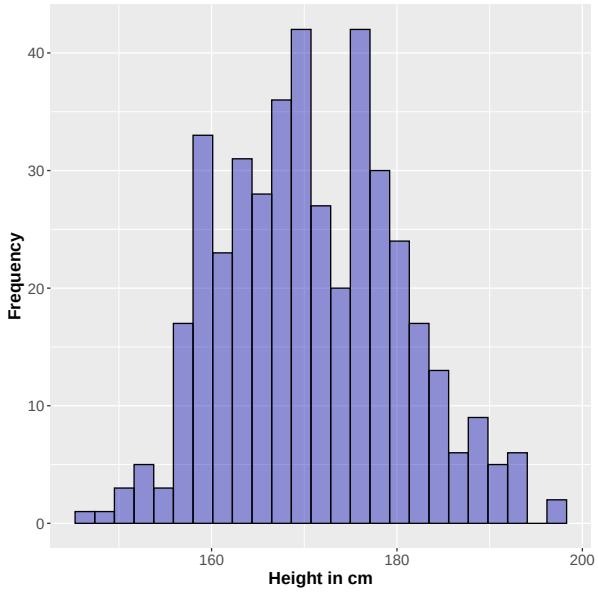


Figure 1: Histogram for the Variable Height in Centimeter (cm.)

Figure 3 on page 17 in the appendix, shows the correlation scatter plot for the response variable height against all other variables in the data set. From the plot, it is evident that there exists a strong positive relationship among the covariates weight, wrist, and chest with variable height. Nevertheless, the variable age shows weak correlation with response variable height.

4.2 Multiple Linear Regression Model with all Covariates

Considering the variable height as the response, with all other 11 variables as covariates, a multiple linear regression model is fit to the data. For the categorical variable sex, x_{male} is taken as the dummy variable, for the category "male", while the category "female" is considered as the reference category. The model function is given by:

$$\begin{aligned}\hat{y}_{height} = 211.231 + 0.0176 \cdot x_{Age} + 6.380 \cdot x_{Male} - 0.145 \cdot x_{Chest} - 0.674 \cdot x_{Belly} \\ -0.588 \cdot x_{Thigh} - 0.096 \cdot x_{Knee} - 0.673 \cdot x_{Calf} + 0.101 \cdot x_{Ankle} \\ -1.013 \cdot x_{Biceps} + 0.520 \cdot x_{Wrist} + 1.403 \cdot x_{Weight}.\end{aligned}$$

The linear regression model's results, including the estimated coefficients, standard errors, R^2 and adjusted R^2 values, can be found in Table 4 on page 18 of the appendix. The model's AIC value is 2443.9 and BIC value is 2496.5. Estimated $R^2 = 0.801$ and

adjusted $R^2 = 0.796$ indicates that the data is well fitted by the model, as both the values are closer to 1.

4.3 Best Model Selection from AIC and BIC

The subset of covariates, according to AIC and BIC, best fits the model is identified in this subsection. Based on all possible combinations of the 11 covariates, 2^{11} models are fitted. The best subset is selected for each combination of the covariate based on the AIC and BIC values. Table 1 tabulates the values of adjusted R^2, AIC and BIC, and every model that were fitted as best subset selected model. From all the models fit, the model with lowest AIC and lowest BIC is selected as the best subset of covariates. The linear model with lowest AIC value 2438.82 is model 8 with 8 covariates: *Sex*, *Chest*, *Belly*, *Thigh*, *Calf*, *Biceps*, *Wrist*, and *Weight*. Analogously, the linear model with lowest BIC value 2474.511 is the model 6 with 6 covariates: *Sex*, *Belly*, *Thigh*, *Calf*, *Biceps*, and *Weight*. By comparing both the models, model 6 have adjusted R^2 value approximately equal to the model 8, even after removing two covariates and the BIC value is also lower.

Table 1: Best subset selection with adjusted R^2, AIC and BIC.

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

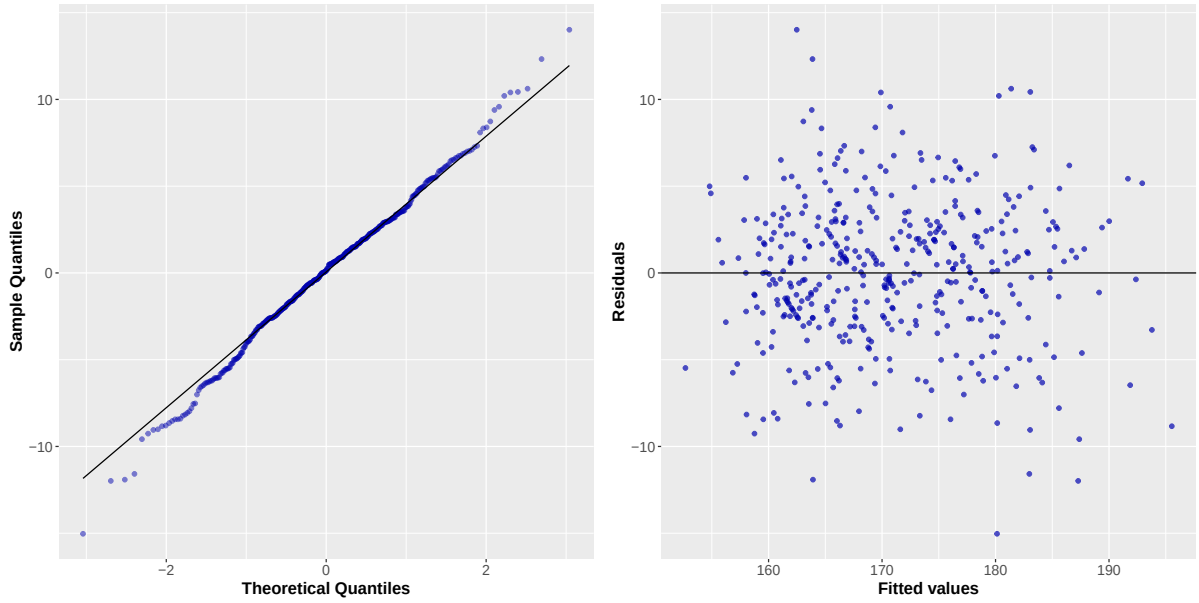
4.4 Interpretation of Model Coefficients, Confidence Intervals and Significance

Based on the subset of covariates based on the BIC criterion in previous section, a multiple linear regression model is fitted. The linear model assumptions are verified for the fitted model: the assumption of independence of observations states that the observations in the dataset should be independent of each other, and should not be influenced by other observations in the dataset. A simple approach is to examine the correlation of the data. Since the covariance matrix is assumed to be diagonal matrix, correlation between the observations is low, this implies the observations are independent. The design matrix is verified to have a full column rank and linear independence among covariates, indicating no multicollinearity.

Table 2: Estimated Coefficient, p-values and Confidence Intervals

	Estimated coefficients	Confidence intervals	P-values
(Intercept)	213.721	[205.194 , 222.247]	0.00000e+00
Sexm	6.609	[4.746 , 8.472]	0.00000e+00
Belly	-0.727	[-0.832 , -0.623]	0.00000e+00
Thigh	-0.598	[-0.78 , -0.416]	3.00000e-10
Calf	-0.528	[-0.778 , -0.278]	4.07118e-05
Biceps	-1.108	[-1.34 , -0.877]	0.00000e+00
Weight	1.416	[1.302 , 1.53]	0.00000e+00
Residual standard error: 4.265 Adjusted R-squared: 0.7941 F-statistic: 272.8			

The residuals from the fitted linear model are then used to create residual and Q-Q normality plots shown in Figure 2, which show that the residuals are homoscedastic and normally distributed. The residual plot in Figure 2(b) shows that there is no pattern in the residuals and they are scattered randomly around the expected value of zero. The Q-Q plot in Figure 2(a) demonstrates that the empirical quantiles of the residuals fit close to the theoretical quantiles of a normal distribution, with the exception of a few points in the tails. This confirms the assumption of normality for the residuals. Table 2 provides information on the estimated coefficients of the parameters along with their corresponding p-values and confidence intervals.



(a) Q-Q normality plot (b) Residual plot

Figure 2: Residual and Q-Q plots for the fitted linear model.

Based on the estimated coefficient the predicted response is given by the model function:

$$\hat{y}_{height} = 213.72 + 6.61 \cdot x_{Male} - 0.73 \cdot x_{Belly} - 0.60 \cdot x_{Thigh} - 0.53 \cdot x_{Calf} \\ -1.11 \cdot x_{Biceps} + 1.42 \cdot x_{Weight}.$$

The effect of each covariate on the predicted response variable is determined by keeping all other covariates constant. The coefficient for the dummy variable "sexm" suggests that, on average, males are predicted to be taller than females by 6.61 cm. Additionally, increasing a person's weight by 1 kilogram is associated with a predicted increase in height of 1.42 cm. The result shows that increasing the circumference of the belly, thigh, calf, and biceps by 1 cm is associated with a decrease in predicted height of 0.73 cm, 0.60 cm, 0.53 cm, and 1.11 cm respectively. All of the t-tests have p-values less than the signficance level $\alpha = 0.05$, indicating that all covariates have a statistically significant effect on the response variable.

5 Summary

Healthcare professionals including developmental psychologists use growth charts to track the development of individuals over a certain period of time. They compare the height, weight, age, sex, and measurements of body parts such as head circumference, biceps, and chest to the expected parameters of other healthy individuals of the same age and sex. The main objective of the project was to study the height of an individual based on the other body measurements using a multiple linear regression model.

At first, the data set was subjected to descriptive analysis. Average value of response variable height was found to be 170.88 cm. After the descriptive analysis, the linear regression model was fitted using the response variable height against 11 covariates. Goodness of fit was measured for the fitted linear model using adjusted R2. The adjusted R2 value was 0.79 which indicates that the model has fitted the data better. Using the best subset selection method, next task was carried out to find the suitable subset of covariates based on AIC and BIC values. The model chosen based on lowest AIC value was fitted using the covariates sex, belly, wrist, thigh, calf, chest, biceps, and weight. By taking the lowest BIC value the variables sex, belly, thigh, calf, biceps, and weight were fitted against height. The goodness of fit measured for both the models using adjusted R^2 indicates that both the models are fitting the data better.

The model which was fitted with the six selected covariates and the response variable height after determining the best subset of covariates using the BIC criterion. The linear model's assumptions were then validated. The data in the study was determined to be independent and without any issues of multicollinearity. The residuals were found to be evenly distributed, with no patterns, and the mean value was zero. The normality assumption for the residuals was also supported. The analysis of the linear model revealed that, on average, men were taller than women. As weight increased, there was a corresponding increase in predicted height. However, an increase in other body measurements such as belly size, thigh size, calf size, and biceps size tended to result in a decrease in predicted height. It was found that all six covariates fitted in the model had a significant impact on predicted height through t-tests.

The fitted regression model limits the flexibility of the model based on various assumptions. For further research on this topic, a much more flexible model with less assumptions, such as a generalised linear model or non-linear models, can be a better fit to the data.

Bibliography

Aravind, H. (2020), *olsrr: Tools for Building OLS Regression Models*. R package version 0.5.3.

Auguie, B. (2017), *gridExtra: Miscellaneous Functions for "Grid" Graphics*. R package version 2.3.

Fahrmeir, K. (2013), *Regression, models,methods and applications*, Springer. ISBN 978-3-642-34333-9.

Fernando Perez, B. G. (2015), *Jupyter: Interactive data science and scientific computing in all programming languages URL: https://jupyter.org/*. version 6.3.0.

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Wickham, H. (2016), *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag. R package version 3.4.0.

Wickham, H. (2022), *dplyr: A Grammar of Data Manipulation*. R package version 1.0.10.

Wikipedia (2022), *Growth chart, URL: https://en.wikipedia.org/wiki/Growth_chart*. (visited on 27th January 2023).

Appendix

A Additional Figures

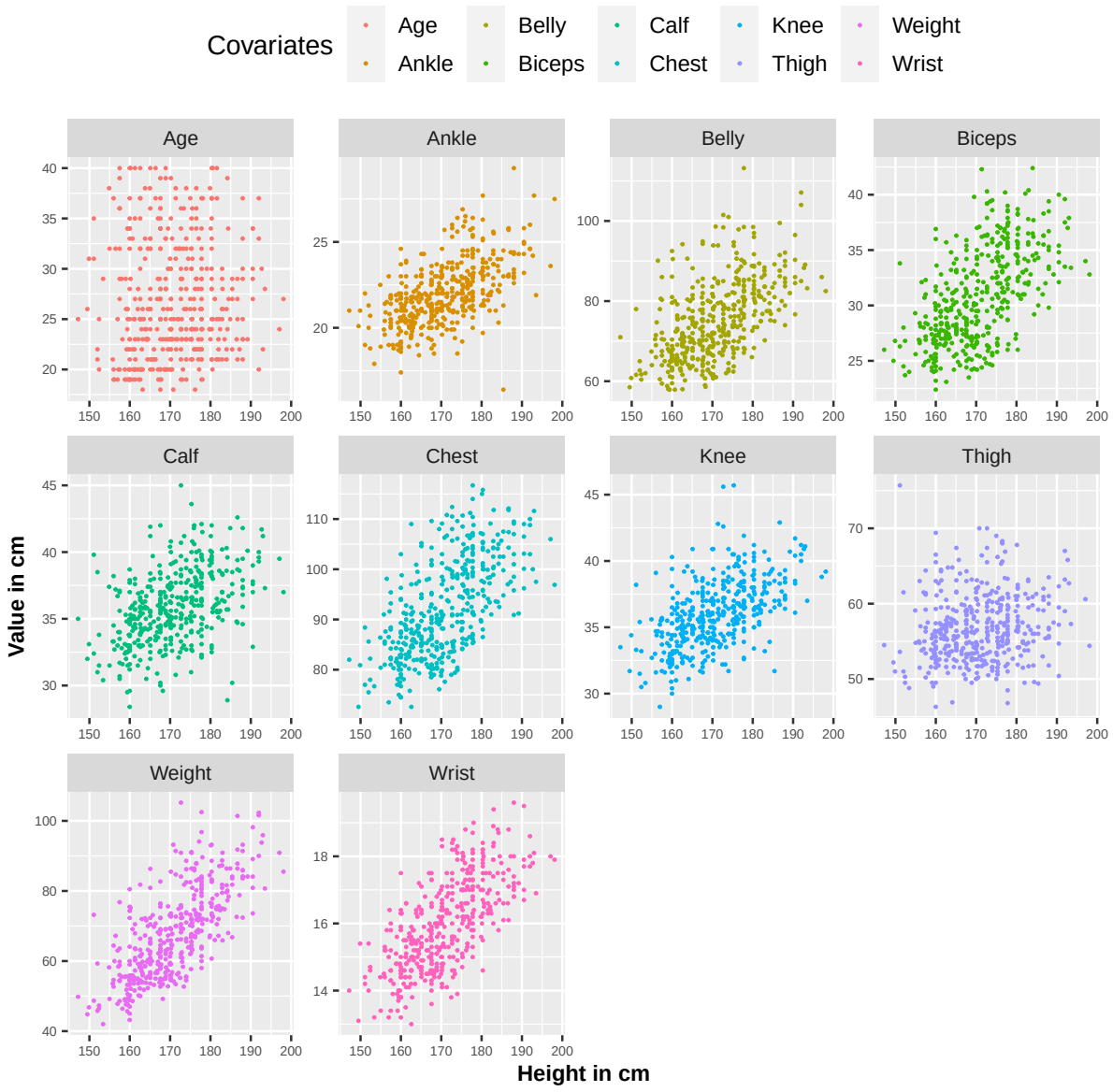


Figure 3: Correlation Scatter Plot for the Variable Height against all 11 Covariates.

B Additional Tables

Table 3: Statistical descriptive analysis of all continuous Variables

Variable	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 4: Estimated Coefficient, Std Errors and P-Values

	Estimated Coefficients	Std. Error	t-values	P-value
(Intercept)	211.23128	6.96030	30.348	$<$ 2e-16
Age	0.01769	0.03802	0.465	0.6420
Sexm	6.38037	0.98056	6.507	2.23e-10
Chest	-0.14534	0.06964	-2.087	0.0375
Belly	-0.67402	0.05946	-11.335	$<$ 2e-16
Thigh	-0.58815	0.10001	-5.881	8.45e-09
Knee	0.09616	0.17084	0.563	0.5738
Thigh	-0.58815	0.10001	-5.881	8.45e-09
Calf	-0.67324	0.14570	-4.621	5.12e-06
Ankle	0.10198	0.20798	0.490	0.6242
Biceps	-1.01355	0.15004	-6.755	4.88e-11
Wrist	0.52087	0.39333	1.324	0.1861
Weight	1.40384	0.06858	20.469	$<$ 2e-16
Residual standard error: 4.249				
Adjusted R-squared: 0.7956				
F-statistic: 150.6				