Applications of Linear Algebra in Linear Models

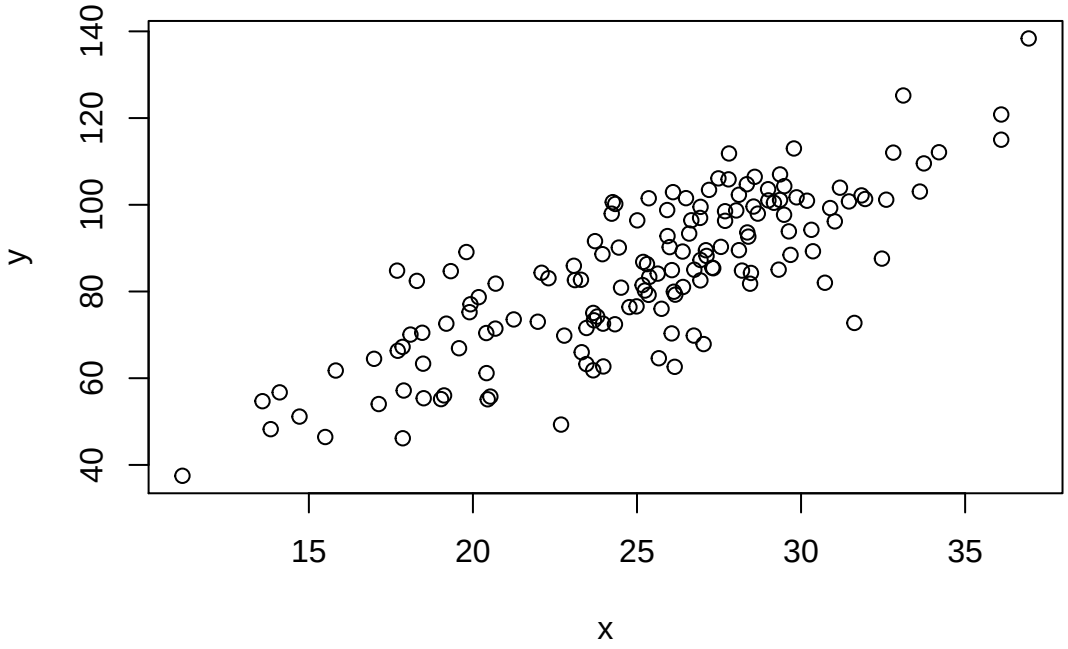
Joshua D. Ingram *New College of Florida*

This is the final project for the Spring 2020 linear algebra course. This document contains an overview of the use of linear algebra in linear models. First, the matrix form of a multiple linear regression equation is given, followed by the derivation and application of the OLS estimator.

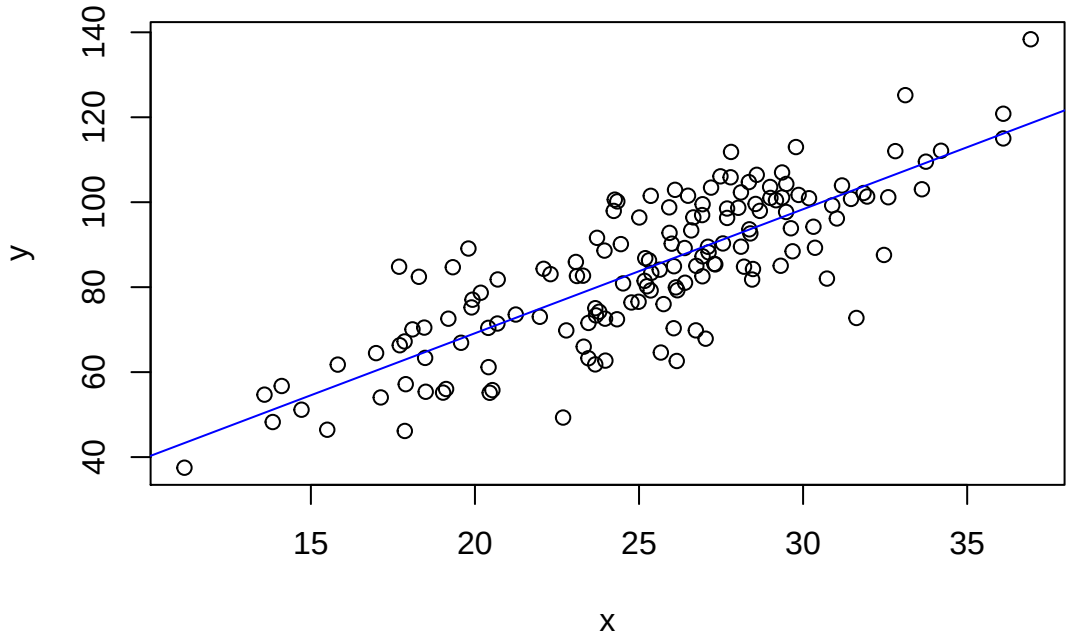
Introduction to Ordinary Least Squares Regression

In many applications of linear algebra, the equation A**x**=**b** is inconsistent. With situations like this, one can use a least squares solution to approximate **b**. If A is an m x n matrix (m > n) and **b** is in $\mathbf{R}^m$, a least squares solution is an $\hat{\mathbf{x}}$ in $\mathbf{R}^n$ that minimizes the distance between $A\hat{\mathbf{x}}$ and **b**. Least squares solutions are especially useful in the field of statistics. Rather than using $A\mathbf{x} = \mathbf{b}$, the form $X\vec{\beta} = \mathbf{y}$ is used. X is referred to as the "design matrix", $\vec{\beta}$ is the "parameter vector", and **y** is the "response vector." Ordinary least sqaures (OLS) regression allows statisticians to use randomly collected data to model relationships between several variables of interest.

When given a set of data with n observations that contains two variables, say y (the response variable) and x_1 (the explanatory variable), we often want to model the relationship between x_1 and y as a linear equation. This is written in the form $y_i = \beta_0 + \beta_1 x_{1,i} + \epsilon_i$, with β_0 being our intercept, β_1 the coefficient for x_1, ϵ_i as our error term, and the subscript i denoting the *ith* observation from our n observations. However, when looking at a random cloud of data points like in the graph below, how exactly do we find a line that fits this data?



The goal of a statistician is to find the β_0 and β_1 that estimates y_i the best. The line created from these estimated coefficients, denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$, is called the least-squares linear regression line because it minimizes the sum of squared residuals. A squared residual is the square of the difference between the observed value of our response (y_i) at $x_{1,i}$ and our predicted value of y_i at $x_{1,i}$ ($\hat{y}_i$). We find the OLS regression line by finding a least squares solution to $X\vec{\beta} = \mathbf{y}$ and once we fit the regression line, it will look like the line in the graph below.



We can rewrite $y = \beta_0 + \beta_1 x_{1,i} + \epsilon_i$ in matrix form as $\mathbf{y} = X\vec{\beta} + \vec{\epsilon}$. Where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{1,1} \\ 1 & x_{1,2} \\ \vdots & \vdots \\ 1 & x_{1,n} \end{bmatrix}, \vec{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \vec{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

The least squares solution to this matrix form equation that gives the beta estimates is given by

$$\hat{\vec{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$$

This solution allows us to find the estimated modeling equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i}$. It's use can be extended to a situation where we want to model the relationship between a response variable y and k explanatory variables. The scalar form for a multiple linear regression model with k explanatory variables is written as $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_k x_{k,i} + \epsilon_i$. The expanded matrix form for a model with k predictors looks like this

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \dots & x_{k,1} \\ 1 & x_{1,2} & x_{2,2} & \dots & x_{k,2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,n} & x_{2,n} & \dots & x_{k,n} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Derivation of the Ordinary Least Squrares Estimator

Before deriving the formula for the least squares solution, we need to be familiar with orthogonality, orthogonal projections, the *Orthogonal Decomposition Theorem*, and the *Best Approximation Theorem*. Stating that two vectors in $\mathbf{R}^n$ are orthogonal is the same as stating that these vectors are perpendicular. For two vectors $\mathbf{u}$ amd $\mathbf{v}$ to be orthogonal, their dot products must be equal to 0. An orthogonal set is a set of vectors in $\mathbf{R}^n$ where any pair of distinct vectors in the set is orthogonal and this set is an orthogonal basis for a subspace W of $\mathbf{R}^n$ if it is a basis for W.

To understand orthogonal projections and their usefulness, say we have a vector $\mathbf{y}$ in $\mathbf{R}^n$ and a vector $\mathbf{x}$ in a subspace W of $\mathbf{R}^n$. We can represent $\mathbf{y}$ as a combination of a vector $\vec{\epsilon}$ that is orthogonal to $\mathbf{x}$, where $\vec{\epsilon} = \mathbf{y} - \hat{\mathbf{y}}$, and a vector $\hat{\mathbf{y}}$ that is a multiple of $\mathbf{x}$, where $\hat{\mathbf{y}} = \beta\mathbf{x}$. Written as

$$\mathbf{y} = \hat{\mathbf{y}} + \vec{\epsilon}$$

$\vec{\epsilon}$ is orthogonal to $\mathbf{x}$ if and only if the dot product of $\mathbf{x}$ and $\vec{\epsilon} = 0$. Under this condition, we can find the values of β and $\hat{\mathbf{y}}$. Shown below

$$\vec{\epsilon} \cdot \mathbf{x} = 0$$

$$(\mathbf{y} - \beta\mathbf{x}) \cdot \mathbf{x} = 0$$

$$\mathbf{y} \cdot \mathbf{x} - (\beta\mathbf{x}) \cdot \mathbf{x} = 0$$

$$\mathbf{y} \cdot \mathbf{x} - \beta(\mathbf{x} \cdot \mathbf{x}) = 0$$

$$\mathbf{y} \cdot \mathbf{x} = \beta(\mathbf{x} \cdot \mathbf{x})$$

$$\beta = \frac{\mathbf{y} \cdot \mathbf{x}}{(\mathbf{x} \cdot \mathbf{x})}$$

$$\hat{\mathbf{y}} = \beta\mathbf{x}$$

$$\hat{\mathbf{y}} = \frac{\mathbf{y} \cdot \mathbf{x}}{(\mathbf{x} \cdot \mathbf{x})}\mathbf{x}$$

We refer to $\hat{\mathbf{y}}$ as the orthogonal projection of $\mathbf{y}$ onto $\mathbf{x}$ and $\vec{\epsilon}$ as the component of $\mathbf{y}$ orthogonal to $\mathbf{x}$. $\hat{\mathbf{y}}$ now has an easily readible formula: $\hat{\mathbf{y}} = \frac{\mathbf{y} \cdot \mathbf{x}}{(\mathbf{x} \cdot \mathbf{x})}\mathbf{x}$. This means we can "project" any vector in $\mathbf{R}^n$ onto a subspace W of $\mathbf{R}^n$. The *Orthogonal Decomposition Theorem* shows us that we can write any $\hat{\mathbf{y}}$ in a subspace W of $\mathbf{R}^n$ as a linear combination of $\{x_1, x_2, \dots, x_k\}$if that set of vectors forms an orthogonal basis of W. That is,

$$\hat{\mathbf{y}} = \frac{\mathbf{y} \cdot \mathbf{x_1}}{(\mathbf{x_1} \cdot \mathbf{x_1})}\mathbf{x_1} + \frac{\mathbf{y} \cdot \mathbf{x_2}}{(\mathbf{x_2} \cdot \mathbf{x_2})}\mathbf{x_2} + \ldots + \frac{\mathbf{y} \cdot \mathbf{x_k}}{(\mathbf{x_k} \cdot \mathbf{x_k})}\mathbf{x_k}$$

$$\hat{\mathbf{y}} = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

$$\mathbf{y} = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \vec{\epsilon}$$

While the notation for the hats on the β's are not the same, the results of this theorem come right back to our multiple linear regression equation. However, we aren't to our matrix form formula for our beta estimates just yet.

When finding our ordinary least squares linear regression equation, we want an equation that has the "best" estimate. Meaning, we want to minimize the sum of squared residuals. Well, the *Best Approximation Theorem* tells us that for a subspace W of $\mathbf{R}^n$ and any vector $\mathbf{y}$ in $\mathbf{R}^n$, $\hat{\mathbf{y}}$, the orthogonal projection of $\mathbf{y}$ onto W, is the closest point in W to $\mathbf{y}$.

Let's write $\mathbf{y} = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \vec{\epsilon}$ into a matrix form equation $\mathbf{y} = X\vec{\beta}$, where

$$X = \begin{bmatrix} x_1 & x_2 & \ldots & x_k \end{bmatrix}, \vec{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

This now brings us to finding our least squares solution. If X is an m x n matrix (m > n), $\mathbf{y}$ is in $\mathbf{R}^m$, and $\hat{\vec{\beta}}$ is in $\mathbf{R}^n$, then the least squares solution of an equation $\mathbf{y} = X\vec{\beta}$ is a vector $\hat{\vec{\beta}}$ that makes $\mathbf{y} - X\hat{\vec{\beta}}$ the smallest value possible for all $\vec{\beta}$ in $\mathbf{R}^n$.

From our earlier examples, we can write the projection of $\mathbf{y}$ onto the subspace spanned by the columns of X as $\hat{\mathbf{y}}$. In other words, $\hat{\mathbf{y}}$ is the projection of $\mathbf{y}$ onto the column space of X. Now we know that $\hat{\mathbf{y}} = X\hat{\vec{\beta}}$.

By the *Orthogonal Decomposition Theorem*, $\mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to the column space of X. Thus $\mathbf{y} - X\hat{\vec{\beta}}$ is orthogonal to the column space of X. For each column x_k of X, $x_k \cdot (\mathbf{y} - X\hat{\vec{\beta}}) = 0$. Since the dot product of two vectors in $\mathbf{u}$ and $\mathbf{v}$ in $\mathbf{R}^n$ equals $\mathbf{u}^T\mathbf{v}$, for all columns x_k of X, $x_k^T(\mathbf{y} - X\hat{\vec{\beta}}) = 0$. This can be rewritten as

$$X^T(\mathbf{y} - X\hat{\vec{\beta}}) = 0$$

This form allows us to reach the formula for the ordinary least squares estimator.

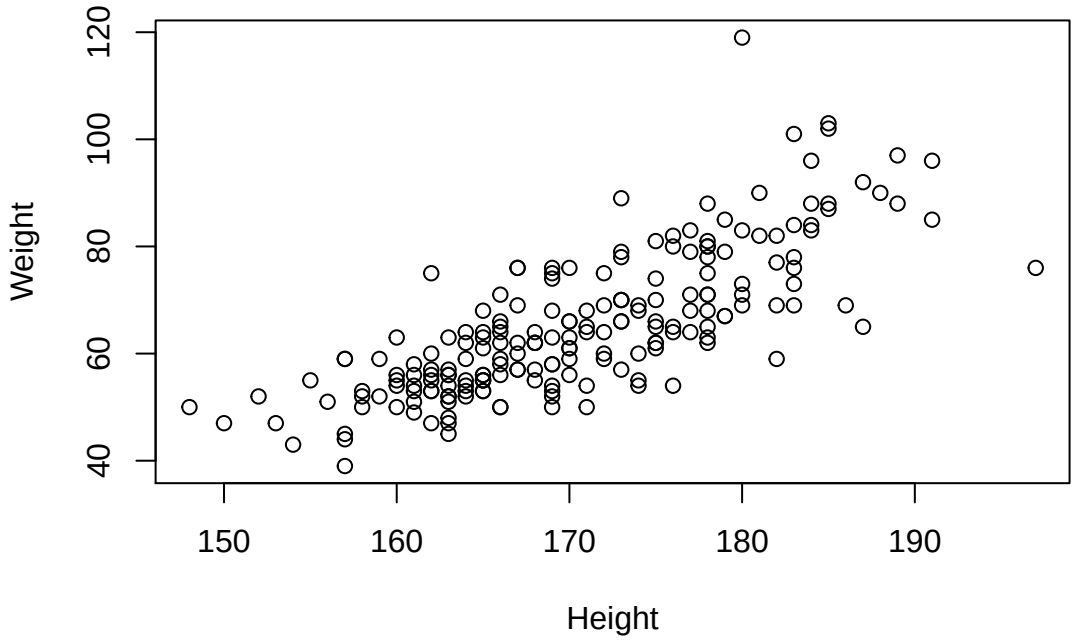
$$X^T\mathbf{y} - X^T X\hat{\vec{\beta}} = 0$$

$$X^T X\hat{\vec{\beta}} = X^T\mathbf{y}$$

$$\hat{\vec{\beta}} = (X^T X)^{-1} X^T\mathbf{y}$$

Applicaiton of the Ordinary Least Squares Estimator

We have now derived the matrix form solution for the OLS estimates, so let's apply this formula to a situation a statistician might find themselves faced with. We will be working with a subset of the "Davis" dataset that contains 199 observations and 5 variables, such as sex, weight, and height of the observations. We want to model the relationship where height (cm) predicts weight (kg), so let's look at the scatterplot to see what we are working with.



It looks like a simple linear relationship would be appropriate to model this data, so our theoretical model will take the form of $y_i = \beta_0 + \beta_1 x_{1,i} + \epsilon_i$, with $x_{1,i}$ being the height and y_i being the weight of the *ith* observation. We can place our data into the design matrix X and the response vector $\mathbf{y}$ so that we can set ourselves up to find the OLS beta coefficient estimates. The first column of X will consist of only 1's so that we can estimate the intercept β_0. The second column of X will consist of the height of each of the 199 observations. The response vector $\mathbf{y}$ will contain the weight corresponding to the height of the observations in X. Only the first few and last observations are given.

$$\mathbf{y} = X\vec{\beta} + \vec{\epsilon}$$

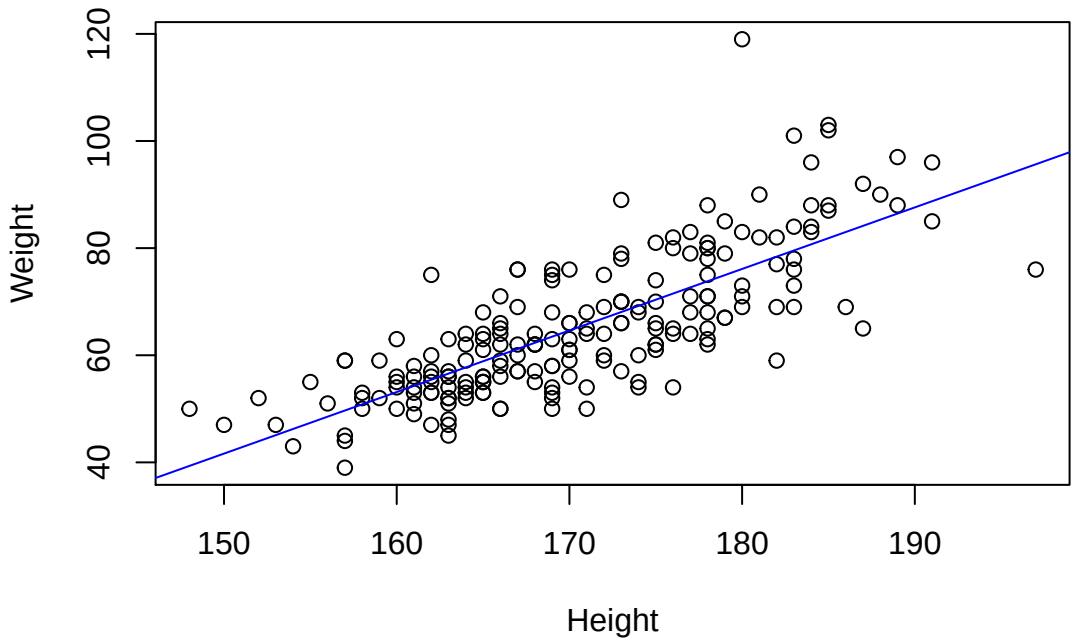
Where

$$\mathbf{y} = \begin{bmatrix} 77 \\ 58 \\ 53 \\ \vdots \\ 79 \end{bmatrix}, X = \begin{bmatrix} 1 & 182 \\ 1 & 161 \\ 1 & 161 \\ \vdots & \vdots \\ 1 & 177 \end{bmatrix}, \vec{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \vec{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_{199} \end{bmatrix}$$

Now that we have this in matrix form, we can can use the OLS estimator formula to find $\hat{\beta}_0$ and $\hat{\beta}_1$. We plug in X and $\mathbf{y}$ into $\hat{\vec{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$ and will get a vector that contains our estimates. Once the calculations are completed, we get

$$\hat{\vec{\beta}} = \begin{bmatrix} -130.747 \\ 1.149 \end{bmatrix}$$

This allows us to write the estimated modeling equation as $\hat{y}_i = -130.747 + 1.149x_{1,i}$. We can now graph the simple linear regression line from this equation.



Now that we have a graph of the relationship and our model to work with, we can interpret the relationship between height and weight, as well as make predictions using our model. Our line has a slope of 1.149, meaning that for every one cm increase in height, we predict the weight to increase by 1.149 kg, on average. What will we predict the weight of a subject to be given they have a height of 170 cm? All we have to do is plug in 170 for $x_{1,i}$.

$$\hat{y} = -130.747 + 1.149(170) = 64.583$$

Given a subject with a height of 170 cm, we would predict their weight to be 64.583 kg.

After deriving the formula for the OLS estimator and going through an example, hopefully one can see the usefulness of linear algebra in linear models. This project only scratched the surface of it's applications, but much more can be explored.

Sources

Lay, D. C., Lay, S. R., & McDonald, J. J. (2015). Least-Squares Problems. In Linear Algebra and Its Applications (5th Edition) (5th ed., pp. 362–367). Pearson.