Patrick Ekel
Austin Purdie

Linear Algebra is a highly applicable area of mathematics and is useful in measuring linear relationships between data. Linear regression is an approach for modeling the relationship between different data, and least squares is an application of linear regression for overdetermined systems. Our group decided to solve for the least squares in order to identify the correlation of different variables to a country's nominal GDP.

The two independent variables we looked at were amount of arable land and perceived level of government corruption. These were chosen partially on a whim, but also because it was hypothesized that countries with the potential for large agricultural industries would be most likely to satiate their citizens' hunger, and therefore be more focused on other revenue producing industries. Corruption was analyzed because my partner and I were curious about how citizens' perception of governmental corruption aligns with that country's GDP.

In the end our analysis showed that a country's amount of arable land was more highly correlated with GDP than a country's perceived level of corruption, but that the two taken together yielded the most accurate least square estimate.

We used a standard linear model to perform three regressions using data obtained from Transparency International, the CIA World Fact Book, and the World Bank. The first two regressions were generated by plotting GDP against the Corruption Perception Index (CPI) and total area of arable land individually, with GDP being the dependent variable. It is important to note that a higher CPI reflects more confidence in the integrity of a government and that a lower CPI reflects greater perceived corruption. CPI is measured on a 0-100 scale. The final regression is a multiple regression displaying the best fit plane by plotting both independent variables against GDP and is the regression which yielded the most insight. We chose to analyze the ten countries with the highest amount of arable land. Our data is as follows:

Country	GDP (millions USD)	CPI	Arable Land (km^2)
United States	18561934	74	1669302
India	2250990	40	1535063
China	11391619	40	1504350
Russia	1267750	29	1192300
Brazil	1769601	40	661299
Canada	1532343	82	474681
Australia	1256640	79	474550
Ukraine	87198	29	333847
Indonesia	940953	37	330037
Nigeria	415080	28	329334

More specifically, when performing these regressions, we are solving an equation of the form

$$\mathbf{y} = X\,\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{y}$ = GDP column vector, $\boldsymbol{\beta}$ = regression coefficients, and $\boldsymbol{\epsilon}$ = residual vector or error. X is our design matrix, the matrix of data for the applicable independent variable(s). The first column of a design matrix is a column of 1's. The subsequent columns are data. By finding a least squares solution for $\boldsymbol{\beta}$, we minimize the length of the residual vector $\boldsymbol{\epsilon}$ and find a model that best represents the data.

To find the least squares solutions for the following regressions, we first calculated the normal equations $X^TX\,\boldsymbol{\beta} = X^T\mathbf{y}$. Our least squares solution is given by solving for $\boldsymbol{\beta}$ such that $\boldsymbol{\beta} = (X^TX)^{-1}\,X^T\mathbf{y}$. Our analysis follows.

Linear Regression of Arable Land vs. GDP: The GDP column vector and design matrix for this regression are

$$\mathbf{y} = \begin{bmatrix} 18561934 \\ 2250990 \\ 11391619 \\ 1267750 \\ 1769601 \\ 1532343 \\ 1256640 \\ 87198 \\ 940953 \\ 415080 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 1669302 \\ 1 & 1535063 \\ 1 & 1504350 \\ 1 & 1192300 \\ 1 & 661299 \\ 1 & 474681 \\ 1 & 474550 \\ 1 & 333847 \\ 1 & 330037 \\ 1 & 329334 \end{bmatrix}$$

First, we find the normal equations by finding X^TX and $X^T\mathbf{y}$.

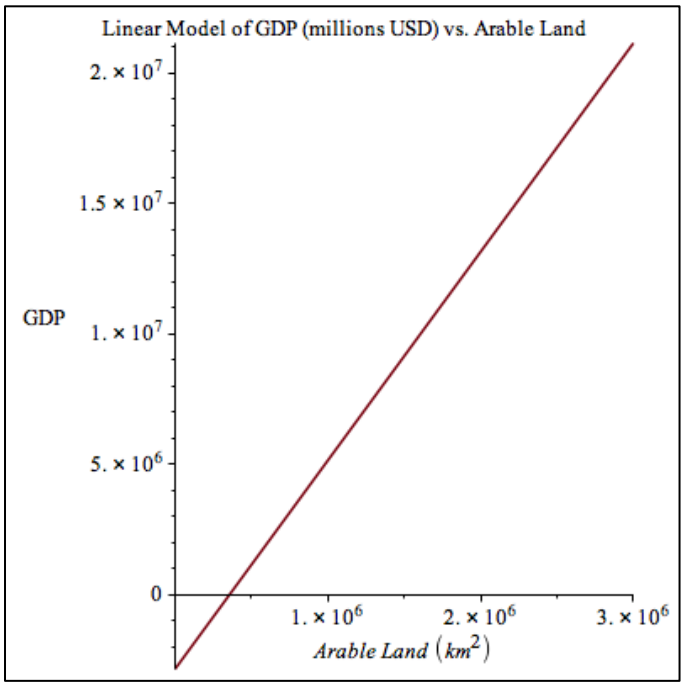
$$X^TX = \begin{bmatrix} 10 & 8504763 \\ 8504763 & 10044311039669 \end{bmatrix}$$

$$X^T\mathbf{y} = \begin{bmatrix} 39474108 \\ 56059713424057 \end{bmatrix}$$

Then, we find the least squares solution to the system.

$$\boldsymbol{\beta} = (\mathbf{X^TX})^{-1}\,\mathbf{X^T}y = \begin{bmatrix} -2855863.13 \\ 8 \end{bmatrix}$$

The regression coefficients in the least squares solution generate the following line of best fit.



Linear Regression of Corruption Perception Index vs. GDP: The GDP column vector and design matrix for this regression are

$$\mathbf{y} = \begin{bmatrix} 18561934 \\ 2250990 \\ 11391619 \\ 1267750 \\ 1769601 \\ 1532343 \\ 1256640 \\ 87198 \\ 940953 \\ 415080 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 74 \\ 1 & 40 \\ 1 & 40 \\ 1 & 29 \\ 1 & 40 \\ 1 & 82 \\ 1 & 79 \\ 1 & 29 \\ 1 & 37 \\ 1 & 28 \end{bmatrix}$$

First, we find the normal equations by finding X^TX and $X^T\mathbf{y}$.

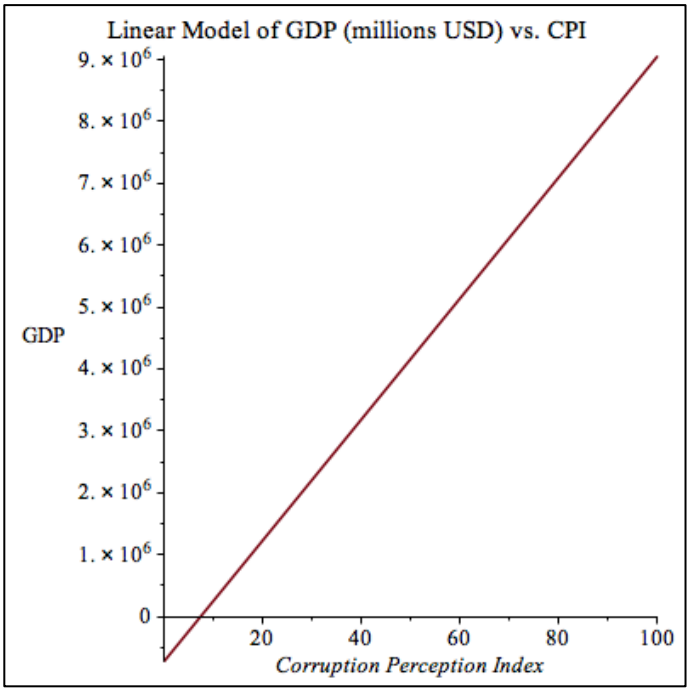
$$X^TX = \begin{bmatrix} 10 & 478 \\ 478 & 27076 \end{bmatrix}$$

$$X^T\mathbf{y} \begin{bmatrix} 39474108 \\ 2300729195 \end{bmatrix}$$

Then, we find the least squares solution to the system.

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{y} = \begin{bmatrix} -732037.25 \\ 97896.4 \end{bmatrix}$$

The regression coefficients in the least squares solution generate the following line of best fit.



The regressions above confirm some of the things we suspected from the beginning. First, the amount of land a country has for food production is crucial to its economic success, and this assertion is clearly indicated by the direct correlation above. Additionally, countries that have a higher Corruption Perception Index appear to be more economically successful. There are many reasons this may be the case, but the analysis becomes more interesting when we perform a multiple regression to look at all of the variables at once.

Multiple Regression for Total Arable Land and CPI vs. GDP: The GDP column vector and design matrix for this regression are

$$\mathbf{y} = \begin{bmatrix} 18561934 \\ 2250990 \\ 11391619 \\ 1267750 \\ 1769601 \\ 1532343 \\ 1256640 \\ 87198 \\ 940953 \\ 415080 \end{bmatrix} \quad X \begin{bmatrix} 1 & 1669302 & 74 \\ 1 & 1535063 & 40 \\ 1 & 1504350 & 40 \\ 1 & 1192300 & 29 \\ 1 & 661299 & 40 \\ 1 & 474681 & 82 \\ 1 & 474550 & 79 \\ 1 & 333847 & 29 \\ 1 & 330037 & 37 \\ 1 & 329334 & 28 \end{bmatrix}$$

First, we find the normal equations by finding X^TX and $X^T\mathbf{y}$.

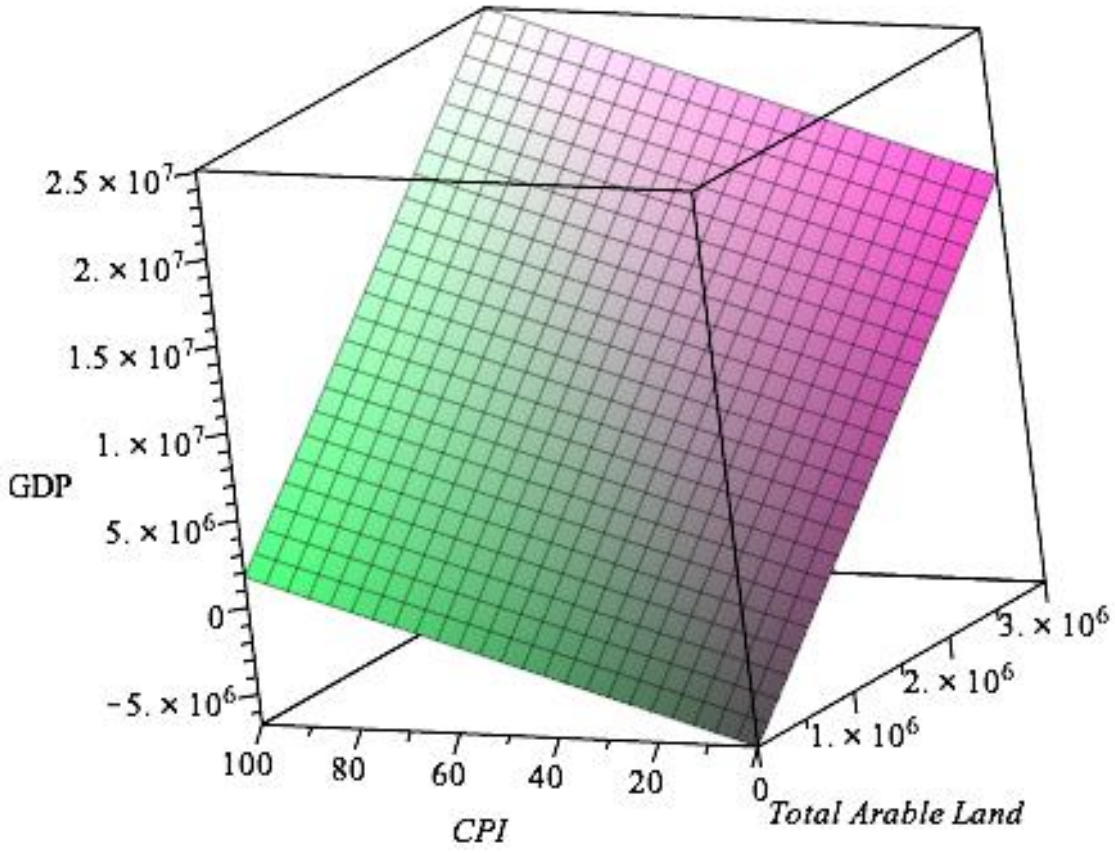
$$X^TX = \begin{bmatrix} 10 & 8504763 & 478 \\ 8504763 & 10044311039669 & 413661104 \\ 478 & 413661104 & 27076 \end{bmatrix}$$

$$X^T\mathbf{y} = \begin{bmatrix} 39474108 \\ 56059713424057 \\ 2300729195 \end{bmatrix}$$

Then, we find the least squares solution of the system.

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{y} = \begin{bmatrix} -6724545.48 \\ 7.78 \\ 84761.60 \end{bmatrix}$$

The regression coefficients in the least squares solution generate the following plane:



This regression provides a richer analysis of how these figures are related. Firstly, we see clearly that our conclusions from the previous linear regressions remain true: high CPI and arable land is a good indicator predictor of high GDP. A more interesting insight is that arable land is a stronger predictor of a country's GDP than perceived corruption. If we look at the slope of the plane relative to each independent variable, we find that both arable land and perceived corruption have positive correlations with GDP, but that arable land is the more powerful predictor of the two. It is, however, interesting to note that extremely high CPIs make up considerably for a lack of arable land.

Works Cited

"Corruption Perceptions Index 2016." www.transparency.org. Transparency International, n.d. Web. 18 April 2017.

"GDP ranking." GDP ranking | Data. World Bank, n.d. Web. 18 April 2017.

"Land Use." CIA World Factbook. Central Intelligence Agency, n.d. Web. 18 April 2017