GDP Case Study

Jacob Mortensen

Stat536 - Dr. Heaton

February 3, 2015

1 Introduction

Gross Domestic Product (GDP) is the market value of all officially recognized final goods and services produced within a country in a year (or some other period of time). It is generally considered to act as a good indicator of the economic health of a country. GDP growth has been associated with low unemployment and increasing wages, while declining GDP is associated with high unemployment and low profit. For these reasons it is of great interest to economists to understand which factors are associated with economic growth or decline.

In an effort to provide some greater understanding on this topic we are working with a dataset consisting of 67 potential geographical, socio-economic, and political predictors of growth. Unfortunately, this data exhibits many problems. One is that we only have 60 observations, which means that we are struggling with the curse of high dimension. Additionally, many of these variables are highly collinear, which is problematic because our estimates will be significantly more sensitive to fluctuations in the data and less accurate as a consequence. Throughout this paper we will discuss how each of these issues were addressed.

Our goal for this analysis is to identify and describe key relationships between GDP and the other covariates in our dataset. In order to proceed with this inference, we have elected to use a Lasso method of regression, as it will help us to account for aforementioned problems in our data.

2 Model

Our model for this analysis is

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

where $\mathbf{y}$ is growth or decline in GDP, $\mathbf{X}$ is the matrix of covariates from our data set, including geographic, political, and socio-economic factors such as what fraction of the population belongs to a specific religion or how densely populated the country is, $\boldsymbol{\beta}$ is a vector of coefficients indicating the size and direction of the effect of each covariate on GDP growth, and $\epsilon \overset{iid}{\sim} N(0, \sigma^2)$ and represents random error.

We have chosen to use Lasso regression, which is a variation of multiple linear regression that introduces a shrinkage parameter, penalizing the model in order to restrict the number of covariates that are included. Because it is a variation of multiple linear regression, it is subject to the same constraints: (1) that there is a linear relationship between the variables, (2) that the errors have constant variance, (3) that the errors are independent from one another, and (4) that the errors are normally distributed.

3 Model Justification

We decided to use Lasso for three reasons: it solves the problem of the curse of dimensionality because it has a form of built-in variable selection; it addresses the issue of multicollinearity by selecting meaningful variables; and it generates a model that is easy to interpret.

Lasso is useful in high-dimension situations because it penalizes models with too many variables and can help to identify the variables that are most influential in the model. By changing the value of a tuning parameter λ, it is possible to adjust how many variables are included in the model in order to minimize Mean Square Error. In this situation, where we have more variables than observations and normal least squares regression can produce an infinite number of nonsensical models that fit the data perfectly, it is convenient to impose a constraint that creates a model that is both simple and unique. It is important to note that its criteria is based primarily on size, and so it is necessary to standardize the variables before using this method.

For much of the same reasons, Lasso is good in instances where data exhibits a high degree of collinearity. Typically collinearity will result in inflated estimates for our coefficients, but the constraint imposed by Lasso will prevent that.

Last of all, we are interested in Lasso because it generates a model that is easy to interpret. Particularly in this case, where we are interested in inference, it is important that our coefficients are easy to understand within the context of the problem. This method does not combine variables, like principle components analysis, which makes them easier to interpret, and it also pares many variables out of the model, unlike ridge regression, making the model easier to explain.

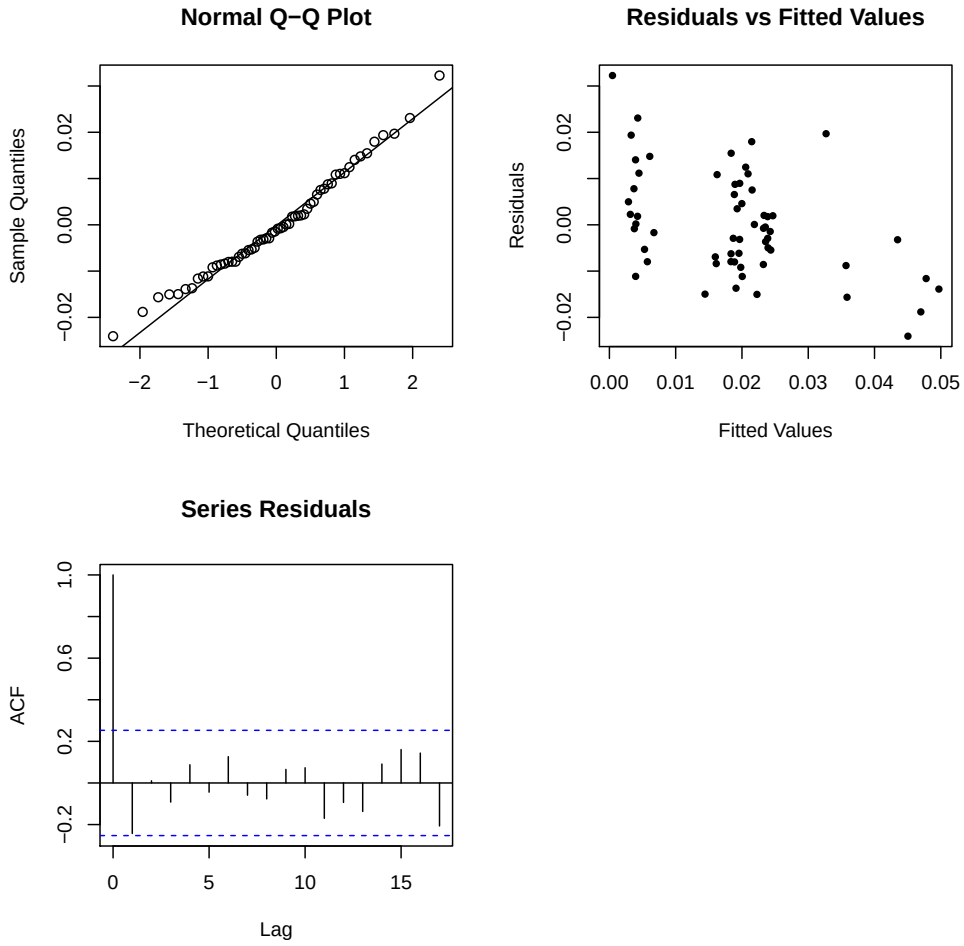

Figure 1: Plots to Examine Assumptions of Linear Regression

Before proceeding with our analysis we first had to scale and center our predictor variables. Then, in order to select which covariates to include in our model we used a 3-fold cross validation technique. The small size of our sample prevented us from splitting our data into more test groups. Our criterion for selecting a tuning parameter was MSE. We conducted Lasso regression for a variety of tuning parameters and selected the one that minimized the MSE in our test data set. The MSE for this particular model was 0.000438 and consisted of only 8 variables, a vast improvement over the 67 that we started with.

Looking at Figure 1, we can see how well our linear regression assumptions were met. If we examine the Normal Q-Q Plot we can see that the majority of the points fall along the line, indicating that the residuals are approximately normal in their distribution. If we examine the Residuals vs Fitted Values we can see some strange clustering in the middle, as well as a drop off in the lower right corner that may suggest that the data is not exactly linear, but with such sparse data, it is difficult to say anything conclusive. Examining

Variable	Description	Estimate	95% CI Lower	95% CI Upper
(Intercept)		0.0223	0.0192	0.0255
CONFUC	Fraction Confucian	0.1237	0.0911	0.1562
EAST	East Asian Country	0.0571	0.0456	0.0686
IPRICE1	Investment Price	-0.0000002	-0.0000002	-0.0000002
MALFAL66	Malaria Presence in 1960	-0.0294	-0.0369	-0.0219
PRIEXP70	Primary Exports in 1970	-0.0082	-0.0141	-0.0024
RERD	Real Exchange Rate Distortions	-0.0000001	-0.0000001	0.0000001
SAFRICA	South African Country	-0.0027	-0.0093	0.0038
YRSOPEN	Years Open (1950 - 1994)	0.0128	-0.0133	0.0388

Table 1: Regression Coefficients

that same plot, we see that the residuals maintain approximately the same thickness throughout, although a few outliers on either end make it difficult to say for sure, suggesting that our residuals do have constant variance. Last of all, if we look at the autocorrelation plot in the bottom left of Figure 1 we can see that there does not appear to be dependence between the residuals as none of the lines indicating correlation with the previous residual extend beyond -0.2 or 0.2. Overall, our assumptions regarding linear regression appear to be met, and I am comfortable proceeding with this analysis.

After generating our model, we backtransformed the coefficients to their original scale, in order to provide a more reasonable interpretation of each coefficient in the context of the problem. Additionally, we calculated standard errors and confidence intervals using a bootstrap technique where we generated a thousand estimates for each coefficient and then found the standard error of them in order to estimate the standard error of the coefficients.

4 Results

Our final model consisted of 8 variables, listed in Table 1, along with their estimates and 95% confidence intervals. These variables help us complete the goal of our analysis, which is to identify which variables are strongly associated with a growth or decline in GDP. These 8 variables are the ones that we found to have the strongest relationship with GDP growth by using the Lasso methodology. By examining the 95% confidence intervals we can see that the last three contain 0, indicating that they may not actually be

significant. However, we stand by our selection technique and will keep them in the model. We can also notice by looking at the table that each of the coefficients are very small. This is appropriate considering the response variable, where the difference between the maximum and minimum value was only 0.1009.

Because we have no interactions in our model the interpretations of each coefficient is fairly straightforward:

CONFUC For every one unit increase in the fraction Confucian, GDP grew by 0.1237%.

EAST If a country was an East Asian Country, its GDP grew by 0.0571%.

IPRICE1 If a country's investment price went up by one unit then their GDP declined by 0.0000002%.

MALFAL66 For every one unit increase in malaria presence in 1960, GDP declined by 0.029%.

PRIEXP70 For every unit increase in primary exports, GDP declined by 0.0082%.

RERD This variable does not appear to be significant since its confidence interval spans zero, but we would interpret its effect as being a 0.0000001% decline in GDP for every one unit increase in real exchange rate distortion.

SAFRICA This variable also does not appear to be significant, but we would interpret its effect as a 0.0027% decline in GDP if a country was in South Africa.

YRSOPEN This also does not appear to be significant but we would interpret its effect as a 0.0128% growth in GDP for every additional year open.

This analysis provides us with some good insight into which variables were associated with GDP growth from 1960 to 1996. We can see that both the fraction of a country that was Confucian and being an East Asian country were associated with GDP growth, which is logical as China and other East Asian countries have become manufacturing powerhouses and experienced great growth. We can also see that countries that were more expensive to invest in typically experienced a decline in GDP. We can also see that the presence of malaria was associated with a decline in GDP. These variables point to the many ways in which geographic factors can play a role, as the East Asian countries tended to exhibit the strongest growth, and areas with high malaria rates, most likely African countries, experienced GDP decline.

5 Conclusion

From this analysis we were able to glean some information about how specific geographic and socio-economic variables were associated with GDP growth and decline. We saw that in the time period from 1960 to 1996, East Asian countries and countries with a high fraction of the population that practiced Confucianism experienced higher rates of GDP growth than the rest of the world, all else constant. We also saw that countries that were expensive to invest in experienced a decline in GDP, as well as countries that had higher rates of malaria. These patterns will be valuable in helping economists recognize what factors make an economy move forward or backward.

We used Lasso regression and it performed well, providing us with a sparse model that had low MSE and allowed us to make some valuable inference about which variables were associated with GDP growth. However, because economic theory suggests that an economy is pushed forward by many small effects rather than a few large ones, a different method such as principle components, which accounts for the effects of all variables, or ridge regression, which does not decrease variable coefficients to zero, may have had more power to explain GDP growth.

The next steps for this analysis would be to use one of these other methods and compare its performance to Lasso. In the face of high collinearity, principle components could be especially effective, because it could account for a large amount of the available information without overfitting the model.