Inquiry to Insight: Google Transparency Reports

Zakariah Teffahi

I. Introduction

I.A. Background

Since 2010, Google has released transparency reports to "shed light on how the policies and actions of governments and corporations affect privacy, security, and access to information online".[1] More specifically, the data released gives critical statistics on various governments' inquiries to view user data and remove Google services' content. A subset of the information released by Google pertains specifically to criminal investigations. In this area, police can request information from Google to aid in suspect identification and sentencing. According to the Washington Post and Bloomberg, while access to user information expedites the crime-solving process, the practice threatens the public's privacy and nears being unconstitutional.[2,3] Amid this controversial practice lies potentially valuable statistical insights. This report delves into Google's transparency data from the first half of 2011, focusing on requests related to criminal investigations.

I.B. Problems Addressed & Research Goals

Google's transparency report provides data on government requests for criminal investigations, which can be analyzed across six variables. By examining these variables alongside request figures, I can uncover insights into regional trends, government types, and motivations behind the requests. Before delving into analysis, it's essential to consider potential regression models and their implications.

Firstly, I can explore the relationship between request numbers and population size. This helps determine if larger populations correlate with more requests and if the relationship is linear or complex. Additionally, comparing request volumes with subjective variables like the Human Development Index (HDI) reveals patterns such as less developed countries making fewer requests due to legal system limitations or making more requests due to higher crime rates. Similarly, examining the Free Press Index (FPI) may indicate whether restrictive governments make more requests for control or fewer requests due to alternative means of censorship. The relationship between request count and internet access rates may suggest increased requests in response to internet-related crimes or due to limited investigation resources. Lastly, analyzing the correlation between request count and a country's Democracy Index (DI) can provide insights into the enforcement of just laws in democratic nations or stricter internet regulation in less transparent regimes.

Overall, the transparency report's variables offer avenues to model and understand government request volumes, incorporating factors such as population size, HDI, FPI, internet access, and DI to gain insights into real-world implications.

II. Data Description

Seven data variables are present in this data set—two qualitative and five quantitative. Our qualitative variables are the country filing requests and the democracy index ranking. The country variable is nominal, while the democracy index is ordinal, ranking democracy from lowest to highest in the following order: hybrid, flawed, or full. Our quantitative variables are the count of requests, population size, HDI, internet access rate, and the Free Press Index (FPI). The request count is a discrete count variable that quantifies the number of times a country has requested data from Google. Request count is the dependent variable I hope to analyze, thus exploring if and how it is influenced by the other six variables. Population is a discrete count variable describing the number of people in a nation. The HDI is a continuous variable that considers a country's life expectancy, education, and income and assigns a country a value between 0-1, with higher values indicating a country with long life expectancy, great education, and wealth. The percentage of internet users is a continuous variable that describes what percentage of the country's population has access to the internet. FPI is a continuous variable representing a measure of press freedom scaled 1-100, with higher values indicating greater press freedom.

III. Preliminary Data Analysis

III.A. Scatter Plots Between Request Count and Each Variable

As seen in all scatterplots in section III.A., the number of requests for the United States consistently acts as a potential outlier. The number of requests for the United States takes on 300% of the value of the second-highest request count.

III.A.1. Request Count vs. Name of Country

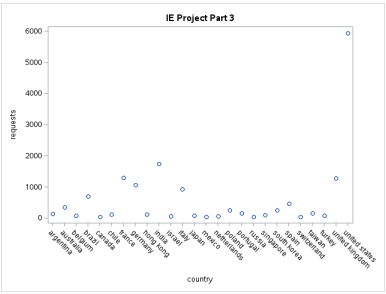


Figure 01. Requests vs. Country

A majority of countries made [0, 1000] requests. Only four countries had request counts exceeding this range, with one of them being the aforementioned potential outlier—The United States.

III.A.2. Request Count vs. Population

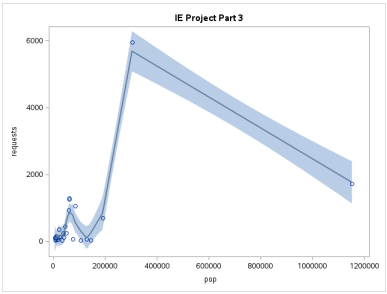


Figure 02. LOESS: Requests vs. Population Size

There does not seem to be a linear relationship between the number of requests and the population size. The points and LOESS exhibit an oscillating trend, especially within the domain of population sizes [0, 200,000]. Thereafter, the amplitude of the LOESS oscillation seems to increase, potentially due to outliers. Because both nonlinearity and the variance of ϵ are concerns, transformations on both request count and population size may be necessary. The positive correlation coefficient is 0.40693, suggesting a moderate positive correlation between the number of data requests and the population size, but since the relationship is nonlinear, this coefficient may be invalid.

III.A.3. Request Count vs. HDI

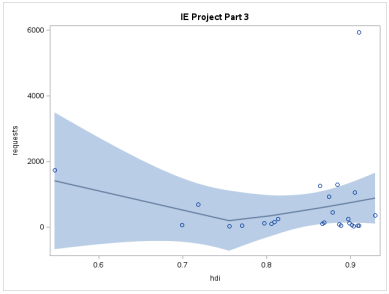


Figure 03. LOESS: Requests vs. HDI

There does not seem to be a linear relationship between the number of requests and the human development index (HDI) value. The plotted points and LOESS exhibit a convex parabolic shape, and the variance of request counts seemingly increases as the human development index values increase. Because both nonlinearity and the variance of ϵ are concerns, transformations on both request count and HDI may be necessary. With a correlation coefficient of 0.01727, there is a very weak positive correlation between the number of data requests and the HDI, but since the relationship is nonlinear, this coefficient may be invalid.

III.A.4. Request Count vs. Democracy Index

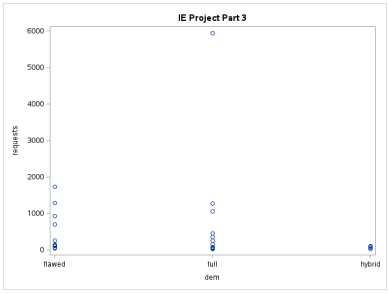
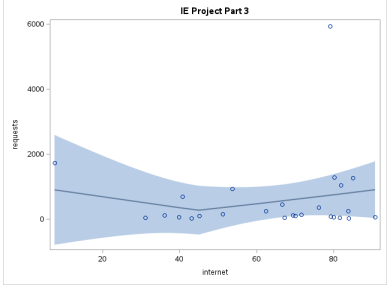


Figure 04. Requests vs. Dem. Index

There seems to be a relationship between a nation's democracy index value and its request count. The democracy index is ordinal, with the order of highest to lowest democracy rates being as follows: full, flawed, and hybrid. As the democracy index increases from hybrid to flawed, the variance of request counts increases drastically, and when the democracy index increases from flawed to full, the variance decreases slightly with the exception of an outlier. This increase and then decrease in variance may be systematic, indicating a relationship between the variance of request count and democracy index values.

III.A.5. Request Count vs. Internet Access Percentage



There does not seem to be a linear relationship between the number of requests and the internet access rate. The plotted points and LOESS exhibit a convex parabolic shape, and the variance of request counts seemingly increases as the internet access rate increases. Because both nonlinearity and the variance of ϵ are concerns, transformations on both request count and internet access may be necessary. With a correlation coefficient of 0.06830, there is a very weak positive correlation between the number of data requests and the internet access rate, but since the relationship is nonlinear, this coefficient may be invalid.

Figure 05. LOESS: Requests vs. Internet Access	

III.A.6. Request Count vs. Free Press Index

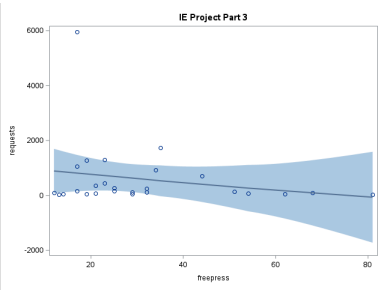


Figure 06. LOESS: Requests vs. Free Press Index

There does seem to be a weak linear relationship between the number of requests and the free press index value of a nation. The LOESS line for the plot has a near-unchanging slope and intercept, indicating near-linearly. The variance of request count seems to have a relationship with free press, as points deviate less from the LOESS line for higher free press values. Because whether the variance of ϵ is constant is of concern, but non-linearity is not a concern, transformations on free press may be necessary. Because linearity *may* be a concern, providing transformations on free press and request count may improve results further. With a correlation coefficient of -0.21601, there is a weak negative correlation between the number of data requests and the free press index.

III.B. Linear Regressions Between Request Count and Each Variable

III.B.1. Request Count vs. Population

The untransformed data had a Q-Q plot that deviates from the linear line, indicating non normal error terms. The studentized residual plot shows a value over |3|, suggesting a significant outlier. The residual vs $\hat{Y}$ and residual vs X plots exhibited non random scattering in a fanning out shape, indicating non-linearity and non-random residuals. I initially applied a square root transformation to the population variable, which did not resolve non-normality or outliers. Subsequent reciprocal and exponential transformations also failed to address these issues. Applying a log transformation removed the model's outliers, and had a linear Q-Q plot indicating normally distributed error terms. The residual vs $\hat{Y}$ and residual vs X plots, still observe a scattering of points below $e = 0$ that fans outward, thus implying that the residuals are not randomly distributed and that the relationship might not be linear, and the residual plot still contained some systematic patterning, however, I attributed that to the rather small sample size of the data. The log transformation kept some concerns of nonlinearity and changing variance due to non-randomness in residual plots, but the results were approved compared to before the transformation.

III.B.2. Request Count vs. HDI

Similarly, the untransformed data has a Q-Q plot that deviates from the linear line, indicating non-normal error terms. The Residual vs. Y and residual vs. X plots show a high concentration of points below the $e = 0$ line, suggesting non-random residuals, and the studentized residual plot contains a value above 10, indicating a significant outlier. Applying a log transformation to Y and X fixed our 5 plots, indicating that all 6 assumptions were met and allowing for SLR analysis.

III.B.3. Request Count vs. Internet Access Percentage

The untransformed data had a residual vs. $\hat{Y}$ plot showing a concentration of points below the $e = 0$ line, indicating non-random residuals and potential non-linearity. Additionally, the studentized residual plot contained a value well over |3|, indicating a significant outlier and the Q-Q plot deviated from the linear line, suggesting non-normal error terms. Applying a log transformation to Y as well as all of our predictors fixed our 5 plots, indicating that all 6 assumptions were met and allowing for SLR analysis.

III.B.4. Request Count vs. Free Press Index

The untransformed data had a Q-Q plot that deviated from the provided linear line, implying the error terms are not distributed normally. The studentized residual plot showed a value over |3|, suggesting an outlier. The residual vs. Y and residual vs. X plots show a scattering of points below $e = 0$, indicating non-random residuals and potential non-linearity. Applying a log transformation to Y as well as all of our predictors fixed our 5 plots, indicating that all 6 assumptions were met and allowing for SLR analysis.

III.C. Statistical Inference

III.C.1 Log Request Count vs. Log Population

The log Request Count vs log Population regression yielded the equation $\widehat{\log Y} = -0.68698 + 0.57231 \cdot \log X$, but its graphs did not meet all assumptions. As stated in section III.B.1, the residual vs. $\hat{Y}$ and residual vs. X plots fanned outwards implying that the residual was not randomly distributed and thus the relation is non-linear. I tried to apply log, exp, square root, and reciprocal

transformations to X and/or Y for this plot but no transformations resulted in the satisfaction of all 6 assumptions—and by extension, no transformation of the data met the prerequisites for standard linear regression. For this reason, the regression provided is of no value.

III.C.2 Log Request Count vs. Log HDI

The log Request Count vs log HDI regression yielded the equation $\widehat{\log Y} = 5.13829 - 1.32153 \cdot \log X$and met all assumptions following preliminary graphical analysis. This indicates that for every 1% increase in a country's HDI value, there is a 1.322% decrease in the number of requests made by that country on average. Similarly, this indicates that when the log HDI is zero, the log number of requests equals 5.138 on average. The 95% confidence interval for log(requests) has a large width when log(HDI) is less than -0.4, likely because of the sparse amount of data in this region. This width narrows as I approach the larger concentration of data points, such as between -0.4 and -0.1. When conducting a hypothesis test for linearity (H_0: $B_1 = 0$, and H_A: $B_1 \neq 0$) with a type I error of 0.05, the P-value of 0.586 indicates that I fail to reject the null hypothesis. Thus, the hypothesis test supports the conclusion that no linear relationship exists between log Regression Count and log HDI.

III.C.3 Log Request Count vs. Log Internet Access Percentage

The log Request Count vs log Internet regression successfully met all assumptions enumerated in section III.C.3. The resulting regression function is $\widehat{\log Y} = 6.70133 - 0.32661 \cdot \log X$, indicating that a 1% increase in log internet causes a 0.3266% decrease in log request count on average. The plot 95% confidence interval can be characterized by a wide span of about 4 log(requests) units that gradually contract until log(internet) = 4, then the interval expands once more. Many values do not fall within the 95% confidence interval, possibly due to the dataset's small sample size. When conducting a hypothesis test for linearity (H_0: $B_1 = 0$, and H_A: $B_1 \neq 0$) with a type I error of 0.05, the P-value of 0.5466 indicates that I fail to reject the null hypothesis. Thus, the hypothesis test supports the conclusion that no linear relationship exists between log Regression Count and log Internet.

III.C.4 Log Request Count vs. Log Free Press Index

The log Request Count vs Log Free Press regression successfully met all assumptions enumerated in part III.C.4. The resulting regression function is $\widehat{\log Y} = 6.93834 - 0.47252 \cdot \log X$, indicating that a 1% increase in log(freepress) causes a 0.4725% decrease in log request count on average. The 95% confidence interval can be characterized by a wide span that gradually contracts, only to expand again with the smallest interval being around when log(freepress) is 3.4. Many data points do not fall within the 95% confidence interval which I hypothesize is due to the dataset having a small sample size. When conducting a hypothesis test for linearity (H_0: $B_1 = 0$, and H_A: $B_1 \neq 0$) with a type I error of 0.05, the P-value of 0.38 indicates that I fail to reject the null hypothesis. Thus, the hypothesis test supports the conclusion that no linear relationship exists between log Regression Count and log Free Press.

IV. Addressing Research Questions Using MLR Model Analysis

IV.A. Additive MLR Using All Quantitative Predictors

IV.A.1. Checking Regression Assumptions 1-5 & 7

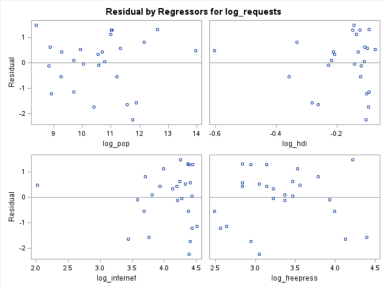


Figure 07. Residual vs Additive MLR Predictors

Using all quantitative predictors—log_population ($\log X_1$), log_HDI ($\log X_2$), log_Internet Access ($\log X_3$), and log_Free Press Index ($\log X_4$)—in an additive MLR with log_requests (logY) takes the form $Y = \beta_0 + \beta_1 \log X_1 + \beta_2 \log X_2 + \beta_3 \log X_3 + \beta_4 \log X_4 + \epsilon$. The resulting regression is $\widehat{\log Y} = 2.654 + 0.732 \cdot \log X_1 + 4.055 \cdot \log X_2 - 0.504 \cdot \log X_3 - 0.682 \cdot \log X_4$. The residual plots for $\log X_1$ and $\log X_4$ show random scattering thus I believe the relationship between these predictors and $\log Y$ is linear. The residual plots of $\log X_2$ and $\log X_3$ show random vertical scattering as well, but the values are concentrated to the right of $\log X_2 =$ -0.2 for the residual plot of $\log X_2$ and $\log X_3 = 4$ for the residual plot of $\log X_3$. This inconsistent horizontal distribution can be attributed to a low sample size and a left-ward outlier in the residual plots. Thus, I believe the relationship between these predictors and $\log Y$ is linear. The residual plot for $\widehat{\log Y}$ exhibit random scattering, so I conclude that the error terms have constant variance. Since none of the predictors are sequential such as 'Time', I need not test whether the error terms are

	independent. The Q-Q plot of the additive MLR is randomly and closely distributed around the linear line, hence I can conclude that the error terms are approximately normal in distribution. The studentized residual plot contains no values more extreme than |3|, thus I can conclude there are no outliers. The predictors $\log X_1$, $\log X_2$, $\log X_3$, and $\log X_4$ have VIF values of 1.531, 9.774, 9.244, and 1.384 respectively, and since all VIF values are below 10, the VIF values suggest that predictors are linearly independent.

The only assumption for a linear regression that still requires checking is that no important predictors are omitted. I cannot diagnose this assumption by including new independent variables from the base dataset because the MLR thus far has included all quantitative variables from the dataset. However, it is possible to diagnose the assumption by including interaction terms and polynomial terms.

IV.A.2. Check for Missing Predictors Using Interaction Terms

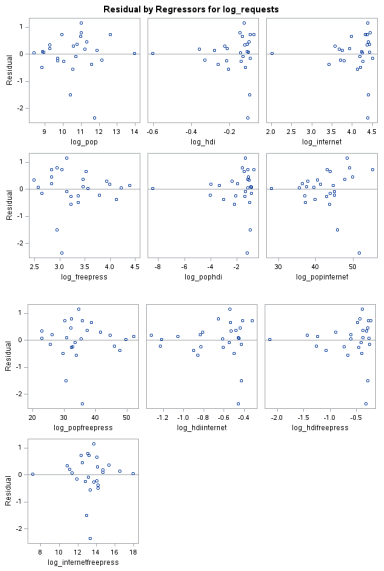


Figure 08. Residual vs Interaction MLR Predictors

I must generate a pairwise interaction-term model to determine whether important interaction-term predictors have been omitted. The model takes the form as follows:

$$\hat{\log Y} = \beta_0 + \beta_1(\log X_1) + \beta_2(\log X_2) + \beta_3(\log X_3) + \beta_4(\log X_4) + \beta_5(\log X_1)(\log X_2) + \beta_6(\log X_1)(\log X_3) + \beta_7(\log X_1)(\log X_4) + \beta_8(\log X_2)(\log X_3) + \beta_9(\log X_2)(\log X_4) + \beta_{10}(\log X_3)(\log X_4) + \epsilon.$$

To analyze the regression model, I must first confirm that it satisfies all important assumptions. The residual plots for $\log X_2$, $\log X_3$, $\log X_1 \log X_2$, $\log X_1 \log X_3$, $\log X_2 \log X_3$, and $\log X_2 \log X_4$ exhibit a slight rightward-opening funnel shape, with the variance of the error increasing as the predictors increase in value. The residual plots for $\log X_1$, $\log X_1 \log X_4$, and $\log X_3 \log X_4$ seem to increase and then decrease in variance, forming a near rhombus shape. This lack of randomness in the residual plots may suggest that the relationship between these predictors and $\log Y$ is not linear. It is unclear how much of this deviation from randomness is due to the small sample size. The residual plot for $\hat{\log Y}$ exhibits random scattering, so I conclude that the error terms have constant variance. Since none of the predictors are sequential such as 'Time', I need not test whether the error terms are independent. The Q-Q plot of the regression is randomly and closely distributed around the linear line, hence I can conclude that the error terms are approximately normal in distribution. The studentized residual plot contains one value more extreme than |3|, thus I can conclude there is an outlier. The VIF values for the predictors vary from 326.297 to 5356.773, and thus are significantly higher than 10, indicating that multicollinearity is a serious concern.

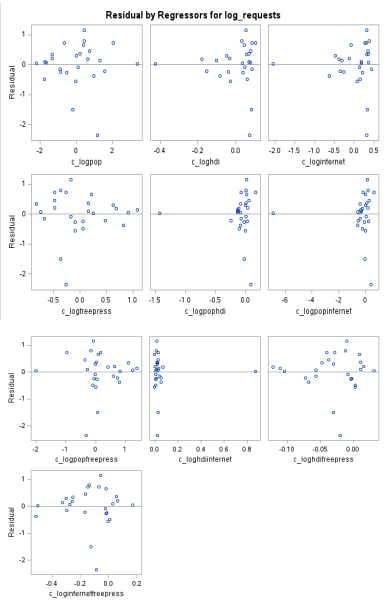


Figure 09. Residual vs Centered Interaction MLR Predictors

To reduce the significance of multicollinearity in the model, I must mean-center all predictors. That is, I must substitute each $\log X_i$ with $\log X_i^\star = \log X_i - \overline{\log X}$ and each $\log X_i \log X_j$ with $\log X_i^\star \log X_j^\star$. The resulting regression model is

$$\widehat{\log Y} = 6.122 + 1.141(\log X_1^\star) + 0.953(\log X_2^\star) - 1.589(\log X_3^\star) - 1.969(\log X_4^\star) - 7.183(\log X_1^\star)(\log X_2^\star) + 0.054(\log X_1^\star)(\log X_3^\star) - 1.074(\log X_1^\star)(\log X_4^\star) - 14.303(\log X_2^\star)(\log X_3^\star) + 6.34(\log X_2^\star)(\log X_4^\star) + 1.597(\log X_3^\star)(\log X_4^\star).$$

To analyze the regression model, I must first confirm that it satisfies all important assumptions. The residual plots for $\log X_2^\star$ and $\log X_3^\star$ exhibit a slight rightward-opening funnel shape, with the variance of the error increasing as the predictors increase in value. The residual plots for $\log X_1^\star$, $\log X_4^\star$, $\log X_1^\star \log X_4^\star$, $\log X_2^\star \log X_4^\star$, and $\log X_3^\star \log X_4^\star$ seem to increase and then decrease in variance, forming a near rhombus shape. The residual plots for $\log X_1^\star \log X_2^\star$, $\log X_1^\star \log X_3^\star$, and $\log X_2^\star \log X_3^\star$ are highly concentrated around one predictor value, and thus are difficult to interpret, yet they seem randomly distributed. The lack of randomness in some of the residual plots may suggest that the relationship between these predictors and $\log Y$ is not linear. It is unclear how much of this deviation from randomness is due to the small sample size. The residual plot for $\widehat{\log Y}$ exhibits mostly random scattering with a slight rhombus shape, so if I attribute the deviation from randomness to the small sample size, I conclude that the error terms have constant variance. Since none of the predictors are sequential such as 'Time', I need not test whether the error terms are independent. The Q-Q plot of the regression is randomly and closely distributed around the linear line, hence I can conclude that the error terms are approximately normal in distribution. The studentized residual plot contains one value more extreme than |3|, thus I can conclude there is an outlier. The VIF values for the predictors, ranging from 3.892 to 249.777, have drastically decreased, and thus multicollinearity has decreased. However, since the VIF for $\log X_2^\star$, $\log X_3^\star$, $\log X_1^\star \log X_2^\star$, $\log X_1^\star \log X_3^\star$, and $\log X_2^\star \log X_3^\star$ are larger than 10, multicollinearity is still a concern.

Now that I know at least one of the pairwise interaction terms is statistically significant, I can conduct further hypothesis tests to investigate which should be kept in the model. Thus, I will proceed with partial *F*-tests using a Type-I error $\alpha = 0.05$.

First, I sort our predictors by their individual SSR values, least to greatest. I then conduct a partial F-test to determine whether $\beta_k = 0$, where k corresponds to the predictor with the smallest SSR value. If I reject the null hypothesis, then I keep β_k's predictor in the model and conduct a new partial F-test using the interaction-term predictor with the next smallest SSR. Otherwise, I will construct a new model without β_k's predictor, and conduct a new partial F-test using the interaction-term predictor with the smallest SSR in the new model. I repeat this process until I have a final model.

To determine if $\log X_3^\star \log X_4^\star$ is statistically significant, I run a partial *F*-test to determine whether $\beta_{10} = 0$. This yields an observed $F_{obs}^\star$ value of 0.21683, and a F_{crit} value of 4.54308, and thus I fail to reject. I remove $\log X_3^\star \log X_4^\star$ from the model and continue.

To determine if $\log X_2^\star \log X_4^\star$ is statistically significant, I run a partial *F*-test to determine whether $\beta_9 = 0$. This yields an observed $F_{obs}^\star$ value of 2.28735, and a F_{crit} value of 4.49400, and thus I fail to reject. I remove $\log X_2^\star \log X_4^\star$ from the model and continue.

To determine if $\log X_1^\star \log X_2^\star$ is statistically significant, I run a partial *F*-test to determine whether $\beta_5 = 0$. This yields an observed $F_{obs}^\star$ value of 1.766726, and a F_{crit} value of 4.45132, and thus I fail to reject. I remove $\log X_1^\star \log X_2^\star$ from the model and continue.

To determine if $\log X_1^\star \log X_3^\star$ is statistically significant, I run a partial F-test to determine whether $\beta_6 = 0$. This yields an observed $F_{obs}^\star$ value of 1.306336, and a F_{crit} value of 4.41387, and thus I fail to reject. I remove $\log X_1^\star \log X_3^\star$ from the model and continue.

To determine if $\log X_2^\star \log X_3^\star$ is statistically significant, I run a partial F-test to determine whether $\beta_8 = 0$. This yields an observed $F_{obs}^\star$ value of 0.15885, and a F_{crit} value of 4.38075, and thus I fail to reject. I remove $\log X_2^\star \log X_3^\star$ from the model and continue.

To determine if $\log X_1^\star \log X_4^\star$ is statistically significant, I run a partial F-test to determine whether $\beta_7 = 0$. This yields an observed $F_{obs}^\star$ value of 10.714096, and a F_{crit} value of 4.35124, and thus I reject it. I keep $\log X_1^\star \log X_4^\star$ in the model and continue.

Thus, our model generated from looking at all of the possible interaction terms is: $log(Y) = -32.22008 + 4.04226 \cdot \log X_1^\star - 0.60966 \cdot \log X_2^\star - 0.30441 \cdot \log X_3^\star + 9.57339 \cdot \log X_4^\star - 1.01624 \cdot \log X_1^\star \log X_4^\star$.

Since the plot of predicted Y vs residuals has no systematic pattern around 0 line, I conclude that the regression function is linear and that the error terms have constant variance. There are no sequential variables such as the residuals vs time, so assumption 3 is satisfied. The Q-Q plot of residuals follows a linear pattern and shows no other systematic pattern, so the error terms are normally distributed. There is only 1 outlier based on the studentized residual plot. I have narrowed down the model to only the necessary interaction term predictors, so no important interaction term predictors are missing from the model, but I may be missing polynomial predictors. The VIF values for each predictor are less than 10, so our predictors are linearly independent, and multicollinearity is not an issue for our model. Therefore, our model satisfies the MLR assumptions thus far, so long as a polynomial term does not improve the results.

IV.A.3. Check for Missing Predictors Using Polynomial Terms

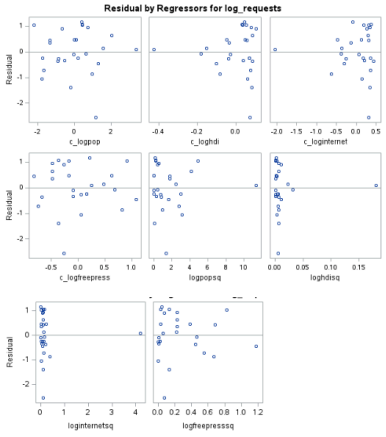


Figure 10. Residual vs Polynomial MLR Predictors

I must generate a polynomial model to determine whether important polynomial predictors have been omitted. For polynomial models, all predictors must be mean-centered. That is, I must substitute each $\log X_i$ with $\log X_i^\star = \log X_i - \overline{\log X}$. Thus, the model takes the form as follows: $\widehat{\log Y} = \beta_0 + \beta_1(\log X_1^\star) + \beta_2(\log X_2^\star) + \beta_3(\log X_3^\star) + \beta_4(\log X_4^\star) + \beta_5(\log X_1^\star \log X_4^\star) + \beta_6(\log X_1^\star)^2 + \beta_7(\log X_2^\star)^2 + \beta_8(\log X_3^\star)^2 + \beta_9(\log X_4^\star)^2$.

To analyze the regression model, I must first confirm that it satisfies all important assumptions. The residual plots for $\log X_2^\star$ and $\log X_3^\star$ exhibit a slight rightward-opening funnel shape, with the variance of the error increasing as the predictors increase in value. The residual plots for $\log X_1^\star$, $\log X_4^\star$, $(\log X_1^\star)^2$ and $(\log X_4^\star)^2$ exhibit a random scattering. The residual plots for $(\log X_2^\star)^2$ and $(\log X_3^\star)^2$ are highly concentrated around one predictor value, and thus it is difficult to interpret the scattering, but it seems mostly random. The lack of randomness in some of the residual plots may suggest that the relationship between these predictors and $\log Y$ is not linear. It is unclear how much of this deviation from randomness is due to the small sample size. The residual plot for $\widehat{\log Y}$ exhibits a random scattering, thus I conclude that the error terms have constant variance. Since none of the predictors are sequential such as 'Time', I need not test whether the error terms are independent. The Q-Q plot of the regression is randomly and closely distributed around the linear line, hence I can conclude that the error terms are approximately normal in distribution. The studentized residual plot contains one value more extreme than |3|, thus I can conclude there is an outlier. The VIF values for the predictors range from 1.81 to 47.486, thus exhibiting a medium concern of multicollinearity. Because predictors $\log X_2^\star$, $\log X_3^\star$, $(\log X_2^\star)^2$, and $(\log X_4^\star)^2$ have VIF values over 10, they are particularly concerning for multicollinearity.

To determine if any of the newly included polynomial predictors are statistically significant, I can conduct a hypothesis test

with the null hypothesis $\forall i \in [6, 9]$: $\beta_i = 0$ and Type I error $\alpha = 0.05$. This test will be conducted using the partial SS and associated F-test of $SSR((\log X_1^\star)^2, (\log X_2^\star)^2, (\log X_3^\star)^2, (\log X_4^\star)^2 \mid \log X_1^\star \log X_4^\star, \log X_1^\star, \log X_2^\star, \log X_3^\star, \log X_4^\star)$. This test results in an observed $F_{obs}^\star$ value of 0.3787, and an associated *P*-value of 0.20047. Because the *P*-value is above $\alpha = 0.05$, I fail to reject the null hypothesis and thus conclude that $\forall i \in [5, 8]$: $\beta_i = 0$. Therefore, the data supports the claim that the coefficient for each polynomial term is 0, and thus that the polynomial terms can be removed from the model.

Due to this conclusion, when incorporating polynomial predictors into the model, not all linear regression assumptions are met, none of the interaction terms are statistically significant, and all but two polynomial terms had VIF values over 10, so I chose to drop the polynomial terms from the model.

IV.A.4. Determining Significance of Remaining Terms

An overall *F*-test can be undertaken to test the null hypothesis $\forall i \in [1, 5], : \beta_i = 0$, and test whether there is no regression relationship between $\log Y$ and the four quantitative predictors. Using an alternate hypothesis of $\exists i \in [1, 5]$: $\beta_i \neq 0$ and a Type I error fixed at $\alpha = 0.05$, I can execute the test. The overall F-test results in a p-value of 0.0009, which is below α and thus I reject the null hypothesis, suggesting that at least one coefficient is non-zero, and thus further analysis is necessary. The resulting linear regression acquired is $\hat{\log Y} = 5.52487 + 0.67814 \cdot \log X_1^\star - 0.60966 \cdot \log X_2^\star - 0.30441 \cdot \log X_3^\star - 1.18917 \cdot \log X_4^\star - 1.01624 \cdot \log X_1^\star \log X_4^\star$. Note that, because the predictor $\log X_1^\star \log X_4^\star$ is statistically significant and of degree 2, I cannot remove predictors $\log X_1$ or $\log X_4$ due to the hierarchy principle.

Knowing at least one of the pairwise interaction terms has a non-zero coefficient, I can conduct further hypothesis tests to investigate which should be kept in the model. Thus, I will proceed with partial *F*-tests using a Type-I error $\alpha = 0.05$. Due to previous interaction term testing and the hierarchy principle, I need only test $\log X_2^\star$ and $\log X_3^\star$.

To determine if $\log X_2^\star$ is statistically significant, I run a partial *F*-test to determine if $\beta_2 = 0$. This yields an observed $F_{obs}^\star$ value of 0.266849 as opposed to the critical $F_{crit} = 4.35124$. Since $F_{obs}^\star < F_{crit}$, I reject the null hypothesis, and thus include $\log X_2^\star$ in the model.

To determine if $\log X_3^\star$ is statistically significant, I run a partial *F*-test to determine if $\beta_3 = 0$. This yields an observed $F_{obs}^\star$ value of 1.51357 as opposed to the critical $F_{crit} = 4.35124$. Since $F_{obs}^\star < F_{crit}$, I reject the null hypothesis, and thus include $\log X_3^\star$ in the model.

Thus, $\log X_1^\star \log X_4^\star$ was found to be a statistically significant interaction term, $\log X_1^\star$ and $\log X_4^\star$ must be included in the model due to the hierarchy principle, and $\log X_2^\star$ and $\log X_3^\star$ were found to be statistically significant in the most recent partial *F*-tests. The resulting model is $\hat{\log Y} = 5.52487 + 0.67814 \cdot \log X_1^\star - 0.60966 \cdot \log X_2^\star - 0.30441 \cdot \log X_3^\star - 1.18917 \cdot \log X_4^\star - 1.01624 \cdot \log X_1^\star \log X_4^\star$. The residual plot for $\hat{\log Y}$ as well as each predictor exhibits a random scattering, thus I conclude that the error terms have constant variance and a linear relation. Since none of the predictors are sequential such as 'Time', I need not test whether the error terms are independent. The Q-Q plot of the regression is randomly and closely distributed around the linear line, hence I can conclude that the error terms are approximately normal in distribution. The studentized residual plot contains one value more extreme than |3|, thus I can conclude there is an outlier. The VIF values for the predictors range from 1.32993 to 10.54176, and thus multicollinearity is of minimal concern because the VIF values do not significantly exceed 10.

IV.B. Coefficient Interpretation & Implications

I interpret the estimated coefficients as follows; for every 1% increase in population, request count on average increases by (0.67814 -1.01624 $\cdot X_4^{\star}$)% holding all other predictors constant. This is considered an interaction effect, the effect of population on request count varies depending on the value of free press. Specifically, this is an interference interaction effect, meaning it is subtracted from the rest of the model. For every 1% increase in HDI, request count on average decreases by 0.60966% holding all other predictors constant. For every 1% increase in Internet Access, request count on average decreases by 0.30441% holding all other predictors constant. For every 1% increase in Free Press, request count on average increases by (-1.01624+0.67814$X_1^{\star}$)% holding all other predictors constant. This is considered an interaction effect, the effect of Free press on request count varies depending on the value of Population. Specifically, this is an interference interaction effect.

Our results find that our model contains one interaction term that contains two predictors, Population and Free Press Index. Our group overlooked this estimate when hypothesizing the importance of each predictor but, upon further reflection, it is intuitive. The greater the size of a nation, the more criminal cases and the greater the pool of potential suspects for investigation, thus resulting in more information requests to Google to expedite the prosecution process. Free press as a predictor for an increase in request count was also an interesting finding. I assume that this is the case due to increased press freedom leading to more pressure to convict criminal activity pushing nations to resort to requesting information from Google. As for our other predictors, they were all found to have an inverse effect of decreasing the request count. This makes sense for HDI as the more developed the nation, the less crackdown and crime there may be. It is also logical that increased internet access would decrease the need for requests as there would already be a lot of information available online. However it is interesting as with low internet access, how are the requests going to provide much information. The democracy index was not found to have an effect on request count.

V. Diagnosing Outliers & Influential Points

Before concluding our analysis of the model $\widehat{\log Y} = 5.52487 + 0.67814 \cdot \log X_1^{\star} - 0.60966 \cdot \log X_2^{\star} - 0.30441 \cdot \log X_3^{\star} - 1.18917 \cdot \log X_4^{\star} - 1.01624 \cdot \log X_1^{\star} \log X_4^{\star}$, I must first diagnose which data points are outliers and/or influential.

To identify outlying X observations, I must note which data points have leverages larger than $\frac{2p}{n}$, which in this instance is $\frac{12}{26} = 0.461$. A larger leverage value indicates that the case is both farther away from other X values and has notable leverage in determining its fitted value. To this end, observations #10 (India), #18 (Russia), and #19 (Singapore) had values 0.82982, 0.49896, and 0.66175 respectively. To identify outlying Y observations, I must note which data points have studentized deleted residuals larger than 3 or less than -3. To this end, observation #13 (Japan) had value -3.66726.

The remedial measures to address potential outliers #10, #18, #19, and #13 depend on whether the outlying observations are influential. The first means through which to diagnose influential points is using DFFITS values. DFFITS values measure the influence that cases have on their own fitted values. Because the dataset consisted of only 26 observations, I classified it as 'small', and thus decided that DFFITS values larger than 1—as opposed to $2\sqrt{p/n}$—suggested an observation was possibly influential. To this end, observation #13 had value -1.5927. I also utilized Cook's Distance (D_i), which measures the influence of cases on all fitted values, to screen for possibly influential points. The Cook's Distance value for an observation being larger than 1 or the probability of Cook's distance is less than $F(n, n-p)$ being larger than 0.5 signals influentiality. However, none of our observations satisfied any of these criteria.

There exist no obvious data errors in the dataset, such as problems with data collection, data recording, data that is out of the scope of our research, etc. Thus, I cannot drop the outliers and influential cases. In addition, the fitted regression model's adequacy has already been assessed, and the outliers and influential cases were not found to be caused by the omission of important predictors, interaction terms, or functional forms. In the final attempt to address the model's outliers and influential data points, I can recreate the model *without* using the influential and outlying data and compare the results.

By generating the model while excluding observations #10, #13, #18, and #19, some predictors experienced rather drastic changes in their β values. While β_3 and β_4 changed by less than approximately 0.1 when dropping the outlying and influential observations, β_0 changed by 0.161, β_1 changed by 0.147, β_2 changed by 0.763, and β_5 changed by 0.138. Thus, I can see that these four points are drastically influencing the model's coefficients, especially β_2 which corresponds to the mean-centered log HDI. The exclusion of these data points increased the R^2 from 0.62 to 0.73, indicating that the model accounts for a higher ratio of the variance in Y. Removal of the four points also decreased the range width of VIF values from $[1.323, 10.542]$ to $[1.537, 4.640]$, suggesting that multicollinearity is potentially less of a concern for at least one of the predictors after I remove the outlying and

influential points. The final major change is that $\log X_4^{\star}$, which is statistically significant according to the individual t-test before the removal of the four points, is found to be insignificant after their removal as its p-value increases from 0.016 to 0.125. However, this does not alter the model, because $\log X_4^{\star}$ is used in a significant degree-2 interaction term, and thus cannot be removed according to the hierarchy principle.

It is worth noting that, because the dataset only consists of 26 observations, each representing vastly diverse nations, influential and outlier observations are to be expected, and for that reason I will keep them in the model.

VI. Discussion of Data Analyses

Based on the results I found applying different transformations, I found that our model generated the best results when a log transformation was applied to request count and all of our predictors. Using this base additive model, I generated residual vs predictor and response plots, a Q-Q plot, e* vs Y plot, and an ANOVA table. Based on the results of these plots, I were able to verify the 7 assumptions of linear regression. Our goal with this research was to determine how our different predictors affect a country's request count, and in building an additive MLR model, I reduced the proportion of unexplained error variance in request count and generated more precise results compared to our SLR model.

I then proceeded to generate an MLR model with interaction terms between our quantitative predictors. Interactive models, in general, allow the relationship between the response and one predictor to vary with the values of other predictors. For our data, I initially generated a model including all the interactions. Our partial F-test analysis, however, revealed that the only significant interaction term was between population and FPI. Our model with this interaction term satisfied the requisite assumptions of linear regression, allowing us to reach more precise results when compared to our additive MLR model.

I also attempted to apply a polynomial MLR model to our data. In general, polynomial models are used if I believe the response function is something other than linear. However, when I applied a 2nd order model, I found that our predictors had significant issues of multicollinearity. In addition, the partial F tests revealed none of the polynomial terms were significant, and thus I will not include any polynomial terms in our final model.

In building our polynomial and interaction models, I ran partial F tests using their extra sum of squares values to determine what predictors were significant and thus needed to be kept in the model. For the interaction model, our hypothesis tests showed that every interaction term was insignificant to our MLR model with the exception of the interaction between population and free press, and therefore, that was the only interaction term included in our final model. For the polynomial model, I also conducted partial F-tests to determine the significance of our 2nd-order terms. Our hypothesis tests showed that every polynomial term was insignificant to our model, and thus I did not include any polynomial terms in our final model.

To determine if our data contained any outliers or influential points, I conducted the analysis using hat matrix leverage values, DFFITS, and Cook's Distance values. The term hii quantifies how far away an observation xi is from the rest of the x values, and an observation is deemed outlying if hii > 2p/n. For our data, 2p/n = 0.461. I found 3 points that were considered outlying with respect to x: Russia, Singapore, and India. The term ti uses the studentized deleted residuals to identify outlying y observations, and a point is deemed outlying if |ti| > 3. I found one point that was deemed an outlier with respect to y: Japan. A case is deemed influential if its exclusion from the model causes significant changes in the regression model. DFFITS is one metric used to identify influentials, measuring the influence an observation has over its fitted value yi. Using the criterion |DFFITS| > 1 for small-medium data sets, I found one point deemed influential, and this point was also identified as outlying with respect to y using ti: Japan. Cook's Distance D_i is another metric that is used, and it measures the influence an observation has on overall fitted values. Using the criterion $D_i > 1$ to identify influentials, no observations were found to be influenced by Cook's distance.

VII. Summary of Final Model

After attempting to apply a second-order polynomial model, as well as a model including interaction terms between all of our predictors, our final model was generated as follows:

$$\widehat{\log Y} = 5.52487 + 0.67814 \cdot \log X_1^{\star} - 0.60966 \cdot \log X_2^{\star} - 0.30441 \cdot \log X_3^{\star} - 1.18917 \cdot \log X_4^{\star} - 1.01624 \cdot \log X_1^{\star} \log X_4^{\star}.$$

The residual plot for $\widehat{\log Y}$ as well as each predictor exhibits a random scattering, thus I conclude that the error terms have constant variance and a linear relation. Since none of the predictors are sequential such as 'Time', I need not test whether the error terms are independent. The Q-Q plot of the regression is randomly and closely distributed around the linear line, hence I can conclude that the error terms are approximately normal in distribution. The studentized residual plot contains one value more extreme than |3|, thus I can conclude there is an outlier. The VIF values for the predictors range from 1.32993 to 10.54176, and thus multicollinearity is of minimal concern because the VIF values do not significantly exceed 10. Taking all of this into account, I can conclude that our final model satisfies the assumptions of linear regression.

The 95% confidence interval for logPop is [5.12739, 5.92235]. 95% of generated confidence intervals over a large sample will contain the true coefficient value for logPop. The 95% confidence interval for logHDI is [-11.62139, 10.40207]. 95% of generated confidence intervals over a large sample will contain the true coefficient value for logHDI. The 95% confidence interval for logInternet is [-2.61227, 2.00309]. 95% of generated confidence intervals over a large sample will contain the true coefficient value for logIntenet. The 95% confidence interval for logFreepress is [-2.13598, -0.24235]. 95% of generated confidence intervals over a large sample will contain the true coefficient value for logFreepress. The 95% confidence interval for logPop*logFreepress is [-1.66387, -0.36861]. 95% of generated confidence intervals over a large sample will contain the true coefficient value for logPop*logFreepress. The 95% confidence interval for our mean response, logRequests is [4.8215833, 5.926657]. 95% of generated confidence intervals over a large sample will contain the true coefficient value for logRequests.

I can gather meaningful interpretations from the coefficients of our predictors. Based on our confidence intervals and coefficients, areas with higher populations may experience higher rates of crime leading to more data requests. Law enforcement could potentially use this to more efficiently allocate resources. I observed a negative relationship with HDI, and government agencies seeing this data could attempt to implement community-driven clean-up events in areas with lower HDI in an attempt to lower criminal activity. I observe a negative relationship with free press index, meaning countries that enjoy more freedom of press have less data requests. Policymakers in areas that enjoy lower levels of freedom of press could potentially attempt to implement more transparency so as to lower crime rates and data requests.

VIII. Conclusion

My goal for this study was to find the factors that influence government request counts to Google for criminal investigations. I did this by accessing data from Google's transparency reports. I investigated the relationship between request count and a number of predictor variables that included population size, human development index, internet access rates, democracy index, and free press index. Through linear regression models and subsequent analysis, a positive relationship between request counts and population size was found. Conversely, the human development index, free press index, and internet access rates were found to have an inverse relationship with the number of requests. Additionally, the interaction term between population and free press revealed an interference interaction effect, suggesting it modifies the influence of the other predictors in the model. My analysis was conducted entirely based on Google's data, and while their data is certainly comprehensive, it is likely it does not cover the full collection of data inquiries across the world. With that in mind, it is possible the data and subsequent conclusions are not representative. Due to the nature of this investigation being an observational study, I am unable to establish a causal relationship between our response and predictor variables.

Works Cited

1. “Google Transparency Report” *Google Transparency Report*, https://transparencyreport.google.com/about?hl=en

2. Ovide, Shira. “Police Love Google’s Surveillance Data. Here’s How to Protect Yourself.” *Washington Post*, 26 Oct. 2023, www.washingtonpost.com/technology/2023/10/24/google-privacy-police-geofence.

3. Love, Julia & Alba, Davey. “Google User Data Has Become a Favorite Police Shortcut” *Bloomberg*, 28 Sept. 2023, www.bloomberg.com/news/features/2023-09-28/google-user-data-is-police-s-top-shortcut-for-solving-crimes.