Political Methodology
Committee on Concepts and Methods
Working Paper Series

6

November 2005

Interpreting Long Tails

Qualitative Analysis of Regression Residuals

Arun Agrawal

University of Michigan

(arunagra@mich.edu)

Ashwini Chhatre

Duke University

(ac43@duke.edu)

C&M
The Committee on Concepts and Methods
www.concepts-methods.org

IPSA
International Political Science Association
www.ipsa.ca

CIDE
Teaching and Research in the Social Sciences
www.cide.edu

The C&M working paper series are published by the Committee on Concepts and Methods (C&M), the Research Committee No. 1 of the International Political Science Association (IPSA), hosted at CIDE in Mexico City. C&M working papers are meant to share work in progress in a timely way before formal publication. Authors bear full responsibility for the content of their contributions.

The Committee on Concepts and Methods (C&M) promotes conceptual and methodological discussion in political science. It provides a forum of debate between methodological schools who otherwise tend to conduct their deliberations on separate tables. It publishes two series of working papers: "Political Concepts" and "Political Methodology."

Political Concepts contains work of excellence on political concepts and political language. It seeks to include innovative contributions to concept analysis, language usage, concept operationalization, and measurement.

Political Methodology contains work of excellence on methods and methodology in the study of politics. It invites innovative work on fundamental questions of research design, the construction and evaluation of empirical evidence, theory building and theory testing. The series welcomes, and hopes to foster, contributions that cut across conventional methodological divides, as between quantitative and qualitative methods, or between interpretative and observational approaches.

Submissions. All papers are subject to review by either a member of the Editorial Board or an external reviewer. Only English-language papers can be admitted. Authors interested in including their work in the C&M Series may seek initial endorsement by one editorial board member. Alternatively, they may send their paper to workingpapers@concepts-methods.org.

The C&M webpage offers full access to past working papers. It also permits readers to comment on the papers.

www.concepts-methods.org

1. Introduction

Comparative analytical studies of political and social phenomena have a long history of tensions between proponents of quantitative and qualitative approaches. In sociology for example, debates between those reposing a faith in quantitative analytical tools and others expressing doubts about the efficacy of quantitative methods can be traced back at least to the 1920s (Cobb 1926, 1927; Ogburn 1934). More recent discussions around such tensions are common as well (Bryman 1984, Collins 1984). Within political science, similar tensions and efforts at conversations across the quantitative-qualitative divide abound (Bates et al. 1998, Coppedge 1999). Thus, much recent writing has elaborated the potential contributions of case studies, QCA (qualitative comparative analysis), and structured-focused comparisons (Gerring 2004, Mahoney and Rueschemeyer 2003). But proponents of quantitative techniques maintain that testing and validation of theoretical propositions is best accomplished by statistical analysis using appropriate safeguards (King, Keohane, and Verba 1994). At best qualitative techniques are viewed as handmaidens to quantitative arguments: useful in generating contingent empirical generalizations to aid theory development.

This paper contributes to the fertile debates in the discipline on the use and relative merits of qualitative and quantitative approaches for theory development and validation. We discuss how a deep knowledge of cases and contexts, the most powerful aspect of qualitative methods, can powerfully be combined with quantitative and statistical methods to strengthen causal analysis, and promote both theory development and validation. Our approach departs from many existing writings that treat qualitative and quantitative approaches as being opposed to each other and existing in an inevitable tension. It builds instead on writings that view qualitative and quantitative approaches as quite distinct and highlight either their characteristic strengths, or focus on the ways in which they are both necessary to a better understanding of a given problem. To do so, we outline a strategy whereby integrated use of qualitative and quantitative methods improves theoretical understanding and empirical validation.

Our strategy hinges on careful analysis of regression residuals as a source of information. Quantitative studies rarely analyze residuals beyond their utility as indicators of goodness of model fit. Even a cursory examination of recent papers in the top-ranked journals in the discipline validates this point. We suggest, in the spirit of a robust qualitative tradition, that focusing on extreme outlier residuals can illuminate aspects of theory that would likely be inaccessible without careful attention to and deep knowledge of cases corresponding to the outlier residuals. A focus on residuals, especially on those residuals that are outliers from the fitted regression model, is inherently an attempt to understand why particular observations depart from explicitly articulated hypotheses. The cases that correspond to outlier residuals are typically the ones for which the theory corresponding to a regression model is least able to account. Closer investigation and intimate knowledge of the specific cases that correspond to these residuals, we suggest, can help in the development of better theoretical arguments, help refine the measurement of specific variables, gain a deeper awareness of the functional form of the relationship between dependent and causal variables, and improve the understanding of causal mechanisms at play: in short, help in advancing knowledge in ways that are the objective of most scientific investigations. Conversely, inattention to residuals and the potential information they contain can hinder the full development of theoretical models that motivate regression analyses, and prevent a better understanding of the data under investigation.

We set up the argument of the paper in the next section by referring to existing approaches to analyze and use regression residuals. We describe how the strategy developed in this paper differs from existing approaches and outline the utility of our strategy in the context of studies of collective action and resource management. We outline the criteria along which improvements would mean that the suggested strategy of analyzing residuals and their information content has a degree of generalizability beyond our case and empirical materials. We concretely illustrate our argument in sections 3 and 4 using a dataset of 205 cases of resource governance from the Indian Himalaya. Section 3 presents the results of our initial OLS regression model which uses existing theoretical advances in the field of common property to analyze how socio-political, institutional, and bio-physical variables affect changes in resource condition – the common objective of much analysis in the study of common property. The model yields a set of outlier residuals that correspond to specific observations in our data set that we examine in section 4. The examination of the observations corresponding to these residuals yields new insights related to the theoretical model such as identifying possible omitted variables, operationalization of specific causal variables and processes in statistical terms, and potential ways of refining measures of variables. We also discuss how the qualitative analysis of residuals may not contribute sometimes to general theoretical development. Section 5 presents the new model that we build, using the information revealed from a closer investigation of the cases corresponding to the outlier residuals. It also discusses the extent to which the results of the new model are an improvement upon the initial analysis, and the ways in which the new model is a departure from existing insights related to studies of the commons. Section 6 discusses the scope of our findings for studies of resource governance, and also more generally for comparative political analyses that use regression models to analyze data and advance the understanding of political phenomena.

2. Residuals and Common Property Theory

In most regression models, even where the statistical model fits the data quite well, several cases are not predicted well by the model: these constitute extreme residuals, falling in the tails of the distribution of errors. Typically, statistical analyses treat residuals, including the extreme cases, as stochastic elements in relation to the regression equation, and therefore not of much consequence. However, observations corresponding to extreme residuals can have high information content, depending on how well the regression model represents underlying causal processes, whether all the relevant causal variables have been included in the model, how well the functional form of causal relationships matches the model, and whether specific variables have been measured along an appropriate scale. To the extent that the residuals are not in fact random, qualitative analysis of large residuals can throw light on the systematic component generating these residuals. They can help check for possible measurement error, aggregation bias, nonlinearities in posited relationships or other misspecification of functional forms, and possible interaction effects. A careful, qualitative examination of observations (cases) corresponding to extreme residuals can also assist in analyzing whether posited causal mechanisms apply to extreme values of independent variables, and thereby help specify the range within which theoretical predictions concerning specific variables are valid. Finally, an examination of residuals can help improve knowledge about potentially omitted variables and conjunctural causation.

Existing discussions of residuals, and how they are to be understood in the context of regression models focus primarily on defining them, and using them to assess violations of assumptions underlying the regression analysis on the basis of some test statistics and examining them to locate the extent of bias in parameter estimates (Cox and Snell 1968, 1971). Some analysts who have focused on residual analysis have suggested one potential strategy to be the removal of extreme outliers that exert a strong influence on parameter estimates. Such a strategy may be justified if there are strong grounds for believing that the cases corresponding to the influential outliers do not belong to the same population as the rest of the sample (Chatterjee and Wiseman 1983, Stevens 1984). But instead of such a radical strategy to create a better fit between the regression model and its underlying theory, we build upon the arguments advanced by Bollen and Jackman (1990: 286-87). They strongly advocate for a careful examination of outliers by using substantive information about cases to understand why they are outliers and highlight the payoffs to such conversations between knowledge of cases and the statistical analysis of data.

Our approach is in some contrast then to more statistically oriented analyses of outliers and efforts to understand why specific cases exert a large effect on coefficient estimates. We focus on gaining additional information from the cases that correspond to the more extreme residuals, even if they do not exert a large effect on coefficient estimates (Kahn and Udry 1986: 734-37).[1] We should note that our objective in examining the residuals is to identify patterns that may apply to the entire set of observations, not to the specific observations per se. Our analysis also contrasts with typical criticisms of quantitative approaches that focus on the difficulty of deriving causal inference from statistical correlations or on the lack of texture that is viewed as being characteristic of statistical analyses of a large number of observations. We accept that statistical approaches are the best available means for testing and validating theory by

[1] It is useful to distinguish between outliers and influential observations. According to Bollen and Jackman (1990: 258), outliers are observations that are distinct from most other data points in a sample. Outliers may be cases with extremely high or low values on specific variables, or ones where the residual is large compared to other residuals from a regression analysis. But an outlier observation is influential only when its deletion causes a pronounced change in one or more of the estimated parameters. See Andrews and Pregibon 1978.

using large amounts of data. Rather than try to undermine the force of quantitative arguments by drawing attention to the odd cases that do not fit a regression equation, we outline an integrated analytical approach in which the greatest strengths of qualitative and quantitative strategies can be combined to produce better causal accounts. We develop a systematic strategy to analyze residuals qualitatively that can also be used by other scholars who rely on regression techniques for data analysis.

The strategy of examining outlier residuals that we propose is especially valuable in the context of the current state of the field of common property. The early literature on the subject had focused primarily on case studies and descriptions of particular examples of the functioning of common property systems (Alexander 1982, Berkes 1989, McCay and Acheson 1987, NRC 1986, Netting 1981). These early analyses enhanced the understanding of how common property systems work in many different locations, and drove home the point that it is possible to govern resources in a sustainable fashion without resorting to market or state-based solutions. Ostrom's (1990) influential analysis of institutional features of common property arrangements used earlier studies and strategic-choice analytical techniques to make new theoretical advances and identify a number of causal variables that are potentially critical to successful collective action around the management of the commons. In the wake of her work, a number of other scholars also identified other causal variables that could explain observed successes in the governance of the commons (Baland and Platteau 1996, McKean 1992, Wade 1994; see also Pinkerton 1989, an earlier contribution). More recent work has begun to use different quantitative techniques to address the multiplicity of hypotheses and conjectures about what makes for successful governance of common pool resources (Agrawal and Chhatre 2005, Bardhan and Dayton-Johnson 2002, Gibson and Lehoucq 2003, Lise 2000). In the process, we are beginning to see an interesting conversation between quantitative data, theory, and knowledge of specific cases.

A number of potential difficulties beset efforts to understand the causal mechanisms that underpin successful functioning of commons institutions. One of the most important such problems stems from the sheer number of causal variables that have been identified in the literature. Agrawal (2001) reviews the literature on the commons to identify more than 35 potential causal variables that can influence outcomes. However, in any given analysis, it is extremely difficult to know a priori which of these 35 variables are the ones upon which to focus analytical attention. Omitted variable bias is therefore a real problem in both quantitative and qualitative case studies of the commons. In addition, scholars of commons also confront problems related to the functional form of the relationship between a given causal variable and the dependent variable, and identifying whether particular causal processes work in the hypothesized manner across the range of observed values in the sample. The strategy for analyzing residuals that we identify and advance in this paper, therefore, has clear implications for students of the commons.

3. Common Property Governance in Himachal Pradesh: Findings and Anomalies

Our sample of 205 cases concerns the governance of forest resources in Himachal Pradesh in northern India, a region where we have been conducting research for more than five years. Our regression model is built around five classes of potential causal influences: biophysical, demographic, economic, institutional, and socio-political, to explain change in the condition of local resources as these changes are reported by our respondents. Existing studies have recognized each of the above categories as being instrumental in influencing resource governance outcomes (Brown and Pearce 1994, Rocheleau and Edmunds 1997). For each category of influence, we examine several variables that have been highlighted in the literature and which appear relevant to the Himachal Pradesh context (see table 1 below).

The state of Himachal Pradesh is ecologically highly diverse owing to distinct climatic and physiographic factors. More than two thirds of the state is administratively classified as forest (37,600 square kilometers). The state's population was 6.01 million in 2001, with an average density of more than 100 persons per square kilometer. The overwhelming majority of the population is rural: more than 90 percent, and lives in nearly 16,000 villages (DOP 1997). Forests are critical to hill agriculture and also contribute directly to livelihoods. The state has a number of different institutional arrangements through which forests are governed. Most generally, these conform to the standard classification often used in the common literature – private, communal, and state or publicly owned.

We selected villages for data collection by sampling across the altitudinal gradient in the state, equally from the lower hills (<900 meters above mean sea level), middle hills (between 900 and 1800 meters), and high hills (> 1800 meters). Within each altitude class, we selected cases to represent different institutional regimes, in proportion to their distribution across the three altitude classes. This strategy ensures that all major types of forests and institutional regimes are represented in our sample, and that the cases are not picked on the basis of variations in the value of the dependent variable. We should note that it is near impossible to identify a fully random sample for local institutional or forest types – not just in India, but for much of the developing world – owing to the non-existence of any comprehensive lists that contains the relevant information.[2]

Table 1 presents the descriptive statistics on the variables on which we collected data during our field research and which we used for initial analysis to identify the causal mechanisms that explain changes in forest condition.

[Table 1 here]

The dependent variable in our analysis is "Change in Forest Condition". It is measured by an index, based on group responses on a five-point scale for changes in the condition of the forest from 1) upper caste men, 2) upper caste women, 3) lower caste men, 4) lower caste women, and 5) forest department guards. The measure is continuous and symmetric around the mean, which leads us to use an OLS model for the regression analysis. Observations vary between 1 for forests whose condition has changed significantly for the worse to 3 for forests that have registered little change, and 5 for forests that have improved substantially over the past.

Although we collected data on a number of different variables regarding the condition of local forests, respondents' assessments of changes in forests they see daily and know intimately are more likely to be accurate than assessments about condition of a local forest in terms of its productivity or biodiversity. While biological measures of changes in forest conditions are likely to be a better measure of what is actually going on in forests, they would require long term data collection that has seldom been attempted in

---

[2] The questions we used during our field work come from the set of data collection instruments developed by the International Forestry Resources and Institutions (IFRI) Program at Indiana University (Poteete and Ostrom 2004; the full set of instruments can be obtained from the Workshop in Political Theory and Policy Analysis at Indiana University). We used two instruments: one focusing on information related to the village and the other on the management institutions in place.

conjunction with data on socio-economic, demographic, and political variables. We believe that respondents' assessments of changes in forest conditions are more reliable than of forest condition itself because of the inherent difficulties associated with comparing forests across ecological contexts and forest types. Indeed, even silviculturists do not have single measures to represent forest conditions or biodiversity that can be taken as appropriate bases for comparison across forest and ecological types.

The independent variables in the table are classified under different categories of influences that a number of analyses of common property have identified as being important to understand sustainable governance of resources. Our decision to include these variables in the analysis was driven by the existing literature on the subject as well as factors that seem to be particularly relevant in the context of Himachal Pradesh. The last column of table 1 refers to works that highlight the causal significance of particular categories of variables in explaining forest condition. The references are only a selection from the vast literature on the commons; there are many other writings that also focus on the variables we use in the model. The specific variables we have chosen for the analysis reflect the contextual particularities of Himachal Pradesh as discussed above. Table 2 provides information on how different variables have been operationalized for the ensuing regression analysis.

[Table 2 here]

The results of the regression analysis are presented in table 3. The table indicates that most of the biophysical, demographic, and sociopolitical variables we included in our analysis are statistically significant. Number of fires in village forests is associated with negative changes in forest condition, number of households, and grazing by goats and cattle also have negative signs and statistically significant coefficients. However, the variable "number of cattle" has a significant and positive association with changes in forest condition. In our sample, the variable represents all the cattle owned by villagers, including those that are not grazed openly. In the presence of the grazing variables, the number of cattle captures the general level of wealth in the village. Therefore, the positive coefficient is in conformity with the hypothesis that higher overall wealth, controlling for the levels of grazing in the forest and for homogeneity in the village, would contribute positively to changes in forest condition.

[Table 3 here]

When we come to socio-political variables, we find that higher levels of education, greater homogeneity in landholding, and presence of conflict resolution mechanisms in the surveyed villages all contribute to improvements in forest condition. On the other hand, higher levels of college education are associated with negative changes in forests. It seems clear that the education variable has a relatively complex relationship with changes in forest condition (see also Agrawal and Gupta 2005). Whereas overall education levels are associated with positive changes, higher education has a negative effect. This points toward a potentially nonlinear relationship between education and changes in forest condition.

Contrary to our expectation, and in contrast to results of some existing empirical research, the economic/dependence variables, and most of the institutional variables turn out not to be statistically significant. Among the economic variables, subsistence use of forests is associated with improvements in condition. But other measures of dependence we used have no statistically significant effects. Similarly, for institutional variables, we find statistically significant effects for only two of the nine variables we used. Decentralized control over forests and the undertaking of plantation activities both have statistically significant positive effects on forest condition, but a range of other institutional variables that the literature identifies as being important do not. Especially puzzling is that none of the three variables -- illegal use, fines, and sanctions -- have clear effects.

Despite the fact that our model does not confirm some of the existing results in the literature on the commons, it does have a good fit with the data. The residuals are normally distributed with the expected number of observations in the tails. We checked for heteroskedasticity using the Breusch-Pagan Test, and the Ramsey Regression Specification Error Test to check for possible nonlinearities in the relationships. We also examined whether individual observations have strong influence on the coefficients by using the Cook's Distance and Leverage statistics. These basic diagnostics confirm that the model is well specified. In addition, we also checked for multicollinearity by calculating the Variance Inflation Factor for each variable. The scores indicate that none of the variables are strongly correlated.

4. Analysis of Residuals

After having confirmed that the model (see table 2) is indeed a good fit for the data, and checked for possible violations of assumptions through purely quantitative methods, we turned our attention to the few observations that are predicted poorly by the model. We did so in the belief that although the model performs well in terms of standard regression diagnostics, the outliers could contain information about patterns in the data that the model does not adequately capture. This section analyzes the five cases each that are most over- or under-predicted by the model. The basic information about these observations is provided in tables 4 and 5.

[Table 4 here]

The first thing we note in the table is the evidence for the classic regression towards the mean. The overpredictions are all low observed values of the dependent variable, and the underpredictions are high values. Given that the unconditional distribution of the dependent variable is symmetric around the mean and concentrated in the middle, it is to be expected that the mass of observations in the middle of the distribution would exert a centripetal force on predictions for the extreme observed values. Therefore, a straightforward explanation for the failure of the model to predict these observations correctly could be that there is less information on the edges of the distribution to incorporate into the model. This is also partially reflected in the slightly higher values of the studentized residuals compared to the standardized residuals for all the ten cases.

Overall, the ten observations do not display any clustering patterns. They are located in six out of the ten districts covered by the study, and represent all the three altitudinal zones that were part of the sample selection criteria. There is great diversity in terms of biophysical conditions across the ten cases, along slope, aspect, altitude, and species composition of the forests under study. At first glance, there are only two clearly discernable biophysical or ecological factors that might be correlated to the dependent variable in these ten cases. First is the size of the forest, which could have an implication on the direction of change in the condition of a particular forest under study. From table 4, it appears that the cases are on the lower side of the mean for the sample (mean = 131 hectares, standard deviation = 405) although only three observations are far from the median (30 hectares). The absence of any cases with large forests in the tails of the distribution of residuals suggests that the size of the forest may be an omitted variable that is correlated with other exogenous variables and exert a causal effect on the dependent variable.

[Table 5 here]

Ecologists have pointed to the importance of the relationship between the size of an ecosystem and its ability to maintain its structure and functions in response to exogenous disturbances. Thus there is indirect evidence from the ecology literature that size might be a relevant variable that we had left out of the original model. However, there is another pattern to the cases as well. All the forests that are underpredicted by the model are forests that are in very good condition (see table 5). Forest condition is measured as a composite average of six group responses (similar to the dependent variable) to questions regarding three measures of total biomass in the forest – stem density, crown cover, and understory. It ranges from 1 for very poor condition to 5 for excellent condition. While the forests that are overpredicted do not display any clear pattern in this regard, the converse is not true. Thus, if the condition of the forest was positively related to the dependent variable (change in forest condition), it is likely that these cases would have been predicted better. That is, if better forests are more likely to have improved (rather than degraded), then the condition of the forests represents another omitted variable to the extent that it is correlated with other exogenous variables.

Other relevant facts also emerge from an examination of the cases corresponding to the extreme residuals. Three of the overpredicted forests experienced multiple fires in the last ten years – Heb, Deol, and Madaro. While they are not far from the mean (3.27) or the median (2.00) for the sample, the variable "Number of fires" does not capture the intensity of any of the fires in the forest. Two of these cases, Heb and Madaro, witnessed devastating fires for extended periods at least twice each in the ten years preceding data collection. The intensity of these fires might explain the overprediction for these cases, since the causal mechanism posits a negative relationship between number of fires and change in forest condition. If we had better instruments to capture the intensity of fires, the model would likely predict lower values for these forests.

All the three overpredicted forests with high intensity of fires also share two other characteristics. All of these are small forests, ranging from five to thirteen hectares, and all three forests have a high proportion of conifer tree species. It is likely that small forests are more vulnerable to fire damage, suggesting an interaction effect between size of forest, and number or intensity of fires. Another possibility suggested by the pattern is that conifer forests are more vulnerable to fires. In the model, the variable “subsistence use” captures the proportion of conifer trees in the forest (1 = pure conifer, 3 = pure broad-leaved). The three cases suggest a possible interaction effect between “subsistence use” and “number of fires”.

The high over-prediction for the Thalli Demarcated Protected Forest raises an interesting issue. The dataset consists of 205 forests, but these are distributed over only 103 villages. Every village has approximately two forests in the dataset. Thalli village has two forests in the sample, one of them being a self-initiated forest management institution. Thalli Mahila Mandal has been protecting this forest actively since 1992, but the village has witnessed several episodes of conflict between the women and a group of goat-herding families that periodically protested against and broke rules regarding restrictions on grazing in the forest protected by the women. Over the decade before the study was undertaken, the conflict was partially settled by allowing the goat-herders higher grazing privileges in the neighboring Demarcated Protected Forest of the same village. Over the five years before the data collection, it is likely that the degradation of the forest (as evident from its low observed value on the dependent variable) was a result of the shift of pressure from the forest protected by the women’s group. The model fails to capture this dependency between two observations from the same village. In other words, it represents a violation of the assumption that all observations are independent of each other. Rather, it appears to be the case that observations might be clustered by village.

There are two private forests in the ten cases being analyzed. Chakmoh is overpredicted, while Kanoh is underpredicted. Several of the variables in the model are measured at the level of the village (for example, “education”, “number of cattle”, and “homogeneity”). Many of these represent causal mechanisms that have little or no interpretation for privately owned forests. For example, the “subsistence use” variable captures the hypothesis that broad-leaved forests are more useful for subsistence purposes than conifer forests, representing the broader argument that the degree of dependence on the forest is an important factor in explaining the change in its condition over time. However, to the extent that most villagers are excluded from decisions regarding the management (or even use) of private forests, it is not meaningful to assume such a relationship. The dummy variable “decentralized control” represents the hypothesis that forests will improve when users are able to make collective decisions regarding management of a forest in relation to their level of dependence and other factors. The variable includes private forests in its ambit, and the above discussion suggests that these forests perhaps belong to the other category. In effect, the dummy variable needs to represent community management rather than decentralized management.

Three of the five underpredicted cases are self-initiated forest management institutions – Dhimkataru, Pangaon, and Bagotla. All of these are forests that have improved far more than predicted by the model, and also have forests in good condition. Bagotla represents a successful case of self-initiated community management, having started protection in 1960 in response to increasing scarcity of subsistence goods such as fuelwood and fodder. The forest was closed to all grazing for a decade between 1990 and 2000, to allow for regeneration and the survival of plantation of saplings undertaken by the community. However, the forest was opened to grazing in the year 2000. Consequently, the dataset has high values on the grazing variables for Bagotla community-managed forest. The model confirms the general hypothesis that high level of grazing is negatively associated with change in forest condition. But the negative coefficients on the grazing variables bring down the predicted value for Bagotla much more than warranted by the hypothesis, since it assumes that the level of grazing recorded in 2000 is representative of the intensity of grazing over the last several years. This is an implicit assumption in the measurement of several variables, but particularly true for causal mechanisms that are played out over time but measured only cross-sectionally.

In a similar vein, the underprediction of Pangaon self-initiated community forest can be attributed to the failure of the measurement of a variable to adequately capture the causal mechanism. Pangaon is close to the city of Manali and subject to the depredations of a timber mafia engaged in illegal felling of precious cedar trees from neighboring forests to meet urban demand. Pangaon villagers, led by its Yuvak Mandal and Mahila Mandal and supported by the district administration, initiated an intensive patrolling operation in the forest in 1990 in collaboration with the Forest Department. Villagers contributed several hundred mandays in the first few years in patrolling the two forests in the village (and the dataset), successfully curbing the incidence of illegal timber harvests. However, by the mid-nineties, the villages shifted to a system of third-party monitoring through the seasonal hiring of guards to protect the forest. The variables that represent monitoring and enforcement in the model – "voluntary labor" and "illegal users" – do not sufficiently represent the process in Pangaon. "Voluntary labor" is a measure of the total number of person-days of voluntary labor in the last five years, but does not capture the effect of high values of voluntary labor preceding the last five years, as in the case of Pangaon.

Heb is a village with high levels of "education", with a value of 18 compared to the sample mean of 13.7. The self-initiated forest management system in Heb is also crumbling under a series of collective action problems. Villagers blame the high level of education as a source of the problem, whereby 'over-educated' people find it below their dignity to participate in monitoring and enforcement activities planned by the community institution. On the other hand, Baggi has a relatively low score on "education", at 11 compared to the mean of 13.7. Heb is overpredicted by the model, whereas Baggi is underpredicted. The combination of the two suggests that there might be a non-linear effect of "education" on the dependent variable, with a negative effect for higher values of "education". In other words, if there was an inverse U-shaped relationship between education and change in forest condition, the inclusion of a quadratic term in the model would make the prediction of Heb slightly lower, and the prediction of Baggi slightly higher than the present model.

To sum up, the qualitative analysis of the cases that were least well predicted by our initial model has led to the following additional hypotheses: (1) Size of the forest might have a positive relationship with the change in forest condition. Further, the effect of "subsistence use" and "number of fires" might be conditional on the size of the forest; (2) The condition of the forest is another possible omitted variable; (3) "Number of fires" fails to adequately capture the causal mechanism linking the incidence of fire and change in forest condition. It might be profitable to use an alternative measurement. Additionally, the effect of fires might be moderated by the composition of the forest in terms of the proportion of conifer trees, represented by an interaction between "number of fires" and "subsistence use"; (4) The variable "decentralized control" does not capture the underlying causal mechanism for private forests. A better alternative would be to use a dummy variable for community-managed forests; and finally, (5) Two observations per village might violate the independence assumption. A possible correction might be to use the Huber-White 'sandwich' estimator for the variance of the errors in a regression with robust standard errors.

5. Presentation and Discussion of New Model

To examine these hypotheses, we reanalyzed our data. We incorporated three new variables (Forest area, forest condition, and education squared) into the model. We also included the interaction terms identified through the analysis of extreme residuals. We rescaled the "number of fires" variable to represent the intensity of fires by down-weighting the observations with a high number of fires. We did this through a categorization of the continuous variable into a five-level categorical variable, so that the distance between the lowest and highest values was reduced from 30 to 5. The qualitative argument informing this transformation relates to the possibility that if a forest experiences a high-intensity fire, it is unlikely to experience another fire at least for a few years. Thus, the forests with lower number of fires might be those that have sustained the greatest damage. The rescaled variable takes this possibility into account without ignoring the negative impact of a high number of low-intensity fires. Finally, we estimated the model with (and without) the Huber-White sandwich estimator to correct for possible violation of the independence assumption through clustering. The basic statistics for the new variables are available in table 6 below.

[Table 6 here]

Table seven presents the results of the new regression analysis. The variables are listed according to the same set of categories of influence as in table 3. The results of the analysis indicate that the new model has improved in terms of several measures of goodness of fit. The statistical significance of several variables included in the first model has improved. The average error in prediction (Root MSE) has declined by approximately 8 percent. The adjusted R-squared has also improved, as has the F-statistic. Our diagnostic tests for heteroskedasticity, multicollinearity, and nonlinearities show that at a general level the second model performs as well as the first one.

The final model, whose coefficients are presented in table 7 does not include size of the forest and the interactions discussed earlier. We found that size has no effect in any of the specifications, and neither do any of the interactions. After the incorporation of the quadratic term for education, the variable "college" becomes insignificant, and has been subsequently dropped from the model. The switch makes substantive sense; the proportion of people who have been to college would be highly correlated with the square of the measure of education, but if there is non-linearity in the relationship, it would be better captured by the quadratic term. Finally, we also estimated the model with the sandwich estimator of variance. The last two columns of table 7 report the results for the standard errors and p-values resulting from the sandwich estimator. The results suggest that the clustering of the data by village is not distorting the parameter estimates of the model.

[Table 7 here]

Perhaps the most important improvement, of the several different areas in which the model improved, is the greater statistical significance of the institutional and economic variables. In the earlier model, both these sets of variables had performed poorly in predicting changes in forest condition. The inclusion of Forest Condition as a control variable, the redefinition of the effects of fires from Number of Fires to Intensity of Fires, and Decentralized Control to Communal Control has led to a much improved model fit. Similarly, the incorporation of a non-linear term to account for effects of education shows that the intuition regarding how education affects changes in forest condition and the information from the inspection of the residuals were on target.

The results of the new model in terms of the coefficients for the institutional variables are interesting to note. The signs of all the variables remain unchanged from model 1, but their statistical significance improves as either the coefficients become larger with standard errors remaining the same, or the standard errors become smaller. The negative coefficients for illegal use, external users, and the positive coefficients of labor contributions, community control, and plantations are in conformity with what the existing literature on the commons says about these variables. However, the negative coefficient for fines, sanctions, and traditional institutions are important to examine further.

A number of studies of the commons have shown that regular monitoring and enforcement, and fines and sanctions associated with such enforcement are an important correlate of better forest condition (Gibson, Williams, and Ostrom 2005, McKean 1992). Our interpretation of the statistical relationship in our model is that in the studied cases, forests are likely to change in a worse direction when forest management associations use fines as a way of improving the association's income or if they resort to sanctions too frequently.

An alternative explanation of the finding also deserves mention. That is, villagers are more likely to impose fines and sanctions if they perceive their forests to be changing for the worse. Thus, the causal arrow suggested by our data runs in the reverse direction from what we had hypothesized. Two different underlying mechanisms may be at work in this explanation. Fines and sanctions are imposed either when decision-makers see adverse changes occurring in forests that were earlier in good condition, or when they believe that the adverse trends in forest conditions can be reversed.

The negative sign for traditional institutions is also in some tension with writings on indigenous knowledge and institutions that suggest that the presence of such institutions can be used to improve contemporary governance. Our data suggests that at least in the context of Himachal Pradesh, the presence of non-forestry related traditional institutions may not work to help the governance of resources. To a great extent, it can be attributed to the hierarchical and caste-based character of traditional institutions, which inevitably puts them in conflict with new social norms of equality of opportunity and access. Villages with strong traditional institutions might have greater difficulty solving collective action problems, especially when the problem involves restrictions on access to forests for collective benefits. Thus, it would appear that a general reliance on indigenous institutions is not necessarily conducive to improvements in conditions of resources.

Similarly, the positive coefficient for encroachments deserves additional commentary. Himachal Pradesh witnessed an aggressive land reforms campaign beginning in the mid-sixties, and extending until the late-seventies. Simultaneously, there was expansion of state support for horticulture, particularly apple orchards. The two processes resulted in a spate of encroachments on public and communal forests. This shock to existing systems of forest management institutions elicited varying responses, but very commonly, it galvanized both the forest department and local communities in improving the monitoring and enforcement of rules in the forests. The knowledge of this specific process in Himachal Pradesh has informed the inclusion of this variable in the model. We test the hypothesis that the incidence of encroachments in an early period prompted village communities suffering encroachments to improve their protection efforts, so that by the time of data collection these villages had forests that had changed in an improving direction. The fact that the encroachments had taken place nearly 30 years ago provides sufficient time for forests to reflect the impact of greater protection.

6. Scope of Findings and Conclusion

This paper has attempted to show, in the spirit of a robust qualitative tradition, that knowledge of particular cases and insights drawn from a thorough awareness of causal processes in specific contexts can help contribute to general theory building. We have applied this insight to the analysis of extreme residuals that are present in nearly all regression analyses. By focusing on specific cases corresponding to these extreme residuals and examining them more carefully to see whether they hold lessons that can be used to provide new hypotheses that can be tested using both the existing data and collecting new data, this paper points toward fruitful new avenues for investing analytical energies for scholars interested in an intimate conversation across the common quantitative-qualitative divisions in political science as well as other social science disciplines.

We show that new hypotheses drawn from an inspection of the residuals can help improve the fit of regression models even to existing data. The improvements we have in mind relate directly to the common diagnostics that most analysts use to evaluate the robustness of their statistical analysis. But equally important to our mind are improvements in the theoretical understanding of causal processes that may be at work in the field of an analyst's interest, in terms of model specification and identification of omitted variables, identification of the proper functional form of the relationship between causal and dependent variables, and refinements of measurements of specific variables. The regression diagnostics associated with our first model indicate that the usual quantitative assessments of regression models are useful first steps to take in evaluating the goodness of fit of particular models, but that they should not be taken as the last word on the subject. Equally, and perhaps more, important is knowledge of the data itself so that a more intimate investigation of cases can be carried out for the dataset in question. In our analysis, we have focused on a dataset that we created ourselves on the basis of new field research. But the lesson of the analysis is applicable more widely since all regression analyses are likely to have some cases that are associated with extreme residuals and inspection of which can potentially yield rich analytical and theoretical payoffs.

Further, we have concentrated our attention in this paper on common property resources and studies of the common that try to posit causal mechanisms that may be common across many different situations. In part we were motivated to focus on the commons because of our own training and familiarity with the literature on the subject. But another reason for our focus is also the emerging quantitatively oriented scholarship in relation to the commons that uses large amounts of data and sophisticated statistical techniques to advance theory building and validation. However, we also believe that our general argument has nothing that restricts its relevance to the scholarship on the commons, and that it is applicable far more broadly.

Acknowledgments: We would like to thank our field research team for the immense effort and time they invested in the data collection for the research project on which the paper is based. Special thanks are due to Satya Prasanna, Akshay Jasrotia, and Vishaal Sharma. High levels of cooperation in the villages we visited made the field research by the team so successful. We appreciate the difficulties and inconvenience to which our respondents were put many times by our visits. Comments, suggestions, and constructive criticisms from Giacomo Chiozza are gratefully acknowledged. Finally, we are grateful for the research support provided by the National Science Foundation under its grant

#SBR 9905443, and to the Ford Foundation for support to the International Forestry Resources and Institutions Program through its grant #950-1160-2.

References:

Agrawal, Arun. 2001. Common Property Institutions and Sustainable Governance of Resources. *World Development* 29(10): 1649-72.

Agrawal, Arun and Ashwini Chhatre. Forthcoming. Explaining success on the commons: Community forest governance in the Indian Himalaya. *World Development*.

Agrawal, Arun and Krishna Gupta. 2005. Decentralization and participation: The governance of common pool resources in Nepal's Terai. *World Development* 33(7): 1101-14

Alexander, Paul. 1982. *Sri Lankan Fishermen: Rural Capitalism and Peasant Society*. Canberra: Australian National University.

Andrews, David F., and Daryl Pregibon. 1978. Finding the outliers that matter. *Journal of the Royal Statistical Society B* 40(1): 85-93.

Angelsen A. and D. Kaimowitz. 1999. Rethinking the causes of deforestation: Lessons from economic models. *World Bank Research Observer* 14(1): 73-98.

Armitage, D. 2002. Socio-institutional dynamics and the political ecology of mangrove forest conservation in Central Sulawesi, Indonesia. *Global Environmental Change* 12(3): 203-17.

Baland, Jean-Marie, and Jean-Philippe Platteau. 1996. *Halting Degradation of Natural Resources: Is There a Role for Rural Communities?* Oxford: Clarendon Press.

Bardhan, Pranab, and Jeff Dayton-Johnson. 2002. Unequal irrigators: Heterogeneity and commons management in large-scale multivariate research. In *The Drama of the Commons*. Edited by Elinor Ostrom, Thomas Dietz, Nives Dolsak, Paul C. Stern, Susan Stonich, and Elke U. Weber. Washington DC: National Academy Press.

Bates, Robert, Avner Greif, Margaret Levi, Jean-Laurent Rosenthal, and Barry Weingast. 1998. *Analytical Narratives*. Princeton, NJ: Princeton University Press.

Berkes, Fikret, ed. 1989. *Common Property Resources: Ecology and Community-Based Sustainable Development*. London: Belhaven Press.

Berkes, F. 2004. Rethinking community-based conservation. *Conservation Biology* 18(3): 621-30.

Bollen, Kenneth A. and Robert W. Jackman. 1990. Regression diagnostics: An expository treatment of outliers and influential cases. In John Fox and J. Scott Long (eds). *Modern methods of data analysis.* Newbury Park, Calif.: Sage Publications.

Brown, K. and Pearce, D. (Eds). (1994). *The Causes of Tropical Deforestation: The Economic and Statistical Analysis of Factors Giving Rise to the Loss of the Tropical Forests*. London: UCL Press.

Bryman, Alan. 1984. The debate about quantitative and qualitative research: A question of method or epistemology? *The British Journal of Sociology* 35(1): 75-92.

Chatterjee, S., and F. Wiseman. 1983. Use of regression diagnostics in political science research. *American Journal of Political Science* 27: 601-13.

Cobb, John Chandler. 1926. Quantitative analysis and the evolution of economic science. *American Economic Review* 16(3): 426-33.

Cobb, John Chandler. 1927. Quantitative restating of sociological and economic problems. *The American Journal of* Sociology 32(6): 921-30.

Collins, Randall. 1984. Statistics versus words. *Sociological Theory* 2: 329-62.

Coppedge, Michael. 1999. Thickening thin concepts and theories: Combining large N and small in comparative politics. *Comparative Politics* 31(4, July): 465-76.

Cox, D. R. and E. J. Snell. 1968. A general definition of residuals. *Journal of the Royal Statistical Society B* 30(2): 248-75.

Cox, D. R. and E. J. Snell. 1971. On test statistics calculated from residuals. *Biometrika* 58(3): 589-94.

Dietz, T., Ostrom, E., and Stern, P. 2003. The Struggle to Govern the Commons. *Science* 302 (special issue, December 12): 1907-12.

Dion, Douglas. 1998. Evidence and inference in the comparative case study. *Comparative Politics* 30:127-45.

DOP (Department of Planning). 1997. *The Ninth Five-Year Plan for Himachal Pradesh, 1997-2002*. Department of Planning. Shimla: Government of Himachal Pradesh.

Ehrlich, Paul and Anne Ehrlich. 1991. *The Population Explosion*. New York: Touchstone, Simon and Schuster Inc.

Gerring, John. 2004. What is a case study and what is it good for? *American Political Science Review* 98(2, May): 341-54.

Gibson, Clark and Fabrice Lehoucq. 2003. The local politics of decentralized environmental policy in Guatemala. *Journal of Environment and Development* 12(1): 28-49.

Gibson, Clark, Margaret A. McKean, and Elinor Ostrom (eds). 2000. *People and Forests: Communities, Institutions, and Governance*. Cambridge: MIT Press.

Gibson, C. C., Williams, J., and Ostrom, E. (2005). Local enforcement and better forests. *World Development*. 33(2) (February): 273-284

Johnson, C. and Forsyth, T. 2002. In the eyes of the state: Negotiating a "rights-based approach" to forest conservation in Thailand. *World Development* 30(9, September): 1591-1605.

Kahn, J. R. and J. R. Udry. 1986. Marital coital frequency: Unnoticed outliers and unspecified interactions lead to erroneous conclusions. *American Sociological Review* 51: 734-37.

Kant, S. 2000. A dynamic approach to forest regimes in developing economies. *Ecological Economics* 32(2, February): 287-300.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.

Lise, W. 2000. Factors influencing people's participation in forest management in India. *Ecological Economics* 34(3, Sept.): 379-92.

McCay, Bonnie J., and James Acheson, eds. 1987. *The Question of the Commons: The Culture and Ecology of Communal Resources*. Tucson: University of Arizona Press.

McKean, Margaret. 1992. Success on the commons: A comparative examination of institutions for common property resource management. *Journal of Theoretical Politics* 4(3): 247-81.

Mahoney, James and Dietrich Rueschemeyer (eds). 2003. *Comparative Historical Analysis in the Social Sciences*. New York: Cambridge University Press.

NRC (National Research Council) 1986. *Proceedings of the Conference on Common Property Resource Management*. Washington DC: National Academy Press.

Ogburn, William. 1934. Limitations of statistics. *The American Journal of Sociology* 40(1): 12-20.

Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. New York: Cambridge University Press.

Ostrom, Elinor, Thomas Dietz, Nives Dolsak, Paul Stern, Susan Stonich, and Elke Weber (eds). 2002. Institutions for Managing the Commons. Washington DC: NAS Press

Pinkerton, Evelyn, ed. 1989. *Cooperative Management of Local Fisheries: New Directions for Improved Management and Community Development*. Vancouver: University of British Columbia Press.

Poteete, A. R. and Ostrom, E. (2004). Heterogeneity, group size and collective action: The role of institutions in forest management. *Development and Change* 35(3): 437-61.

Rocheleau, D. and Edmunds, D. (1997). Women, men, and trees: Gender, power, and property in forest and agrarian landscapes. *World Development* 25(8): 1351-71.

Rudel T.K., Bates, D., and Machinguiashi, R. 2002. A tropical forest transition: Agricultural change, out-migration, and secondary forests in the Ecuadorian Amazon. *Annals of the Association of American Geographers* 92(1): 87-102.

Sarin, M. 1995. Regenerating India's forests: Reconciling gender equity with joint forest management. *IDS Bulletin: Institute of Development Studies* 26(1): 83-91.

Southworth J., and Tucker, C. (2001). The influence of accessibility, local institutions, and socioeconomic factors on forest cover change in the mountains of western Honduras. *Mountain Research and Development* 21(3): 276-83.

Stevens, J. P. 1984. Outliers and influential data points in regression analysis. *Psychological Bulletin* 95: 334-44.

Tiffen, M., Mortimore, M., and Gichuki, F. (1994). *More People–Less Erosion: Environmental Recovery in Kenya*. London: John Wiley.

Varughese, George, and Elinor Ostrom. 2001. The Contested Role of Heterogeneity in Collective Action: Some Evidence from Community Forestry in Nepal. *World Development* 29(5) (May): 747-765.

Wade, Robert. 1994. *Village Republics: Economic Conditions for Collective Action in South India*. Oaklando: ICS Press.

Wickramasinghe, Anoja. 1997. Anthropogenic factors and forest management in Sri Lanka. *Applied Geography* 17(2): 87-110.

Table 1: Descriptive Statistics for Variables in Model 1 (n=205)

Variable Name	Mean	Standard Deviation	Min.	Max.	References
Dependent: Change in forest condition	2.97	0.837	1	5	
Biophysical Variables					
Number of fires	3.27	4.21	0	30	
Demographic Variables					Ehrlich and Ehrlich, 1990, Rudel et al., 2002; Tiffen et al., 1994.
Number of households	182.07	322.12	6	2,700	
Number of cattle	389.6	521.5	19	2,600	
Grazing by goats	0.74	1.51	0	10	
Cattle grazing	3.4	2.32	0	7.6	
Sociopolitical Variables					Agrawal and Gupta 2005; Johnson and Forsyth, 2002; Kant, 2000; Sarin, 1995; Varughese and Ostrom, 2001.
Education	13.74	3.30	5	18	
College	1.67	0.8	1	3	
Homogeneity	0.2	0.2	0	0.88	
Conflict resolution	96.59	NA	0	1	
Economic and Dependence Variables					Angelsen and Kaimowitz, 1999, Ehrlich and Ehrlich, 1990; Poteete and Ostrom, 2004; Southworth and Tucker, 2001; Wickramsinghe, 1997.
Subsistence use	1.93	0.74	1	3	
Utility	3.26	0.8	1	5	
Adverse effect	1.46	0.63	1	3	
Institutional Variables					Armitage, 2002; Baland and Platteau 1996, Berkes, 2004; Dietz et al., 2003; Gibson, McKean, and Ostrom 2000; Gibson, Williams, and Ostrom 2005; McKean, 1992; Ostrom, 1990.
Traditional institutions	26.83	NA	0	1	
Voluntary labor	143.09	348.69	0	2,000	
Illegal users	18.83	35.66	0	100	
External users	39.51	NA	0	1	
Encroachments	0.47	0.98	0	5.29	
Decentralized control	38.05	NA	0	1	
Sanctions	0.06	0.17	0	1.16	
Income from fines	11.71	NA	0	1	
Plantation	53.66	NA	0	1	

Table 2: Description of Variables

Variable	Description
Change in forest condition	Index measuring change in the condition of a forest over the last five years, ranging from 1 for "degraded significantly" to 5 for "improved significantly", based on responses by five categories of individuals in the village (upper and lower caste men and women, and forest department guards.
Number of fires	Number of fires in the forest in the last 10 years
Number of households	The total number of households in the village according to official records
Number of cattle	Total number of cattle owned by families residing in the village
Education	Index of the level of education amongst villagers, generated by adding literacy (1= less than 40%, to 3 = more than 65%), female literacy (1= less than 10% to 4= more than 40%), percent with 8 years of school education (1= less than 10%, 4= more than 40%), percent with upto 12 years of education (1= less than 5%, 4= more than 20%), and percent educated till college (1= less than 5%, 3= more than 10%)
College	Proportion of the village population educated till college (1= less than 5%, 3= more than 10%)
Homogeneity	Proportion of households with less than 0.4 hectares of agricultural land.
Conflict resolution	Dummy variable; 1 if intra-village conflicts around forests have been resolved
Traditional institutions	Dummy variable; 1 if traditional community institutions are present for non-forestry issues
Voluntary labor	Person-days of labor contributed towards forest management in the last five years
Subsistence use	Composition of forest in terms of tree species; 1= pure conifer, 2= mixed, 3= pure broadleaved
Utility	Index of six group responses on the level of utility of the forest for villagers. Ranges from 1 for very low utility to 5 for very high utility
Adverse effect	Measure of the impact of the loss of the forest on local livelihoods; 1= very adverse, 2= somewhat adverse, 3= no impact
Grazing by goats	Number of months of grazing in the forest by migratory herds of sheep and goats
Cattle grazing	Log of the number of cattle grazed in the forest
Illegal users	Proportion of users of the forest not entitled
External users	Dummy variable, 1= residents of other villages have entitlement to use the forest
Encroachments	Log of the number of encroachments in the forest for agriculture 30 years ago
Decentralized control	Dummy variable, 1= managed decisions partially or fully in hands of local users
Sanctions	The number of people fined in the last two years for rule infringement, divided by total number of households
Income from fines	Dummy variable, 1= income from fines is a significant part of revenues for local institution
Plantation	Dummy variable, 1= plantation activity carried out in the forest in the last ten years

Table 3: Regression Results for Model 1 (n=205)
(Dependent Variable=Change in Forest Condition)

Variable Name	Coefficient	Standard Error	P-value
Biophysical Variables			
Number of fires	-0.034	0.125	0.007***
Demographic Variables			
Number of households	-0.0005	0.00023	0.036**
Number of cattle	0.0006	0.0001	0.000***
Grazing by goats	-0.085	0.037	0.025**
Cattle grazing	-0.085	0.025	0.001***
Sociopolitical Variables			
Education	0.033	0.020	0.098*
College	-0.087	0.085	0.306
Homogeneity	0.791	0.252	0.002***
Conflict resolution	0.667	0.274	0.016**
Economic and Dependence Variables			
Subsistence use	0.119	0.068	0.083*
Utility	-0.034	0.060	0.569
Adverse effect	-0.077	0.085	0.368
Institutional Variables			
Traditional institutions	-0.157	0.119	0.19
Voluntary labor	0.00024	0.00015	0.131
Illegal users	-0.0027	0.0017	0.126
External users	-0.17	0.112	0.132
Encroachments	0.11	0.058	0.61
Decentralized control	0.298	0.126	0.019**
Sanctions	-0.378	0.279	0.177
Income from fines	-0.141	0.179	0.432
Plantation	0.217	0.098	0.029**
$F_{(21, 183)} = 7.27$; $R^2 = 0.4548$; Adj-$R^2 = 0.3922$; Root MSE = 0.6529			

Table 4: Over- and Under-predicted Observations Corresponding to Extreme Residuals

Name of village	**Observed value of dependent variable**	**Fitted value of dependent variable**	**Standardized residual**	**Studentized residual**
Overpredicted Values of Change in Forest Condition				
Heb	2	3.85	-2.95	-3.01
Thalli	1	2.63	-2.70	-2.74
Deol	1	2.43	-2.33	-2.36
Madaro	2	3.42	-2.15	-2.17
Chakmoh	1.66	3.01	-2.05	-2.07
Underpredicted Values of Change in Forest Condition				
Dhimkataru	5	3.72	1.93	1.94
Kanoh	4	2.69	2.13	2.16
Pangaon	4.66	3.27	2.21	2.24
Baggi	4	2.62	2.22	2.25
Bagotla	4.66	3.18	2.46	2.49

Table 5: Information on Relevant Variables for Cases Associated with Extreme Residuals

Name of village	**Type of institution***	**Number of fires**	**Education**	**Area of forest**	**Forest condition**
Overpredictions					
Heb	Self-initiated	3	18	9	3.33
Thalli	Demarcated Protected	0	15	25	1
Deol	Reserved	3	9	5	3
Madaro	Self-initiated	2	7	13	4
Chakmoh	Private	0	17	20	4.75
Underpredictions					
Dhimkataru	Self-initiated	2	15	10	4
Kanoh	Private	0	17	30	4.25
Pangaon	Self-initiated	0	17	35	4.75
Baggi	Demarcated protected	0	11	50	4.66
Bagotla	Self-initiated	0	17	24	4.25

* Demarcated Protected Forest is one of the four classes of public forests in our sample. Others include Reserved Forests, Undemarcated Protected Forests, and Revenue Department Forests. Similarly, classes of decentralized forests include Private, Self-initiated community managed, Forest cooperatives, and Soil conservation cooperatives.

Table 6: New Variables used in Model 2

Variable	**Mean**	**Standard deviation**	**Minimum**	**Maximum**
Forest Condition	3.34	0.87	1	5
Community managed	46.34	NA	0	1
Education squared	199.85	83.06	25	324
Fire intensity	2.01	0.96	1	4

Table 7: Regression Results for Model 2 (n=205)
(Dependent Variable=Change in Forest Condition)

Variable Name	Coefficient	Standard error	P-value	Robust Standard Errors	P-Value
Biophysical Variables					
Intensity of fires	-0.101	0.05	0.048**	0.051	0.053**
Forest Condition	0.304	0.056	0.000***	0.047	0.000***
Demographic Variables					
Number of households	-0.00044	0.00022	0.049**	0.00014	0.003***
Number of cattle	0.00058	0.00012	0.000***	0.00007	0.000***
Cattle grazing	-0.072	0.023	0.002***	0.021	0.001***
Grazing by goats	-0.081	0.035	0.021**	0.031	0.012**
Sociopolitical Variables					
Education	0.183	0.104	0.08*	0.103	0.08*
Education squared	-0.0068	0.0041	0.101	0.0041	0.098*
Homogeneity	0.92	0.235	0.000***	0.247	0.000***
Conflict resolution	0.839	0.257	0.001***	0.236	0.001***
Economic Variables					
Subsistence use	0.111	0.063	0.083*	0.066	0.100*
Utility	-0.117	0.057	0.042**	0.056	0.041**
Adverse effect	-0.15	0.079	0.06*	0.086	0.086*
Institutional Variables					
Traditional institutions	-0.227	0.107	0.036**	0.094	0.018**
Voluntary labor	0.00041	0.00015	0.009***	0.00015	0.009***
Illegal users	-0.0042	0.0015	0.008**	0.0014	0.005***
External users	-0.177	0.105	0.095*	0.106	0.098*
Encroachments	0.13	0.054	0.018**	0.049	0.009***
Community control	0.33	0.122	0.008***	0.124	0.009***
Sanctions	-0.457	0.259	0.079*	0.224	0.045**
Income from fines	-0.461	0.172	0.008***	0.174	0.009***
Plantation	0.275	0.099	0.006***	0.098	0.006***
OLS model: $F_{(22, 182)} = 9.45$; $R^2 = 0.5332$; Adj-$R^2 = 0.4767$; Root MSE = 0.6058; Huber-White Estimator: $F_{(22, 102)} = 15.85$; $R^2 = 0.5332$; Root MSE = 0.6058					