

**Midterm: Answer Key****A) Firms and Bribes in Transition Countries**

- 1) Older firms are predicted to pay a smaller share of their revenue in informal payments. Specifically, an increase of a firm's age with 1 year is predicted to decrease the (reported) share of revenue payments to public officials with 0.028 percentage points, *ceteris paribus*. The coefficient is significantly different from 0 at the 1% significance level ( $t\text{-value} = -0.028/0.009 = 3.11$  which is bigger than 2.576, the critical value at 1% (with more than 120 degrees of freedom). Since we can reject the null-hypothesis at the 1%, we can automatically reject at the 5 and 10%.
- 2) 99% confidence interval for coefficient of firm size:  

$$[-0.005 - 2.576 * 0.001, -0.005 + 2.576 * 0.001] = [-0.00758, -0.00242]$$
- 3) Since column (1) and (2) show 2 non-nested models, we would use the adjusted R-squared to choose between the 2 specifications – in this case because (1) and (2) don't have a different number of variables, the R-squared gives you the same result. Given that the 2 models were estimated with practically the same variables, only a different functional form, there is no need to worry about introducing bias or the like (a priori it is not clear that the linearity assumption in the specification in column (1) is more reasonable than the one in column (2).
- 4) To evaluate whether the regressions are BLUE we want to consider whether the 5 Gauss-Markov assumptions are reasonable for this regression:

**Linearity assumption:** While we could potentially hypothesize that there might be decreasing effects of age and size, linearity can be argued to be reasonable as a first approximation. Moreover, the second regression allows for decreasing returns in size.

**Random sampling assumption:** Although the survey was conducted on a random sample of 4000 firms, the regression only shows 1921 observations (1915 in second column). This makes us suspicious, as it is highly likely that the reason for this drop in observations, is the unwillingness of many firms to answer the question about informal payments. Given that those that don't want to answer is probably not a random sample, the remaining sample is probably not random.

**Zero Conditional Mean Assumption:** One can imagine many variables that might affect both the independent and the dependent variables. E.g. informal payments to officials might be more common in some sectors/countries than in other, and these sectors/countries might also happen to have bigger firms, or more firms with foreign capital, etc

**No Perfect Collinearity Assumption:** None of the independent variables is a linear combination of the other. Indeed, given that an estimation of the model was obtained, we know there can't have been perfect collinearity.

Homoscedasticity assumption: One might doubt whether the variance of the error terms is constant for different levels of age. It is not impossible to imagine that all young firms might report rather similar shares of informal payment, but there is more variation for older firms (e.g. if some of them manage to establish a relationship with the officials that allows them to avoid payments, and others not...).

In conclusion, the random sampling, and the zero conditional mean assumption are likely to be violated in these regressions, and possibly, the homoscedasticity assumption is too. Hence the regressions are not BLUE.

NOTE:

- 1) The linearity assumption is an assumption. You need to argue (even if shortly) why it is ok to assume that these variables affect the dependent variable in a linear way.
- 2) Always pay attention to the number of observations!
- 3) When arguing why there might be a violation of the zero conditional mean assumption because of an omitted variable, you need to discuss an example of something that might be in the error term that is correlated with both the dependent and the independent variable.

- 5) No, since we just argued (in Q 4) that we have many reasons to believe these estimates are biased, we cannot conclude anything from them.
- 6) That would be a model with an interaction term:

$$pay = \beta_0 + \beta_1 age + \beta_2 size + \beta_3 sharefor + \beta_4 sharestate + \beta_5 (sharefor * size) + u$$

(with the definitions of the variables as in the midterm). You should keep all the original variables in the model (otherwise you make the violation of ass. 3 even worse), or if you don't think that is a good idea, argue why you left them out.

## B) Explaining Conflict in Nepal

- 1) For each additional year of average life-expectancy, the district-level number of conflict related incidents is predicted to decrease with 4%, ceteris paribus. The coefficient is significantly different from 0 at the 5% level (or to be precise at the 3.6% significance level (see p-value in stata output).
- 2) Given that beta2 is the coefficient of the square term of slope, and slope itself is also in the regression, there is no ceteris paribus interpretation. Instead we can calculate the derivative at each level of the independent variable.

$$\frac{d \log(nrinc)}{d(slope)} = \hat{\beta}_1 + 2 * \hat{\beta}_2 slope = .02355 - .0006 * slope$$

Given that the coefficient of *slope* is positive and the one of *slope*<sup>2</sup> is negative, we know that for districts with relatively small shares of territory with more than 30% slope, the number of conflict-related incidents will increase as the terrain becomes more mountainous, but at a decreasing rate. In fact the negative coefficient on the square term means that after beyond a certain point, the relationship between slope and incidents becomes negative (this

is intuitive – insurgencies might benefit from mountainous terrain (more easy to move unnoticed), but only upto a certain level (if the terrain becomes too mountainous, the difficulties in moving around will become too large – moreover, the population will become scarce and hence the likelihood of incidents will automatically decrease). Note however that the coefficient of the square term is only significant at the 10% level – hence we do not have strong evidence to conclude that such a quadratic relationship might exist.

- 3)  $H_0: \beta_4 = 0, \beta_5 = 0, \beta_6 = 0,$   
 $H_1: H_0$  does not hold

$$F \equiv \frac{(R_{ur}^2 - R_r^2) / q}{(1 - R_{ur}^2) / (n - k - 1)} = \frac{(0.1552 - 0.0608) / 3}{(1 - 0.1552) / 67} = \text{approximately } 2.49.$$

The critical value for an F-test at the 5% significance level with  $q=3$  and  $(n-k-1)=67$  is approximately 2.74, and approximately 2.17 at the 10% significance level. Hence we cannot reject the null-hypothesis at the 5% level, but we can reject at the 10% level. The variables capturing the level of development are jointly significant at the 10% level.

- 4) False. We can never accept hypotheses in econometrics
- 5) There can be different valid opinions. Some people might think it is more sensible to assume increasing returns in this regression (e.g. effect of low levels of development will have a bigger effect at very low levels of development), other might argue that the interpretation in % is not very intuitive when we talk about the number of conflict-related incidents. One of the big advantages of the log in this case is that it reduces the impact of potential outliers (a few district with very high number of incidents (e.g. in the Maoist heartland) that might otherwise drive the regression outcomes (given that we have a relatively small number of observations). It also makes the homoscedasticity assumption more reasonable.
- 6) a) Because this is a multivariate regression, we cannot be sure about the direction of the bias. We can however make an educated guess based on what we would expect to be the bias in the bivariate case: given that we expect a positive correlation between income and life expectancy, and given that life-expectancy is negatively correlated with the dependent variable, we expect the coefficient of income to be negatively biased when we exclude life expectancy. That implies the coefficient would be more negative.
- b) Omitting the life expectancy variables might have 2 offsetting effects on the precision of the coefficient estimate of income. Given that we might expect district-level average life expectancy (as a measure of welfare/health) to be positively correlated with district-level income per capita, we might suspect there is a lot of multicollinearity in the regression when both are included. Removing the multicollinearity by taking life expectancy out of the model might hence increase the precision of income per capita. On the other hand, it will increase the noise, which has an offsetting effect on the variance of the estimate of the coefficient. The net effect is unclear.

You can also see this by looking at the formula:

$$se(\hat{\beta}_j) = \frac{\hat{\sigma}}{[(1 - R_j^2) \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2]^{1/2}}$$

Excluding the life expectancy variable will increase sigma-hat, but decrease  $R_j^2$ .

- 7) We might have theoretical reasons to believe that inequality might help to explain conflict related incidents. Nevertheless, including a variable measuring inequality in 2004 is not a good idea because it would lead to a likely violation of the zero conditional mean assumption. In particular, this is a case of potential reverse causality. Inequality might have increased (or decreased) because of the conflict (that started in 1996), and hence including this variable in the model would cause a bias.

NOTE: having independent variables that measure outcomes in different years itself is not a problem. The problem arises if the independent variable is from the same year as the dependent variable and there might be reasons to believe reverse causality (or simultaneity – i.e. a third factor that might affect both).

- 8) 0.2193 (this is the P-value related to the F-test for overall significance of the model for the second regression, i.e. the regression that only includes the geography and forest variables).
- 9) No – given that we only have district level data for Nepal (and we might even worry about whether we counted appropriately for natural resources in this model), we definitely cannot conclude anything about the role of natural resources in affecting the likelihood of civil conflict in other countries. (The regression tells us something about the intensity of the conflict in different locations in Nepal, it does not tell us anything about the likelihood of civil conflict per se).

### C) Your Own Hypothesis.

1) Obviously, many different answers were acceptable here. Note however that you were asked to formulate a hypothesis. Many of you instead told me they wanted to analyze the (many different) determinants of a certain outcome.

2) In specifying a model, don't forget to add the error term.

In motivating why you think your model would give you unbiased estimates, you need to argue why it is reasonable to assume each of the 4 assumptions for your model (linearity, random sampling, zero conditional mean assumption, no perfect collinearity). In doing so, you needed to say a bit about what type of data you envisioned having (in particular to motivate the random sampling assumption, and the general model specification and interpretation.). The most tricky assumption is usually the zero conditional mean assumption. Not only should you discuss why certain control variables might make that assumption more reasonable, but you should also discuss any potential remaining doubts on that assumption (and recognize that, if existing, these might imply a biased estimate).

Homoscedasticity is not an assumption we need for unbiasedness.

Note also that some of you wanted to test a model with time-series data. Since we hadn't covered this yet, I ignored the time-series specific issues, but, for future purposes, pay attention in the time-series class to learn about the more stringent assumptions you need for such data

3) You were asked one hypothesis. This needed to be consistent with your description in point 1. If you want to have more than one hypothesis, make sure to be very clear about what you test, what's the alternative hypothesis, etc... Very often it is hard enough to test one hypothesis in a satisfactory way, yet alone multiple hypotheses at once.

4) Make sure your test allows you to test specifically the hypothesis you formulate in point 3 (and described in point 1). You were also asked to describe how you would do the test – which implies explaining which restricted model you would estimate if you wanted to do an F-test, etc.