Econ Thesis Data Analysis

Yeshwant Chillakuru

# I. Exploratory Data Analysis

colnames(df)

[1] "id" "state"  
[3] "district" "village"  
[5] "year" "roadPaved"  
[7] "distanceToPavedRoad" "roadMonsoonUsability"  
[9] "yearsWithPavedRoad" "ImmunizationCampaignsNumber" [11] "distanceToNearestTown" "distanceToNearestDistrictHQ" [13] "area" "populationCategories"  
[15] "householdNumbersOf" "mbCataract"  
[17] "mbTuberculosis" "mbHighBP"  
[19] "mbHeartDisease" "mbDiabetes"  
[21] "mbLeprosy" "mbCancer"  
[23] "mbAsthma" "mbPolio"  
[25] "mbParalysis" "mbEpilepsy"  
[27] "mbMentalIllness" "mbSTDorAIDS"  
[29] "mbOtherLongTerm" "mbDaysIncapacitated"  
[31] "mbTreatmentRecieved" "mbTreatmentWho1"  
[33] "mbTreatmentWhere1" "mbTreatmentWho2"  
[35] "mbTreatmentWhere2" "mbDaysHospitalized"  
[37] "mbCostOfTreatment" "mbCostIncludesMedication"  
[39] "mbCostOfMeds" "mbTravelExpenses"  
[41] "smokeTobacco" "chewTobacco"  
[43] "drinkAlcohol" "attendedSchool"  
[45] "gradeCompleted" "readingScore"  
[47] "mathScore" "writingScore"  
[49] "urban" "income"  
[51] "caste" "yearsLivingHere"  
[53] "ownToilet" "electricity"  
[55] "electricityHours" "drinkingWaterSource"  
[57] "healthSubCenter" "primaryHealthCenter"  
[59] "communityHealthCenter" "districtHospital"  
[61] "govtMaternityCenter" "pvtClinicTrainedDoc"  
[63] "pvtClinicUntrainedDoc" "pvtHospital"  
[65] "pvtPharm" "otherGovtMedFacility"  
[67] "distHealthSubCenter" "distPrimaryHealthCenter"  
[69] "distCommunityHealthCenter" "distDistrictHospital"  
[71] "distGovtMaternityCenter" "distPvtClinicTrainedDoc"  
[73] "distPvtClinicUntrainedDoc" "distPvtHospital"  
[75] "distPvtPharm" "distOtherGovtMedFacility"

# Conduct Diff of Differences analysis

Between 2005 and 2012, \* 344 got their first paved road(s). \* 950 saw no change. \* 47 lost their paved road(s).

# Output Summary Stats for Road and Morbidity

## distanceToPavedRoad roadPaved   
## Min. :-3.000 Min. :0.0000   
## 1st Qu.: 0.000 1st Qu.:0.0000   
## Median : 0.000 Median :1.0000   
## Mean : 1.559 Mean :0.6476   
## 3rd Qu.: 2.000 3rd Qu.:1.0000   
## Max. :50.000 Max. :1.0000

## distanceToPavedRoad roadPaved   
## Min. : 0.0000 Min. :0.000   
## 1st Qu.: 0.0000 1st Qu.:1.000   
## Median : 0.0000 Median :1.000   
## Mean : 0.5131 Mean :0.869   
## 3rd Qu.: 0.0000 3rd Qu.:1.000   
## Max. :35.0000 Max. :1.000   
## NA's :43 NA's :2

## mbCataract mbTuberculosis mbHighBP mbHeartDisease   
## Min. :0.00000 Min. :0.000000 Min. :0.00000 Min. :0.000000   
## 1st Qu.:0.00000 1st Qu.:0.000000 1st Qu.:0.00000 1st Qu.:0.000000   
## Median :0.00000 Median :0.000000 Median :0.00000 Median :0.000000   
## Mean :0.00860 Mean :0.004073 Mean :0.01215 Mean :0.004438   
## 3rd Qu.:0.01053 3rd Qu.:0.000000 3rd Qu.:0.01504 3rd Qu.:0.006211   
## Max. :0.18182 Max. :0.119048 Max. :1.00000 Max. :0.111111   
## mbDiabetes mbLeprosy mbCancer   
## Min. :0.000000 Min. :0.0000000 Min. :0.0000000   
## 1st Qu.:0.000000 1st Qu.:0.0000000 1st Qu.:0.0000000   
## Median :0.000000 Median :0.0000000 Median :0.0000000   
## Mean :0.006537 Mean :0.0006073 Mean :0.0007687   
## 3rd Qu.:0.007194 3rd Qu.:0.0000000 3rd Qu.:0.0000000   
## Max. :0.250000 Max. :0.0769231 Max. :0.0952381   
## mbAsthma mbPolio mbParalysis   
## Min. :0.000000 Min. :0.00000 Min. :0.000000   
## 1st Qu.:0.000000 1st Qu.:0.00000 1st Qu.:0.000000   
## Median :0.000000 Median :0.00000 Median :0.000000   
## Mean :0.006618 Mean :0.00129 Mean :0.001715   
## 3rd Qu.:0.010101 3rd Qu.:0.00000 3rd Qu.:0.000000   
## Max. :0.131579 Max. :0.04286 Max. :0.075949   
## mbEpilepsy mbMentalIllness mbSTDorAIDS   
## Min. :0.000000 Min. :0.000000 Min. :0.0000000   
## 1st Qu.:0.000000 1st Qu.:0.000000 1st Qu.:0.0000000   
## Median :0.000000 Median :0.000000 Median :0.0000000   
## Mean :0.001282 Mean :0.001613 Mean :0.0007416   
## 3rd Qu.:0.000000 3rd Qu.:0.000000 3rd Qu.:0.0000000   
## Max. :0.100000 Max. :0.222222 Max. :0.1052632   
## mbOtherLongTerm mbDaysIncapacitated  
## Min. :0.00000 Min. : 0.0000   
## 1st Qu.:0.00000 1st Qu.: 0.2857   
## Median :0.01333 Median : 1.6923   
## Mean :0.02273 Mean : 3.5212   
## 3rd Qu.:0.03488 3rd Qu.: 4.7473   
## Max. :0.20000 Max. :61.0870

## mbCataract mbTuberculosis mbHighBP mbHeartDisease   
## Min. :0.00000 Min. :0.000000 Min. :0.00000 Min. :0.00000   
## 1st Qu.:0.00000 1st Qu.:0.000000 1st Qu.:0.00000 1st Qu.:0.00000   
## Median :0.00000 Median :0.000000 Median :0.02985 Median :0.00000   
## Mean :0.02209 Mean :0.006872 Mean :0.05048 Mean :0.01332   
## 3rd Qu.:0.03247 3rd Qu.:0.009950 3rd Qu.:0.07292 3rd Qu.:0.01942   
## Max. :0.35484 Max. :0.173913 Max. :0.58065 Max. :0.23077   
## mbDiabetes mbLeprosy mbCancer mbAsthma   
## Min. :0.00000 Min. :0.000000 Min. :0.000000 Min. :0.00000   
## 1st Qu.:0.00000 1st Qu.:0.000000 1st Qu.:0.000000 1st Qu.:0.00000   
## Median :0.00000 Median :0.000000 Median :0.000000 Median :0.01020   
## Mean :0.02412 Mean :0.001184 Mean :0.001256 Mean :0.02042   
## 3rd Qu.:0.03030 3rd Qu.:0.000000 3rd Qu.:0.000000 3rd Qu.:0.03200   
## Max. :0.51613 Max. :0.125000 Max. :0.125000 Max. :0.50000   
## mbPolio mbParalysis mbEpilepsy   
## Min. :0.000000 Min. :0.000000 Min. :0.000000   
## 1st Qu.:0.000000 1st Qu.:0.000000 1st Qu.:0.000000   
## Median :0.000000 Median :0.000000 Median :0.000000   
## Mean :0.002048 Mean :0.007779 Mean :0.004689   
## 3rd Qu.:0.000000 3rd Qu.:0.000000 3rd Qu.:0.000000   
## Max. :0.084507 Max. :0.406780 Max. :0.250000   
## mbMentalIllness mbSTDorAIDS mbOtherLongTerm   
## Min. :0.000000 Min. :0.0000000 Min. :0.00000   
## 1st Qu.:0.000000 1st Qu.:0.0000000 1st Qu.:0.02247   
## Median :0.000000 Median :0.0000000 Median :0.06250   
## Mean :0.006675 Mean :0.0007053 Mean :0.08091   
## 3rd Qu.:0.000000 3rd Qu.:0.0000000 3rd Qu.:0.11538   
## Max. :0.285714 Max. :0.1111111 Max. :0.64000   
## mbDaysIncapacitated  
## Min. : 0.0000   
## 1st Qu.: 0.9375   
## Median : 2.9118   
## Mean : 4.4673   
## 3rd Qu.: 6.3542   
## Max. :75.0000

|  |
| --- |
| # II. Regressions |
| ## 1. Univariate models |
| ## Warning in rm(lm.mb1, lm.mb2, lm.mb3): object 'lm.mb3' not found |
| ## ## Calls: ## MB 1: lm(formula = mbDaysIncapacitated ~ roadPaved, data = df) ## MB 2: lm(formula = mbDaysIncapacitated ~ distanceToPavedRoad, data = df) ## ## ============================================= ## MB 1 MB 2 ## --------------------------------------------- ## (Intercept) 3.359\*\*\* 4.061\*\*\* ## (0.197) (0.105) ## roadPaved 0.812\*\*\* ## (0.227) ## distanceToPavedRoad -0.085\*\* ## (0.029) ## --------------------------------------------- ## R-squared 0.0 0.0 ## adj. R-squared 0.0 0.0 ## F 12.8 8.3 ## p 0.0 0.0 ## N 2844 2759 ## ============================================= |
| ## Warning in rm(mtableMulti): object 'mtableMulti' not found |
| ## 2. Multivariate Models ### Independent: Distance To Paved Road |
| ## ## Calls: ## MB 1: lm(formula = mbDaysIncapacitated ~ distanceToPavedRoad + electricity + ## income, data = df) ## MB 2: lm(formula = mbDaysIncapacitated ~ distanceToPavedRoad + distanceToNearestTown + ## writingScore + smokeTobacco, data = df) ## MB 3: lm(formula = mbDaysIncapacitated ~ distanceToPavedRoad + electricity + ## income + distanceToNearestTown + writingScore + smokeTobacco, ## data = df) ## ## ========================================================== ## MB 1 MB 2 MB 3 ## ---------------------------------------------------------- ## (Intercept) 2.684\*\*\* 2.940\*\*\* 2.322\*\*\* ## (0.278) (0.288) (0.354) ## distanceToPavedRoad -0.054 -0.050 -0.039 ## (0.030) (0.030) (0.030) ## electricity 1.370\*\*\* 0.968\*\* ## (0.353) (0.339) ## income 0.000\* 0.000 ## (0.000) (0.000) ## distanceToNearestTown -0.011 -0.009 ## (0.009) (0.009) ## writingScore 0.605\*\* 0.357 ## (0.209) (0.224) ## smokeTobacco 0.393\*\*\* 0.330\*\* ## (0.114) (0.118) ## ---------------------------------------------------------- ## R-squared 0.0 0.0 0.0 ## adj. R-squared 0.0 0.0 0.0 ## F 13.4 9.2 8.3 ## p 0.0 0.0 0.0 ## N 2759 2545 2545 ## ========================================================== |
| ### Independent: Road Paved |
| ## ## Calls: ## MB 1: lm(formula = mbDaysIncapacitated ~ roadPaved + electricity + ## income, data = df) ## MB 2: lm(formula = mbDaysIncapacitated ~ roadPaved + distanceToNearestTown + ## writingScore + smokeTobacco, data = df) ## MB 3: lm(formula = mbDaysIncapacitated ~ roadPaved + electricity + ## income + distanceToNearestTown + writingScore + smokeTobacco, ## data = df) ## ## ========================================================== ## MB 1 MB 2 MB 3 ## ---------------------------------------------------------- ## (Intercept) 2.303\*\*\* 2.393\*\*\* 1.924\*\*\* ## (0.286) (0.316) (0.362) ## roadPaved 0.467\* 0.772\*\*\* 0.631\*\* ## (0.236) (0.225) (0.229) ## electricity 1.411\*\*\* 0.905\*\* ## (0.347) (0.337) ## income 0.000\* 0.000 ## (0.000) (0.000) ## distanceToNearestTown -0.010 -0.008 ## (0.009) (0.009) ## writingScore 0.517\* 0.316 ## (0.207) (0.220) ## smokeTobacco 0.371\*\* 0.327\*\* ## (0.113) (0.117) ## ---------------------------------------------------------- ## R-squared 0.0 0.0 0.0 ## adj. R-squared 0.0 0.0 0.0 ## F 14.0 11.1 9.1 ## p 0.0 0.0 0.0 ## N 2844 2586 2586 ## ========================================================== |
| ## 3. Fixed Effects Regression Model |
| ## This series is constant and has been removed: districtHospital |
| ## Oneway (individual) effect Within Model ## ## Call: ## plm(formula = mbDaysIncapacitated ~ distanceToPavedRoad + electricity + ## income + distanceToNearestTown + writingScore + smokeTobacco, ## data = df, model = "within", index = c("id", "year")) ## ## Unbalanced Panel: n=1435, T=1-2, N=2545 ## ## Residuals : ## Min. 1st Qu. Median 3rd Qu. Max. ## -17.60 -1.09 0.00 1.09 17.60 ## ## Coefficients : ## Estimate Std. Error t-value Pr(>|t|) ## distanceToPavedRoad 1.2919e-02 4.4828e-02 0.2882 0.77325 ## electricity 8.1819e-01 7.4260e-01 1.1018 0.27079 ## income 2.3872e-06 2.9222e-06 0.8169 0.41415 ## distanceToNearestTown -2.5526e-02 1.5371e-02 -1.6606 0.09708 . ## writingScore 4.5683e-01 3.3735e-01 1.3542 0.17596 ## smokeTobacco 4.9458e-01 2.2414e-01 2.2066 0.02755 \* ## --- ## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1 ## ## Total Sum of Squares: 22028 ## Residual Sum of Squares: 21344 ## R-Squared: 0.031054 ## Adj. R-Squared: -1.2328 ## F-statistic: 5.89702 on 6 and 1104 DF, p-value: 4.5265e-06 |
| ## This series is constant and has been removed: districtHospital |
| ## Oneway (individual) effect Within Model ## ## Call: ## plm(formula = mbDaysIncapacitated ~ roadPaved + electricity + ## income + distanceToNearestTown + writingScore + smokeTobacco, ## data = df, model = "within", index = c("id", "year")) ## ## Unbalanced Panel: n=1438, T=1-2, N=2586 ## ## Residuals : ## Min. 1st Qu. Median 3rd Qu. Max. ## -17.50 -1.13 0.00 1.13 17.50 ## ## Coefficients : ## Estimate Std. Error t-value Pr(>|t|) ## roadPaved 5.0946e-01 3.7383e-01 1.3628 0.17322 ## electricity 8.5687e-01 7.3591e-01 1.1644 0.24452 ## income 1.7142e-06 2.8916e-06 0.5928 0.55342 ## distanceToNearestTown -2.6504e-02 1.5179e-02 -1.7461 0.08106 . ## writingScore 4.5573e-01 3.2852e-01 1.3872 0.16565 ## smokeTobacco 4.0945e-01 2.2104e-01 1.8524 0.06423 . ## --- ## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1 ## ## Total Sum of Squares: 22784 ## Residual Sum of Squares: 22072 ## R-Squared: 0.031255 ## Adj. R-Squared: -1.1928 ## F-statistic: 6.14072 on 6 and 1142 DF, p-value: 2.3825e-06 |
| ## ## F test for individual effects ## ## data: mbDaysIncapacitated ~ distanceToPavedRoad + electricity + income + ... ## F = 1.3066, df1 = 1434, df2 = 1104, p-value = 1.395e-06 ## alternative hypothesis: significant effects |
| ## ## F test for individual effects ## ## data: mbDaysIncapacitated ~ roadPaved + electricity + income + distanceToNearestTown + ... ## F = 1.3026, df1 = 1437, df2 = 1142, p-value = 1.418e-06 ## alternative hypothesis: significant effects |

# Regression Sandbox

# To Do:

* Natural Experiment of Rural Road Expansion Program --> Evan and Owens --> COPS program
  + **Can differences in differences method be used?**
  + Use roadPaved as Treatment
* Granger Test for determining simultaneous causality?