Econometrics Project

AADIT KANT JHA (202001)
ABHIK S BASU (2020165)
MADHAVA KRISHNA (2020217)
MOHIT JAIN (2020221)
RITWICK PAL (2020461)

Project Analysis

KEY FINDINGS

First, we would be discussing our findings that we had made through the data assignment.

CREATIVE COMPONENT

Second, we elucidate more on what extra can be done by our group on the given objective.

CONCLUSION

We move on to concluding remarks, and the questions and answer session

Key findings will be used directly in deriving our creative component part of the project.

Objective

- Measles is an airborne disease spread through the air, saliva, touching contaminated surfaces, skin-to skin contact and from mother to child through pregnancy, labour or nursing.

- The objective of this assignment is to study the percentages of children with Measles from age 0 to 5, and to understand how this quantity varies over various parameters related to agricultural sector.

Spread of Measles

- Measles is a highly contagious virus that lives in the nose and throat mucus of an infected person. It can spread to others through coughing and sneezing. If other people breathe the contaminated air or touch the infected surface, then touch their eyes, noses, or mouths, they can become infected

- MMR vaccine in a two dose schedule has successfully eliminated measles, mumps and rubella from many developed countries. In India, it is not a part of national immunization programme but is included in the State immunization programme of Delhi as a single dose between 15-18 months.

Regression Model

The initial model included a lot of regressors, primarily focused on mothers' healthcare. **v1** (pregnant women registered for ANC), v4(pregnant women receiving the second dose of tetanus-toxoid vaccine or booster), **v5** (pregnant women received 100 IFA tablets), **v8** (home deliveries), **v12** (percentage of women discharged in less than 48 hours of delivery), **v16** (percentage of safe deliveries), **v20** (percentage of women received postpartum checkup within 48 hours of delivery), **v23** (percentage of total reported live births), v26 (newborns having weight less than 2.5 kg), **v29** (newborns breastfed within 1 hour), v34 (fully immunised children in the age gruop of 9 to 11 years), **v35** (percentage of children with polio in the age group of 0 to 5), **v37** (percentage of children with diarrhoea and dehydration in the age group of 0 to 5 years), **v38** (percentage of children with Malaria in the age gruop of 0 to 5 years), **v39** (infant deaths reported), index (yield index), gdp (GDP of the state), beds (hospital bed availability), and tap (percentage of households fully connected with tap water supply). However...

Regression Model

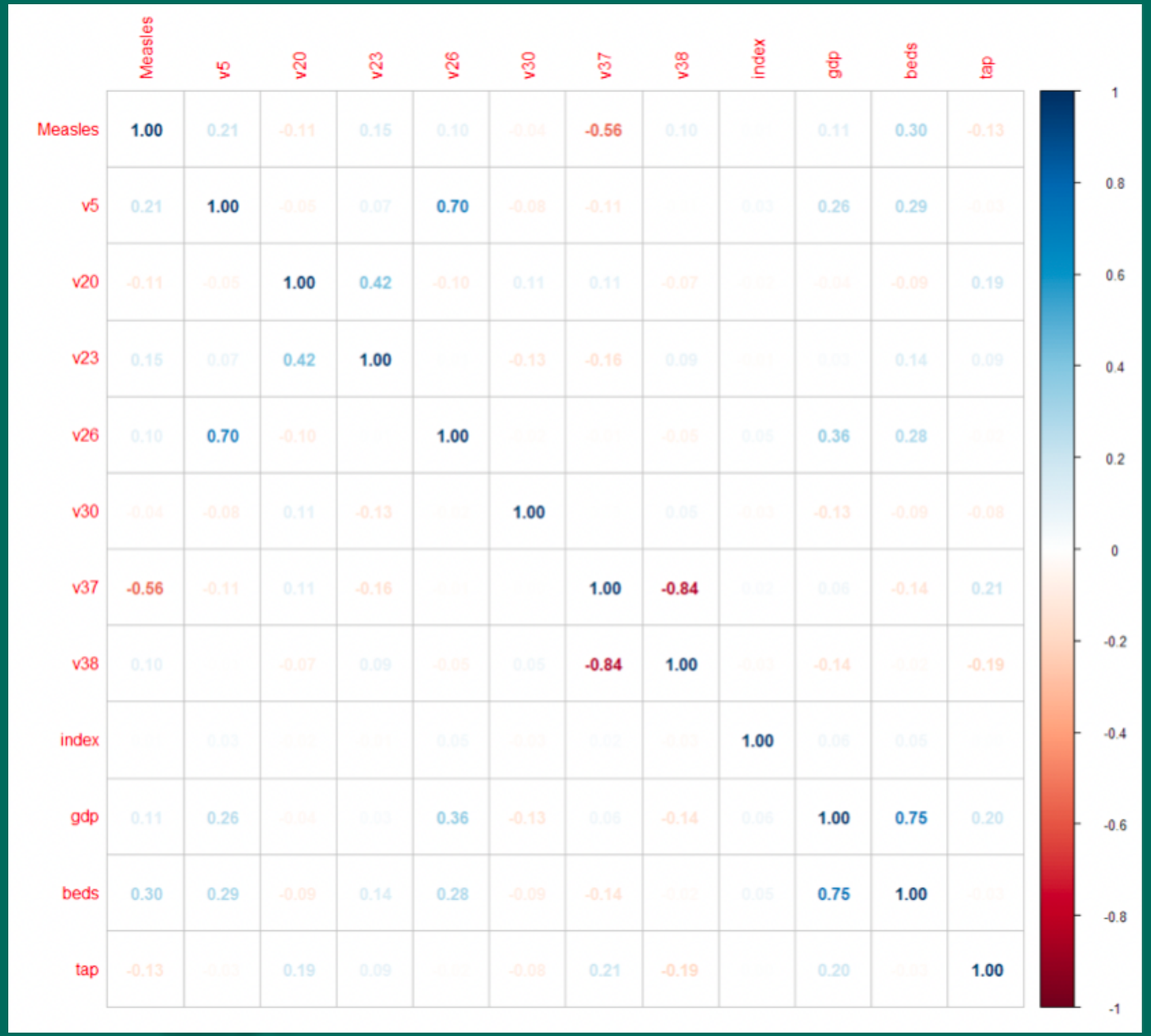
On grounds of correlation and after verifying that the chosen variables are not multicollinear (using the VIF test), the regressors were reduced to the following:

v5: Pregnant women received 100 Iron and Folic Acid (IFA) tablets
v20: Percentage of women who received postpartum checkups within 48 hours of delivery (to total reported deliveries)
v23: Percentage of total reported live births (to total reported deliveries)
v26: Newborns weighed at birth
v30: Percentage of new borns breastfed within 1 hour (to total live births)
v37: Percentage of children with Diarrhoea and Dehydration in the age group of 0 to 5 years.
v38:
index: Yield Index
GDP: GDP of the state
beds: hospital bed availability in the state
tap: district wise tap water access (percentage of households)

Variables : Description and Summary

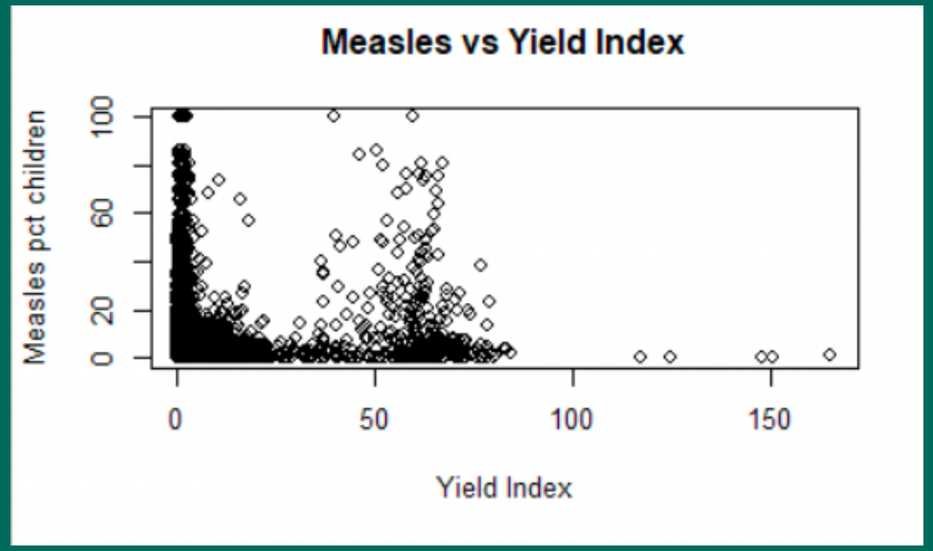
Using Data for 35 States and Union Territories for 2011-2019

Variable	Description	Mean	Median	Standard Deviation	Minimum	Maximum
v5	Pregnant women received 100 Iron and Folic Acid (IFA) tablets	33245	24697	34872.04	0	586631
v20	Percentage of women who received postpartum checkups within 48 hours of delivery (to total	284.5	80.2	959.3107	0	9836.4
v23	Percentage of total reported live births (to total reported	102.1	99.3	18.52958	39.1	512.7
v26	Newborns weighed at bir	29051	23887	23477.94	0	170700
v30	Percentage of new borns breastfed within 1 hour (to total live births)	88.24	92.8	13.00521	0	120.4
v37	Percentage of children with Diarrhoea and Dehydration in the age group of 0 to 5 years.	87.73	95.5	18.1507	0	100
v38	Percentage of children with Malaria in the age group of 0 to 5 years.	8.347	1.9	14.53038	0	100
index	Yield Index	5.7268	1.3523	15.15087	0	311
gdp	GDP of the state	56473411	48727384	42026165	397843	203331431
beds	Hospital bed availability in the state	70578	43064	72973.3	1294	205142
tap	District wise tap water access (percentage of households)	20.3	8.3	24.95925	0	100

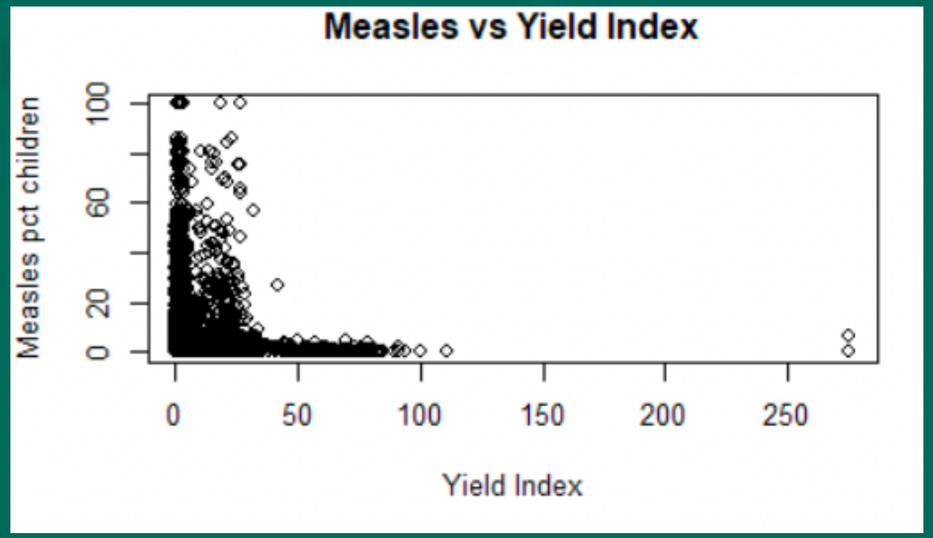

Measles v5 v20 v23 v26 v30 v37 v38 index gdp beds tap
Measles 1.00 0.21 -0.11 0.15 0.10 -0.04 -0.56 0.10 0.00 0.11 0.30 -0.13
v5 0.21 1.00 -0.05 0.07 0.70 -0.08 -0.11 0.07 0.03 0.26 0.29 -0.03
v20 -0.11 -0.05 1.00 0.42 -0.10 0.11 0.11 -0.07 0.02 -0.04 -0.09 0.19
v23 0.15 0.07 0.42 1.00 0.00 -0.13 -0.16 0.09 0.01 0.03 0.14 0.09
v26 0.10 0.70 -0.10 0.00 1.00 0.05 0.01 -0.05 0.05 0.36 0.28 -0.03
v30 -0.04 -0.08 0.11 -0.13 0.05 1.00 0.00 0.05 -0.03 -0.13 -0.09 -0.08
v37 -0.56 -0.11 0.11 -0.16 0.01 0.00 1.00 -0.84 0.02 0.06 -0.14 0.21
v38 0.10 0.07 -0.07 0.09 -0.05 0.05 -0.84 1.00 -0.03 -0.14 -0.02 -0.19
index 0.00 0.03 0.02 0.01 0.05 0.03 0.02 -0.03 1.00 0.06 0.05 0.00
gdp 0.11 0.26 -0.04 0.03 0.36 -0.13 0.06 -0.14 0.06 1.00 0.75 0.20
beds 0.30 0.29 -0.09 0.14 0.28 -0.09 -0.14 -0.02 0.05 0.75 1.00 -0.03
tap -0.13 -0.03 0.19 0.09 -0.03 -0.08 0.21 -0.19 0.00 0.20 -0.03 1.00
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Measles vs Yield Index

It is observed that at low yield indices and at intermediate yield indices, the Measles percentage is considerably high, while it is relatively low in between, and has a few outliers upwards of YI of 100, in case of Kharif. For the Rabi season, the regressand has very high values concentrated around lower values of yield index. In summary, both cases, Rabi and Kharif, depict a high concentration of Measles percentage for lower values of yield indices.



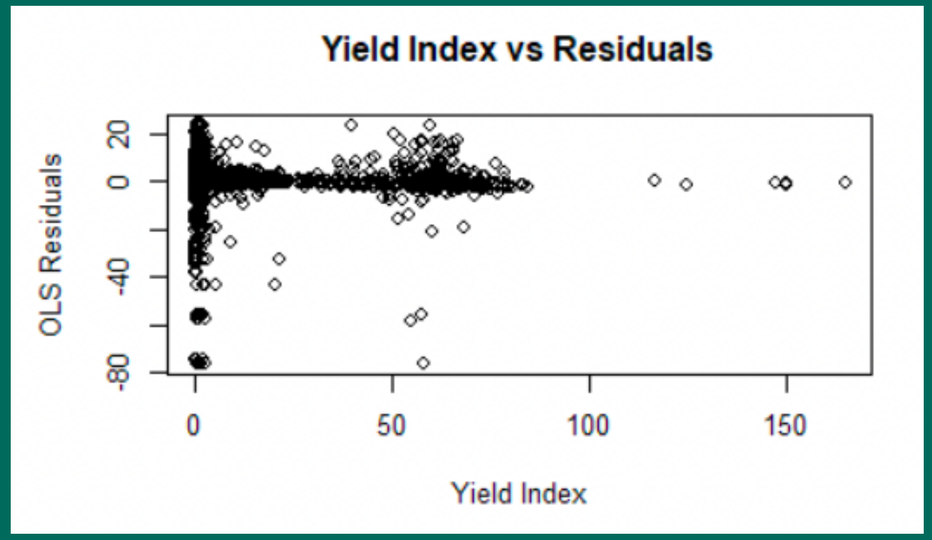
Kharif



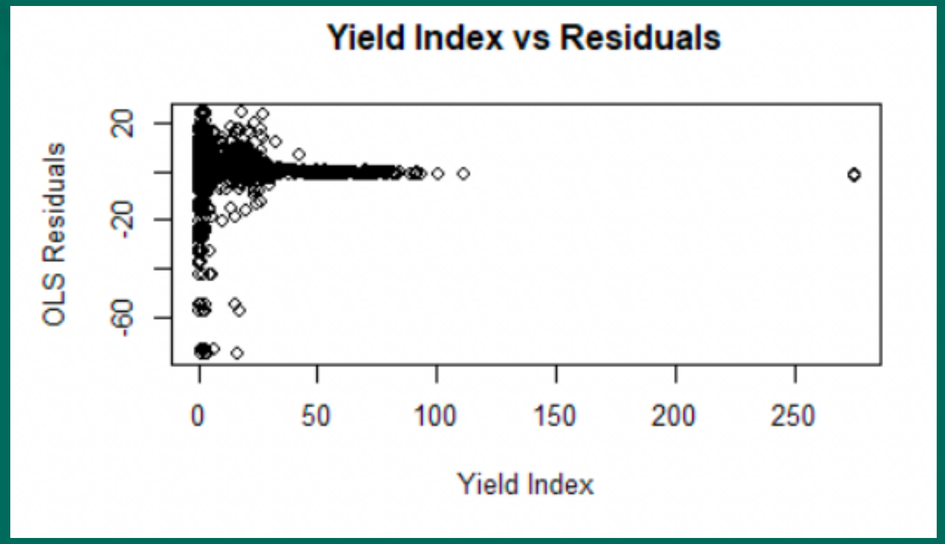
Rabi

Yield Index vs Residuals

Yield Index vs Residuals explains how residuals from running OLS regression vary depending on the yield-index. The graph seems to be aligned around the 0 mark for residuals, which aligns with the OLS property that sum of residuals is 0 for all points in the dataset. Otherwise, we observe that for both Rabi and Kharif, the absolute value of residuals are high for lower values of yield indices, indicating a higher error when yield indices are low.



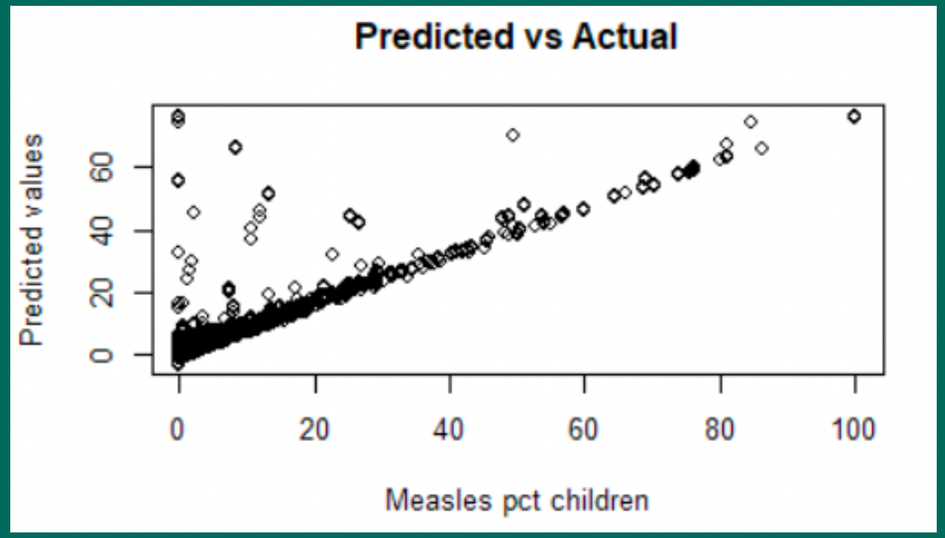
Kharif



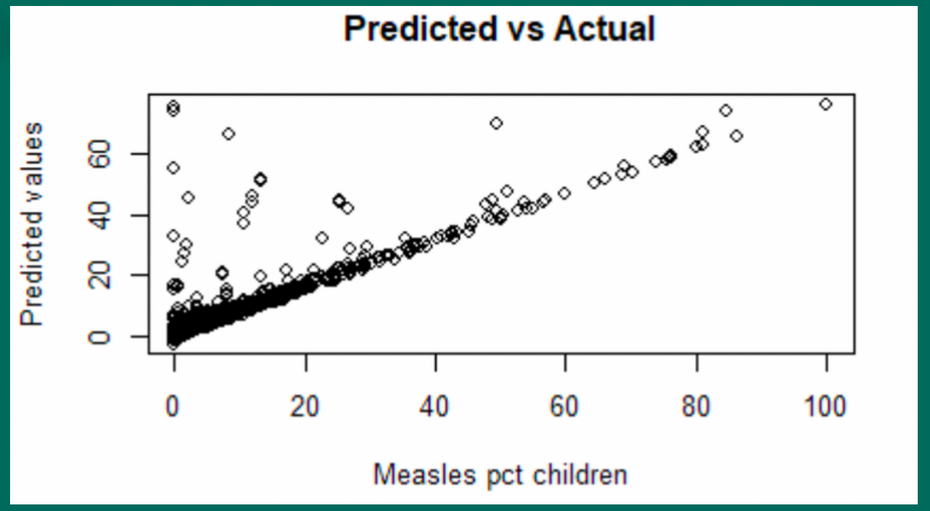
Rabi

Predicted vs Actual

In the third plot, Predicted vs Actual values of Measles, we observe that they are related along a straight line, which is the line y = x, with residuals prohibiting a straight line path.



Kharif



Rabi

Sum of Residuals and u*x

```
> #numerically showing that sum of residuals is 0
> sum(kharif_resid)
[1] 0.00000000000006244874
> sum(rabi_resid)
[1] 0.00000000003591221
```

```
> #sum for kharif is -0.0003767827, and for rabi is 0.00001958485,w hich are very
> #close to 0.
> sum(as.numeric(kharif_xi_ui))
[1] -0.0003767827
> sum(as.numeric(rabi_xi_ui))
[1] 0.00001958485
```

The sum for all 4 values is almost zero, and the presence of numerical in the end can be attributed to floating point error. Hence our regression model is following the 2 assumptions as well.

Hypothesis Testing (Structural outbreak (south zone))

- We introduce a dummy variable which is 1 if the district is south, and 0 otherwise.
H0 (null hypothesis): Bsouth = 0
Ha (alternate) : Bsouth != 0

We perform our regression with this dummy variable and check for confidence intervals (p values) from our model summary.

Structural outbreak in Southern region (t-test)

Testing for Kharif, Rabi and all seasons, we notice that p values are:

1. Rabi: 0.000000000022822
2. Kharif: 0.00000000344165
3. Total: < 0.0000000000000002

All of which are significantly low, implying the dummy variable to be statistically significant.

Structural outbreak in Southern region (t-test)

For the T-test, if Pr(>|t|) is less than 0.05 (confidence interval) then we can reject the null hypothesis.

All of the pvalues were considerably low implying probability that they play a meaningful part in the OLS regression in predicting the model is high. (as H0 that Beta(Dsouth) = 0 is rejected)

Since the probability that beta(Dsouth) is 0 is low, we note that it implies a structural break in mean outcome levels (Measles percentage) across southern and the rest of the groups.

Structural break in Southern region (F-test)

F-test to check validity of null hypothesis produced the following result:

Performing ANOVA on both Kharif and Rabi for restricted (without the dummy variable DSouth), and unrestricted model (with DSouth).

```
F value (kharif) = 32.24
F value (rabi) = 35.6

p value (kharif) = 0.00000001381
p value (rabi)   = 0.000000002473
```

A low P-value confirms the rejection of null hypothesis as was verified by our T-test as well.

Clustering for districts

```
Coefficients:
                  Estimate    Std. Error  t value            Pr(>|t|)
(Intercept)   74.2833724263  0.2591359152  286.658 < 0.0000000000000002 ***
v5             0.0000076058  0.0000007301   10.417 < 0.0000000000000002 ***
v20           -0.0084521753  0.0009016014   -9.375 < 0.0000000000000002 ***
v23            0.0034577886  0.0009382829    3.685             0.000229 ***
v26           -0.0000120238  0.0000011553  -10.407 < 0.0000000000000002 ***
v30            0.0108182760  0.0014789521    7.315    0.000000000000262 ***
v37           -0.7546074170  0.0019212326 -392.773 < 0.0000000000000002 ***
v38           -0.7445304491  0.0023308026 -319.431 < 0.0000000000000002 ***
index          0.0002804818  0.0012957818    0.216             0.828632
gdp           -0.0000001054  0.0000001154   -0.913             0.361100
beds           0.0000052940  0.0000004911   10.780 < 0.0000000000000002 ***
tap           -0.0061351488  0.0009915611   -6.187    0.000000000616989 ***
year2012       0.3591844599  0.0609669960    5.891    0.000000003854910 ***
year2013       0.8635215437  0.0615013902   14.041 < 0.0000000000000002 ***
year2014       0.4511210566  0.0629800574    7.163    0.000000000000802 ***
year2015       0.5252898150  0.0637425901    8.241 < 0.0000000000000002 ***
year2016       0.5959881071  0.0658364308    9.053 < 0.0000000000000002 ***
seasonRabi     0.0011774719  0.0392619802    0.030             0.976075
seasonSummer  -0.0149169729  0.0596526734   -0.250             0.802539
seasonWhole.Year  0.0030282489  0.0657133368    0.046             0.963245
south         -0.3867741439  0.0730163303   -5.297    0.000000118207986 ***
north          0.3982861415  0.0756792891    5.263    0.000000142507701 ***
east           0.1923287030  0.0744703966    2.583             0.009808 **
west          -0.0217903242  0.0772813783   -0.282             0.777975
central        0.0991847816  0.0999337980    0.993             0.320957
northeast      0.4983979427  0.0806291149    6.181    0.000000000640881 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.656 on 45493 degrees of freedom
Multiple R-squared:  0.8165,    Adjusted R-squared:  0.8164
F-statistic:  8099 on 25 and 45493 DF,  p-value: < 0.00000000000000022
```

Hypothesis Testing:

H0 (null) => B(region) = 0

Ha(alternate) => B(region) != 0

If we can reject the null hypothesis, then that would imply structural breaks for different regions.

Inference: Performing the T-test, we found the dummy variables for the south, north, east, and northeast district to be statistically significant. (Structural clustering in these regions)

Creative Component: Hypotheses

Objective: Yearly Variation in Measles and its effect on the regression model. The following hypotheses are to be tested:

a) Measles infections depend on yield index

b) Measles infections depends on the GDP

c) There are seasonal and yearly variations for measles infections.

The analysis will be conducted using the same variables as for the data assignment regression but with dummy variables for years and seasons.

Model Summary

```
> summary(time_analysis_model);

Call:
lm(formula = measles_model, data = test_analysis)

Residuals:
    Min      1Q  Median      3Q     Max
-76.156  -0.526  -0.057   0.502  23.693

Coefficients:
                      Estimate     Std. Error  t value             Pr(>|t|)
(Intercept)      74.79684527701  0.24951005663  299.775 < 0.0000000000000002 ***
v5                0.00000771307  0.00000072390   10.655 < 0.0000000000000002 ***
v20              -0.01178974895  0.00086061631  -13.699 < 0.0000000000000002 ***
v23               0.00383433323  0.00093367777    4.107  0.00004020548991345 ***
v26              -0.00001253602  0.00000111174  -11.276 < 0.0000000000000002 ***
v30               0.01488644764  0.00138978880   10.711 < 0.0000000000000002 ***
v37              -0.75878844742  0.00190087535 -399.178 < 0.0000000000000002 ***
v38              -0.74868829758  0.00230072032 -325.415 < 0.0000000000000002 ***
index             0.00157492109  0.00128445217    1.226              0.22015
gdp              -0.00000030584  0.00000008812   -3.471              0.00052 ***
beds              0.00000502024  0.00000033261   15.094 < 0.0000000000000002 ***
tap              -0.00874061056  0.00079446462  -11.002 < 0.0000000000000002 ***
year2012          0.37525232450  0.06104214224    6.147  0.00000000079400838 ***
year2013          0.87476137500  0.06123874305   14.284 < 0.0000000000000002 ***
year2014          0.48793479383  0.06240813213    7.818  0.00000000000000546 ***
year2015          0.59055903589  0.06264095550    9.428 < 0.0000000000000002 ***
year2016          0.66643658396  0.06371902800   10.459 < 0.0000000000000002 ***
seasonRabi       -0.00451959871  0.03922606086   -0.115              0.90827
seasonSummer      0.02414359647  0.05829216968    0.414              0.67874
seasonWhole.Year -0.09602525975  0.06495595819   -1.478              0.13933
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.664 on 45499 degrees of freedom
Multiple R-squared:  0.8157,    Adjusted R-squared:  0.8156
F-statistic: 1.06e+04 on 19 and 45499 DF,  p-value: < 0.00000000000000022
```

Yield Index and Measles Infections

Using the model summary obtained, we observe that the p-value for index is 0.2215, indicating with 95% confidence, that the slope for that variable is 0.

This means that yield index is not significant towards the analysis of measles infections in children.

GDP and Measles Infections

```
> summary(time_analysis_model);

Call:
lm(formula = measles_model, data = test_analysis)

Residuals:
    Min      1Q  Median      3Q     Max
-76.559  -0.486  -0.059   0.486  23.298

Coefficients:
                      Estimate     Std. Error  t value             Pr(>|t|)
(Intercept)      74.80966586992  0.24916342554  300.243 < 0.0000000000000002 ***
v5                0.00000885920  0.00000071942   12.314 < 0.0000000000000002 ***
v20              -0.01120199856  0.00085636125  -13.081 < 0.0000000000000002 ***
v23               0.00511801319  0.00092728883    5.519          0.0000000342 ***
v26              -0.00001419404  0.00000110538  -12.841 < 0.0000000000000002 ***
v30               0.01861725965  0.00136068770   13.682 < 0.0000000000000002 ***
v37              -0.75906762112  0.00190254943 -398.974 < 0.0000000000000002 ***
v38              -0.74996589144  0.00230425419 -325.470 < 0.0000000000000002 ***
index             0.00213086423  0.00128626785    1.657                0.0976 .
gdp              -0.00000005026  0.00000008405   -0.598                0.5498
beds              0.00000432468  0.00000032514   13.301 < 0.0000000000000002 ***
tap              -0.00992126514  0.00078535393  -12.633 < 0.0000000000000002 ***
seasonRabi        0.00378026661  0.03931271830    0.096                0.9234
seasonSummer      0.02482163847  0.05843238910    0.425                0.6710
seasonWhole.Year -0.12088248968  0.06500939102   -1.859                0.0630 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.673 on 45504 degrees of freedom
Multiple R-squared:  0.8148,    Adjusted R-squared:  0.8147
F-statistic: 1.43e+04 on 14 and 45504 DF,  p-value: < 0.00000000000000022
```

The model suggests, through the p-value, that GDP is significant towards analysis of Measles. However, if we run linear regression without creating dummy variables for the different years, the p-value increases to 0.54, indicating that it isn't significant for that case.

Variation across Harvest Seasons

Looking at p-values of dummy variables for seasons, we notice that they are not significant. A major reason for this is that the variable of interest (v36, measles percentage for children) is an aggregate over the entire year. Therefore, no seasonal changes can be observed with this dataset.

Yearly Variation

All of the yearly dummy variables are significant. With the base year taken to be 2011, we note that the intercepts for all the dummy variables are positive, indicating that all else held constant, there was a net yearly increase in the measles infections in children as compared to the base year.

The infections were maximum for 2013, followed by 2016, 2015, 2014, 2012, and 2011.

Additional Component

While doing the data project, we were allowed to think freely about how we can decrease the rate of death of children. We feel that agricultural growth and maternal care and GDP do make an important factor in determining death. However, it is also important to consider the application of MMR vaccine in India across different states and use that as a regressand as well to determine and to work upon decreasing the death of children aged (0 to 5) dying from measles.

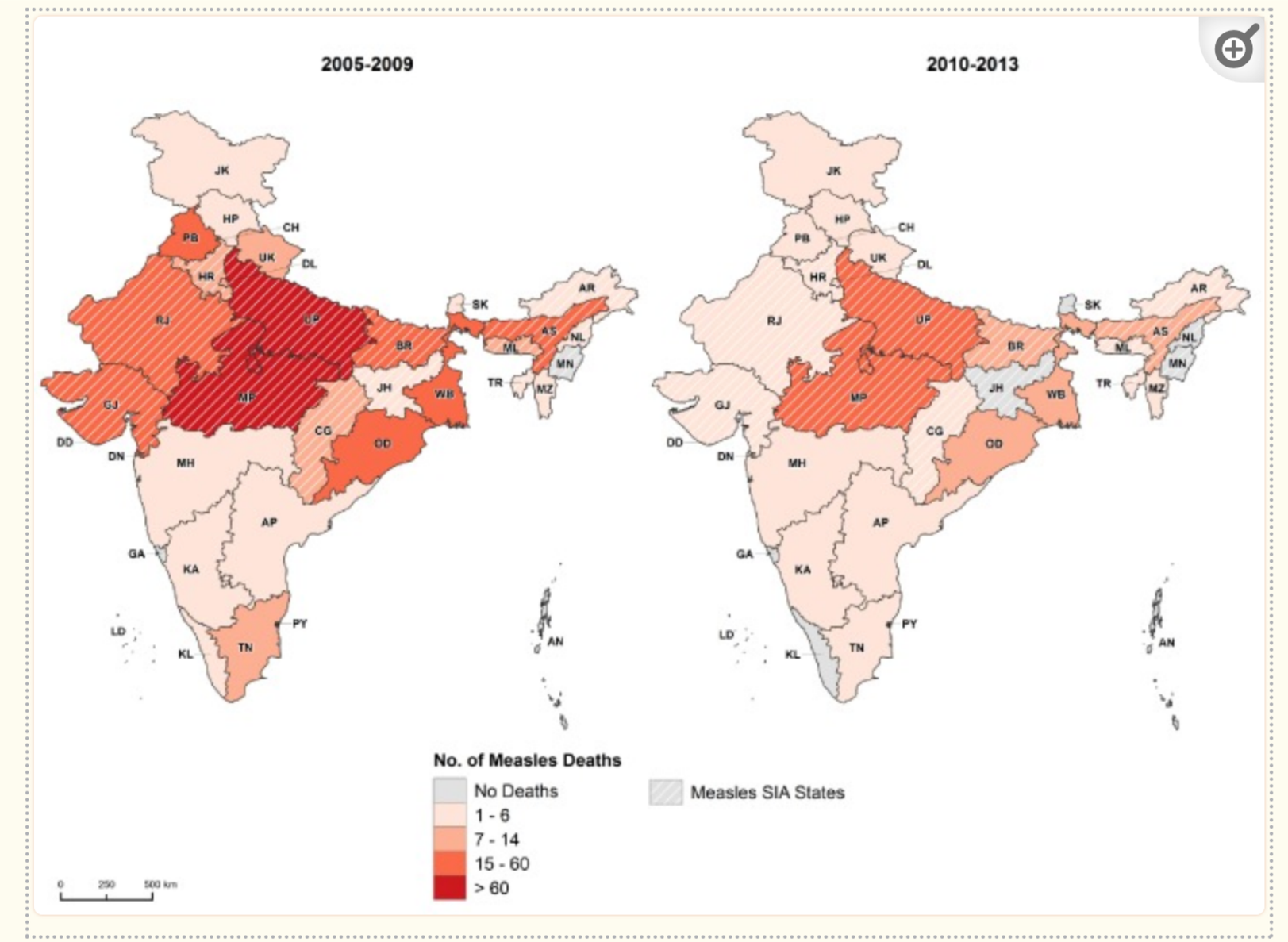
Additional Component

Citing 'The impact of measles immunization campaigns in India using a nationally representative sample of 27,000 child deaths' wherein interrupted time series and multi-level regression has been used to quantify the campaign's impact on measles mortality using the nationally representative Million Death Study.

We can add a new index as follows:

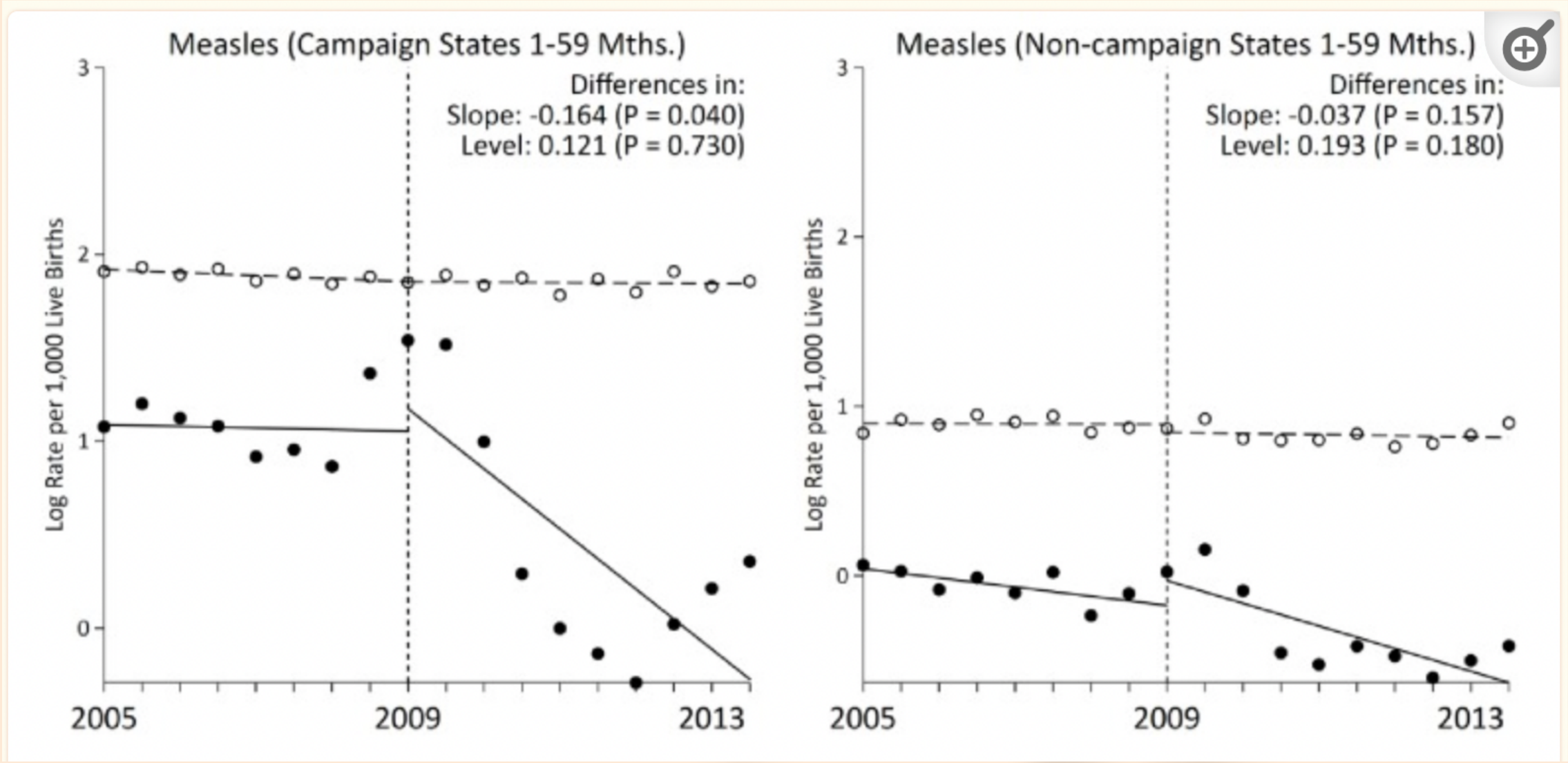
impact of National measles immunization coverage (defined as the percentage of children aged 12 to 23 months receiving any dose of measles vaccine from routine immunization)

Past studies have shown the following impacts: We can add a new index as follows:



State-level distribution of 1–59 month measles deaths before and after measles campaign launch, India, 2005–2013.

Past studies have shown the following impacts: We can add a new index as follows:



Past studies have shown the following impacts: We can add a new index as follows:

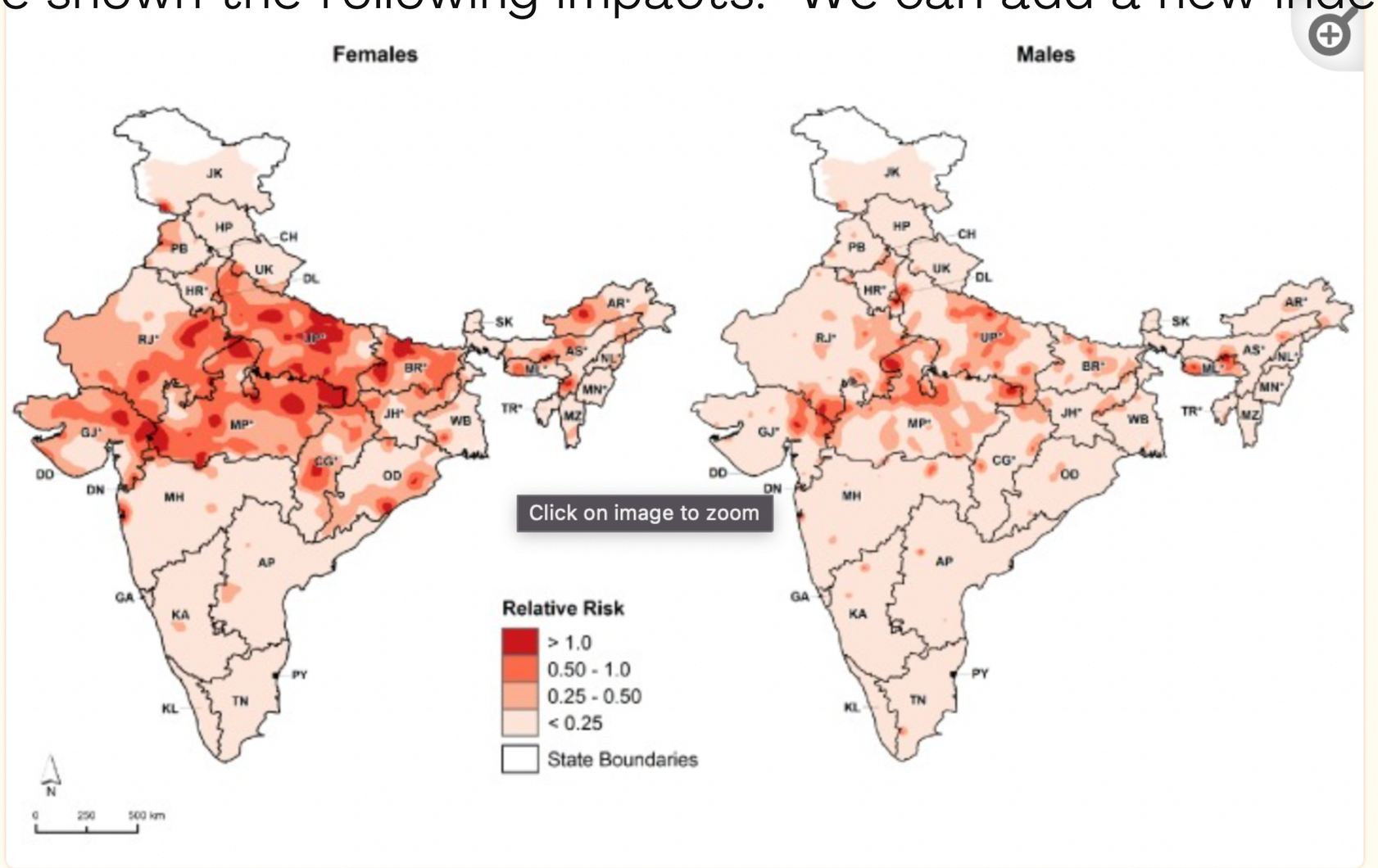
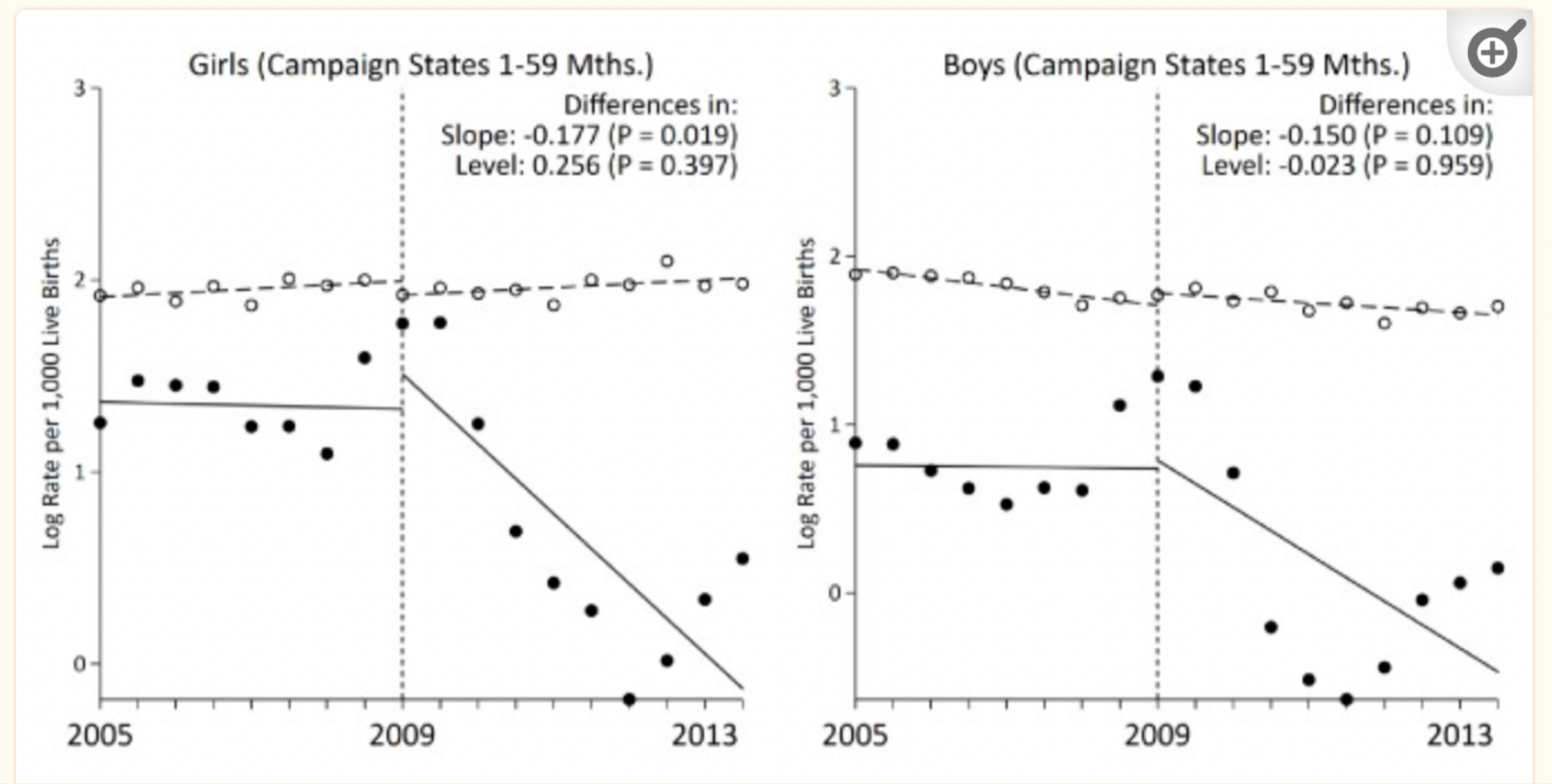


Figure 7.

Distribution of 1–59 month measles mortality risk (relative to all-cause mortality) by sex, India, 2005–2013.

Past studies have shown the following impacts: We can add a new index as follows:



Additional Component

Apart from that, the yield index and its growth were in many cases statistically insignificant (data in previous slides). An interpretation of this is that agricultural growth in the current era is measured on the basis of net profit, instead of the nutritional value being available to the general populous. More emphasis on increasing the affordability of staple foods and increasing access to micronutrients is essential to impact the health sector.

Reference: https://www.linkedin.com/pulse/data-science-vs-academia-perspectives-from-economists-rose-tan?trk=news-guest_share-article