IMPACT OF STATE-LEVEL SOCIO-ECONOMIC INDICATORS ON STARTUP SURGE IN INDIA.

Name – Aakash Vashishtha (02)

Debasmita Poddar (13)

Course – Ma Economics

Subject – Basic Econometrics

Professor – Dr. NACHIKETA CHATTOPADHYAY

Problem Statement

Impact of state-level socio-economic indicators on startup surge in India.

This study aims to estimate the effect of state-level GDP per capita, higher education institutions, and public capital outlay on the number of startups across Indian states and Union Territories for the year 2021 using cross-sectional data.

Data Set

The data of all the variables have been taken from Department for Promotion of Industry and Internal Trade (DPIIT), RBI and All India Survey on Higher Education (AISHE). The set has 4 variables (1 dependent and 3 independent) and 32 observations. They are as follows

Dependent Variable

1. Number of Startups per State/UT (startup_count) – This variable measures the total number of DPIIT-recognized startups in each Indian state and Union Territory for the year 2021. It serves as an indicator of formal entrepreneurial activity at the state level.

Independent Variables

1. GDP per capita (₹, current prices) (gdp_pc) – GDP per capita represents the average income level of a state and is used as a proxy for the level of economic development and market potential.
2. Number of Higher Education Institutions (uni_per_state) – This variable captures the total number of higher education institutions, including universities, colleges, and standalone institutions, in each state/UT. It is used as a proxy for the institutional support.
3. Capital Outlay (₹ crore) (capital_outlay) – Capital outlay refers to state government expenditure on the creation of physical assets such as infrastructure. It is used to represent public investment and infrastructure development at the state level.

The study uses a small sample of 32 observations. Given the limited data availability, a missing value for capital outlay of Chandigarh was replaced with zero to preserve the sample size.

Data Exploration

Descriptive Statistics:

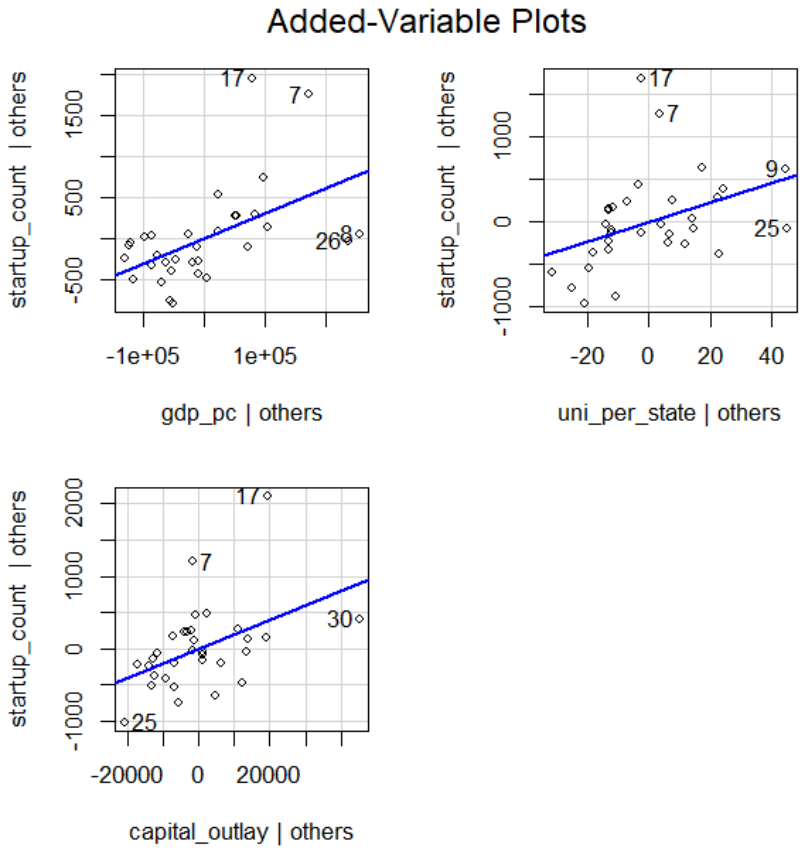
Statistic	startup_count	gdp_pc	uni_per_state	capital_outlay

Min	2.0	47,498	3.00	0
1st Qu.	49.5	121,393	10.00	6,998
Median	210.0	195,775	32.00	13,208
Mean	604.6	198,176	36.44	19,814
3rd Qu.	907.8	244,526	56.50	29,477
Max	3,552.0	472,070	91.00	96,481

1. **Linearity**:

We created added-variable (AV) plots to examine the relationship between each predictor and startup_count after accounting for the effects of other variables.

The added-variable plots show a positive partial relationship between startup activity and each explanatory variable—GDP per capita, universities per state, and capital outlay—after controlling for the other regressors. The roughly linear fitted lines and absence of strong curvature support the linearity assumption of the multiple regression model, with universities per state exhibiting the strongest marginal effect.



2. Heteroskedasticity:

The residuals are randomly distributed around zero across the range of fitted values, with no evident systematic pattern or curvature, indicating that the linear functional form is appropriate. Moreover, the dispersion of residuals appears broadly constant, with no clear funnel-shaped pattern, suggesting that heteroskedasticity is not severe.

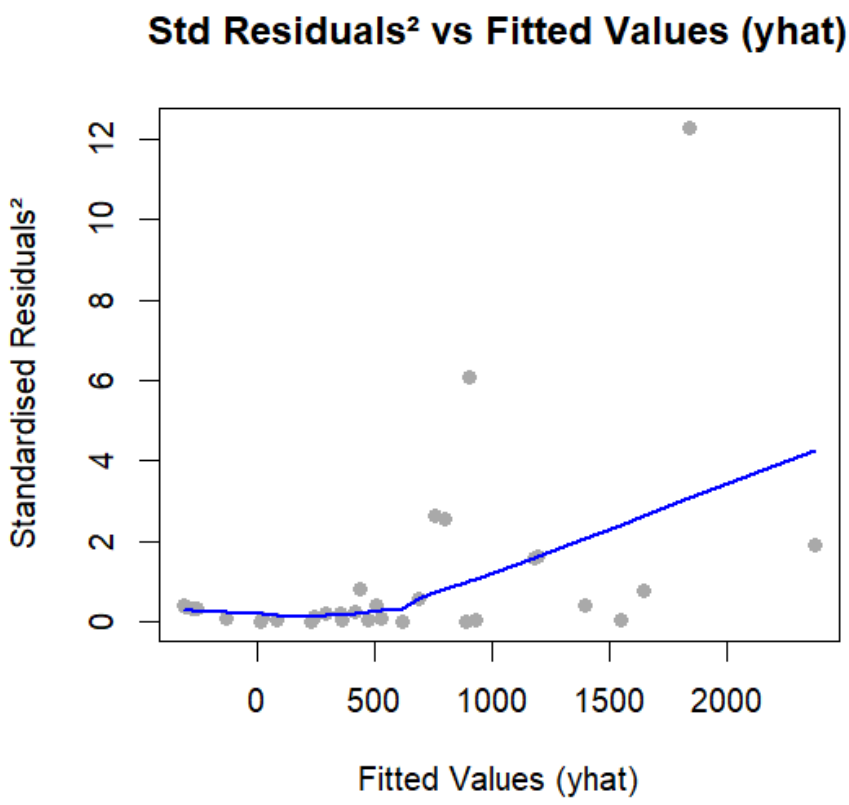
We conducted Breusch-Pagan test and White test to confirm for heteroscedasticity. It gave the following results:

<u>BP Test:</u>

BP = 7.7647

df = 3

p-value = 0.05113



The test yields a test statistic of 7.76 with a p-value of 0.051, which is slightly above the conventional 5% significance level. Hence, the null hypothesis of homoscedastic errors cannot be rejected at the 5% level, although the result is borderline and suggests weak evidence of heteroskedasticity at the 10% level.

White Test:

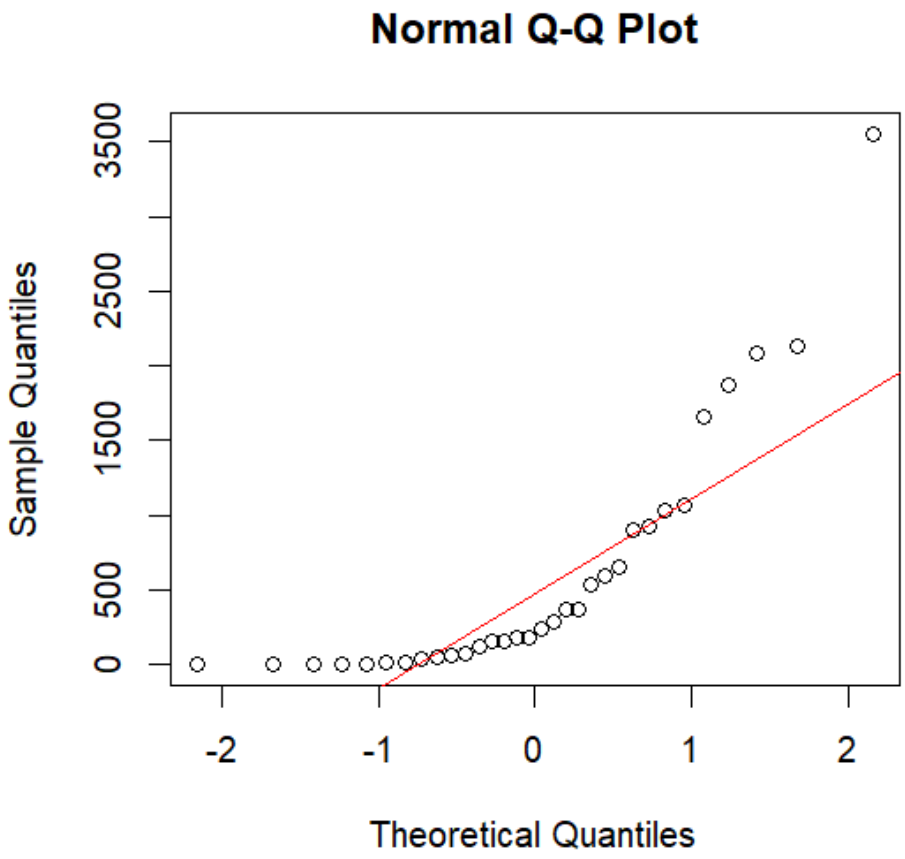
BP = 5.9585

df = 2

p-value = 0.05083

The White test produces a test statistic of 5.96 with a p-value of 0.051, which is marginally above the 5% significance level. Therefore, the null hypothesis of homoskedastic errors cannot be rejected at the 5% level, although the result is borderline and provides weak evidence of heteroskedasticity at the 10% level.

3. Normality:



The Q–Q plot shows approximate normality in the central portion of the distribution, but noticeable deviations in the tails, indicating mild departures from the normality assumption.

We then conducted the Shapiro – Wilk Test and Anderson Darling (AD) Test to check for normality. These were the results

Shapiro-Wilk Test:

W = 0.73828

p-value = 3.426e-06

The Shapiro–Wilk test strongly rejects the null hypothesis of normality (W = 0.74, p <0.001), confirming that the regression residuals are not normally distributed.

Anderson – Darling Test:

A = 0.88595

p-value = 0.02068

The Anderson–Darling test rejects the null hypothesis of normality at the 5% level (A =0.886, p = 0.021), providing evidence that the regression residuals are not normally distributed and exhibit tail departures from the normal distribution.

Since both the tests indicated that our data is not normally distributed, we did a Box – Cox transformation to treat the non – normality in the model. The transformation gave us 0.263 as the value of λ.

$$y^{(\lambda)} = \begin{cases} \frac{y_i^{\lambda}-1}{\lambda} & if\ \lambda = 0 \dots (i) \\ lny_i & if\ \lambda \neq 0 \dots (ii) \end{cases}$$

Since the estimated value of $\lambda$is not equal to zero, the model qualifies for case (i) Accordingly, the dependent variable was transformed using the above expression. After the transformation, the prior conducted tests for normality give the following results:

Shapiro-Wilk Test:

W = 0.97206

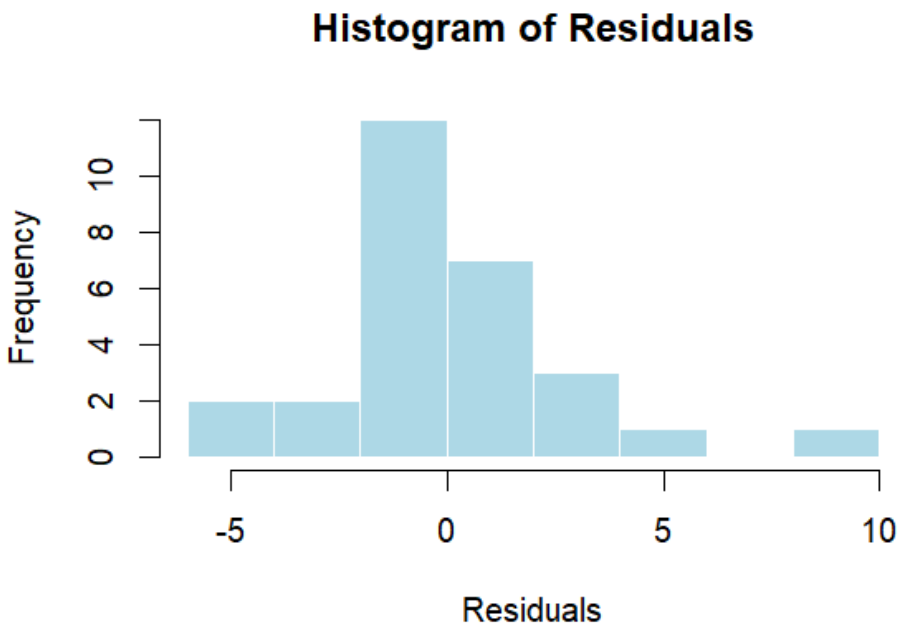
p-value = 0.5581

Anderson – Darling Test:

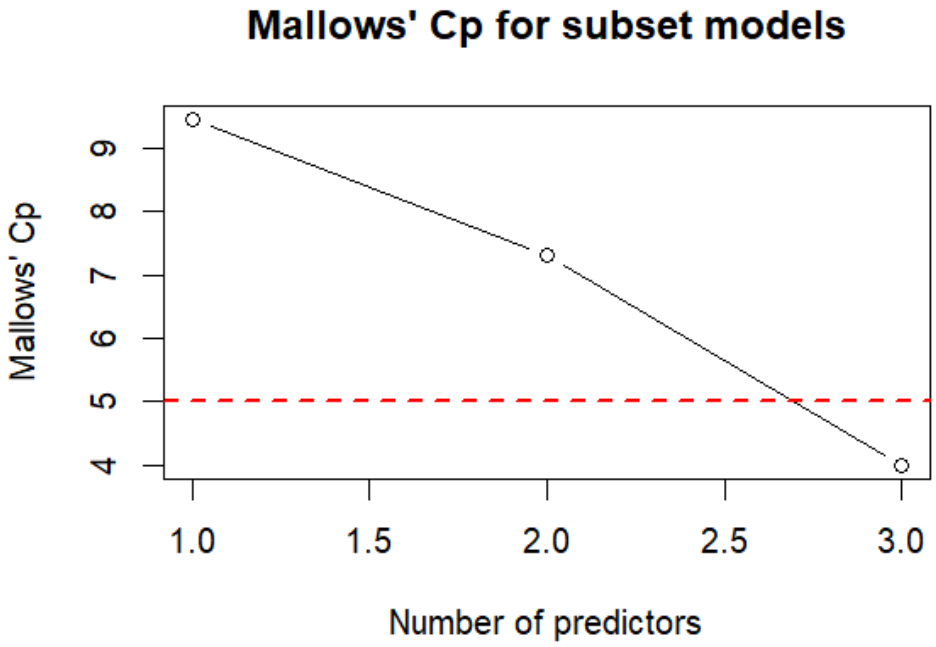
A = 0.32561

p-value = 0.5087

After applying a Box-Cox transformation, both the Shapiro-Wilk ($p = 0.5581$) and Anderson Darling ($p = 0.5087$) tests indicate that the transformed data is approximately normally distributed.



4. Model Selection:



For model selection first we did Mallows CP and this is the values we got 9.440184 7.308016 4.000000 which suggests that Mallows' Cp decreases with the inclusion of additional predictors, and the model with three predictors (Cp = 4) is selected as it best balances fit and parsimony.

After this we did AIC/BIC and this is the result we got

Model Specification	**AIC**	**BIC**
OLS (Linear Model)	498.61	505.94
Log-Linear Model	496.36	503.69
Box-Cox Model	188.78	196.11

The result suggests that the Box–Cox transformed model performs better than the other models, as it has the lowest AIC and BIC values

5. Multicollinearity:

For checking multicollinearity, we did Variance Inflation Factor (VIF), Inverse Variance Inflation Factor (IVIF) and Correlation Matrix Eigenvalue Test and these are the results we got

Variance Inflation Factor (VIF):

Variable	**VIF Value**
log(gdp_pc)	1.173969
uni_per_state	2.303809
capital_outlay	2.564251

The Variance Inflation Factor (VIF) values for all explanatory variables are low and fall within acceptable limits. This indicates that the variables are not highly correlated with each other. Therefore, multicollinearity is not a concern in the model. The estimated results can be considered stable and reliable.

Inverse Variance Inflation Factor (IVIF):

Variable	**IVIF Value**
log(gdp_pc)	0.8518116
uni_per_state	0.4340638
capital_outlay	0.3899774

To be further sure, the Inverse Variance Inflation Factor (IVIF) test was also conducted. All IVIF values are well above the benchmark level, indicating that the explanatory variables do not suffer from multicollinearity. This further confirms that the model is stable and the results are reliable.

Correlation Matrix:

	ln_gdp_pc	**uni_per_state**	**capital_outlay**
ln_gdp_pc	1.0000000	-0.1614181	-0.3535194
uni_per_state	-0.1614181	1.0000000	0.7443841
capital_outlay	-0.3535194	0.7443841	1.0000000

The correlation matrix indicates a low negative correlation between universities per state and log (GDP per capita). A relatively strong positive correlation is observed between universities per state and capital outlay; however, it is not sufficiently high to invalidate the analysis.

Eigenvalue Test:

2.524887e+10

1.583781e+04

1.686079e+03

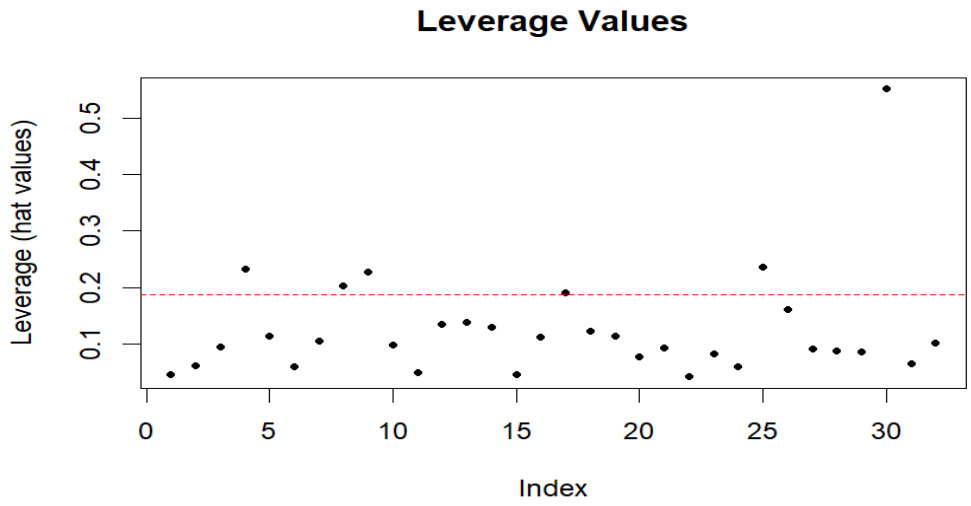
5.072437e-02

Eigen value test was also conducted and the eigenvalues do not indicate any strong dependency among the explanatory variables. This suggests that the variables provide independent information in the model. Hence, multicollinearity is not a concern.

6. Influence and Outlier Analysis:

For influence analysis we did the following :

Leverage

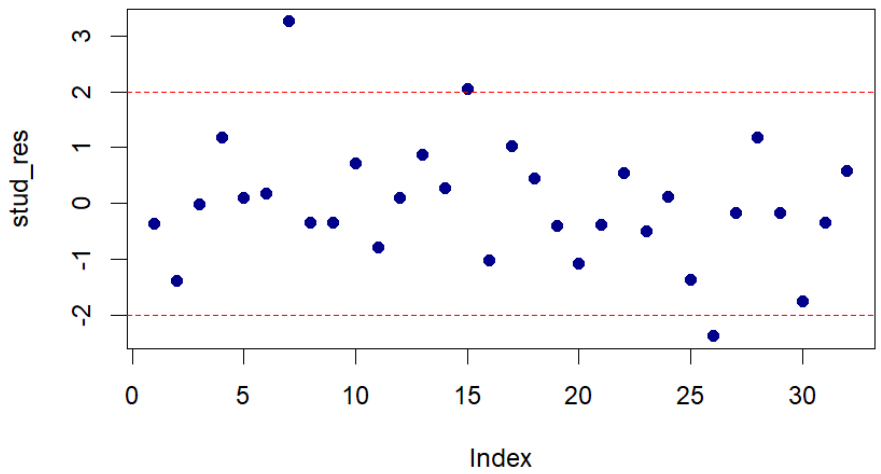


Leverage values were examined to assess the influence of individual observations. A few observations lie above the reference cutoff, indicating the presence of high-leverage points. However, these are limited in number and do not indicate a serious influence problem in the model.

Using the cutoff rule of 2k/n, several observations were identified as high-leverage points.

Outliers

Observations with absolute standardized residuals greater than 2 were classified as outliers. Observations 7, 15, and 26 were identified as outliers.



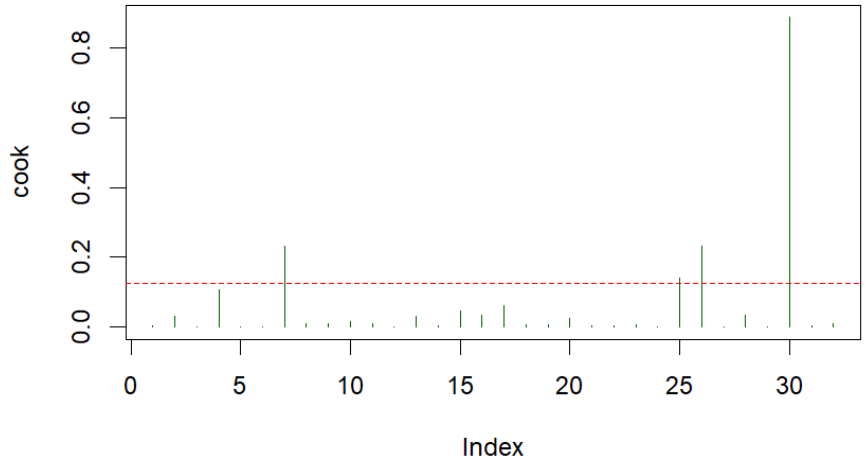
DFBETAS

7, 26, 39, 58, 89, 94, 126

DFBETAS were examined to identify observations that have a noticeable impact on the estimated coefficients. A small number of observations were flagged, indicating that these points may influence specific parameter estimates.

Cook's Distance

Cook's Distance was also analyzed to assess the overall influence of observations on the



model. Based on the cutoff criterion using the cutoff of 4/n, observations 7, 25, 26, and 30 exceeded the threshold. These observations were therefore removed to improve model performance and ensure more robust and reliable estimation results.

Model Refinement:

After removing the identified outliers, a refined model (Model_BC2) was estimated and R^2 increased from 0.74 to 0.85 and the coefficient significance improved substantially. Influence diagnostics showed fewer problematic observations.

Covariance Ratio Analysis:

Observation	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Covariance Ratio	1.19075 71	0.73112 38	1.30538 92	1.65584 28	1.34981 59	1.26052 12	1.64304 05	1.56779 89	1.07095 65	1.10327 49	1.45445 43	1.09634 03
Observation	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
Covariance Ratio	1.42843 30	0.25994 45	0.69683 94	1.73747 63	1.27740 81	1.30259 99	0.99130 91	1.27692 10	1.20001 82	1.30781 37	1.26779 63	1.25990 96

Most covariance ratio values were close to 1, indicating limited influence on the variance–covariance structure.

Post-Refinement Influence Diagnostics:
Cook's distance, residual, and leverage plots confirmed reduced influence after refinement.

7. ANOVA:

The *Type I ANOVA* results indicate that **log(gdp_pc)** ($F = 6.51$, $p = 0.0176$), **uni_per_state** ($F = 118.44$, $p < 0.001$), and **capital_outlay** ($F = 12.73$, $p = 0.0016$) all significantly explain variation in the transformed startup outcome when entered sequentially into the model. This suggests that each variable contributes meaningfully to the model in the specified order of inclusion.

Type II ANOVA was conducted to assess the significance of each explanatory variable while controlling for the presence of the others. The results show that **log(gdp_pc)** ($F = 8.43$, $p = 0.0078$), **uni_per_state** ($F = 19.12$, $p < 0.001$), and **capital_outlay** ($F = 12.73$, $p = 0.0016$) remain statistically significant. This confirms that the effects are robust and not influenced by the order of variable entry.

(*For results check appendix)

Conclusion

This study sets out to examine the impact of key state-level socio-economic indicators on startup activity across Indian states and Union Territories for the year 2021. Using cross-sectional data, the analysis focused on GDP per capita, the number of higher education institutions, and public capital outlay as determinants of the number of DPIIT-recognized startups. The empirical framework was carefully constructed and subjected to extensive diagnostic testing to ensure the robustness and reliability of the results.

Initial estimation using an ordinary least squares framework revealed meaningful associations between startup activity and all three explanatory variables. Diagnostic checks indicated that while the assumptions of linearity and homoskedasticity were broadly satisfied, the normality assumption was violated. This issue was effectively addressed through a Box–Cox transformation of the dependent variable, which substantially improved the distributional properties of the residuals. Model selection criteria—including Mallows' Cp, AIC, and BIC—consistently favoured the Box–Cox transformed specification, confirming its superior fit relative to alternative functional forms.

Further diagnostics showed no evidence of serious multicollinearity, as supported by VIF, IVIF, correlation matrix, and eigenvalue tests. Influence and outlier analyses identified a small number of observations exerting disproportionate leverage on the estimates. After removing these influential points, the refined model exhibited a notable improvement in explanatory power, with the coefficient of determination increasing from 0.74 to 0.85, alongside enhanced statistical significance of the regressors. Post-refinement diagnostics confirmed that the model was more stable and less sensitive to individual observations.

The final results provide strong empirical evidence that economic development, institutional capacity, and public investment play a significant role in shaping startup ecosystems at the state level. In particular, the number of higher education institutions emerged as the most influential factor, underscoring the importance of human capital formation and knowledge infrastructure in fostering entrepreneurial activity. GDP per capita and capital outlay also exhibited statistically significant effects, highlighting the complementary roles of market size and public infrastructure in supporting startup growth.

Despite these insights, the study is subject to certain limitations. The analysis relies on a relatively small cross-sectional sample and pertains to a single year, which restricts the ability to capture dynamic effects or causal relationships. Additionally, the substitution of missing data for capital outlay, while necessary to preserve sample size, may introduce minor measurement error.

Overall, the findings of this study contribute to a better understanding of the structural factors underlying regional disparities in startup activity in India and offer useful policy implications for fostering balanced and sustainable entrepreneurial development across states and Union Territories.

Appendix

a) model_ols

```
> summary(model_ols)

Call:
lm(formula = startup_count ~ gdp_pc + uni_per_state + capital_outlay,
    data = dataset_ecotrix_final_)

Residuals:
    Min      1Q  Median      3Q     Max
-753.54 -272.78  -72.36  192.82 1713.15

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -8.084e+02  2.685e+02  -3.011  0.00547 **
gdp_pc          3.048e-03  9.585e-04   3.180  0.00358 **
uni_per_state   1.139e+01  5.087e+00   2.239  0.03326 *
capital_outlay  1.988e-02  7.319e-03   2.716  0.01119 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 535 on 28 degrees of freedom
Multiple R-squared:  0.628,     Adjusted R-squared:  0.5882
F-statistic: 15.76 on 3 and 28 DF,  p-value: 3.398e-06
```

b) model_log2

```
> summary(model_log2)

Call:
lm(formula = log(startup_count) ~ gdp_pc + uni_per_state + capital_outlay,
    data = dataset_ecotrix_final_[-c(7, 26, 30), ])

Residuals:
     Min       1Q   Median       3Q      Max
-2.20872 -0.57004  0.03544  0.66871  2.30435

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.722e+00  5.609e-01   3.071  0.00509 **
gdp_pc         4.911e-06  2.269e-06   2.165  0.04017 *
uni_per_state  3.915e-02  1.067e-02   3.669  0.00115 **
capital_outlay 5.313e-05  1.962e-05   2.708  0.01202 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.052 on 25 degrees of freedom
```

c) model_bc

```
> summary(model_bc)

Call:
lm(formula = startup_bc_transformed ~ log(gdp_pc) + uni_per_state +
    capital_outlay, data = dataset_ecotrix_final_)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5167 -1.7286 -0.4148  2.2759 11.2496

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -4.634e+01  1.870e+01  -2.478 0.019517 *
log(gdp_pc)     4.163e+00  1.532e+00   2.717 0.011170 *
uni_per_state   1.639e-01  4.070e-02   4.027 0.000391 ***
capital_outlay  1.384e-04  6.008e-05   2.304 0.028861 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.226 on 28 degrees of freedom
Multiple R-squared:  0.7401,    Adjusted R-squared:  0.7122
F-statistic: 26.58 on 3 and 28 DF,  p-value: 2.422e-08
```

d) model_bc2

```
> summary(model_bc2)

Call:
lm(formula = startup_bc_transformed ~ log(gdp_pc) + uni_per_state +
    capital_outlay, data = dataset_ecotrix_final_[-c(7, 25, 26,
    30), ])

Residuals:
    Min      1Q  Median      3Q     Max
-5.5016 -1.6036 -0.3871  1.5718  8.2700

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -4.011e+01  1.491e+01  -2.689 0.012815 *
log(gdp_pc)     3.586e+00  1.235e+00   2.903 0.007811 **
uni_per_state   1.527e-01  3.493e-02   4.372 0.000205 ***
capital_outlay  2.138e-04  5.993e-05   3.568 0.001558 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.046 on 24 degrees of freedom
Multiple R-squared:  0.8516,    Adjusted R-squared:  0.833
F-statistic: 45.89 on 3 and 24 DF,  p-value: 4.286e-10
```

e) Dfbeta

```
> dfb
    (Intercept)   log(gdp_pc) uni_per_state capital_outlay
1   0.004474253 -0.0058962863  -0.041798870    0.030811063
2   0.035922347 -0.0586115134   0.190033527   -0.049594145
3  -0.003230585  0.0029973676   0.004635938   -0.004541367
4   0.517097807 -0.5093144810  -0.156566259    0.128689201
5  -0.018055423  0.0200066580  -0.015853577    0.004444420
6   0.028253272 -0.0269351442   0.009940802   -0.018858601
7  -0.829412854  0.8634125263   0.008842730   -0.002731969
8   0.127194430 -0.1344273400   0.093135850   -0.069928857
9   0.037197092 -0.0329650870  -0.164366685    0.087532827
10 -0.073307986  0.0741042228   0.164132853   -0.119251751
11 -0.007210624 -0.0004068356  -0.051880240    0.093904570
12  0.006142398 -0.0046889238  -0.031771532    0.027940664
13  0.295670631 -0.2884516219   0.107765552   -0.196757540
14 -0.058808582  0.0572436733   0.040816115    0.017150924
15 -0.162351275  0.1855487966  -0.099670836    0.026562110
16 -0.107326405  0.1144597333  -0.249431384    0.098914884
17 -0.245189092  0.2392406714  -0.059807536    0.338146567
18  0.126612531 -0.1193378406  -0.042832411   -0.030105660
19 -0.106052378  0.0996166801   0.014786750    0.053694297
20 -0.008141671 -0.0124242854   0.155480778    0.001573714
21 -0.070452999  0.0635320417   0.037938402    0.027623531
22  0.054303279 -0.0506819431  -0.007316874    0.003518553
23  0.035951445 -0.0452948552   0.066064220   -0.001356020
24  0.009045300 -0.0082789852   0.017921030   -0.021587730
25 -0.186546189  0.2031410661  -0.690717376    0.464348538
26  0.766942191 -0.8107088981   0.360691946   -0.199572882
27  0.032235784 -0.0320286994  -0.006410846   -0.021767840
28 -0.235746696  0.2469571338  -0.176212205    0.225135291
29 -0.026228883  0.0229668398   0.020176512    0.009361082
30 -0.002043847  0.0332491572   0.617847970   -1.545468116
31 -0.000186219 -0.0021916664  -0.048222921    0.058169052
32  0.095944173 -0.0960787927   0.144081047   -0.126823304
```

f) Anova

Type I

```
Analysis of Variance Table

Response: startup_bc_transformed
               Df  Sum Sq Mean Sq  F value    Pr(>F)
log(gdp_pc)     1   60.36   60.36   6.5052  0.017555 *
uni_per_state   1 1099.03 1099.03 118.4436 9.181e-11 ***
capital_outlay  1  118.10  118.10  12.7280  0.001558 **
Residuals      24  222.70    9.28
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Type II

```
Anova Table (Type II tests)

Response: startup_bc_transformed
                Sum Sq Df F value   Pr(>F)
log(gdp_pc)     78.179  1  8.4254 0.007811 **
uni_per_state  177.384  1 19.1167 0.000205 ***
capital_outlay 118.102  1 12.7280 0.001558 **
Residuals      222.695 24
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```