1. Data Summary

Summary table (Year: 2001 and 2011; Urban and Rural)

Variable	Obs	Mean	Std. Dev.	Min	Max
Urban (1=Urban, 0=Rural)	2308	.5	.5	0	1
Outcome variables:					
Population	2308	955105.99	1108032.9	0	16368899
Employment, Total	2308	376069.21	436762.08	0	5456822
Wage, all sector average	2136	208.93	116.485	35.714	790.908
Wage, formal manufacturing	1756	140.743	67.286	41.721	588.708
Wage, informal manufacturing	2100	63.525	33.212	13.333	339.591
Other important variables:					
Rental price, housing	2242	7.123	.532	4.305	9.234
Distance to Straight Line Connect	2308	255.725	219.807	1	1584.172
Distance to GQ Highway	2308	236.01	211.97	.288	1583.745
Employment, Agricultural	1136	206277.92	281417.87	8.766	2181390.4
Employment, Industrial	1136	42114.877	70545.661	64.244	1094715.2
Employment, Services	1136	96366.503	157242.65	475.203	3107962.6

Outcomes to look at: To assess the impact of highway improvements on urbanization in India, I will analyze several key outcomes. Population growth in urban area is a direct measure of urbanization and will indicate how changes in infrastructure influence demographic shifts. Total employment is another important indicator of urbanization; an increase in employment in urban areas would suggest increase urbanization. Moreover, I am interested in look into Wage changes, particularly the average wages across all sectors. As more people move into cities, they may not all enter the formal job market, so it is interesting to look at the changes in informal and formal wages respectively.

Sectoral employment and housing rental prices are included in the table as they may serve as important control variables. Furthermore, I include the distance metrics, such as the straight-line distance between nodal cities, as control variables, and the proximity to the GQ highway to create the treatment variable. These aspects will be explained in more detail below.

2. Issues with the data

Urban 2011

Variable	Obs	Mean	Std. Dev.	Min	Max
Population	577	642091.27	1184825.9	0	16368899
Employment, Total	577	225824.54	437030.64	0	5456822
Employment, Agricultural	0	.	.	.	.
Employment, Industrial	0	.	.	.	.
Employment, Services	0	.	.	.	.
Wage, all sector average	563	275.643	117.27	65.162	790.908
Wage, formal manufacturing	474	175.925	65.507	45.065	588.708

Wage, informal manufacturing	562	87.486	26.903	34.518	339.591
Rental price, housing	563	7.33	.492	5.469	9.234

Rural 2011

Variable	Obs	Mean	Std. Dev.	Min	Max
Population	577	1423781.9	1066353	0	9693932
Employment, Total	577	594225.15	431172.32	0	3973608
Employment, Agricultural	0	.	.	.	.
Employment, Industrial	0	.	.	.	.
Employment, Services	0	.	.	.	.
Wage, all sector average	563	275.643	117.27	65.162	790.908
Wage, formal manufacturing	474	175.925	65.507	45.065	588.708
Wage, informal manufacturing	562	87.486	26.903	34.518	339.591
Rental price, housing	563	7.33	.492	5.469	9.234

Urban 2001

Variable	Obs	Mean	Std. Dev.	Min	Max
Population	577	487028.89	931117.26	0	12905780
Employment, Total	577	156267.57	318187.28	0	4244170
Employment, Agricultural	566	12871.688	15091.258	8.766	125234.48
Employment, Industrial	566	37285.947	85431.51	64.244	1094715.2
Employment, Services	566	103577.14	208116.06	475.203	3107962.6
Wage, all sector average	505	134.555	53.763	35.714	436.591
Wage, formal manufacturing	404	99.464	40.711	41.721	324.792
Wage, informal manufacturing	488	35.929	10.848	13.333	106.393
Rental price, housing	558	6.914	.487	4.305	8.383

Rural 2001

Variable	Obs	Mean	Std. Dev.	Min	Max
Population	577	1267521.9	934882.59	0	8626883
Employment, Total	577	527959.6	379598.28	0	3445916
Employment, Agricultural	570	398326.92	289122.72	2280.544	2181390.4
Employment, Industrial	570	46909.92	51319.371	141.175	445057.23
Employment, Services	570	89206.469	78798.786	2137.877	819468.38
Wage, all sector average	505	134.555	53.763	35.714	436.591
Wage, formal manufacturing	404	99.464	40.711	41.721	324.792
Wage, informal manufacturing	488	35.929	10.848	13.333	106.393
Rental price, housing	558	6.914	.487	4.305	8.383

There are several issues with the data that should be noted. Firstly, there is missing data for sectoral employment in 2011, the employment data for agricultural, industrial, and service sectors are only available for 2001. The limitation constrains the analysis of sector-specific employment changes over time and may restrict the inclusion of these variables as controls. Secondly, there are inconsistent observations across variables, with fewer observations for wage data compared to population and total employment data. Lastly, the large standard deviations and wide ranges for some variables, such as population and employment, suggest the potential presence of outliers.

3. Balance table and initial differences

Treatment group: Districts closer than 75km for the highway

Control group: Districts farer than 75km for the highway

Balance table 1
Two-sample t test with unequal variances

	obs1	obs2	Mean1	Mean2	dif	St Err	p value
Population	850	304	732519.637	1282020.1	-549500.42	85860.622	0
Employment, Total	850	304	289488.421	489256.31	-199767.89	33089.059	0
Employment, Agricultural	839	297	188775.702	255720.23	-66944.523	21336.122	.002
Employment, Industrial	839	297	30690.782	74386.985	-43696.204	6799.771	0
Employment, Services	839	297	73817.376	160065.89	-86248.512	15835.121	0
Wage, all sector average	730	280	135.470	132.171	3.3	3.701	.373
Wage, formal manufacturing	548	260	99.142	100.142	-1	2.74	.716
Wage, informal manufacturing	706	270	36.110	35.456	.653	.786	.407
Rental price, housing	816	300	6.921	6.893	.029	.033	.37

Note: the balance table uses data in 2001(urban and rural) since we are interested in the initial differences between districts which are close vs. far from the highway. The 2 groups are classified by whether the districts are closer than or beyond 75km for the highway. Observation 1 is the control group and Observation 2 is the treatment group. same for the mean.

Balance table 2
Two-sample t test with unequal variances

	obs1	obs2	Mean1	Mean2	dif	St Err	p value
Population	425	152	333873.099	915260.53	-581387.43	132063.13	0
Employment, Total	425	152	105161.600	299162.57	-194000.97	45667.392	0
Employment, Agricultural	416	150	12029.181	15208.243	-3179.062	1308.444	.015
Employment, Industrial	416	150	23339.186	75964.964	-52625.778	12275.03	0
Employment, Services	416	150	72068.364	190961.47	-118893.1	29890.349	0
Wage, all sector average	365	140	135.470	132.171	3.3	5.242	.529
Wage, formal manufacturing	274	130	99.142	100.142	-1	3.88	.797
Wage, informal manufacturing	353	135	36.110	35.456	.653	1.114	.558
Rental price, housing	408	150	6.921	6.893	.029	.046	.527

Note: the balance table uses data in 2001(only urban area). Observation 1 is the control group and Observation 2 is the treatment group. same for the mean.

As observed from the balance tables above, the p-values for Population, Total Employment, Employment in Agriculture, Employment in Industry, and Employment in Services are all smaller than 0.05. This indicates that we reject the null hypothesis at the 95% confidence level, suggesting significant initial differences between these variables. To account for potential indirect effects on our outcome urbanization and ensuring that the estimated treatment effect is not biased by these pre-existing disparities, we will include these variables as controls in the subsequent regression analysis.

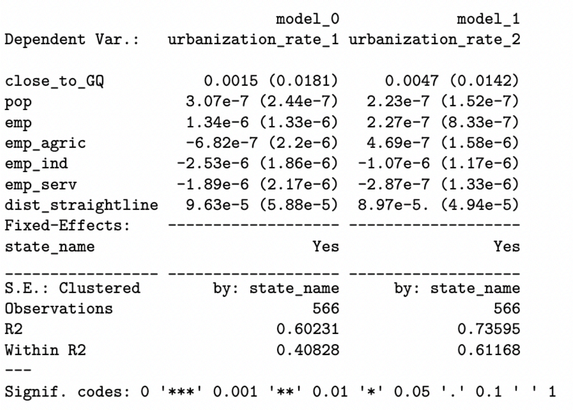
4. Specification(s)

The central assumption is that the treatment is exogenous—that is, not influenced by unmeasured confounders that could simultaneously affect the treatment (highway improvements and district close to the highway) and the outcome (urbanization). The primary concern is ensuring that the factors determining where the highway system was built do not themselves cause increased urbanization level. Therefore, it is essential to understand the determinants of why the highway system was constructed in one area (A) rather than another (B), given that the highway was not built randomly.

One of the primary reasons for the highway's routing is its need to connect the four major cities (Delhi, Mumbai, Chennai, and Kolkata) while minimizing construction costs by following a straight path. The highway generally follows the shortest possible route though there are deviations. To account for this in the analysis, we should include the variable `dist_straightline`, which measures the distance from the straight-line path connecting these cities, as a control variable in the regression model. The rationale is that the factors determining where the highway system was built are not causing improvement in outcome variables (urbanization rates), thus allowing the impact on urbanization be directly attributed to the railway improvement.

In practice, it is important to identify and include other factors that influenced the highway's construction location. These factors may include geographical features, economic conditions, and existing infrastructure. Including these additional controls in our regression model helps ensure that our estimates of the impact on urbanization are directly attributed to the improvement in railway rather than these other factors.

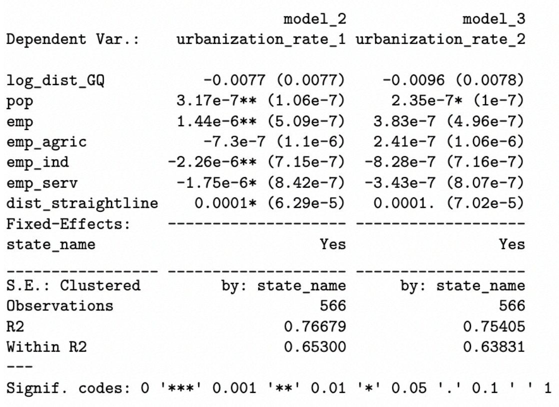
Another concerning issue is pre-existing urbanization trend, for example, we could use data to investigate whether there was any trend in increase in urbanization before railway improvement (before year 2001). We could also test that ‘close_to GQ’ is not correlated with initial urbanization level while controlling same set of controls used in the regression model in part 5. As we could see from the regressions below, the coefficients for ‘close_to_GQ’ are very small and not significant (same for log_dist_GQ), supporting the assumption of no pre-existing trend.



```
                               model_0             model_1
Dependent Var.:   urbanization_rate_1 urbanization_rate_2

close_to_GQ            0.0015 (0.0181)     0.0047 (0.0142)
pop                  3.07e-7 (2.44e-7)   2.23e-7 (1.52e-7)
emp                  1.34e-6 (1.33e-6)   2.27e-7 (8.33e-7)
emp_agric           -6.82e-7 (2.2e-6)    4.69e-7 (1.58e-6)
emp_ind             -2.53e-6 (1.86e-6)  -1.07e-6 (1.17e-6)
emp_serv            -1.89e-6 (2.17e-6)  -2.87e-7 (1.33e-6)
dist_straightline    9.63e-5 (5.88e-5)  8.97e-5. (4.94e-5)
Fixed-Effects:     ------------------- -------------------
state_name                         Yes                 Yes

_________________ ___________________ ___________________
S.E.: Clustered         by: state_name      by: state_name
Observations                       566                 566
R2                             0.60231             0.73595
Within R2                      0.40828             0.61168
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```
                                model_2             model_3
Dependent Var.:     urbanization_rate_1 urbanization_rate_2

log_dist_GQ             -0.0077 (0.0077)    -0.0096 (0.0078)
pop                  3.17e-7** (1.06e-7)     2.35e-7* (1e-7)
emp                  1.44e-6** (5.09e-7)   3.83e-7 (4.96e-7)
emp_agric              -7.3e-7 (1.1e-6)    2.41e-7 (1.06e-6)
emp_ind             -2.26e-6** (7.15e-7)  -8.28e-7 (7.16e-7)
emp_serv             -1.75e-6* (8.42e-7)  -3.43e-7 (8.07e-7)
dist_straightline     0.0001* (6.29e-5)   0.0001. (7.02e-5)
Fixed-Effects:     -------------------- -------------------
state_name                          Yes                 Yes

_________________ ____________________ ___________________
S.E.: Clustered          by: state_name      by: state_name
Observations                        566                 566
R2                              0.76679             0.75405
Within R2                       0.65300             0.63831
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note: the regressions above use 2001 data since we are interested in the pre-existing trend.

5. Empirical strategy and regression

The basic empirical strategy is to compare changes in urbanization for districts that are close to GQ highway relative to those that are not.

Regression model:

$$U_{sdt}=\beta_1 * \ C_{sd} \ +\tau * x_{sdt} \ +\gamma_s + \ \varepsilon_{sdt}$$

Where:

U_{sdt}: The dependent variable representing the urbanization level, for district d, in state s, at time t. This is constructed based on population and employment data, with two measures:

- urbanization_rate_1 = urban population/ (urban population + rural population)
- urbanization_rate_2 = urban employment/ (urban employment + rural employment)

C_{sd}: Indicate whether the district is within 75 km of GQ highway. $C_{sd} = 1$ if district is within 75 km for GQ highway. $C_{sd} = 0$ if district is beyond 75 km for GQ highway.

γs: State-level fixed effect. Control for all time-invariant traits of the state that could influence the outcomes.[1]

x_{sdt}: A series of control variables.[2]

ε_{sdt}: Error term. Standard error is clustered at the state level.

For the continuous effect:

$$U_{sdt}=\beta_1 * \log\left(distance_{sd}\right) + \tau * x_{sdt} + \gamma_{sd} + \varepsilon_{sdt}$$

Where:
$\log\left(distance_{sd}\right)$:the continuous effect of (log) distance

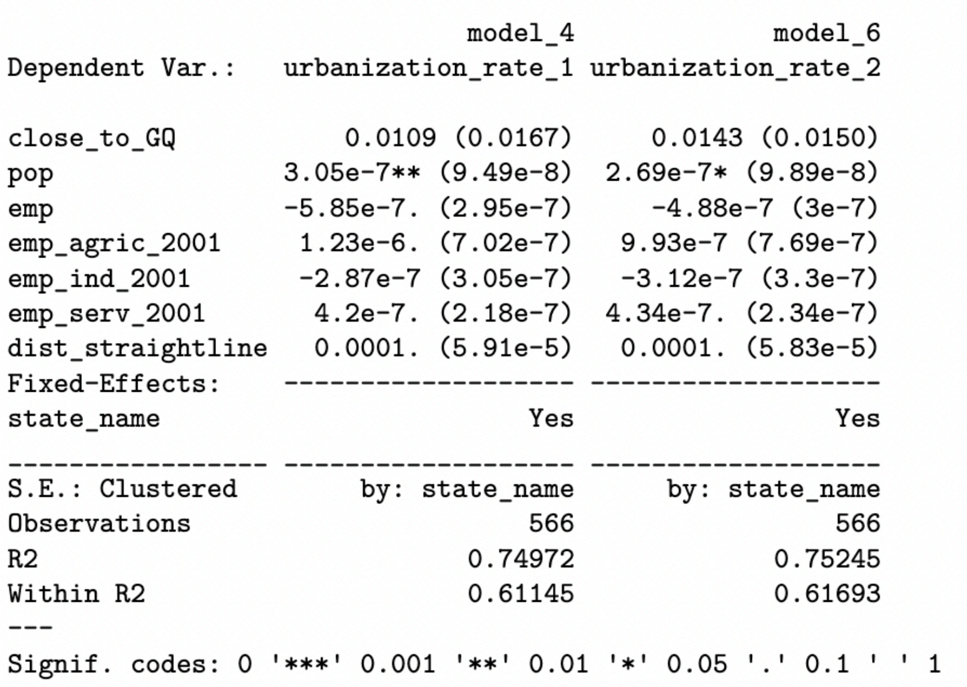
The regressions use 2011 urban area data to capture the impact of highway improvements on urbanization. However, since data on employment in agriculture, industry, and services for 2011 are unavailable, sectoral employment data from 2001 are used to control for initial differences, this could potentially introduce bias to the result.

The coefficient of interest in the regression model is β_1.

[1] While district-level fixed effects could control for all time-invariant traits of the districts that could influence the outcomes, due to data limitations, we only have 1 observation per district level in the 2011 data. Therefore, we cannot apply district-level fixed effects and instead use state-level fixed effects.

[2] The control variables are Population, Total Employment, Employment in Agriculture, Employment in Industry, and Employment in Services, Distance to Straight Line Connecting Node Cities.

Regression result:



```
                              model_4             model_6
Dependent Var.:   urbanization_rate_1 urbanization_rate_2

close_to_GQ           0.0109 (0.0167)     0.0143 (0.0150)
pop               3.05e-7** (9.49e-8)  2.69e-7* (9.89e-8)
emp              -5.85e-7. (2.95e-7)     -4.88e-7 (3e-7)
emp_agric_2001     1.23e-6. (7.02e-7)   9.93e-7 (7.69e-7)
emp_ind_2001      -2.87e-7 (3.05e-7)   -3.12e-7 (3.3e-7)
emp_serv_2001      4.2e-7. (2.18e-7)  4.34e-7. (2.34e-7)
dist_straightline  0.0001. (5.91e-5)   0.0001. (5.83e-5)
Fixed-Effects:    ------------------- -------------------
state_name                        Yes                 Yes
_________________ ___________________ ___________________
S.E.: Clustered        by: state_name      by: state_name
Observations                      566                 566
R2                            0.74972             0.75245
Within R2                     0.61145             0.61693
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

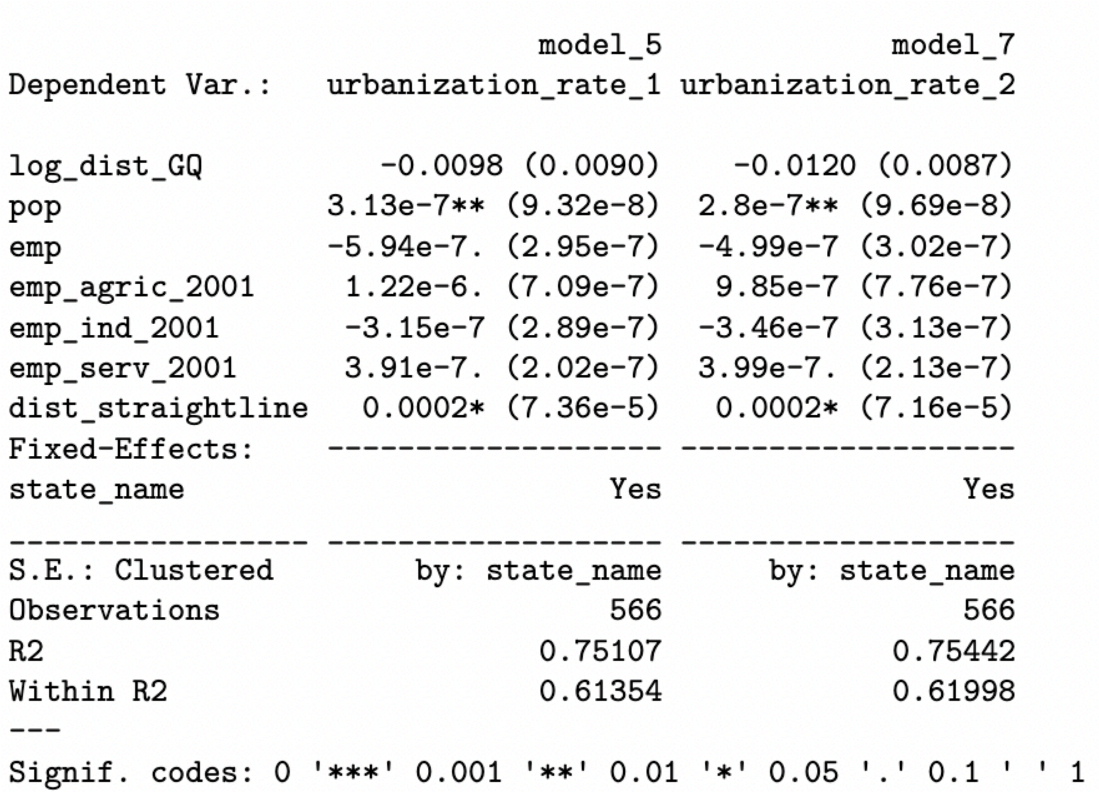
Note: The control variables are Population, Total Employment, Employment in Agriculture, Employment in Industry, and Employment in Services, Distance to Straight Line Connecting Node Cities. Standard error is clustered at state level.

The coefficient 0.0109 suggests that on average, being close to the Golden Quadrilateral (GQ) highway is associated with an increase in 0.0109 units of the urbanization rate (population), holding all else constant. The effect is not statically significant at 0.05 level.

The coefficient 0.0143 suggests that on average, being close to the Golden Quadrilateral (GQ) highway is associated with an increase in 0.0143 units of the urbanization rate (employment), holding all else constant. The effect is not statically significant at 0.05 level.

The treatment group does not show a statistically significant effect on either urbanization rate (both urbanization_rate_1 and urbanization_rate_2).

Continuous effect:



```
                             model_5             model_7
Dependent Var.:   urbanization_rate_1 urbanization_rate_2

log_dist_GQ         -0.0098 (0.0090)    -0.0120 (0.0087)
pop               3.13e-7** (9.32e-8)  2.8e-7** (9.69e-8)
emp               -5.94e-7. (2.95e-7)  -4.99e-7 (3.02e-7)
emp_agric_2001     1.22e-6. (7.09e-7)   9.85e-7 (7.76e-7)
emp_ind_2001       -3.15e-7 (2.89e-7)  -3.46e-7 (3.13e-7)
emp_serv_2001      3.91e-7. (2.02e-7)  3.99e-7. (2.13e-7)
dist_straightline   0.0002* (7.36e-5)   0.0002* (7.16e-5)
Fixed-Effects:    ------------------- -------------------
state_name                        Yes                 Yes

----------------- ------------------- -------------------
S.E.: Clustered        by: state_name      by: state_name
Observations                      566                 566
R2                            0.75107             0.75442
Within R2                     0.61354             0.61998
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

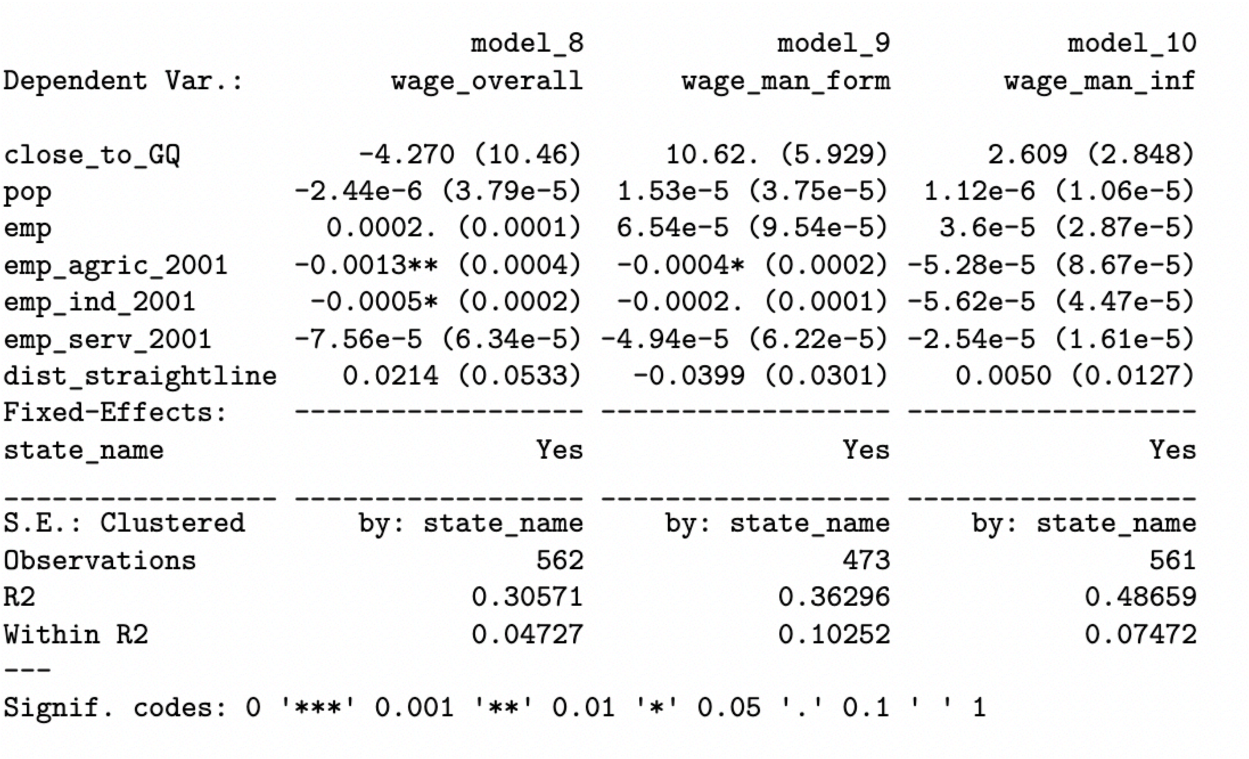
Note: The control variables are Population, Total Employment, Employment in Agriculture, Employment in Industry, and Employment in Services, Distance to Straight Line Connecting Node Cities. Standard error is clustered at state level.

The coefficient -0.0098 suggests that on average, a 1% increase in the distance to the Golden Quadrilateral (GQ) highway is associated with a decrease in the urbanization rate (urbanization_rate_1) by approximately 0.0098 units, holding all else constant. The effect is not statically significant at 0.05 level.

The coefficient -0.0120 suggests that on average, a 1% increase in the distance to the Golden Quadrilateral (GQ) highway is associated with a decrease in the urbanization rate (urbanization_rate_2) by approximately 0.0120 units, holding all else constant. The effect is not statically significant at 0.05 level.

In both models, increased distance from the Golden Quadrilateral highway is associated with a decrease in urbanization rates. However, the effect is not statistically significant.

Effect on wages:



```
                            model_8            model_9           model_10
Dependent Var.:        wage_overall      wage_man_form       wage_man_inf

close_to_GQ          -4.270 (10.46)     10.62. (5.929)      2.609 (2.848)
pop              -2.44e-6 (3.79e-5)  1.53e-5 (3.75e-5)  1.12e-6 (1.06e-5)
emp                0.0002. (0.0001)  6.54e-5 (9.54e-5)   3.6e-5 (2.87e-5)
emp_agric_2001   -0.0013** (0.0004)  -0.0004* (0.0002) -5.28e-5 (8.67e-5)
emp_ind_2001      -0.0005* (0.0002)  -0.0002. (0.0001) -5.62e-5 (4.47e-5)
emp_serv_2001    -7.56e-5 (6.34e-5) -4.94e-5 (6.22e-5) -2.54e-5 (1.61e-5)
dist_straightline   0.0214 (0.0533)   -0.0399 (0.0301)    0.0050 (0.0127)
Fixed-Effects:   ------------------ ------------------ ------------------
state_name                      Yes                Yes                Yes
_________________ __________________ __________________ __________________
S.E.: Clustered      by: state_name     by: state_name     by: state_name
Observations                    562                473                561
R2                          0.30571            0.36296            0.48659
Within R2                   0.04727            0.10252            0.07472
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note: The control variables are Population, Total Employment, Employment in Agriculture, Employment in Industry, and Employment in Services, Distance to Straight Line Connecting Node Cities. Standard error is clustered at state level.

The coefficient -4.270 suggests that on average, being close to the Golden Quadrilateral (GQ) highway is associated with a decrease in the overall wage by approximately 4.27 units, holding all else constant. The effect is not statically significant at 0.05 level.

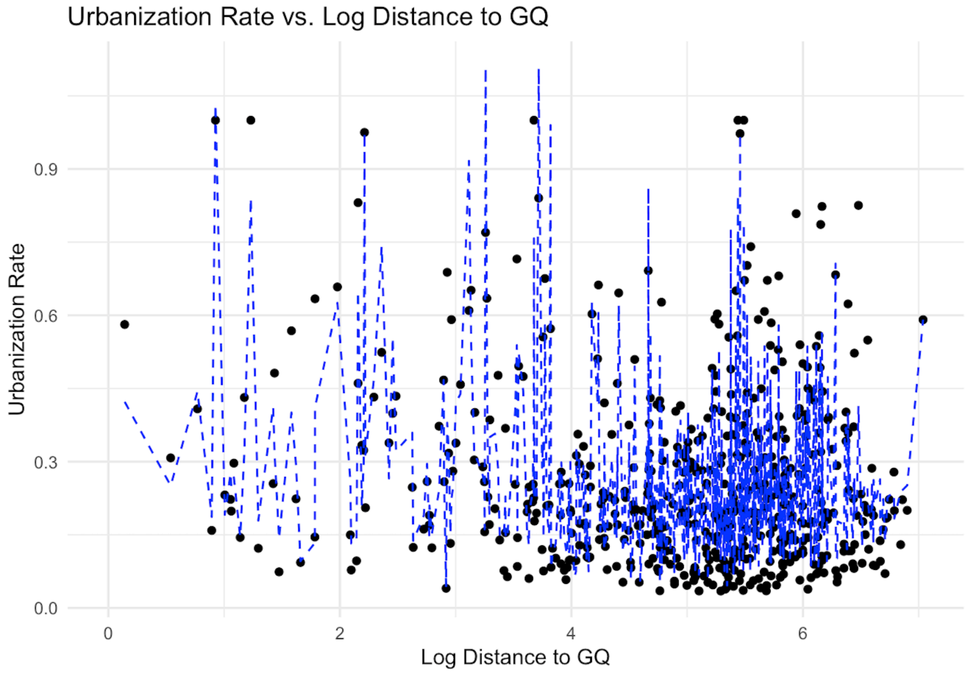
The coefficient 10.62 suggests that on average, being close to the Golden Quadrilateral (GQ) highway is associated with an increase in the wage in the manufacturing formal sector by approximately 10.62 units, holding all else constant. The effect is not statically significant at 0.05 level.

The coefficient 2.609 suggests that on average, being close to the Golden Quadrilateral (GQ) highway is associated with an increase in the wage in the manufacturing informal sector by approximately 2.609 units, holding all else constant. The effect is not statically significant at 0.05 level.

In none of the above models does treatment group show a statistically significant effect on wages (whether overall, in formal manufacturing, or informal manufacturing sectors).

6. Graph

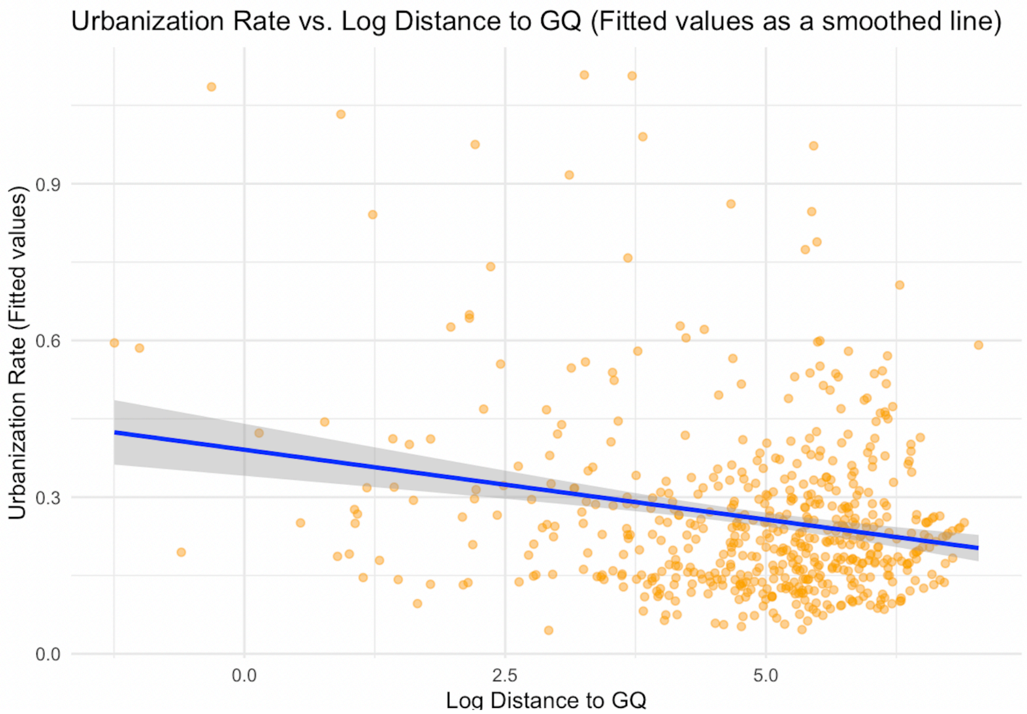
Graph 1



Note: The fitted values are based on regression model 5. The black dots are the actual values. The blue line connected the fitted values.

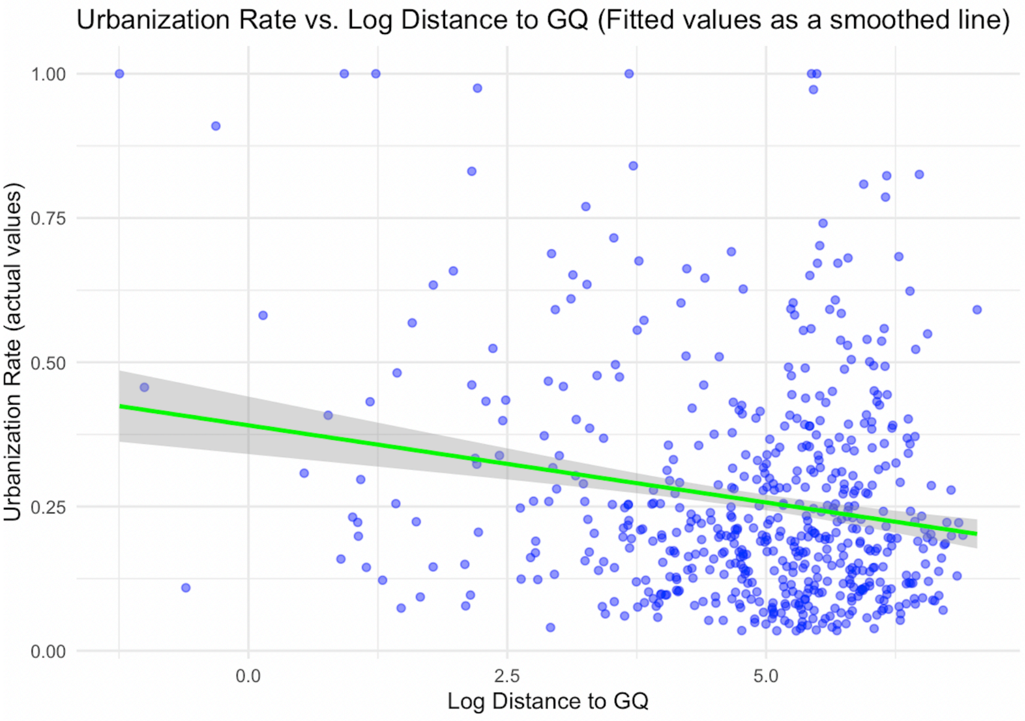
The curve appears to fluctuate, indicating a potential non-linear relationship may exist between log distance to the GQ and urbanization rate.

Graph 2



Note: The orange dots are the fitted values from regression model 5. The blue line shows the fitted values as a linear trend line.

Graph 3



Note: The fitted values are based on regression model 5. The blue dots are the actual values. The green line shows the fitted values as a linear trend line.

The scatter of fitted value data (orange dots, graph 2) around the linear fit suggests that there could be non-linear patterns not fully captured by the linear model. There are clusters of points at various log distances where the urbanization rate varies widely, indicating potential non-linear effects.