Regression Analysis of CPI Data

a. Equations for the various hypotheses:
   1. Zero Order: $y = 8.0714$
   2. Linear: $y = -0.0004x + 8.8689$
   3. Quadratic: $y = 0.0000x^2 - 0.0009x + 9.1720$
   4. Cubic: $y = 0.0000x^3 + 0.0000x^2 - 0.0008x + 9.1223$
   5. Order 8 with 9 points:
      $y = 0.0000x^8 + 0.0000x^7 + 0.0000x^6 + 0.0000x^5 + 0.0000x^4 + 0.0000x^3 - 0.0119x^2 + 1.5796x - 51.1287$

b. The graphs are:

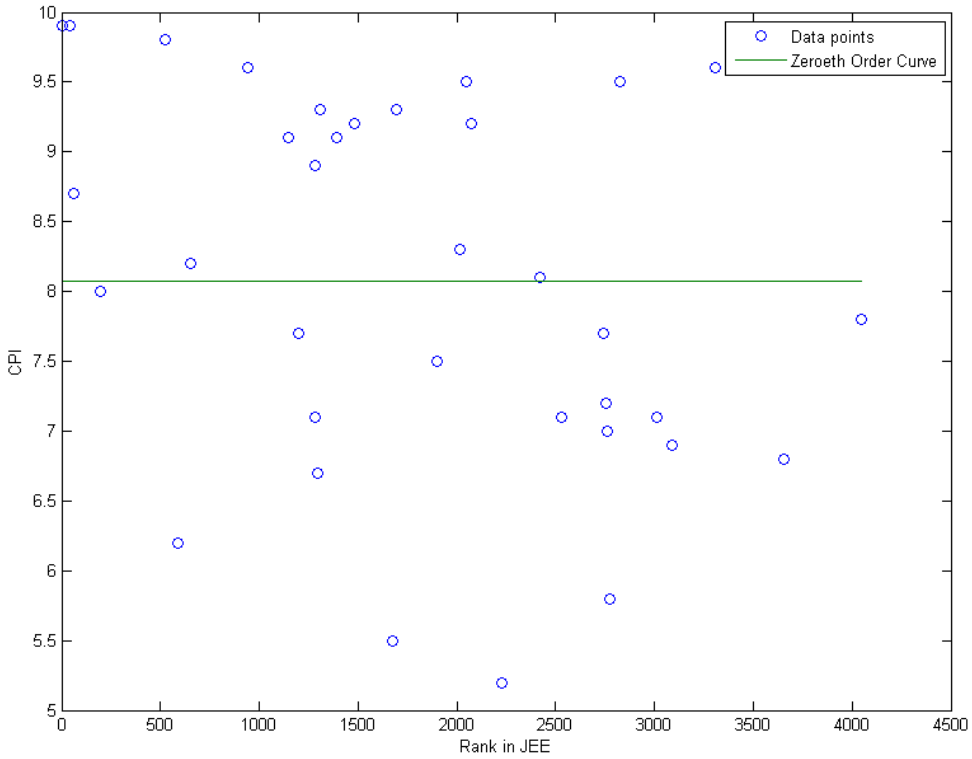

Figure 1: Zeroeth Order Approximation

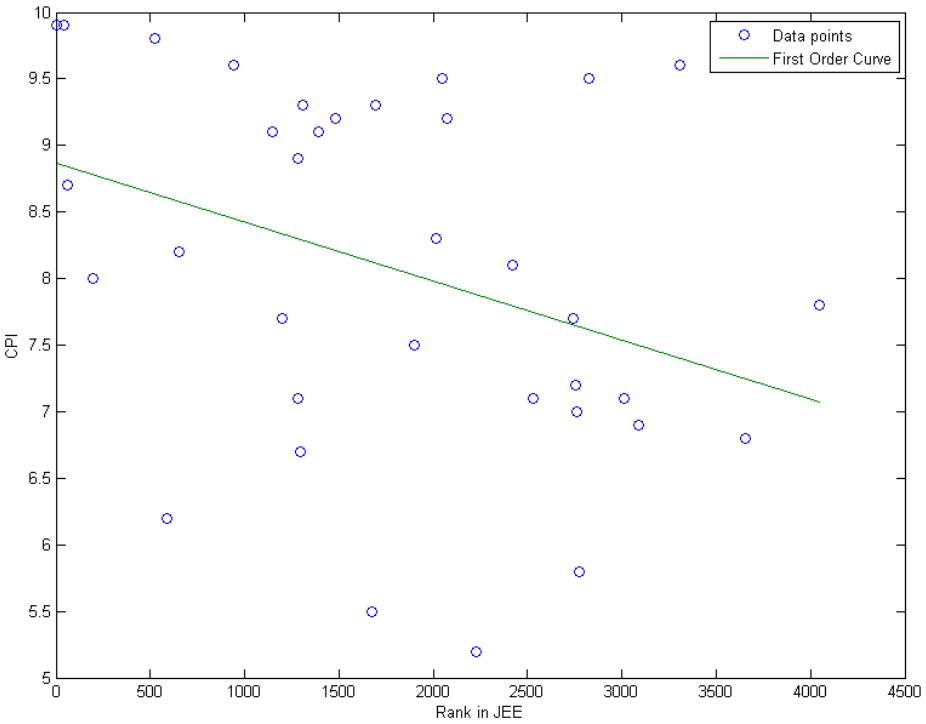

Figure 2: Linear Approximation

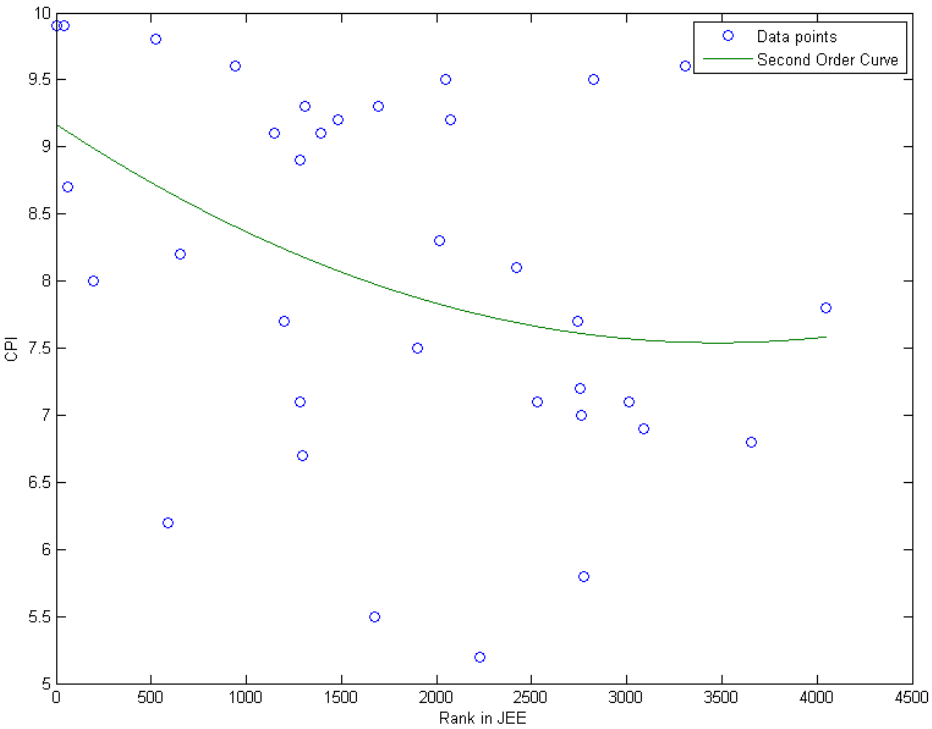

Figure 3: Quadratic Approximation

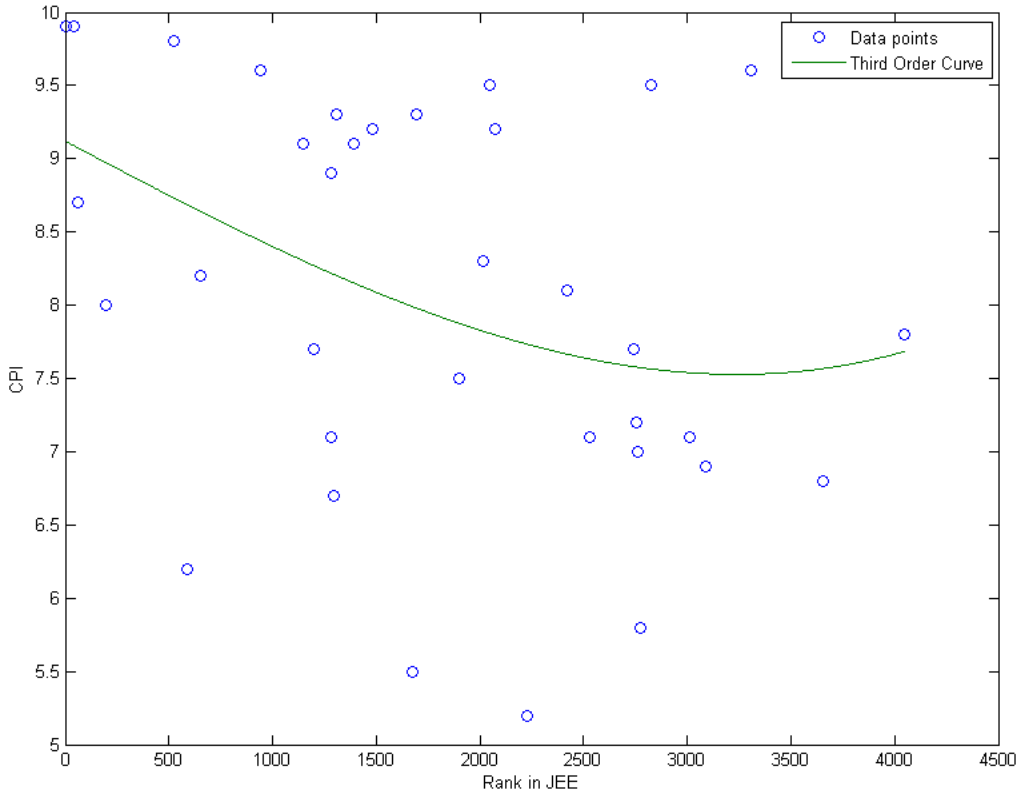

Figure 4: Cubic Approximation

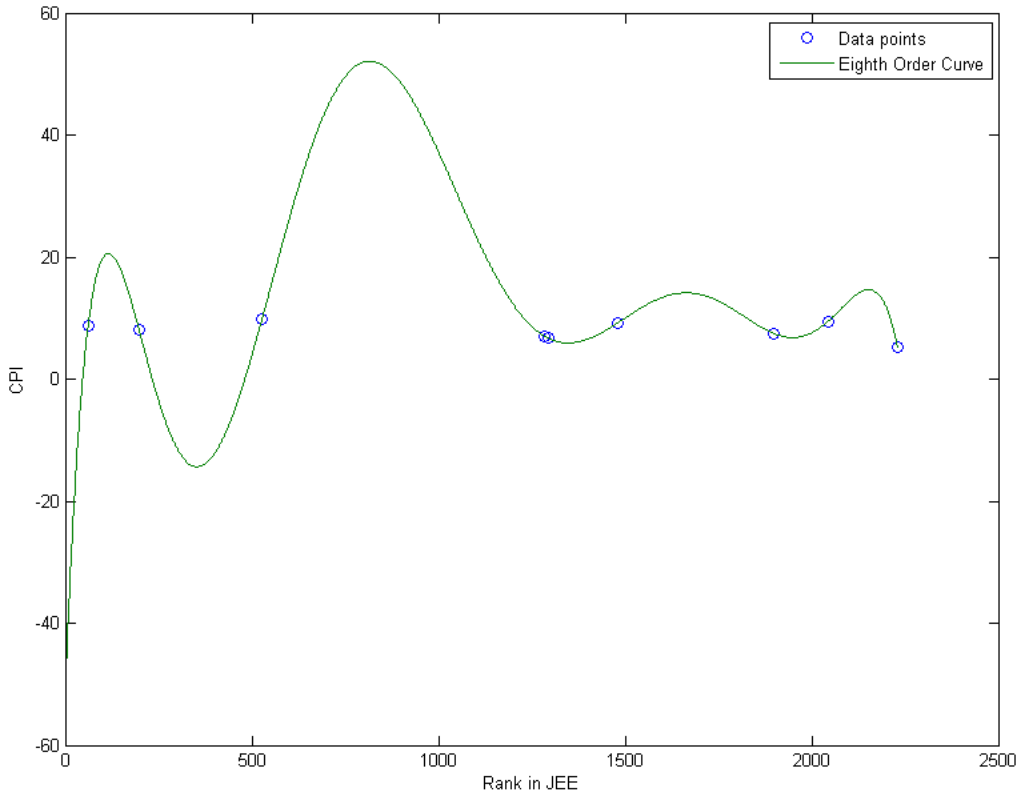

Figure 5: Order 8 Approximation with 9 Points

c. Residual Error:

I have taken the overall residual error as

$\text{Err} = \sum_{i=1}^{n} |\text{y}(x_i) - \text{y}'(x_i)|$,

where x_i is the JEE rank of the i[th] person , $\text{y}(x_i)$ is the corresponding CPI and $\text{y}'(x_i)$ is the estimated CPI

Polynomial Order	Sum of Residual Errors
Zero Order	39.8286
Linear	36.0715
Quadratic	36.1584
Cubic	35.9932

It can be inferred that the residual error sum varies inversely with polynomial order. When we have an exact curve fitting like with a degree eight polynomial to fit nine points, it drops to zero.

d. Coefficients:

Polynomial Order	x^3	x^2	x	1
Zero				8.0714
Linear			−0.0004	8.8689
Quadratic		Approx. 0	−0.0009	9.1720
Cubic	Approx.0	Approx.0	−0.0008	9.1223

The zeroeth order coefficient seems to increase with increase in polynomial order. As the eighth order function is an exact fit for 9 points, the coefficients seem to have no relation and the zeroeth order coefficient is 51.1287.

e. Multivariate Regression Results:

1. Linear: $y = -0.0005 - 0.0578a + 10.0467$
2. Quadratic:

$$y = 0.0000x^2 - 0.1127a^2 + 0.0003ax - 0.0069x + 4.0165a - 25.9279$$

where x =JEE rank and a=age

Polynomial Order	Sum of Residual Errors
Linear	35.9306
Quadratic	35.1004

A minor reduction in error is seen for the multivariate case, probably because the ages lie in [19 20 21] and hence the number of elements is not large. Also, logically one would not expect the age to be related to the CPI.

f. Choice of Regression Scheme:

The residual errors are so close that there is no single scheme I can pick out. Although going for a higher degree polynomial fit in the univariate cases does reduce the residual error, but the higher degree polynomials are badly conditioned. Also, the bivariate regression yields a

similar error although it is computationally more expensive. Based on this data, I would opt for a univariate cubic regression scheme with JEE rank being the predictor variable.