**Statistical Learning Lab**

Assignment - 6

**Non Linear Regression**

Anurag Singha | 22IM30007



Q1) Import the designated data file and Display first few rows of the dataset

**Code**

library(leaps)

data <- read.csv("C:\\Users\\USER\\OneDrive - iitkgp.ac.in\\Desktop\\Stat Lab\\Assignment 6\\poverty.csv")

head(data)

**Output**

Location PovPct Brth15to17 Brth18to19 ViolCrime TeenBrth

1 Alabama 20.1 31.5 88.7 11.2 54.5

2 Alaska 7.1 18.9 73.7 9.1 39.5

3 Arizona 16.1 35.0 102.5 10.4 61.2

4 Arkansas 14.9 31.6 101.7 10.4 59.9

5 California 16.7 22.6 69.1 11.2 41.1

6 Colorado 8.8 26.2 79.1 5.8 47.0

Q2) Data cleaning and pre-processing and Perform preliminary analysis to show how the variables are related to each other. Use scatter plot, box plot etc. to visualize how different variables impact the response variable.

**Code**

# Missing values

colSums(is.na(data))

summary(data)

# Summary statistics

summary(data)

# Pair Plot

pairs(data[, -1], main = "Scatterplot Matrix", pch = 21, bg = "lightblue")

# Box Plot

boxplot(data[, -1], main = "Boxplot of Numerical Variables", col = rainbow(ncol(data)-1), las=2)

**Output**

|  |
| --- |
| > colSums(is.na(data))  Location PovPct Brth15to17 Brth18to19 ViolCrime TeenBrth  0 0 0 0 0 0  > summary(data)  Location PovPct Brth15to17 Brth18to19 ViolCrime TeenBrth  Length:51 Min. : 5.30 Min. : 8.10 Min. : 39.00 Min. : 0.900 Min. :20.00  Class :character 1st Qu.:10.25 1st Qu.:17.25 1st Qu.: 58.30 1st Qu.: 3.900 1st Qu.:33.90  Mode :character Median :12.20 Median :20.00 Median : 69.40 Median : 6.300 Median :39.50  Mean :13.12 Mean :22.28 Mean : 72.02 Mean : 7.855 Mean :42.24  3rd Qu.:15.80 3rd Qu.:28.10 3rd Qu.: 87.95 3rd Qu.: 9.450 3rd Qu.:52.60  Max. :25.30 Max. :44.80 Max. :104.30 Max. :65.000 Max. :69.10  > # Summary statistics  > summary(data)  Location PovPct Brth15to17 Brth18to19 ViolCrime TeenBrth  Length:51 Min. : 5.30 Min. : 8.10 Min. : 39.00 Min. : 0.900 Min. :20.00  Class :character 1st Qu.:10.25 1st Qu.:17.25 1st Qu.: 58.30 1st Qu.: 3.900 1st Qu.:33.90  Mode :character Median :12.20 Median :20.00 Median : 69.40 Median : 6.300 Median :39.50  Mean :13.12 Mean :22.28 Mean : 72.02 Mean : 7.855 Mean :42.24  3rd Qu.:15.80 3rd Qu.:28.10 3rd Qu.: 87.95 3rd Qu.: 9.450 3rd Qu.:52.60  Max. :25.30 Max. :44.80 Max. :104.30 Max. :65.000 Max. :69.10 |
|  |
| |  | | --- | | > | |

**Explanation and Insights:**

We are considering **PovPct** (poverty percentage) as the response variable.

**Scatterplot Matrix:** The scatterplot matrix reveals that nearly all pairs of variables exhibit a linear relationship.

**Boxplots:** The boxplots allow us to examine the distribution of each variable. Notably, **ViolCrime**, **PovPct**, and **Brth15to17** display extreme outliers with widely dispersed values.

**Data Quality:** The dataset contains no missing values, ensuring it is clean and ready for modeling.

**PovPct Range:** Poverty percentage varies between **5.3% and 25.3%**, with a mean value of **13.12%**.

**Variable Relationship:** There is a strong positive correlation between **higher poverty rates** and **higher teenage birth rates**, indicating a significant association between these variables.

Q3,4)Convert categorical inputs or consider it while fitting the data and Fit a linear model first

**#Code**

linear\_model <- lm(PovPct ~ TeenBrth + Brth15to17 + Brth18to19 + ViolCrime, data = data)

summary(linear\_model)

# Scatter Plot: Poverty % vs Teen Birth Rate (Overall)

plot(data$TeenBrth, data$PovPct,

main = "Linear Model: Poverty % vs Teen Birth Rate (Overall)",

xlab = "Teen Birth Rate",

ylab = "Poverty Percentage",

pch = 16, col = "blue")

# Scatter Plot: Poverty % vs Birth Rate (Ages 15-17)

plot(data$Brth15to17, data$PovPct,

main = "Linear Model: Poverty % vs Birth Rate (15-17)",

xlab = "Birth Rate (15-17)",

ylab = "Poverty Percentage",

pch = 16, col = "green")

# Scatter Plot: Poverty % vs Birth Rate (Ages 18-19)

plot(data$Brth18to19, data$PovPct,

main = "Linear Model: Poverty % vs Birth Rate (18-19)",

xlab = "Birth Rate (18-19)",

ylab = "Poverty Percentage",

pch = 16, col = "navyblue")

# Scatter Plot: Poverty % vs Violent Crime Rate

plot(data$ViolCrime, data$PovPct,

main = "Linear Model: Poverty % vs Violent Crime Rate",

xlab = "Violent Crime Rate",

ylab = "Poverty Percentage",

pch = 16, col = "skyblue")

**#Output**

**#A)Summary**

Call:

lm(formula = PovPct ~ TeenBrth + Brth15to17 + Brth18to19 + ViolCrime,

data = data)

Residuals:

Min 1Q Median 3Q Max

-5.5239 -1.9763 -0.1048 1.6729 5.6012

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.22349 1.82549 3.409 0.00136 \*\*

TeenBrth 1.81957 0.66635 2.731 0.00893 \*\*

Brth15to17 -0.45769 0.44681 -1.024 0.31102

Brth18to19 -0.82144 0.27311 -3.008 0.00426 \*\*

ViolCrime -0.07786 0.06683 -1.165 0.24997

---

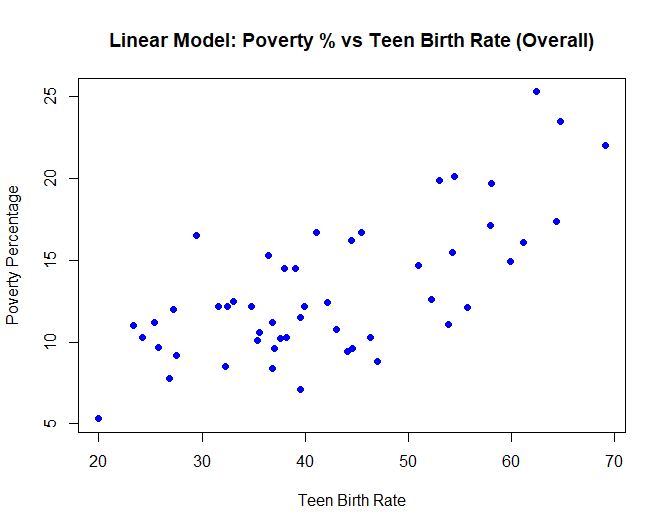
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

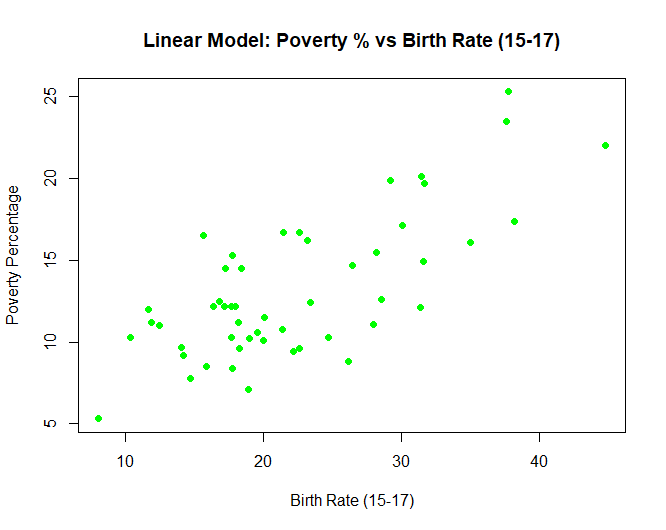
Residual standard error: 2.773 on 46 degrees of freedom

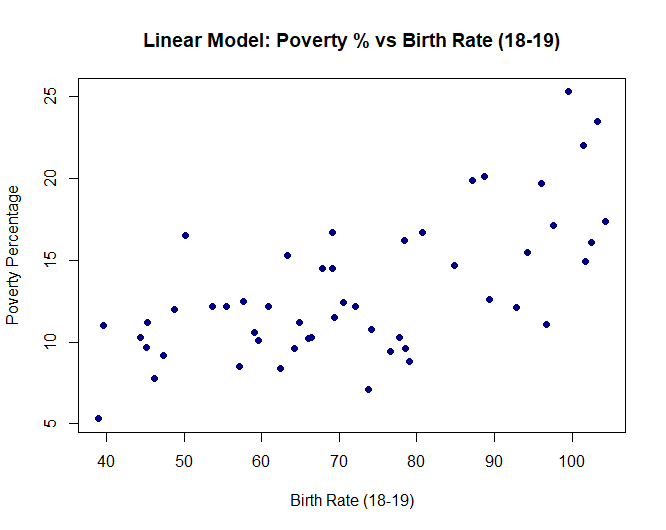
Multiple R-squared: 0.6132, Adjusted R-squared: 0.5796

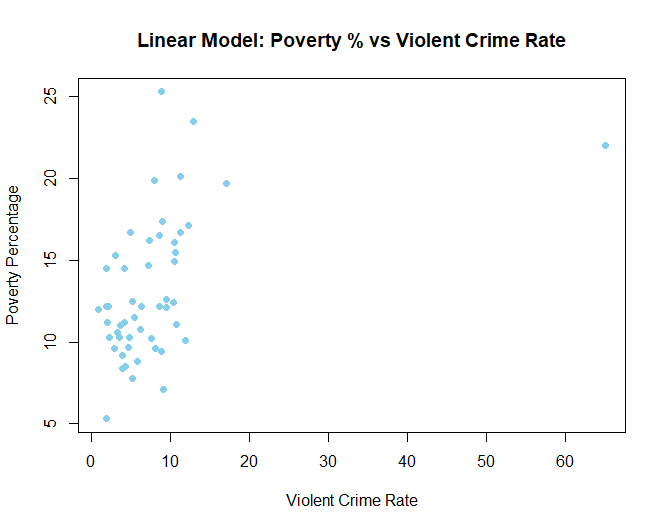
F-statistic: 18.23 on 4 and 46 DF, p-value: 4.916e-09

**#B)Response vs Different features**









In the linear model, **TeenBrth** and **Brth18to19** emerge as significant predictors.

The **p-value (4.916e-09)** indicates strong statistical significance.

The **adjusted R-squared value (0.5796)** suggests the model provides a **moderate fit** to the data.

Q5)Fit a Polynomial Regression Model

**#Code**

poly\_model <- lm(PovPct ~ poly(TeenBrth, 2) + poly(Brth15to17, 2) + poly(Brth18to19, 2) + poly(ViolCrime, 2), data = data)

# Display the summary of the model

summary(poly\_model)

**#Output**

Call:

lm(formula = PovPct ~ poly(TeenBrth, 2) + poly(Brth15to17, 2) +

poly(Brth18to19, 2) + poly(ViolCrime, 2), data = data)

Residuals:

Min 1Q Median 3Q Max

-5.458 -1.797 0.063 1.322 5.407

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 13.1176 0.3832 34.233 < 2e-16 \*\*\*

poly(TeenBrth, 2)1 204.6820 67.7293 3.022 0.00426 \*\*

poly(TeenBrth, 2)2 -24.3704 39.9370 -0.610 0.54500

poly(Brth15to17, 2)1 -47.8460 28.9210 -1.654 0.10551

poly(Brth15to17, 2)2 24.1395 20.5479 1.175 0.24670

poly(Brth18to19, 2)1 -137.9250 45.9011 -3.005 0.00447 \*\*

poly(Brth18to19, 2)2 8.4431 23.0546 0.366 0.71604

poly(ViolCrime, 2)1 -6.3508 5.0363 -1.261 0.21426

poly(ViolCrime, 2)2 -11.1218 5.4067 -2.057 0.04593 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.737 on 42 degrees of freedom

Multiple R-squared: 0.6562, Adjusted R-squared: 0.5907

F-statistic: 10.02 on 8 and 42 DF, p-value: 1.127e-07

Q6) Analyze the fitted model using ANOVA

**#Code**

anova(linear\_model, poly\_model)

**#Output**

Analysis of Variance Table

Model 1: PovPct ~ TeenBrth + Brth15to17 + Brth18to19 + ViolCrime

Model 2: PovPct ~ poly(TeenBrth, 2) + poly(Brth15to17, 2) + poly(Brth18to19,

2) + poly(ViolCrime, 2)

Res.Df RSS Df Sum of Sq F Pr(>F)

1 46 353.83

2 42 314.52 4 39.313 1.3124 0.281

Q7) Select best fit degree polynomial

**#Code**

for (degree in 1:4) {

model <- lm(PovPct ~ poly(TeenBrth, degree) + poly(Brth15to17, degree) + poly(Brth18to19, degree) + poly(ViolCrime, degree),

data = data)

cat("Degree:", degree, "AIC:", AIC(model), "\n")

}

for (degree in 1:4) {

model <- lm(PovPct ~ poly(TeenBrth, degree) + poly(Brth15to17, degree) + poly(Brth18to19, degree) + poly(ViolCrime, degree),

data = data)

anova\_results <- anova(linear\_model, model)

print(anova\_results)

}

**#Output**

Degree: 1 AIC: 255.5184

Degree: 2 AIC: 257.5118

Degree: 3 AIC: 260.7275

Degree: 4 AIC: 262.1818

Analysis of Variance Table

Model 1: PovPct ~ TeenBrth + Brth15to17 + Brth18to19 + ViolCrime

Model 2: PovPct ~ poly(TeenBrth, degree) + poly(Brth15to17, degree) +

poly(Brth18to19, degree) + poly(ViolCrime, degree)

Res.Df RSS Df Sum of Sq F Pr(>F)

1 46 353.83

2 46 353.83 0 1.0232e-12

Analysis of Variance Table

Model 1: PovPct ~ TeenBrth + Brth15to17 + Brth18to19 + ViolCrime

Model 2: PovPct ~ poly(TeenBrth, degree) + poly(Brth15to17, degree) +

poly(Brth18to19, degree) + poly(ViolCrime, degree)

Res.Df RSS Df Sum of Sq F Pr(>F)

1 46 353.83

2 42 314.52 4 39.313 1.3124 0.281

Analysis of Variance Table

Model 1: PovPct ~ TeenBrth + Brth15to17 + Brth18to19 + ViolCrime

Model 2: PovPct ~ poly(TeenBrth, degree) + poly(Brth15to17, degree) +

poly(Brth18to19, degree) + poly(ViolCrime, degree)

Res.Df RSS Df Sum of Sq F Pr(>F)

1 46 353.83

2 38 286.35 8 67.476 1.1193 0.3726

Analysis of Variance Table

Model 1: PovPct ~ TeenBrth + Brth15to17 + Brth18to19 + ViolCrime

Model 2: PovPct ~ poly(TeenBrth, degree) + poly(Brth15to17, degree) +

poly(Brth18to19, degree) + poly(ViolCrime, degree)

Res.Df RSS Df Sum of Sq F Pr(>F)

1 46 353.83

2 34 251.86 12 101.97 1.1471 0.3576

### ****Explanation and Insights****

* **AIC Values:** The model with **degree 1** (linear model) has the **lowest AIC**, indicating it is the **best-fitting model**.
* **ANOVA Comparison:** All **p-values** are greater than **0.05**, suggesting that **higher-degree polynomials** do not significantly improve the model.

**Conclusion:** The **linear model (degree = 1)** is the **optimal choice** due to its simplicity and comparable performance.

Q8) Fit spline with varying knots and GAM model.

**#Code**

# Load required libraries

library(mgcv)

library(splines)

# Fit a Generalized Additive Model (GAM)

gam\_model <- gam(PovPct ~ s(TeenBrth) + s(Brth15to17) + s(Brth18to19) + s(ViolCrime), data = data)

summary(gam\_model)

# Define knot positions (replace with appropriate values)

knots <- c(quantile(data$TeenBrth, probs = c(0.25, 0.5, 0.75)))

# Fit a spline-based Linear Model (LM)

spline\_model <- lm(PovPct ~ bs(TeenBrth, knots = knots) + bs(Brth15to17, knots = knots) +

bs(Brth18to19, knots = knots) + bs(ViolCrime, knots = knots), data = data)

summary(spline\_model)

**#Output**

Family: gaussian

Link function: identity

Formula:

PovPct ~ s(TeenBrth) + s(Brth15to17) + s(Brth18to19) + s(ViolCrime)

Parametric coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 13.1176 0.3777 34.73 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:

edf Ref.df F p-value

s(TeenBrth) 0.9999 0.9999 9.471 0.00359 \*\*

s(Brth15to17) 1.7630 2.2300 1.624 0.20315

s(Brth18to19) 0.9999 0.9999 10.432 0.00235 \*\*

s(ViolCrime) 2.0816 2.1533 1.698 0.16608

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Rank: 36/37

R-sq.(adj) = 0.602 Deviance explained = 64.9%

GCV = 8.4029 Scale est. = 7.2752 n = 51

Call:

lm(formula = PovPct ~ bs(TeenBrth, knots = knots) + bs(Brth15to17,

knots = knots) + bs(Brth18to19, knots = knots) + bs(ViolCrime,

knots = knots), data = data)

Residuals:

Min 1Q Median 3Q Max

-4.4892 -1.6287 0.0157 0.9989 5.7813

Coefficients: (5 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) -44.554 40.797 -1.092 0.283

bs(TeenBrth, knots = knots)1 3.013 38.162 0.079 0.938

bs(TeenBrth, knots = knots)2 13.264 49.554 0.268 0.791

bs(TeenBrth, knots = knots)3 28.410 58.620 0.485 0.631

bs(TeenBrth, knots = knots)4 56.224 54.953 1.023 0.314

bs(TeenBrth, knots = knots)5 79.785 58.265 1.369 0.181

bs(TeenBrth, knots = knots)6 79.264 66.963 1.184 0.246

bs(Brth15to17, knots = knots)1 12.346 33.916 0.364 0.718

bs(Brth15to17, knots = knots)2 -26.252 22.447 -1.170 0.251

bs(Brth15to17, knots = knots)3 -9.973 26.874 -0.371 0.713

bs(Brth15to17, knots = knots)4 -13.941 58.629 -0.238 0.814

bs(Brth15to17, knots = knots)5 -17.302 34.860 -0.496 0.623

bs(Brth15to17, knots = knots)6 NA NA NA NA

bs(Brth18to19, knots = knots)1 50.910 40.510 1.257 0.218

bs(Brth18to19, knots = knots)2 52.083 35.192 1.480 0.149

bs(Brth18to19, knots = knots)3 49.986 30.639 1.631 0.113

bs(Brth18to19, knots = knots)4 34.373 27.613 1.245 0.223

bs(Brth18to19, knots = knots)5 29.627 18.415 1.609 0.118

bs(Brth18to19, knots = knots)6 NA NA NA NA

bs(ViolCrime, knots = knots)1 -15.121 14.592 -1.036 0.308

bs(ViolCrime, knots = knots)2 65.419 47.319 1.383 0.177

bs(ViolCrime, knots = knots)3 -243.739 225.846 -1.079 0.289

bs(ViolCrime, knots = knots)4 NA NA NA NA

bs(ViolCrime, knots = knots)5 NA NA NA NA

bs(ViolCrime, knots = knots)6 NA NA NA NA

Residual standard error: 2.853 on 31 degrees of freedom

Multiple R-squared: 0.7242, Adjusted R-squared: 0.5551

F-statistic: 4.283 on 19 and 31 DF, p-value: 0.0001701

### ****Explanation and Insights****

* **Spline Regression:** Offers **local flexibility** by dividing the function into multiple segments, allowing for better modeling of complex relationships.
* **GAM Models:** Capture **non-linear effects**, providing greater adaptability compared to polynomial regression.
* **Model Comparison:** The **GAM model** fits the data **better** than the spline model, as indicated by its **higher adjusted R-squared** value.