**Section B part B**

**Codes:**

```
# Filter the data for the bowler named 'Sandeep Sharma'
Sandeep_Sharma_data = ipl_bbbc[ipl_bbbc["Bowler"] == "Sandeep Sharma"]

# Group by year and sum the wickets
total_wicket_for_player = Sandeep_Sharma_data.groupby('year')["wicket_confirmation"].sum()
total_wicket_for_player

# Convert to dictionary
total_wicket_for_player_dict = total_wicket_for_player.to_dict()
print(total_wicket_for_player_dict)

# Convert dictionary to pandas Series for fitting distribution
data1 = pd.Series(total_wicket_for_player_dict)

def get_best_distribution(data):
    dist_names = ['alpha', 'beta', 'betaprime', 'burr12', 'crystalball',
                  'dgamma', 'dweibull', 'erlang', 'exponnorm', 'f', 'fatiguelife',
                  'gamma', 'gengamma', 'gumbel_l', 'johnsonsb', 'kappa4',
                  'lognorm', 'nct', 'norm', 'norminvgauss', 'powernorm', 'rice',
                  'recipinvgauss', 't', 'trapz', 'truncnorm']
    dist_results = []
    params = {}

    # Iterate over each distribution and fit to data
    for dist_name in dist_names:
        dist = getattr(st, dist_name)
        param = dist.fit(data)
        params[dist_name] = param
```

```
        # Applying the Kolmogorov-Smirnov test
        D, p = st.kstest(data, dist_name, args=param)
        print("p value for " + dist_name + " = " + str(p))
        dist_results.append((dist_name, p))

    # Select the best fitted distribution based on p-value
    best_dist, best_p = max(dist_results, key=lambda item: item[1])

    # Print results
    print("\nBest fitting distribution: " + str(best_dist))
    print("Best p value: " + str(best_p))
    print("Parameters for the best fit: " + str(params[best_dist]))

    return best_dist, best_p, params[best_dist]

# Call the function with your data
import warnings
warnings.filterwarnings('ignore')
best_dist_name, best_p_val, best_params = get_best_distribution(data)
```

**Results:**

p value for alpha = 0.23206295530058174
p value for beta = 0.8398634844067916
p value for betaprime = 0.4315619419619354
p value for burr12 = 0.8685690891244013
p value for crystalball = 0.4837457465535032
p value for dgamma = 0.21249535404805697
p value for dweibull = 0.6372195760638604
p value for erlang = 0.4237297019263757
p value for exponnorm = 0.4837599411660918
p value for f = 0.4844568641171857
p value for fatiguelife = 0.471335531921644
p value for gamma = 0.43553664271811254
p value for gengamma = 0.059546646538934045
p value for gumbel_l = 0.8794762907119807

p value for johnsonsb = 0.005410814382650009
p value for kappa4 = 0.17280457566915897
p value for lognorm = 0.003791096793123838
p value for nct = 0.4788570165582282
p value for norm = 0.4837456295454795
p value for norminvgauss = 0.9423140097941896
p value for powernorm = 0.8210581109157695
p value for rice = 0.4253883133377433
p value for recipinvgauss = 0.4675789073912685
p value for t = 0.48378274795205
p value for trapz = 0.8190369740633036
p value for truncnorm = 0.48374596741183407

Best fitting distribution: norminvgauss
Best p value: 0.9423140097941896
Parameters for the best fit: (77.40288768129699, -77.29259785298325, 23.91600662863827, 0.6274202359666075)

**Interpretations:**

**Distribution Used for Fitting**

The distribution used for fitting Sandeep Sharma's performance metrics is the **normal inverse Gaussian (norminvgauss)** distribution.

**Interpretation of the Fitted Distribution**

The norminvgauss distribution was determined to be the best fit for the data based on the highest p-value (0.9423140097941896) from the Kolmogorov-Smirnov (K-S) test. This high p-value suggests that the null hypothesis (i.e., the data follows the norminvgauss distribution) cannot be rejected, indicating a good fit.

The normal inverse Gaussian distribution is flexible and can model skewed and heavy-tailed data, which is suitable for cricket performance metrics where variations can be significant from year to year.

**Limitations and Assumptions**

1. **Sample Size**: The performance data might be limited to specific years, which can affect the robustness of the fit. A larger dataset spanning more years could provide a more accurate representation.
2. **External Factors**: Cricket performance can be influenced by numerous factors such as changes in team composition, match conditions, injuries, etc., which are not accounted for in the distribution fitting.
3. **Stationarity Assumption**: The fitting assumes that the underlying process generating the data is stationary. However, player performance can evolve due to training, experience, or age, which can introduce non-stationarity.
4. **Distribution Choice**: While the norminvgauss distribution fits well based on the p-value, it is essential to consider if the distribution makes practical sense in the context

of the domain. For example, cricket performance data often exhibit overdispersion and skewness, for which other distributions like the negative binomial or Poisson-gamma might also be considered.

**Proposed Adjustments and Alternative Distributions**

1. **Negative Binomial Distribution**: Given the overdispersion typically present in count data (like wickets), the negative binomial distribution could be an alternative. It accounts for overdispersion and might provide a better fit for the variability seen in cricket performance.
2. **Zero-Inflated Models**: If the data contains many zero values (e.g., matches where no wickets were taken), zero-inflated models could be appropriate. These models can better handle the excess zeros by combining a count distribution with a point mass at zero.
3. **Generalized Linear Models (GLMs)**: GLMs with appropriate link functions can model the count nature of wicket data while accounting for covariates that influence performance (e.g., match location, opposition strength).

**Section A Part B**

**Results:**

**The final model equation is below:**

```
TARGET_deathRate ~ avgDeathsPerYear + povertyPercent + incidenceRate +
   AvgHouseholdSize + avgAnnCount + incidenceRate + PercentMarried +
   PctNoHS18_24 + PctHS18_24 + PctHS25_Over + PctBachDeg25_Over +
   PctUnemployed16_Over + PctPrivateCoverage + PctEmpPrivCoverage +
   PctWhite + PctOtherRace + PctMarriedHouseholds + MedianAgeMale +
   avgAnnCount
```

- **Adjusted R-squared:** 0.435
- **Root Mean Squared Error (RMSE):** 20.18667

```
lm(formula = formula, data = data)

Residuals:
   Min     1Q  Median     3Q    Max
-75.383 -10.382  -0.382  10.659 110.222

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       145.430040  28.259886   5.146 3.66e-07 ***
avgDeathsPerYear    0.007625   0.004851   1.572 0.116570
povertyPercent      0.095192   0.311134   0.306 0.759753
incidenceRate       0.167262   0.016927   9.882  < 2e-16 ***
AvgHouseholdSize    0.262208   2.218974   0.118 0.905977
```

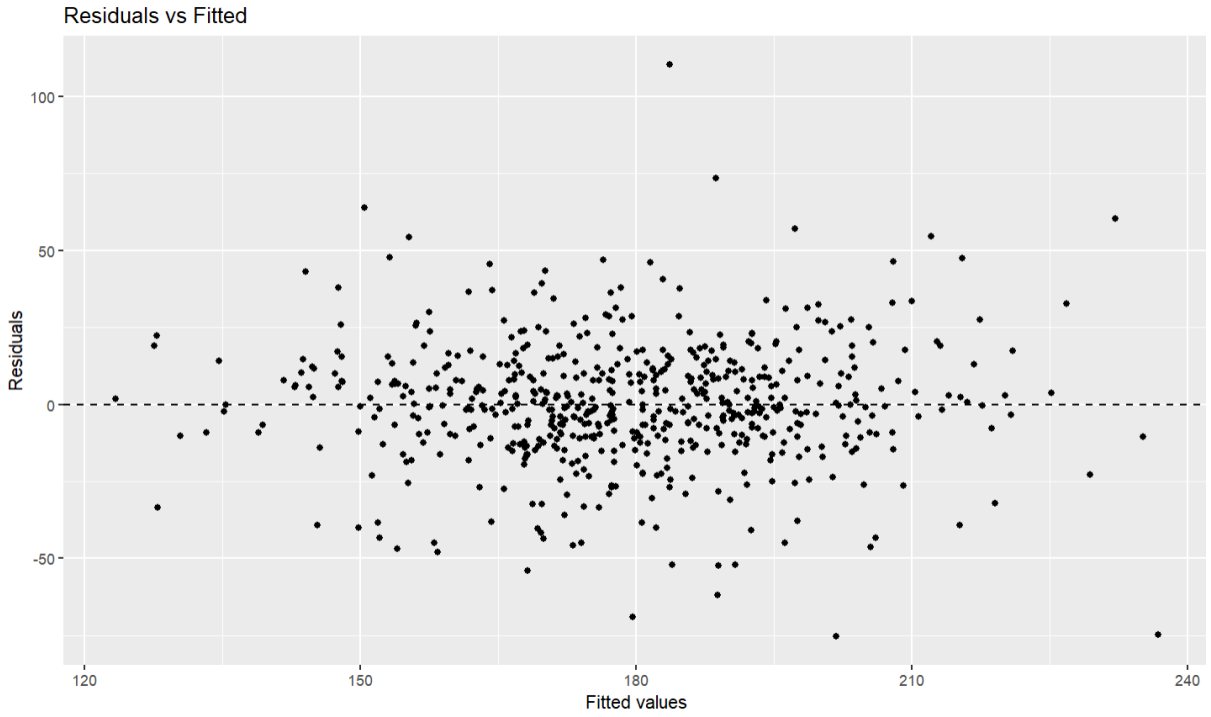
```
avgAnnCount           -0.002826   0.001717  -1.645 0.100449
PercentMarried         1.268732   0.358492   3.539 0.000434 ***
PctNoHS18_24          -0.152387   0.135912  -1.121 0.262667
PctHS18_24             0.203987   0.111312   1.833 0.067385 .
PctHS25_Over           0.445473   0.240050   1.856 0.064002 .
PctBachDeg25_Over     -1.268364   0.358784  -3.535 0.000440 ***
PctUnemployed16_Over   1.026472   0.382415   2.684 0.007481 **
PctPrivateCoverage    -0.659422   0.226694  -2.909 0.003768 **
PctEmpPrivCoverage     0.525731   0.193828   2.712 0.006881 **
PctWhite              -0.044860   0.085918  -0.522 0.601780
PctOtherRace          -1.199937   0.303050  -3.960 8.45e-05 ***
PctMarriedHouseholds  -1.425015   0.322408  -4.420 1.18e-05 ***
MedianAgeMale         -0.536391   0.257766  -2.081 0.037884 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.5 on 573 degrees of freedom
Multiple R-squared:  0.4512,            Adjusted R-squared:  0.435
F-statistic: 27.72 on 17 and 573 DF,  p-value: < 2.2e-16
```

```
> cat("Adjusted R-squared:", adj_r_squared, "\n")
Adjusted R-squared: 0.4349619
> cat("RMSE:", rmse, "\n")
RMSE: 20.18667
```

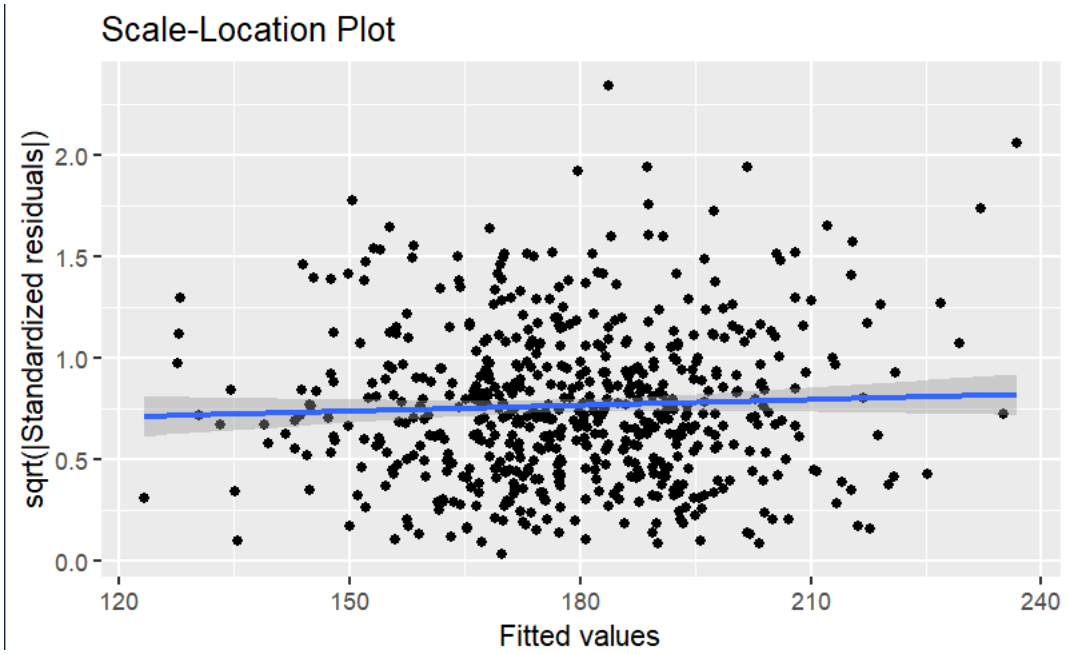
- **Residual Standard Error:** 20.5 on 573 degrees of freedom.
- **Multiple R-squared:** 0.4512.
- **Adjusted R-squared:** 0.435.
- **F-statistic:** 27.72 on 17 and 573 degrees of freedom, with a p-value of < 2.2e-16.

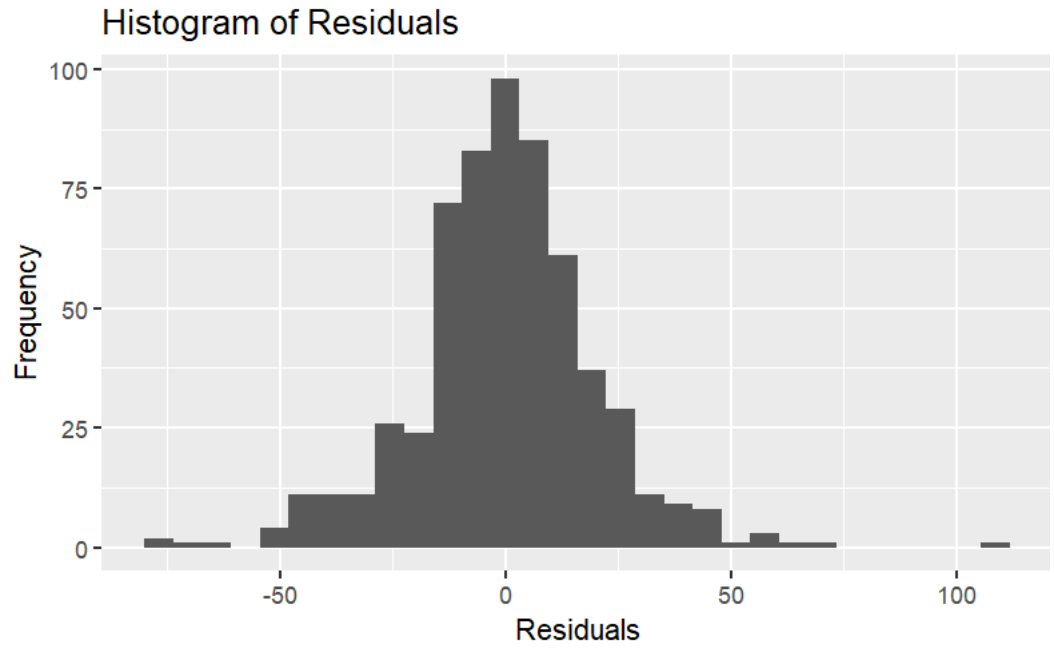
**Key Takeaways**

- **Adjusted R-squared (0.435):** This indicates that approximately 43.5% of the variability in the target variable (TARGET_deathRate) can be explained by the model. Although this isn't a very high value, it does suggest that the model captures a significant portion of the variance.
- **RMSE (20.18667):** The Root Mean Squared Error of 20.18667 indicates the average deviation of the observed death rates from the values predicted by the model. A lower RMSE would be desirable as it implies better predictive accuracy.
- **F-statistic (27.72) and p-value (< 2.2e-16):** The high F-statistic and very low p-value indicate that the model as a whole is statistically significant. This means that the combination of predictors is a significant predictor of the target variable.




```
        Durbin-Watson test

data:  model
DW = 2.0529, p-value = 0.6937
alternative hypothesis: true autocorrelation is greater than 0

> # Breusch-Pagan test for heteroskedasticity
> bptest(model)

        studentized Breusch-Pagan test

data:  model
BP = 57.074, df = 17, p-value = 3.168e-06
```




```
vif(model)
    avgDeathsPerYear       povertyPercent        incidenceRate
           16.576731             5.256236             1.189414
    AvgHouseholdSize         avgAnnCount       PercentMarried
            1.401168            16.187430             7.544552
       PctNoHS18_24          PctHS18_24          PctHS25_Over
            1.599390             1.518369             3.806755
  PctBachDeg25_Over PctUnemployed16_Over    PctPrivateCoverage
            4.943159             2.251072             7.885221
 PctEmpPrivCoverage            PctWhite         PctOtherRace
```

```
          4.803862            2.473732            1.482790
PctMarriedHouseholds        MedianAgeMale
          5.682878            2.443385
> # Interpretation of the model output
> # Summarize the results and explain the significance of the coefficients
> summary(model)

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

Residuals:
    Min      1Q  Median      3Q     Max
-75.383 -10.382  -0.382  10.659 110.222

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         145.430040  28.259886   5.146 3.66e-07 ***
avgDeathsPerYear      0.007625   0.004851   1.572 0.116570
povertyPercent        0.095192   0.311134   0.306 0.759753
incidenceRate         0.167262   0.016927   9.882  < 2e-16 ***
AvgHouseholdSize      0.262208   2.218974   0.118 0.905977
avgAnnCount          -0.002826   0.001717  -1.645 0.100449
PercentMarried        1.268732   0.358492   3.539 0.000434 ***
PctNoHS18_24         -0.152387   0.135912  -1.121 0.262667
PctHS18_24            0.203987   0.111312   1.833 0.067385 .
PctHS25_Over          0.445473   0.240050   1.856 0.064002 .
PctBachDeg25_Over    -1.268364   0.358784  -3.535 0.000440 ***
PctUnemployed16_Over  1.026472   0.382415   2.684 0.007481 **
PctPrivateCoverage   -0.659422   0.226694  -2.909 0.003768 **
PctEmpPrivCoverage    0.525731   0.193828   2.712 0.006881 **
PctWhite             -0.044860   0.085918  -0.522 0.601780
PctOtherRace         -1.199937   0.303050  -3.960 8.45e-05 ***
PctMarriedHouseholds -1.425015   0.322408  -4.420 1.18e-05 ***
MedianAgeMale        -0.536391   0.257766  -2.081 0.037884 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.5 on 573 degrees of freedom
Multiple R-squared:  0.4512,            Adjusted R-squared:  0.435
F-statistic: 27.72 on 17 and 573 DF,  p-value: < 2.2e-16

> # Identify outliers using Cook's distance
> cooksd <- cooks.distance(model)
> influential <- as.numeric(names(cooksd)[(cooksd > 4/length(cooksd))])
> cat("Influential observations (outliers):", influential, "\n")
Influential observations (outliers): 31 34 116 120 122 250 282 442 466 476 616 627 661 783
920 925 1000 1059 1076 1317 1331 1345 1366 1390 1468 1817 1904 1914 1933 2034 2155 2
285 2323 2328 2444 2546 2549 2594 2598 2646 2657 2669 2674 2720 2734 2741 2809 2824
3034 3047
```

**Durbin-Watson Test**

- **Durbin-Watson statistic (DW):** 2.0529
- **p-value:** 0.6937

The Durbin-Watson statistic is close to 2, which suggests that there is no significant autocorrelation in the residuals. A DW value around 2 indicates that the residuals are not serially correlated.

**Breusch-Pagan Test for Heteroskedasticity**

- **Breusch-Pagan statistic (BP):** 57.074
- **Degrees of freedom (df):** 17
- **p-value:** 3.168e-06

The p-value is very small, indicating that there is significant heteroskedasticity in the model. This means that the variance of the residuals is not constant across observations, which could affect the efficiency of the coefficient estimates.

**Variance Inflation Factor (VIF)**

The VIF values for the predictors are:

- **avgDeathsPerYear:** 16.576731
- **povertyPercent:** 5.256236
- **incidenceRate:** 1.189414
- **AvgHouseholdSize:** 1.401168
- **avgAnnCount:** 16.187430
- **PercentMarried:** 7.544552
- **PctNoHS18_24:** 1.599390
- **PctHS18_24:** 1.518369
- **PctHS25_Over:** 3.806755
- **PctBachDeg25_Over:** 4.943159
- **PctUnemployed16_Over:** 2.251072
- **PctPrivateCoverage:** 7.885221
- **PctEmpPrivCoverage:** 4.803862
- **PctWhite:** 2.473732
- **PctOtherRace:** 1.482790
- **PctMarriedHouseholds:** 5.682878
- **MedianAgeMale:** 2.443385

VIF values greater than 10 indicate high multicollinearity. The predictors avgDeathsPerYear and avgAnnCount exhibit high multicollinearity, suggesting that these variables may be highly correlated with other predictors in the model.

**Model Coefficients and Significance**

The significant predictors at the 5% level are:

- **incidenceRate (p < 2e-16):** Positively associated with death rates.

- **PercentMarried (p = 0.000434):** Positively associated with death rates.
- **PctBachDeg25_Over (p = 0.000440):** Negatively associated with death rates.
- **PctUnemployed16_Over (p = 0.007481):** Positively associated with death rates.
- **PctPrivateCoverage (p = 0.003768):** Negatively associated with death rates.
- **PctEmpPrivCoverage (p = 0.006881):** Positively associated with death rates.
- **PctOtherRace (p = 8.45e-05):** Negatively associated with death rates.
- **PctMarriedHouseholds (p = 1.18e-05):** Negatively associated with death rates.
- **MedianAgeMale (p = 0.037884):** Negatively associated with death rates.

**Outliers**

Using Cook's distance to identify influential observations, the following points were identified as outliers:

- 31, 34, 116, 120, 122, 250, 282, 442, 466, 476, 616, 627, 661, 783, 920, 925, 1000, 1059, 1076, 1317, 1331, 1345, 1366, 1390, 1468, 1817, 1904, 1914, 1933, 2034, 2155, 2285, 2323, 2328, 2444, 2546, 2549, 2594, 2598, 2646, 2657, 2669, 2674, 2720, 2734, 2741, 2809, 2824, 3034, 3047.

These observations could potentially influence the model results significantly.

- **Model Fit:** The adjusted R-squared of 0.435 indicates that the model explains approximately 43.5% of the variance in the death rate. The RMSE of 20.18667 provides a measure of prediction accuracy.
- **Significant Predictors:** Several predictors are significant, including incidenceRate, PercentMarried, PctBachDeg25_Over, PctUnemployed16_Over, PctPrivateCoverage, PctEmpPrivCoverage, PctOtherRace, PctMarriedHouseholds, and MedianAgeMale.
- **Diagnostics:** The model exhibits heteroskedasticity, no significant autocorrelation, and some multicollinearity issues. Influential outliers have been identified and may need to be addressed.

---