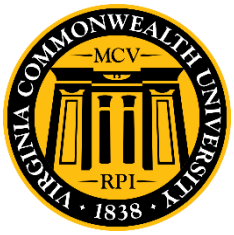

VIRGINIA COMMONWEALTH UNIVERSITY

Statistical analysis and modelling (SCMA 632)

EXAM 1

MICAH ASHADEEP EMMANUEL

V01101166

Date of Submission: 11-07-2024

SECTION A

PART B

Using R

Qn. Build a Multivariate OLS Regression Model to Predict Cancer Mortality Rates in R.

Output

```
Call:
lm(formula = TARGET_deathRate ~ ., data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-82.295 -10.336  -0.503  10.502  92.275

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.780e+02  1.779e+01  10.009  < 2e-16 ***
avgAnnCount             -3.327e-03  8.644e-04  -3.849 0.000122 ***
avgDeathsPerYear         3.221e-02  5.857e-03   5.500 4.21e-08 ***
incidenceRate            2.025e-01  8.361e-03  24.219  < 2e-16 ***
medIncome                1.985e-06  1.006e-04   0.020 0.984261
popEst2015              -4.633e-05  1.023e-05  -4.528 6.25e-06 ***
povertyPercent           9.764e-02  1.899e-01   0.514 0.607221
studyPerCap              9.789e-05  1.072e-03   0.091 0.927239
binnedInc                2.679e-01  1.526e-01   1.755 0.079378 .
MedianAge               -7.968e-02  1.755e-01  -0.454 0.649879
MedianAgeMale           -4.587e-01  2.482e-01  -1.848 0.064678 .
MedianAgeFemale         -2.153e-01  2.491e-01  -0.864 0.387532
Geography               -3.801e-05  4.296e-04  -0.088 0.929503
AvgHouseholdSize         3.637e-01  1.039e+00   0.350 0.726435
PercentMarried           1.184e+00  1.809e-01   6.545 7.25e-11 ***
PctNoHS18_24            -1.347e-01  6.611e-02  -2.038 0.041691 *
PctHS18_24               2.231e-01  5.771e-02   3.866 0.000113 ***
PctSomeCol18_24          4.462e-02  8.324e-02   0.536 0.591975
PctBachDeg18_24         -2.529e-01  1.300e-01  -1.945 0.051859 .
PctHS25_Over             3.875e-01  1.054e-01   3.675 0.000243 ***
PctBachDeg25_Over       -1.001e+00  1.742e-01  -5.744 1.04e-08 ***
PctEmployed16_Over      -6.393e-01  1.076e-01  -5.942 3.23e-09 ***
PctUnemployed16_Over    -2.445e-03  1.865e-01  -0.013 0.989542
PctPrivateCoverage      -5.572e-01  1.511e-01  -3.687 0.000232 ***
PctPrivateCoverageAlone  5.145e-02  8.993e-02   0.572 0.567315
PctEmpPrivCoverage       3.876e-01  1.112e-01   3.487 0.000498 ***
PctPublicCoverage       -1.755e-01  2.460e-01  -0.714 0.475546
PctPublicCoverageAlone   2.465e-01  3.048e-01   0.809 0.418718
PctWhite                -1.366e-01  5.698e-02  -2.397 0.016598 *
PctBlack                -9.068e-03  5.780e-02  -0.157 0.875350
PctAsian                 4.018e-01  3.421e-01   1.175 0.240244
PctOtherRace            -1.037e+00  1.652e-01  -6.276 4.10e-10 ***
PctMarriedHouseholds    -1.145e+00  1.735e-01  -6.602 4.99e-11 ***
BirthRate               -8.989e-01  2.195e-01  -4.096 4.35e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.51 on 2405 degrees of freedom
Multiple R-squared:  0.5373,   Adjusted R-squared:  0.531
F-statistic: 84.63 on 33 and 2405 DF,  p-value: < 2.2e-16
```

INTERPRETATIONS

- Significant Predictors (with p-values < 0.05):
    - avgAnnCount (-3.327e-03): For each additional average annual count, the cancer death rate decreases by 0.003.
    - avgDeathsPerYear (3.221e-02): For each additional average death per year, the cancer death rate increases by 0.032.
    - incidenceRate (2.025e-01): For each unit increase in incidence rate, the cancer death rate increases by 0.202.
    - popEst2015 (-4.633e-05): For each unit increase in the population estimate for 2015, the cancer death rate decreases by 0.000046.
    - PercentMarried (1.184e+00): For each percent increase in the married population, the cancer death rate increases by 1.184.
    - PctNoHS18_24 (-1.347e-01): For each percent increase in individuals aged 18-24 without a high school diploma, the cancer death rate decreases by 0.135.
    - PctHS18_24 (2.231e-01): For each percent increase in high school graduates aged 18-24, the cancer death rate increases by 0.223.
    - PctHS25_Over (3.875e-01): For each percent increase in high school graduates aged 25 and over, the cancer death rate increases by 0.387.
    - PctBachDeg25_Over (-1.001e+00): For each percent increase in bachelor's degree holders aged 25 and over, the cancer death rate decreases by 1.001.
    - PctEmployed16_Over (-6.393e-01): For each percent increase in employed individuals aged 16 and over, the cancer death rate decreases by 0.639.
    - PctPrivateCoverage (-5.572e-01): For each percent increase in individuals with private insurance, the cancer death rate decreases by 0.557.
    - PctEmpPrivCoverage (3.876e-01): For each percent increase in employed individuals with private insurance, the cancer death rate increases by 0.388.
    - PctWhite (-1.366e-01): For each percent increase in the white population, the cancer death rate decreases by 0.137.
    - PctOtherRace (-1.037e+00): For each percent increase in other races, the cancer death rate decreases by 1.037.
    - PctMarriedHouseholds (-1.145e+00): For each percent increase in married households, the cancer death rate decreases by 1.145.
    - BirthRate (-8.989e-01): For each unit increase in birth rate, the cancer death rate decreases by 0.899.

Model Performance

- Multiple R-squared: 0.5373 - This indicates that approximately 53.73% of the variance in cancer mortality rate is explained by the model.
- Adjusted R-squared: 0.531 - Adjusted for the number of predictors in the model, this value is slightly lower than the multiple R-squared.

- Residual Standard Error: 18.51 - This is the standard deviation of the residuals, representing the average distance that the observed values fall from the regression line.
- F-statistic: 84.63 (p-value < 2.2e-16) - This suggests that the model is statistically significant, and at least one of the predictors is associated with the outcome variable.

Interpretation and Considerations

1. Significant Predictors: Variables like incidenceRate, PercentMarried, PctHS18_24, PctHS25_Over, PctBachDeg25_Over, and PctEmployed16_Over significantly influence cancer mortality rates.
2. Non-significant Predictors: Variables such as medIncome, povertyPercent, and studyPerCap have high p-values, indicating they may not be strong predictors of cancer mortality rates.

RMSE: 19.7652
Adjusted R-squared: 0.5309595

```
durbinWatsonTest(model)
 lag Autocorrelation D-W Statistic p-value
 1       0.1073591      1.784823       0
 Alternative hypothesis: rho != 0

Multicollinearity (Variance Inflation Factor)
> vif(model)
            avgAnnCount         avgDeathsPerYear            incidenceRate
               4.974178               26.988813                 1.306824
              medIncome                popEst2015           povertyPercent
               9.764180               27.684716                10.352566
            studyPerCap                binnedInc                MedianAge
               1.094202                1.383486                 6.846081
          MedianAgeMale         MedianAgeFemale                Geography
              11.786849               12.326287                 1.011646
       AvgHouseholdSize          PercentMarried             PctNoHS18_24
               1.364966               11.227277                 1.915199
           PctHS18_24          PctSomeCol18_24          PctBachDeg18_24
               1.884901                1.354167                 2.101733
          PctHS25_Over       PctBachDeg25_Over       PctEmployed16_Over
               3.878843                5.978407                 5.414082
   PctUnemployed16_Over      PctPrivateCoverage PctPrivateCoverageAlone
               2.825412               18.704842                 4.649595
     PctEmpPrivCoverage       PctPublicCoverage  PctPublicCoverageAlone
               7.896974               26.535610                24.491001
               PctWhite                PctBlack                 PctAsian
               6.502488                4.961253                 2.714511
          PctOtherRace    PctMarriedHouseholds                BirthRate
               1.624896                9.186589                 1.227331
```

Interpretation of Results

Model Performance

- **Adjusted R-squared**: 0.5309595
  - This indicates that approximately 53.1% of the variability in the cancer mortality rate is explained by the predictors in the model, adjusted for the number of predictors.
- **RMSE (Root Mean Square Error)**: 19.7652
  - The RMSE indicates the average deviation of the observed cancer mortality rates from the values predicted by the model. A lower RMSE indicates a better fit.

Durbin-Watson Test for Autocorrelation

- **D-W Statistic**: 1.784823
- **p-value**: 0
  - The Durbin-Watson statistic ranges from 0 to 4. A value of 2 indicates no autocorrelation, values approaching 0 indicate positive autocorrelation, and values toward 4 indicate negative autocorrelation. A D-W statistic of 1.78 suggests some positive autocorrelation in the residuals. Given the p-value is 0, it indicates that we reject the null hypothesis of no autocorrelation.

Multicollinearity (Variance Inflation Factor, VIF)

- **High VIF Values (Potential Multicollinearity)**
  - avgDeathsPerYear: 26.988813
  - popEst2015: 27.684716
  - povertyPercent: 10.352566
  - PercentMarried: 11.227277
  - PctPrivateCoverage: 18.704842
  - PctPublicCoverage: 26.535610
  - PctPublicCoverageAlone: 24.491001
  - MedianAgeMale: 11.786849
  - MedianAgeFemale: 12.326287

A VIF value greater than 10 is often considered indicative of multicollinearity. In this model, several predictors exhibit high VIF values, suggesting that they are highly correlated with other predictors.

Interpretation and Considerations

1. **Significant Predictors**:
   - Variables such as avgDeathsPerYear, incidenceRate, PercentMarried, PctHS18_24, PctHS25_Over, PctBachDeg25_Over, and PctEmployed16_Over are significant predictors of cancer mortality rates.
2. **Multicollinearity**:
   - High VIF values indicate multicollinearity among several predictors. This can inflate the standard errors of the coefficients and make it difficult to assess the individual effect of each predictor. Addressing multicollinearity may involve:

        - Removing or combining highly correlated variables.
        - Using Principal Component Analysis (PCA) or Partial Least Squares (PLS) to reduce the dimensionality of the data.
3. **Autocorrelation**:
    - The Durbin-Watson test suggests some positive autocorrelation in the residuals. This could be addressed by adding lagged variables or using time-series models if the data has a temporal structure.

Section B

Part B

Line graph that shows a Poisson distribution fit to the number of wickets taken by R. Ashwin, an Indian cricketer. A Poisson distribution is a probability distribution that represents the likelihood of a certain number of events occurring in a fixed interval of time or space, if those events occur at a constant rate and independently of each other.

In the context of cricket, the Poisson distribution could be used to model the number of wickets a bowler takes in a match. The x-axis of the graph represents the number of wickets taken per match, and the y-axis represents the density. The density function shows the probability of R. Ashwin taking a certain number of wickets in a match.

For example, the graph shows that it is more likely for Ashwin to take 0 or 1 wickets in a match than it is for him to take 3 or more wickets.